TRILOGY: Discovery of sequence–structure patterns across diverse proteins

Philip Bradley*†, Peter S. Kim‡§, and Bonnie Berger*¶

*Department of Mathematics and Laboratory for Computer Science, Massachusetts Institute of Technology, Cambridge, MA 02139; and ‡Howard Hughes Medical Institute, Whitehead Institute, Department of Biology, Massachusetts Institute of Technology, Cambridge, MA 02142

Contributed by Peter S. Kim, April 12, 2002

We describe a new computer program, TRILOGY, for the automated discovery of sequence–structure patterns in proteins. TRILOGY implements a pattern discovery algorithm that begins with an exhaustive analysis of flexible three-residue patterns; a subset of these patterns are selected as seeds for an extension process in which longer patterns are identified. A key feature of the method is explicit treatment of both the sequence and structure components of these motifs: each TRILOGY pattern is a pair consisting of a sequence pattern and a structure pattern. Matches to both these component patterns are identified independently, allowing the program to assign a significance score to each sequence–structure pattern that assesses the degree of correlation between the corresponding sequence and structure motifs. TRILOGY identifies several thousand high-scoring patterns that occur across protein families. These include both previously identified and potentially novel motifs. We expect that these sequence–structure patterns will be useful in predicting protein structure from sequence, annotating newly determined protein structures, and identifying novel motifs of potential functional or structural significance. Further details on 7,768 significant patterns identified by TRILOGY can be found at http://theory.lcs.mit.edu/trilogy.

protein folding | sequence pattern | protein motif discovery

Since the discovery that the three-dimensional structures of proteins are completely determined by their amino acid sequences, considerable effort has been invested in the search for patterns in sequence that correlate with patterns in structure. This search has traditionally proceeded by the identification of a structural motif, followed by alignment of sequences corresponding to the motif and calculation of amino acid frequencies at positions in this structural alignment. Amino acid sequence preferences associated with protein secondary structures (1), β-turns (2, 3), α-helix caps (4–6), and supersecondary motifs such as the coiled-coil (7) have been successfully identified. One limitation of this approach is the fact that the motifs in question must be specified in advance.

In this paper, we introduce an approach to pattern discovery that succeeds in identifying known and unknown sequence–structure patterns across diverse protein families. The algorithm is unsupervised in that the motifs are not specified in advance. A key feature of our approach is the explicit treatment of both the sequence and structure components of these motifs as independent patterns that are identified and extended simultaneously during the search process. This allows the assignment of a statistical score, based on the hypergeometric distribution, measuring the degree of correlation between sequence and structure patterns. The patterns considered by the algorithm are assembled from building blocks with three matched residues (triple patterns, Fig. 1); these patterns are quite flexible, allowing conservative substitutions, and linkers of variable length and conformation between the pattern residues (only the pattern residues themselves are structurally constrained).

The TRILOGY program implements this algorithm and identifies several thousand high-scoring patterns (Table 1) in a set of representative domains taken from the SCOP (8) protein structure classification database. Each pattern is required to span at least three SCOP superfamilies; thus these patterns cannot be easily explained by sequence similarity between the matched proteins. The high-scoring patterns represent known motifs of structural and functional significance (helix capping patterns; an NAD/FAD binding pattern), as well as potentially novel motifs. Novel motifs include helix–strand and helix–helix transitions, interlinked disulfides, and striking similarities between proteins with different overall folds. Additionally, TRILOGY identifies novel variants of existing motifs: a helix–hairpin–helix motif that binds RNA rather than DNA was found in the S13 ribosomal protein from *Thermus thermophilus*; this motif is not found by several existing helix–hairpin–helix sequence signatures. Unsupervised motif-discovery algorithms such as TRILOGY will become increasingly important as the structure databases continue to grow.

Methods

Triple Patterns. The basic pattern objects considered by the TRILOGY algorithm are sequence–structure patterns with three specified residues, which we term triple patterns. A triple pattern P has two components: a structure pattern P_{str}, specifying the relative three-dimensional arrangement and orientation of the three residues, and a sequence pattern P_{seq} that defines the sequence spacing and residue type of the three pattern residues. The matches to a triple pattern P in the structure set are the matches common to both P_{str} and P_{seq}.

The structure pattern P_{str} is defined by a representative arrangement of three residues (e.g., Fig. 1*b*) that determines the pattern C_α–C_α distances and $C_\alpha C_\beta$ vectors (Fig. 1).

Matches to P_{str} are those residue triplets whose C_α–C_α distances and $C_\alpha C_\beta$ vectors agree with those of the pattern to within predefined allowances (1.5 Å for the distances and 60° for the vectors).

A three-residue sequence pattern P_{seq} is an expression of the form

$$R_1\, x^{a-b}\, R_2\, x^{c-d}\, R_3,$$

with three specified residue positions (R_i) and two sequence gaps (x^{a-b} indicates a gap of length between a and b residues). The specified residue positions in P_{seq} may allow a single amino acid or a class of amino acids. Our default values allowed seven predefined residue classes: {V,I,L,M}, {F,Y,W}, {D,E}, {K,R,H}, {N,Q}, {S,T}, and {A,G,S}. Gaps between pattern positions may be of fixed or variable length depending on the total gap length. We used the following empirically chosen set of 20 allowed sequence gaps, indicating the number of residues occurring between pattern positions: 0, 1, 2, 3–4, 4–5, 5–6, 6–7,

†Present address: Department of Biochemistry, University of Washington, Box 357350, Seattle, WA 98195.

§Present address: Merck Research Laboratories, 770 Sumneytown Pike, West Point, PA 19486.

¶To whom reprint requests should be addressed at: Department of Mathematics, Room 2-373, Massachusetts Institute of Technology, Cambridge, MA 02139. E-mail: bab@mit.edu.

 www.pnas.org/cgi/doi/10.1073/pnas.112221999

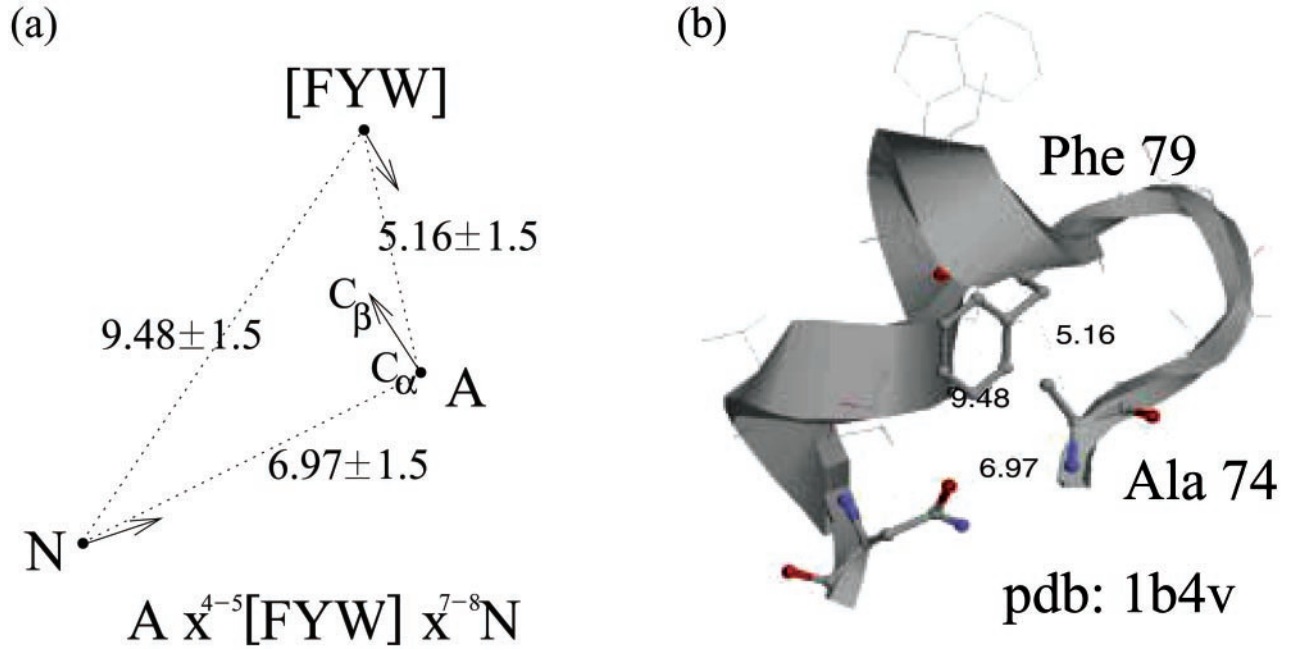

Fig. 1. (*a*) Example of a triple pattern with its component sequence and structure patterns. The sequence pattern consists of an alanine, followed by a gap of 4–5 residues, then either phenylalanine, tyrosine, or tryptophan, followed by a gap of 7–8 residues and an asparagine. The structure pattern is represented by a triangle whose vertices correspond to the C_α atoms of the pattern residues; the edges give the C_α–C_α distance constraints; and the pattern $C_\alpha C_\beta$ vectors are attached to each vertex. (*b*) The structure of the residue triplet on which the structure pattern (*a*) is based.

7–8, 8–9, 9–11, 10–13, 12–15, 14–17, 16–19, 18–21, 20–23, 22–26, 24–29, 27–32, and 30–34. There are thus, in total, $27^3 \cdot 20^2$ or roughly 8 million three-residue sequence patterns.

Longer Patterns. Sequence–structure patterns with more than three residues are formed by "gluing together" several triple patterns. Beginning with a triple pattern, we add a fourth residue by specifying a second triple pattern that matches the new residue together with two of the three residues in the first pattern (Fig. 2). This process is repeated to build longer patterns: at each step we add a new residue by joining it to the existing pattern residues by a triple pattern that constrains its position in sequence and in structure relative to the existing pattern residues. In this way, both the sequence and structure components of a pattern are extended simultaneously. Note that although the total length of triple sequence patterns is limited by the range of allowed gap lengths, patterns of more than three residues can grow to arbitrary length. In Fig. 3, 1–8, the assembled triple patterns are delineated by dotted lines drawn between pattern residues.

Significance Score. The significance score, or P-score, attached to a sequence–structure pattern *P* measures the degree of corre-

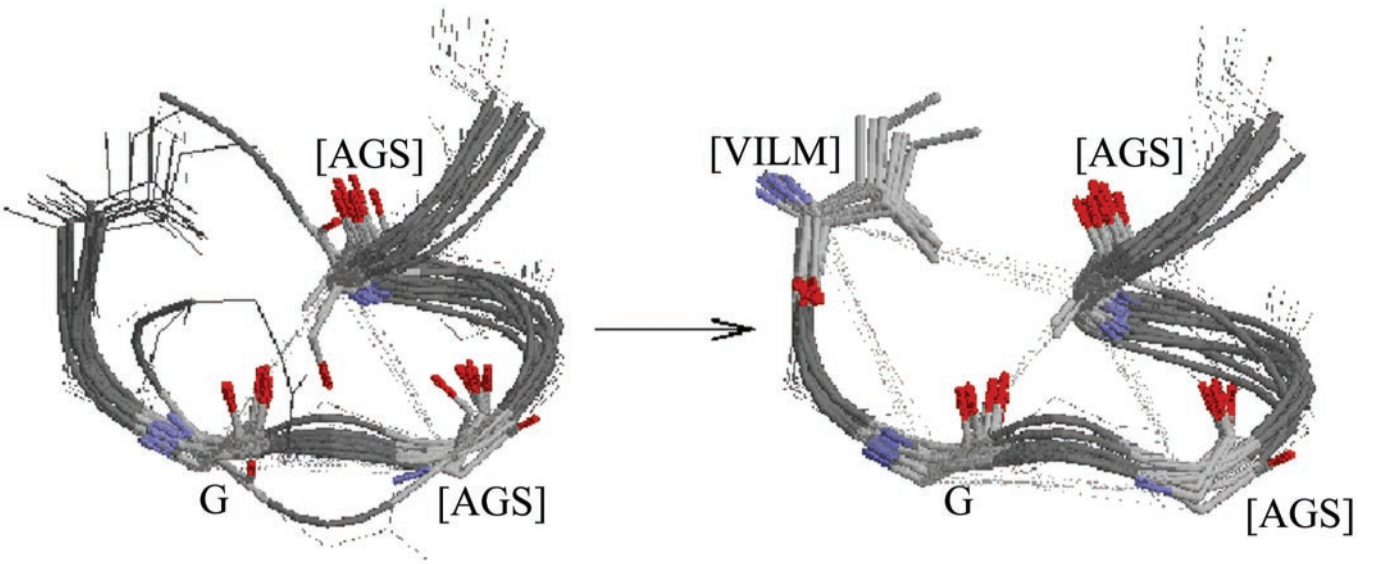

Fig. 2. Pattern extension: a hydrophobic residue [VILM] is added to the pattern on the left by adjoining a triple-pattern connection with the first and third pattern residues. Pattern representation and coloring are as described in the caption to Fig. 3.

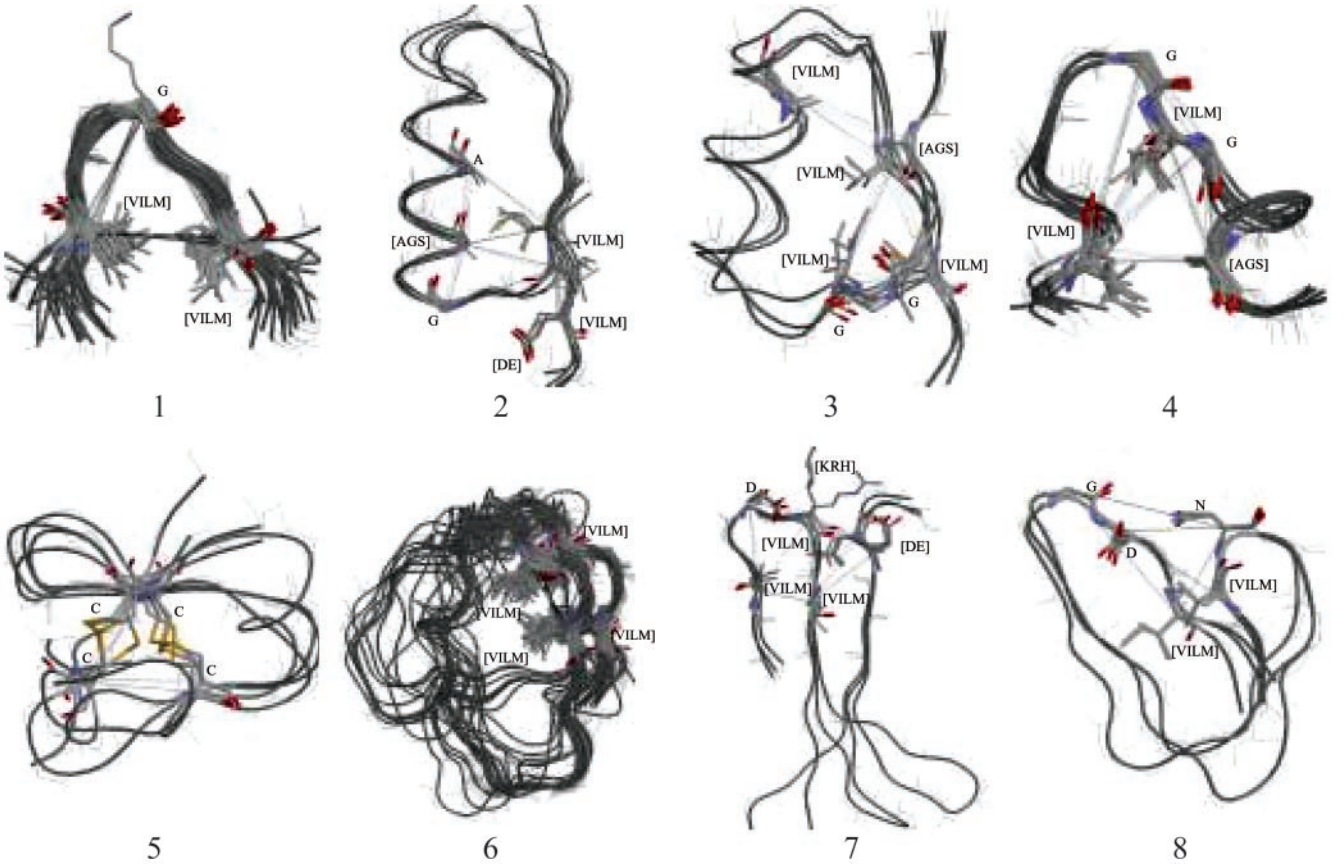

Fig. 3. The eight patterns: their matches, pattern residues, and constituent triple patterns. For each of the eight patterns described in the text, the pattern matches were structurally aligned and an image was generated using the program RASMOL (28). A backbone trace is shown for each match; pattern residues are represented in thick wireframe, with oxygen atoms colored red, nitrogen colored blue, carbon colored light gray, and sulfur colored yellow. Dashed lines are drawn between pairs of C_α atoms whose distance is constrained by a triple pattern. Note that the pattern residues need not be present in each template structure, provided that the correct amino acid (or class) is found at the matched position in 75% of the members of the sequence family represented by that structure; for example, in 1, one lysine residue is found at the position occupied by glycine in the pattern.

lation between the individual sequence and structure patterns P_{seq} and P_{str}. It equals the likelihood of seeing an equal or greater number of matches common to P_{seq} and P_{str} by chance, and is calculated in a straightforward fashion by applying the hypergeometric distribution. The relevant data are: M_{pat}, the number of matches to P; M_{seq}, the number of matches to P_{seq}; M_{str}, the number of matches to P_{str} with the sequence spacing determined by P_{seq}; T, the total number of times that we see a set of conserved residues in the structure set with sequence spacing that matches P_{seq}. Thus,

$$\text{P-score} = \sum_i C(M_{seq}, i)\, C(T - M_{seq}, M_{str} - i)\, C(T, M_{str})^{-1},$$

where the summation over i runs from M_{pat} to $\min(M_{seq}, M_{str})$, and $C(a,b)$ equals a choose b. By restricting the set of potential matches T to those having the same spacing as P, we normalize for the effect that sequence spacing alone has on the relative three-dimensional positioning of the pattern residues. Note that the score is symmetric in M_{seq} and M_{str}; in a sense it measures both the specificity of the sequence pattern for the given structure, as well as the structure for the sequence pattern.

To associate a significance threshold to the P-scores produced by the algorithm, we reran each of the three TRILOGY pattern searches (see *Search Algorithm*) after randomly permuting the residues of each sequence of the proteins in the structure set, thereby producing P-scores for a search that should yield no significant patterns. A threshold for significance was set at 0.01 multiplied by the lowest P-score obtained in the randomized search. Although the P-score alone allows us to attach a significance to each pattern, these scores must be evaluated in relation to the size of the pattern search space; for example, a P-score of 10^{-6} is outstanding in a search of only a handful of patterns, but if several million distinct patterns are examined one might expect P-scores of equal significance to occur by chance. Moreover, the potentially redundant nature of the TRILOGY search algorithm suggests that a simple count of the number of patterns scored may not give an accurate measure of the pattern space that was actually searched. Thus we compared with the results of a randomized search to set significance thresholds.

Search Algorithm. The search for significant sequence–structure patterns proceeds in three stages. In the first stage, all triplets of conserved residues (see *Structure Set*) that are nearby in three dimensions are extracted from the structure set. Three residues are considered to be nearby if the distance of closest approach (measured between nonhydrogen atoms) between each pair of residues is less than a predetermined cutoff (as described in *Results*, we experimented with three cutoffs: 4.5, 6, and 7.5 Å). Triplets spanning fewer than 6 or more than 36 residues are excluded to narrow the search (early experiments indicated diminishing returns in new patterns found with an increase of the length threshold beyond this range), as are triplets that are contained within a single stretch of secondary structure (α-helix or β-strand). This set of residue triplets forms the raw material for the pattern search.

Table 1. Summary of results

			3		4		5		6		7	
	Triplets	Seeds	*N*	P-score	*N*	P-score	*N*	P-score	*N*	P-score	*N*	P-score
4.5	193,613	36,398	1258	10^{-165}	131	10^{-53}	14	10^{-24}	2	10^{-14}	—	—
6.0	480,435	95,593	2158	10^{-152}	311	10^{-109}	37	10^{-27}	8	10^{-16}	1	10^{-15}
7.5	1,346,528	273,277	3004	10^{-150}	679	10^{-110}	117	10^{-49}	40	10^{-33}	8	10^{-21}
4.5R	142,310	17,093	—	10^{-9}	—	10^{-9}	—	10^{-8}	—	—	—	—
6.0R	367,126	53,540	—	10^{-9}	—	10^{-11}	—	10^{-9}	—	10^{-9}	—	—
7.5R	1,074,756	155,682	—	10^{-9}	—	10^{-11}	—	10^{-12}	—	10^{-11}	—	—

Statistics for TRILOGY searches conducted at three residue contact distances (4.5, 6.0, and 7.5 Å) on the native (first three rows) and randomized (last three rows, labeled with "R") structure sets. The first column gives the total number of conserved residue triplets that satisfy the distance constraint. The second column indicates the number of triple-pattern seeds passed to the extension process. For each pattern length 3–7, the total number of significant patterns (*N*) found at that length, as well as the lowest P-score attained, are given in the corresponding columns. A significant pattern is one whose P-score is less than 0.01 times the lowest P-score achieved in the randomized search at the corresponding distance constraint (see *Significance Score*).

In the second stage of the pattern search, we assign significance scores to a large set of triple patterns and choose a subset of these as seeds for the pattern extension process. We consider all triple patterns of the form (P_{seq}, t), where P_{seq} is a three-residue sequence pattern and t is a residue triplet that matches P_{seq} and serves as a template for the structure component. For each sequence pattern P_{seq}, we identify those residue triplets t that match P_{seq}. Each triple pattern pair (P_{seq}, t) is assigned a P-score. For each P_{seq}, we compare P-scores among similar residue triplets to choose their structural representatives. In particular, we choose those pairs (P_{seq}, t) with the property that

$$\text{P-score of } (P_{seq}, t) \leq \text{P-score of } (P_{seq}, t') \text{ for all } t' \sim t,$$

where $\sim$ denotes structural similarity (agreement of residue distances and side chain vectors as in *Triple Patterns*). By comparing P-scores only among similar triples, we permit several selected triple patterns to have the same sequence component provided that they are structurally distinct. These patterns may then differentiate from one another in sequence during the extension stage.

In the final stage of the pattern discovery process, we search for significant patterns with more than three residues by extending the triple patterns identified in the previous stage. The search proceeds by repeatedly adding a new residue to an existing pattern, terminating when no extension can be made without reducing the number of matched SCOP superfamilies to fewer than three. To extend a pattern P, the algorithm searches all proteins with matches to P to identify residue triplets that intersect the pattern in exactly two residues. Each such triplet defines a pattern extension: the type and location of the third triplet residue specify a triple pattern which is "glued" to P (see *Longer Patterns* and Fig. 2) to create a new sequence–structure pattern.

The list of identified patterns is processed to remove obvious redundancies. Any pattern whose match set is completely contained in the match set of a pattern with lower P-score is rejected. This includes cases in which the inferior pattern is an extension of the more significant one and *vice versa*. In this way we attempt to choose patterns that best represent the underlying motif, using the P-score as a guide. Nevertheless, some redundancy remains in the final set of patterns, with a significant number of the motifs represented by multiple pattern variants.

Structure Set. The SCOP database (Version 1.55; ref. 8) was used for the classification of structures and definition of protein domains. The structure set analyzed in this work consisted of a single protein domain from each SCOP family of proteins, chosen using the SCOP sequence resources at the ASTRAL (9) web site (1,557 families; PDB identifiers and sequence ranges of representatives available at the TRILOGY web site). Atomic coordinates for all domains were extracted from the corresponding Protein Data Bank files (10). Secondary structure assignments were made using the DSSP algorithm (11).

For each protein in the structure set, sequence homologs were taken from the HSSP database (12). An alignment was constructed by trimming these homologs in the neighborhood of gaps or in regions of low similarity to the representative protein (from each homolog, those residues were included that were contained in a window of 20 residues with no gaps and similarity to the parent protein of at least 25%). Then positions at which the same amino acid or amino acid class occurred in 75% of the aligned sequences were considered to be conserved. All sequence and structure pattern matches were restricted to conserved residues. (Note, however, that the conserved residue need not necessarily occur in the protein of known structure, provided that it occurs in 75% of the aligned sequences.)

Results

The TRILOGY algorithm identifies a striking number of significant patterns when applied to a representative set of domains taken from the SCOP database. In general, short patterns identified by TRILOGY correspond to supersecondary structures (helix caps, β-turns, β–α–β units), whereas longer patterns represent functional motifs or possible evidence of distant evolutionary relatedness. Table 1 summarizes the results of TRILOGY searches for three values of the inter-residue contact distance threshold. As expected, the number of patterns increases as the distance threshold is increased; however, this comes at the expense of better P-scores attained in the randomized search (i.e., higher noise levels; see *Significance Score*). As a result, some patterns which appear in the 4.5 Å search fall below the significance threshold in later searches.

The most significant (i.e., lowest) P-scores in the three searches are found at the shorter pattern lengths; by contrast, the distribution of P-scores in the randomized searches is fairly level across pattern lengths, indicating that this feature of the native protein set is unlikely to be an artifact of the search procedure. The two most significant scoring three-residue patterns for all distance thresholds are helix capping patterns: [VILM] x^{3-4} [DE] [VILM], found at the N termini of α-helices, and [VILM] x^{3-4} G [VILM], found at the C termini. These correspond to well known capping motifs (4–6). The longest patterns remaining after filtering contain seven residues. Patterns with as many as ten residues were identified in the 7.5 Å search, but these were extensions of more significant patterns and were filtered out as described in *Search Algorithm*.

In the following subsections we describe eight of the patterns in more detail. The selected patterns were chosen for potential interest and to illustrate general features of the set of significant patterns. Table 2 gives summary information for the eight patterns, which are portrayed in Fig. 3, 1–8. Descriptions and

BIOCHEMISTRY

Table 2. Selected significant triple patterns identified

	L	P-score	Positions	Sequence	Structure	M_{pat}	M_{seq}	M_{str}
1	3	4.1e-114	1,2,3	[VILM] x^2 G x^2 [VILM]	2rslA,55,58,61	82	587	255
2	6	5.3e-21	1,5,6	[VILM] x^{20-23} [VILM] x^1 [DE]	1qniA,392,416,418	4	5	4
			1,3,5	[VILM] x^{4-5} [AGS] x^{14-17} [VILM]	3grs_,26,32,48			
			3,4,5	[AGS] x^{3-4} A x^{10-13} [VILM]	1cjcA,18,22,36			
			1,2,3	[VILM] x^2 G x^2 [AGS]	1cjcA,12,15,18			
3	7	2.4e-16	1,4,6	[AGS] x^{3-4} [VILM] x^{14-17} [VILM]	1zfjA,342,346,363	3	3	3
			4,6,7	[VILM] x^{14-17} [VILM] x^1 [VILM]	2tpsA,189,205,207			
			3,4,7	G [VILM] x^{16-19} [VILM]	2tpsA,188,189,207			
			1,5,6	[AGS] x^{10-13} [VILM] x^{5-6} [VILM]	1d3gA,332,346,353			
			2,3,7	G G x^{16-19} [VILM]	1zfjA,344,345,365			
4	5	2.1e-32	1,2,4	[VILM] x^{3-4} G x^1 G	2abk_,111,116,118	7	13	7
			1,4,5	[VILM] x^{6-7} G x^{3-4} [AGS]	2abk_,111,118,122			
			1,3,5	[VILM] x^{4-5} [VILM] x^{3-4} [AGS]	1mun_,111,117,122			
5	4	2.8e-19	1,2,4	C x^{4-5} C x^{10-13} C	2btcl,510,516,528	4	20	4
			1,3,4	C x^{10-13} C x^{4-5} C	2btcl,510,522,528			
6	4	8.1e-70	1,2,3	[VILM] x^2 [VILM] x^{18-21} [VILM]	1cdzA,44,47,67	34	1434	38
			2,3,4	[VILM] x^{18-21} [VILM] x^2 [VILM]	1d0bA,210,229,232			
7	6	4.5e-15	3,5,6	[KRH] x^{16-19} D [VILM]	1qfmA,245,265,266	3	3	3
			4,5,6	[VILM] x^{14-17} D [VILM]	1mrj_,48,65,66			
			1,3,4	[VILM] x^{3-4} [KRH] x^1 [VILM]	1c9lA,32,36,38			
			1,2,3	[VILM] x^2 D x^1 [KRH]	1qfmA,240,243,245			
8	5	1.4e-15	2,3,4	N x^{12-15} G D	1fm0D,56,69,70	3	3	3
			1,2,4	[VILM] N x^{12-15} D	1fm0D,55,56,70			
			2,3,5	N x^{12-15} G x^2 [VILM]	1g6gA,112,127,130			

Summary information for the eight patterns described in the text. The first two columns give the pattern length (*L*) and P-score; the last three columns indicate the number of matches to the pattern (M_{pat}) and to its sequence (M_{seq}) and the structure (M_{str}) patterns independently. TRILOGY patterns are formed by linking together several triple patterns; for each of the eight patterns, the middle three columns describe the assembly of these constituent triple patterns. For each triple pattern, the pattern residues that it links to are given in the ''Positions'' column (pattern residues are numbered in sequential order); the sequence pattern is given in the ''Sequence'' column; the ''Structure'' column lists the structure and residue triplet on which the structure pattern is based (see Fig. 1).

pictures for an additional eight patterns are published as supporting information at the PNAS web site, www.pnas.org. The full set of patterns can be accessed at the TRILOGY web site.

Supersecondary Structural Motifs. TRILOGY identifies many high-scoring patterns of length three and four that highlight residue preferences in links between secondary structure elements. Pattern 1, [VILM] x^2 G x^2 [VILM], corresponds to a type-II β-turn between unpaired β-strands and is the highest scoring β-turn pattern in the 7.5 Å search. This motif corresponds to the novel diverging type-II β-turn discovered in the construction of the I-sites library (13). The consensus backbone angle pattern is $\beta\beta\beta\alpha_L\beta\beta\beta$ with the central glycine residue in α_L conformation. The first and third pattern residues form a hydrophobic contact.

Functional Patterns. TRILOGY identifies several high-scoring patterns with functional significance. Pattern 2 is a member of a cluster of patterns corresponding to a recurring NAD/FAD binding motif found in several different SCOP folds. The binding motif spans a strand–helix–strand unit. ADP is bound parallel to the loop connecting the first strand and the α-helix; side- and main-chain atoms from loop residues interact with the ribose ring and the first phosphate. In addition, the negatively charged residue (aspartic or glutamic acid) at the end of the pattern forms hydrogen bonds to a pair of ribose oxygens. This sequence motif has been described previously and used for prediction purposes (14, 15); in addition, variants of this pattern were discovered by the motif discovery algorithm of Rigoutsos *et al.* (16).

Pattern 3 matches a β–α–β unit in three proteins from different superfamilies in the TIM-barrel fold. In all three matches, the loop between the first strand and the helix is involved in binding a phosphate group: the backbone nitrogen of the third pattern residue forms a hydrogen bond to one of the phosphate oxygens. In addition, the backbone nitrogen of the residue after the last pattern residue forms a hydrogen bond to a second phosphate oxygen. The three matches (with their bound phosphate-containing molecule) are as follows: thiamin phosphate synthase (thiamin phosphate), dihydroorotate dehydrogenase (flavin mononucleotide), and inosine monophosphate dehydrogenase (inosinic acid). This structure pattern highlights a known similarity between the phosphate-binding regions of these proteins.

Pattern 4 reveals a potentially novel variant of the helix–hairpin–helix DNA-binding motif. The matched proteins all bind DNA, with the exception of the S13 protein of the *T. thermophilus* 30S ribosomal subunit (1fjf, chain M), which binds 16S rRNA. In the DNA-binding proteins, the pattern matches correspond to the well known helix–hairpin–helix motif (17, 18), which is involved in non-sequence-specific DNA binding. Examination of the 30S crystal structure (19) reveals that the pattern match in the S13 protein is in contact with the 16S rRNA, and that it interacts with the nucleic acid in the same fashion as do the DNA-binding proteins: by hydrogen bonding between backbone atoms in the hairpin region and RNA phosphate groups. It is noteworthy that this RNA-binding motif in the S13 protein is not detected by helix–hairpin–helix models in the Pfam (20) or Interpro (21) databases.

Cysteine Patterns. The flexibility inherent in TRILOGY patterns allows novel groupings of known motifs to be identified. The four cysteines in pattern 5 form a network of interlinked disulfide bonds. The first pattern residue is disulfide bonded to the third residue, and the second to the fourth. All four matches come from different superfamilies in a single SCOP fold, the knottins, in the class of small proteins. The interlinked disulfides likely play a role in stabilizing these small proteins.

Repeat Patterns. The TRILOGY algorithm is quite successful at identifying residues in sequence–structure repeats. Pattern 6 spans 1.5 complete rungs of the leucine-rich-repeat fold. The

first and third residues are in alignment in a β-sheet, as are the second and fourth residues. The matched proteins are characterized by a processive fold in which repeated β–α or β–β–α units stack to form a cylindrical structure.

Structural Similarities. Many of the patterns identified by the algorithm reflect sequence–structure similarities that cannot be readily explained in terms of known functional or structural motifs. These patterns may reflect distant evolutionary relatedness, or the convergent evolution of sequence and structure to fit a common architectural framework. Pattern 7 spans three strands of an anti-parallel β-sheet. The pattern occurs in two different β-propeller folds (one with six blades and one with seven), at the same position within a single-blade unit in all three matches. Thus the pattern lends support to the notion that different propeller folds are related by duplication of blades.

Fig. 3, 8 shows three matches to a five residue pattern that spans a β-hairpin and a distinctive extended crossover connection to a third β-strand. The conserved asparagine is within hydrogen bonding distance of the aspartic acid (2.94, 3.01, and 3.02 Å), potentially forming a stabilizing interaction. The conserved glycine residue forms a tight turn in α_L conformation. The three matches come from proteins with quite different overall folds; however, two of the proteins (the S4 protein from the *T. thermophilus* 30S ribosomal subunit and molybdopterin synthase subunit 1) display extended structural similarity over a region N-terminal to the pattern.

Discussion

The TRILOGY algorithm successfully identifies known and novel motifs when applied to a representative set of protein domains. Many of these motifs have clear functional and/or structural significance; some may be evidence of distant evolutionary relationships. TRILOGY also identifies sequence–structure patterns for which a clear biological explanation is not apparent; this allows the generation of new hypotheses regarding functional or structural significance that can then be tested experimentally. Central to the algorithm's success are the triple-pattern representation and a statistically well founded significance score that highlights potentially interesting motifs.

The TRILOGY algorithm is automated and unsupervised, and as a result it has the potential to find entirely new patterns, as well as novel variants of existing motifs and new groupings of structurally similar motifs. At the same time, the patterns discovered by unsupervised learning may not be optimized for prediction (13). Patterns of amino acid variation are likely to be more complicated than those allowed for by predefined residue classes. The critical role played by the TRILOGY algorithm is the initial identification of a motif. Once the motif is recognized it can be examined more closely, for example by considering those proteins that matched either the sequence or structure pattern but not both. Some of these hits may be true examples of the underlying motif. Thus an important next step will be clustering (to reduce redundancy) and refinement of the identified patterns, perhaps following the iterative strategy of Bystroff and Baker (13). The resulting library of refined motifs may be useful in structure prediction and functional annotation of experimentally determined crystal structures.

Our approach to motif recognition is complementary to existing methods of pattern discovery. Traditional family-specific sequence motifs, such as those found in the PROSITE (22) database, are not reported by the algorithm because they fail to span the requisite number of SCOP superfamilies. These patterns can often be discovered by analysis of multiple sequence alignments without the need for structural information. On the other hand, patterns of structurally conserved residues in which the sequence spacing and order is not well maintained, such as convergently evolved triads of catalytic residues (23), will not be detected by the algorithm because they lack a sequence–pattern component. Although motifs of this sort have been identified primarily by expert inspection, several methods for more automated construction of structural-motif descriptors have recently been described (24, 25).

The TRILOGY algorithm differs from previously described methods for automated discovery of sequence–structure patterns (13, 16, 26, 27) by handling sequence and structure simultaneously and symmetrically in the search process. This allows identification of motifs with very few occurrences, which are difficult to identify in procedures that begin by clustering and searching in sequence space. The constraint on sequence spacing, while restricting the scope of the algorithm, makes it feasible to search all structures at once and thereby up-weight patterns with multiple occurrences, in contrast to procedures that rely on pairwise comparison of structures (25).

We expect that automated approaches such as the TRILOGY algorithm will become increasingly important as the structural genomics initiatives begin to produce protein structures in high-throughput fashion. This view is supported on a small scale by the identification of several interesting pattern matches in proteins from the recently solved crystal structure of the 30S ribosomal subunit (19).

Thanks to David Akey, Alex Coventry, Amy Keating, Jonathan King, Dan Kleitman, and Bob Sauer for helpful discussions, and to Alex Coventry for generous computational assistance. P.B. was partially supported by a Massachusetts Institute of Technology/Merck graduate fellowship and a Clay Mathematics Institute Liftoff fellowship.

1. Chou, P. Y. & Fasman, G. D. (1978) *Adv. Enzymol.* **47,** 145–148.
2. Sibanda, B. L. & Thornton, J. M. (1985) *Nature (London)* **316,** 170–174.
3. Hutchinson, E. G. & Thornton, J. M. (1994) *Protein Sci.* **3,** 2207–2216.
4. Richardson, J. S. & Richardson, D. C. (1988) *Science* **240,** 1648–1652.
5. Presta, L. G. & Rose, G. D. (1988) *Science* **240,** 1632–1641.
6. Harper, E. T. & Rose, G. D. (1993) *Biochemistry* **32,** 7605–7609.
7. Parry, D. A. (1982) *Biosci. Rep.* **2,** 1017–1024.
8. Murzin, A., Brenner, S., Hubbard, T. & Chothia, C. (1995) *J. Mol. Biol.* **297,** 536–540.
9. Brenner, S. E., Koehl, P. & Levitt, M. (2000) *Nucleic Acids Res.* **28,** 254–256.
10. Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., Shindyalov, I. N. & Bourne, P. E. (2000) *Nucleic Acids Res.* **28,** 235–242.
11. Kabsch, W. & Sander, C. (1983) *J. Mol. Biol.* **297,** 536–540.
12. Sander, C. & Schneider, R. (1991) *Proteins* **9,** 56–68.
13. Bystroff, C. & Baker, D. (1998) *J. Mol. Biol.* **281,** 565–577.
14. Wieranga, R. K., Terpstra, P. & Hol, W. G. (1986) *J. Mol. Biol.* **187,** 101–107.
15. Bork, P. & Grunwald, C. (1990) *Eur. J. Biochem.* **191,** 347–358.
16. Rigoutsos, I., Floratos, A., Ouzounis, C., Gao, Y. & Parida, L. (1999) *Proteins* **37,** 264–277.
17. Doherty, A. J., Serpell, L. C. & Ponting, C. P. (1996) *Nucleic Acids Res.* **24,** 2488–2497.
18. Shao, X. & Grishin, N. (2000) *Nucleic Acids Res.* **28,** 2643–2650.
19. Wimberly, B. T., Broderson, D. E., Clemons, W. M. J., Morgan-Warren, R. J., Carter, A. P., Vonrhein, C., Hartsch, T. & Ramakrishnan, V. (2000) *Nature (London)* **407,** 327–339.
20. Sonnhamer, E., Eddy, S., Birney, E., Bateman, A. & Durbin, R. (1998) *Nucleic Acids Res.* **26,** 320–322.
21. Apweiler, R., Attwood, T. K., Bairoch, A., Bateman, A., Birney, E., Biswas, M., Bucher, P., Cerutti, L., Corpet, F., Croning, M. D. R., *et al.* (2000) *Bioinformatics* **16,** 1145–1150.
22. Hofmann, K., Bucher, P., Falquet, L. & Bairoch, A. (1999) *Nucleic Acids Res.* **27,** 215–219.
23. Dodson, G. & Wlodawer, A. (1998) *Trends Biochem. Sci.* **23,** 347–352.
24. Di Gennaro, J. A., Siew, N., Hoffman, B. T., Zhang, L., Skolnick, J., Neilson, L. I. & Fetrow, J. S. (2001) *J. Struct. Biol.* **134,** 232–245.
25. Russell, R. B. (1998) *J. Mol. Biol.* **279,** 1211–1227.
26. Han, K. F. & Baker, D. (1995) *J. Mol. Biol.* **251,** 176–187.
27. Jonassen, I., Eidhammer, I. & Taylor, W. R. (1999) *Proteins Struct. Funct. Genet.* **34,** 206–219.
28. Sayle, R. A. & Milner-White, E. J. (1995) *Trends Biochem. Sci.* **20,** 374–376.