

Consistency Analysis of Similarity between Multiple Alignments: Prediction of Protein Function and Fold Structure from Analysis of Local Sequence Motifs

Victor Kunin¹, Bob Chan², Einat Sitbon¹, Gila Lithwick¹
and Shmuel Pietrokovski^{1*}

¹*Department of Molecular Genetics, The Weizmann Institute of Science, Rehovot 76100, Israel*

²*Fred Hutchinson Cancer Research Center, Seattle WA 98109, USA*

A new method to analyze the similarity between multiply aligned protein motifs (blocks) was developed. It identifies sets of consistently aligned blocks. These are found to be protein regions of similar function and structure that appear in different contexts. For example, the Rossmann fold ligand-binding region is found similar to TIM barrel and methylase regions, various protein families are predicted to have a TIM-barrel fold and the structural relation between the ClpP protease and crotonase folds is identified from their sequence. Besides identifying local structure features, sequence similarity across short sequence-regions (less than 20 amino acid regions) also predicts structure similarity of whole domains (folds) a few hundred amino acid residues long. Most of these relations could not be identified by other advanced sequence-to-sequence or sequence-to-multiple alignments comparisons. We describe the method (termed CYRCA), present examples of our findings, and discuss their implications.

© 2001 Academic Press

Keywords: protein structure and function prediction; computational biology; blocks analysis

*Corresponding author

Introduction

Rapidly increasing amounts of available sequence data are both the boon and bane of sequence search methods. While the data are now more likely to include sequences relevant for the analysis, they also contains many more unrelated sequences. In a typical analysis, the aim is to identify similarity between protein (or nucleotide) sequences that code for molecules with the same or similar function. The sequences might be of whole genes, of domains that appear in different genes, or of local sequence regions that correspond to specific functional or structural motifs.

Corresponding whole protein molecules (genes), domains or regions are often described as members of a family that usually are also of a common origin (homologous). Once some members of a family are known, all their sequences can be used together to improve the identification of other sequences. Typically, this is done by multiply aligning the sequences and then using the multiple alignments (rather than the sequences) in the analysis.

Multiple alignment-to-sequence searches can identify genuine relations that are not found by sequence-to-sequence searches.^{1–6} This is the result of identifying what residues are preferred in the conserved sequence regions of the family members and of not using sequence regions that are not conserved.

Searching multiple alignment databases with multiple alignment queries is superior in many cases to multiple alignment-to-sequence searches for the same reasons that this last approach is better than sequence-to-sequence searches.^{7,8} In general, more information is used for regions that are assumed to be important (conserved).

Another approach to increase the selectivity of database searches is to use multiple pairwise alignments. This was first suggested by Altschul and Lipman for local ungapped sequence to sequence searches.⁹ The basic idea is that similarity scores that are statistically insignificant in isolated pairwise relations are significant when occurring in a consistent set of relations. Previously, this approach was manually used in the analysis of one type of fold (HTH DNA-binding domains) in multiple alignment-to-multiple alignment

E-mail address of the corresponding author:
pietro@bioinfo.weizmann.ac.il

searches.^{7,10,11} Structure determination of some of the analyzed proteins verified the predictions.^{12,13}

Here, we describe a fully automated method (termed CYRCA) that continues the advance from sequence-to-sequence to multiple alignment-to-multiple alignment searches. CYRCA clusters three or more consistently aligned blocks (multiple-alignment of protein motifs) into sets. Careful inspection of many sets shows that they contain proteins with the same function, specific structural motifs and even global structural fold. Most of these relations cannot be identified by other advanced sequence-to-sequence or sequence-to-multiple alignments comparisons. Our approach assigns previously unknown structure and function features to many regions in diverse protein families. Remarkably, block-to-block sequence similarity across regions shorter than 20 amino acid residues identified structural similarity of whole protein folds, hundreds of amino acid residues long.

Algorithm

Our method is named CYRCA for cyclical relations consistency analysis. The input for CYRCA is the output of the LAMA block-to-block comparison program. LAMA compares pairs of blocks, finding their optimal local alignments and calculating a Z-score for these alignments.⁷ LAMA output consists of aligned pairs of blocks scoring above some Z-score threshold. From these data, pairs with blocks that have biased composition are removed. Similarities to such blocks are often "trivial", due to only the composition of the block sequences and thus do not reflect functional similarity.⁷

CYRCA first identifies all cyclical block-relations (e.g. A similar to B, B similar to C, and C similar to A, for a cycle of three blocks). We used the minimal cycle size of 3, identifying all triangles of blocks, but the algorithm can function with any cycle size. Only triangles where the pairwise aligned regions between all blocks both overlap and are consistent were kept. Triangles of aligned

blocks are consistent when position x in block A (A_x) is aligned with position B_y , B_y is aligned with C_z , and C_z is aligned with A_x (Figure 1(a)). Next, triangles containing identical edges (aligned pairs of blocks) are unified, forming larger connected sets. Such unified sets also are consistent (Figure 2). Following this step, any additional consistent block-to-block relations within each set (those that form cycles larger than 3) are added. Last, pairwise blocks relations of high significance⁷ are added to the sets. These relations are added without a consistency check, and may add linear segments to some cyclical sets (edge E-F in Figure 1(b)).

Despite the non-redundant nature of the Blocks+ database,¹⁴ some of its blocks are quite similar to one another and not independent. This is typically the result of similarity between blocks representing corresponding (orthologous) regions in sub-families. Such dependent blocks are usually found with very high scores to each other by LAMA. This can lead to false CYRCA sets, since the same chance similarities will be found for all dependent blocks. To avoid these false sets, edges with a score above a determined high threshold (mean column comparison score of 0.9⁷) are not considered during the flow of the program and are added in the last stage.

Results

Consistent sets of similar blocks

All blocks from the Blocks+ database¹⁴ were compared with each other using the LAMA program.⁷ A low score threshold was used to increase sensitivity (include most of the true-positives). Cyclical sets of blocks were detected in this data using the CYRCA algorithm. Table 1 summarizes the data condensation in the progression from protein sequences to block sets.

To estimate the number of false CYRCA sets and determine an optimal threshold, we repeated our procedure on a data set where no genuine hits should occur. These data were derived by shuffling

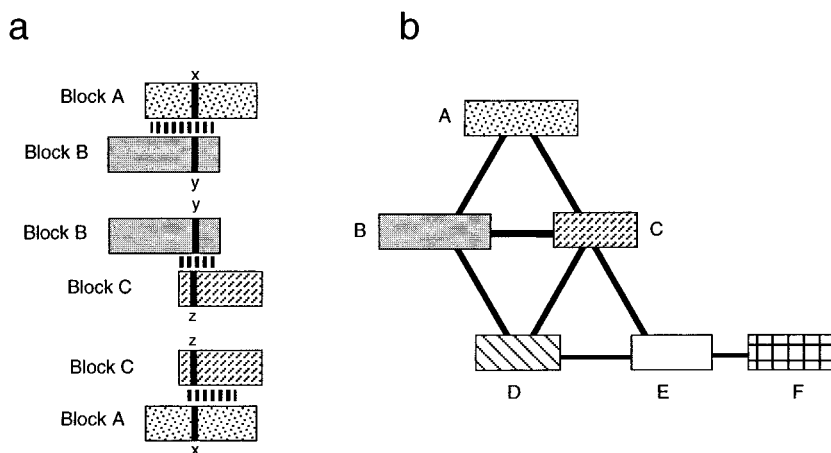


Figure 1. Consistent set of aligned blocks. (a) Three consistently aligned pairs of blocks. Blocks are shown as rectangles with one position marked by a vertical line. The aligned region between the blocks is shown between them. (b) A possible consistent CYRCA set. The top triangle (ABC) results from the block alignments shown in (a). This triangle was joined to triangle BCD and that one to triangle CDE. Finally, segment EF was joined to the set (see the text).

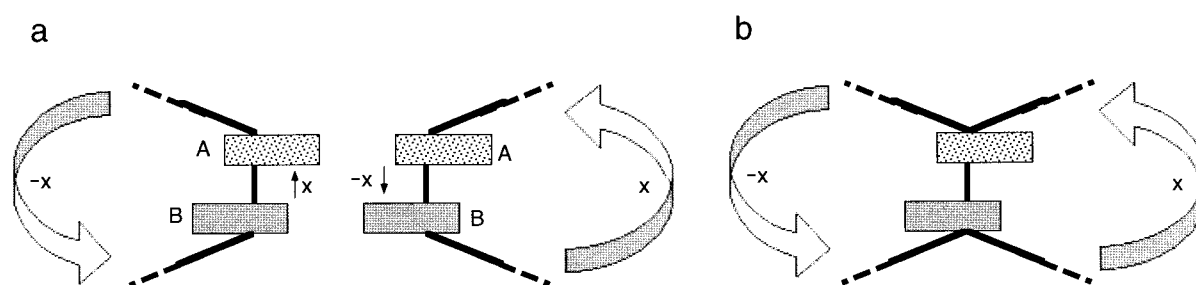


Figure 2. Unifying consistent sets that share a common edge creates a consistent set. Each block alignment (edge) has a shift describing the offset of the first block relative to the second (Figure 1(a)). This shift is an integer whose sign, + or -, depends on the direction the edge is traversed. The sum of these shifts traveling through a consistent cycle of edges is 0, since the aligned region on any block must eventually be aligned back to itself (Figure 1(a)). (a) two consistent cyclic sets that have the same edge A-B are shown. This edge has a shift x on one set and $-x$ on the other when it is traversed in the opposite direction, as indicated by the arrows. The sum of shifts on the rest of the set is either $-x$ or x , indicated by the broad arrows. (b) Unifying the two sets along edge A-B forms a new cyclic set. This set can be traversed along the broad arrows with a zero sum of shifts ($-x + x$). Hence, the new set is consistent.

columns of individual blocks⁷ in the Blocks+ database we used. We found a LAMA Z-score threshold of 5.6 to be the most informative. In the 3295 pairwise relations found above this threshold, CYRCA identified only two small cyclical sets of three and four edges. This is significantly less than the 61 sets observed in analysis of the original Blocks+ database (Table 1). This simulation indicates that we should expect a few false sets with the threshold we used. However, this threshold is a reasonable compromise: increasing it eliminates some genuine sets and lowering it adds apparently false sets (results not shown).

CYRCA sets identified here range in size from the minimal three to 47 blocks. The connectivity within each set is partial (not all blocks are connected to each other above the score threshold), being lower in the larger sets. In order to examine the nature of CYRCA sets, we used protein structures available for some block sequences. The three-dimensional coordinates of pairs of protein regions whose sequences were aligned by CYRCA were superimposed according to the alignment. Thirty-two sets contained two or more blocks with proteins where the structure of the aligned region was determined. For these sets, we could superimpose these regions, calculate structural similarity

and inspect the resemblance of the protein structures. These sets appear true, except one. This one false set contained three blocks with structurally unrelated protein motifs. Five additional sets were identified as true using a transmembrane region prediction method,¹⁵ and five more sets were considered true by their sequence and family annotation. Hence, we believe that at least 67%, $(31 + 5 + 5)/61$, of the sets are due to genuine relations and one set (1.6%) is false. At present, we cannot evaluate the correctness of the other 19 sets and we cannot verify all the pair relations in the genuine sets. However, we believe these relations to be almost all correct, since almost no false pairs were identified in these pairs from structure data. The total number of block pairs in the genuine sets is 87%; 579/663, of the total number of pairs in CYRCA identified sets. Known false block pairs are less than 0.5%, 3/663. Thus, CYRCA is a selective and reliable method; its sensitivity is demonstrated in the examples below.

CYRCA sets include protein active sites, ligand-binding regions and protein-protein interaction domains. Many sets with identified function and/or structure included families that were not known to possess these features. Thus, based on identified function of the sets and superimposition of some

Table 1. Data condensation in identifying of CYRCA block sets

SwissProt database version 38.0 Protein sequences	Blocks+database Jan. 2000 ^a Protein families	Blocks	Possibly related block pairs ^b	Cyclical block sets (block pairs in these sets)
80,000	2334	10,532	5537	61 (663)
Results from shuffled block data:			3295	2 (7)

^a Some sequences appear in more than one family due to the modular nature of many proteins and the presence of protein sub-families and super-families contained or containing other families in the Blocks database.

^b All entries in the Blocks+ database were inter-compared by LAMA with a Z-score threshold of 5.6. A Z-score threshold of 8.1 was used to identify linear segments in cyclical sets (see the algorithm in the text).

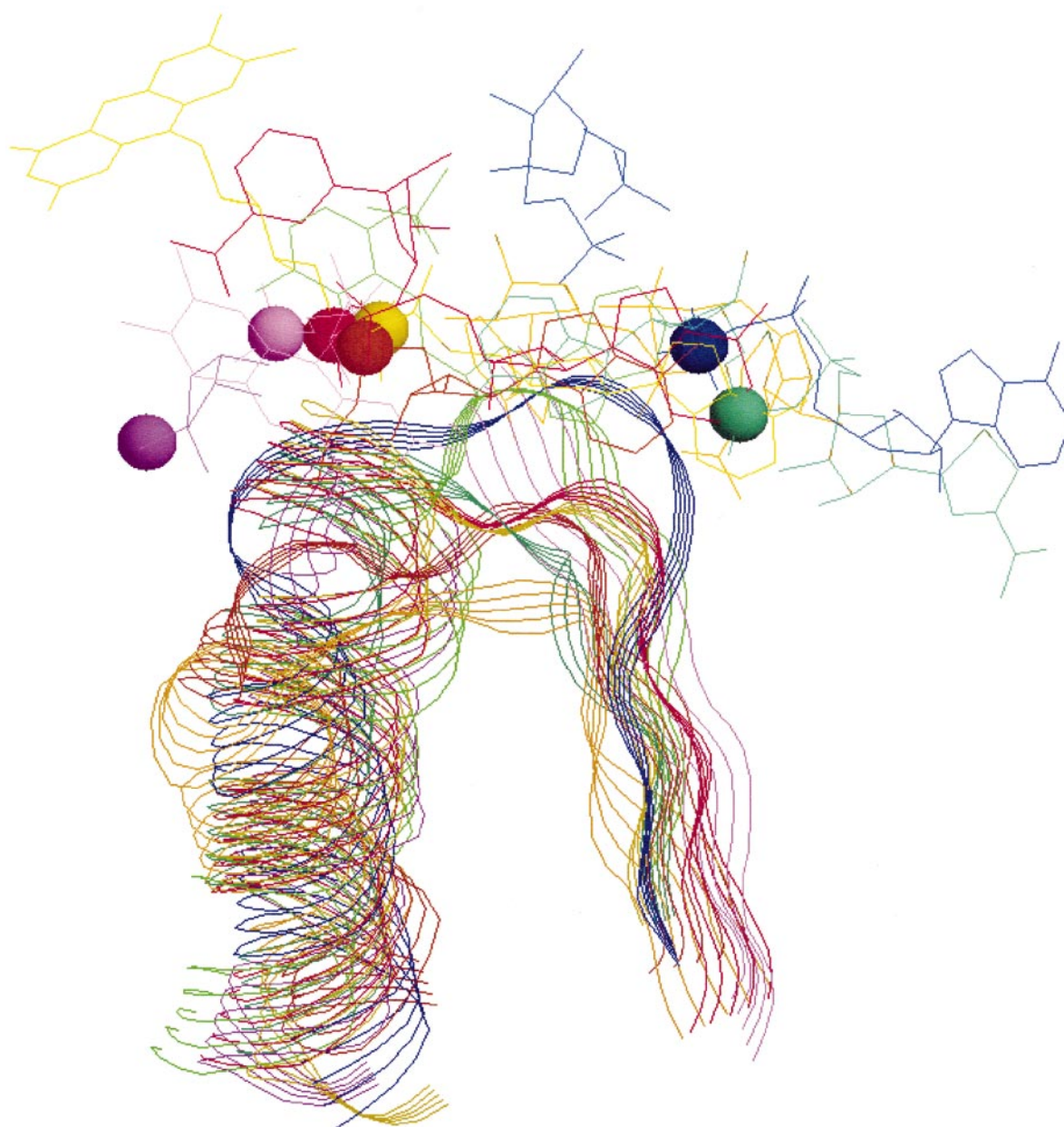


Figure 3. Rossmann-type folds ligand binding motif. Structure superimposition of ligand-binding regions from representative structures. Each structure is colored differently with the protein backbones shown as strands and ligands shown as “sticks” with phosphate and sulfate groups, and an oxygen atom to be phosphorylated shown as spheres. The representative structures for different folds (Table 2) and their aligned region (corresponding to the Rossmann fold structure 1A5Z 23-42) are: blue, phosphofructokinase (phosphofructokinase fold) 1PFK 6-24, with fructose 1,6-diphosphate and ADP, ADP β -phosphate shown. Light green, tryptophan synthase (tryptophan synthase β -chain like fold) 2TYS-B 107-125, with tryptophan bound to pyridoxal-phosphate. Khaki, glutathione reductase (FAD/NAD(P) binding fold) 1GER-A 7-27, with FAD, flavin phosphate shown. Purple, ribokinase (ribokinase-like fold) 1RKD 40-55, with ribose O-5 to be phosphorylated shown. Red, L-lactate dehydrogenase (Rossmann fold) 1A5Z 23-42, with NAD, nicotinamide phosphate shown. Violet, flavodoxin (flavodoxin-like fold) 1FLV 82-88, with FMN, phosphate shown. Dark green, inosine monophosphate dehydrogenase (Tim-barrels fold) 1B3O-A 248-265, with NAD-analog, adenine phosphate shown and in yellow 1B3O-A 299-316, with NAD-analog. Brown, rRNA adenine N-6 methylase (methylases fold) 1QAO 33-51, with SAM sulfur atom shown.

protein structures, we could generate structural and/or functional predictions for these families. To illustrate these results and the general potential of our approach, we next present detailed descriptions of some sets we identified.

Rossmann fold ligand-binding motif

The largest CYRCA set found in this study (47 blocks) includes blocks from proteins with Rossmann-type structure folds. Rossmann fold^{16,17} and

Table 2. Rossmann-type fold ligand binding motif set

Known structures ^a	Unknown structures with possible fold assignments ^b
	<i>A. Rossmann and Rossmann-like folds</i>
Zinc-containing alcohol dehydrogenases	3-Hydroxyisobutyrate dehydrogenases
L-Lactate dehydrogenases	3-Hydroxyacyl-CoA dehydrogenases
Glu/Leu/Phe/Val dehydrogenases	Potassium transport proteins ^d
6-Phosphogluconate dehydrogenases.	Alanine-dehydrogenase/pyridine-nucleotide transhydrogenases
Transketolases	NAD-dependent glycerol-3-phosphate dehydrogenases
Phosphoribosylglycinamide synthetases	Delta 1-pyrroline-5-carboxylate reductases
D-Isomer specific 2-hydroxyacid dehydrogenases	
Glutamyl-tRNA reductases ^c	
	<i>B. FAB/NAD(P)-binding fold</i>
Pyridine nucleotide-disulphide oxidoreductases class I	FAD-dependent glycerol-3-phosphate dehydrogenases
Pyridine nucleotide-disulphide oxidoreductases class II	Flavin-containing amine oxidases
GMC oxidoreductases	Bacterial-type phytoene dehydrogenases
Adrenodoxin reductases	Glucose-inhibited division-proteins A
Rab GDI/REP proteins	
	<i>C. Flavodoxin-like and Flavodoxin-like 3 layers folds</i>
Flavodoxin proteins	
	<i>D. Phosphofructokinase</i>
Phosphofructokinases	
	<i>E. Tryptophan synthase beta subunit-like fold</i>
Tryptophan-synthase beta chains	
	<i>F. Ribokinase-like</i>
pfkB carbohydrate kinases	
	<i>G. TIM barrel fold</i>
Pyruvate kinases	Alpha-isopropylmalate and homocitrate synthases
Tryptophan synthase alpha chains	U32 peptidases
IMP-dehydrogenases/GMP-reductases	Anaerobic coproporphyrinogen III oxidases
KDPG and KHG aldolases	
	<i>H. Methylase fold</i>
rRNA adenine dimethylases	O-Methyltransferases
	D-Aspartate O-methyltransferases
Unknown structures with unassigned folds	
Iron-containing alcohol dehydrogenases	
Homoserine dehydrogenases	
Eukaryotic molybdopterin oxidoreductases	
Cytochrome b5 proteins	
Ribosomal S9 proteins	
Cysteine synthase/cystathionine beta-synthases	
Molybdenum cofactor biosynthesis proteins	

^a Folds of known structures are according to the SCOP database.²¹ Multi-domain proteins that have multiple folds were classified according to the fold in which the set block appeared in.

^b A fold was assigned to a family with unknown structure when the family sequences were similar to another family that had a known structure, when the family blocks formed a distinct sub-set with blocks that had sequence(s) with known structure (see the TIM barrel section) or by the annotation of the blocks (methylases).

^c The structure of a glutamyl-tRNA reductase was predicted to have a Rossmann-like fold domain by theoretical modelling and the motif identified in this work was predicted to bind NADPH.²²

^d This family of potassium transport proteins might include two types of proteins (according to their SwissProt annotation). For one of these groups, TrkA proteins, there is experimental evidence for NAD(+) binding.²³

related structures are found in a number of protein families that bind nucleotide-containing ligands, such as NAD, FAD and FMN, but also phospho-sugars.¹⁸ This super-family contains a phosphate-binding motif conserved in structure and sequence.¹⁷ The blocks from this super-family in the examined set are of the phosphate-binding motif. However, the set includes blocks from proteins with known structures that do not have the Rossmann fold. Our analysis identifies the Rossmann fold phosphate-binding motif in other folds, where it also binds ligands other than phosphate.

Our Rossmann-type fold phosphate-binding motif CYRCA set includes blocks from folds known to be related to the Rossmann fold, such as NAD/NADP/FAD binding, flavodoxin and phosphofructokinase folds, but also blocks from TIM barrel, methylases, tryptophan-synthase β -subunit-like and ribokinase folds (Table 2). Methylases were shown by structure analysis to be similar to the Rossmann fold,¹⁹ and we now show this by sequence analysis alone. While the TIM barrel fold is different from the Rossmann fold, it also often contains a phosphate-binding site.²⁰ Ribokinases

Table 3. Known and predicted TIM barrel families

Families with known structures		Families with no known structures	
	Corresponding strands ^a	Well connected families	Weakly connected families
<ul style="list-style-type: none"> • Tryptophan-synthase alpha chains • Tryptophan-synthase alpha chains • IMP-dehydrogenases/GMP-reductases • IMP-dehydrogenases/GMP-reductases • FMN-dependent alpha-hydroxy acid dehydrogenases • FMN-dependent alpha-hydroxy acid dehydrogenases • Pyruvate kinases • Uroporphyrinogen decarboxylases • Delta-aminolevulinic acid dehydratases • KDPG and KHG aldolases 	<ul style="list-style-type: none"> 1-2-3-4-5-6-7-8 7-8 1-2-3-4-5-6- 3-4-5-6-7-8 1-2- 5-6-7-8 1-2-3-4- 7-8 1-2-3-4-5-6- 5-6-7-8 1-2-3-4- 1-2-3-4-5-6-7-8 3-4-5-6-7-8 1-2- 1-2-3-4-5-6-7-8 3-4-5-6-7-8 1-2- 	<ul style="list-style-type: none"> • U32 peptidases • Anaerobic coproporphyrinogen III oxidases • Alpha-isopropylmalate and homocitrate synthases • Dihydroorotate dehydrogenases • Uncharacterized UPF0019 family • Vitamin B12-independent methionine synthases • HYPB/HUPM hydrogenase isoenzymes formation proteins 	<ul style="list-style-type: none"> • Ribulose-phosphate 3-epimerases • AIR carboxylases • Uncharacterized UPF0034 family
mean RMSD- segments: 1.4±0.5 Å, whole structures: 3.9±0.6 Å ^b			

Each entry corresponds to one block. Three pairs of blocks each are from a single family, corresponding to separate conserved regions in the families. These pairs represent different orientations of protein structures from one family.

^a Corresponding β -barrel strands, according to the structural superposition (see Figure 4(b)), are shown above each other. The strands in the aligned blocks that were used for the structural superposition are in bold. Strands are numbered by the order of their appearance in the protein sequences, N' to C' ends.

^b Root-mean-square deviation (RMSD) between each CYRCA aligned segments and each whole structure (except for the KDPG and KHG aldolases, see the legend to Figure 3). The structures used to represent the families are those shown in Figure 4(b).

and tryptophan-synthase β -subunits also bind phosphate-containing ligands but methylases bind a sulfate-containing ligand (*S*-adenosyl-methionine).

The structure of the motif we found common to these structures is a planar (N') β -strand, loop, α -helix. Exceptions to this are the flavodoxin fold, in which the α -helix is missing, and the ribokinase fold, in which the β -strand is missing. All seem to bind ligands on the same plane as the motif, across from the loop region (Figure 3). Variability in ligand positions is probably due to the heterogeneity of the ligands and folds in which this motif is found (Table 2). These results can be confirmed for proteins with determined structures. Proteins with different folds are found significantly similar to each other by structure-to-structure comparisons (not shown). The validity of this set is reinforced by its distinction from the Walker-type phosphate-binding motif, p-loop, set (the second largest set with 31 blocks). The two structures bind ligands differently and have separate sequence motifs (not shown). PSI-BLAST and block searches among the sequences present in this set did not identify the similarity between families of different folds, except for the known relation between the Rossmann, Rossmann-like and FAD/NAD(P)-binding folds.²¹

Identifying TIM barrel proteins

TIM barrel is an ubiquitous fold appearing in many proteins, mainly in metabolic enzymes. This α/β fold has an eight stranded β -barrel core surrounded by α -helices. TIM barrel proteins catalyze diverse reactions and have extremely low sequence

similarity between members from different families and are thus hard to identify.^{20,24}

Three CYRCA sets include blocks from different families with known TIM barrel folds. Two sets seem to include only TIM barrel fold blocks and the third is the Rossmann ligand-binding motif discussed above (Table 2). In this last set, the TIM barrel fold blocks cluster together in a distinct sub-set (not shown). More blocks join these three sets when lower LAMA thresholds are used and at Z-score ≥ 5.0 all three sets are found together as a sub-set of the Rossmann-type ligand-binding motif set (Figure 4(a)).

The common sequence segments in the TIM barrel families detected by CYRCA are nine to 18 amino acid residues long, corresponding to ten regions of a β -strand, loop, α -helix (the strand being a barrel stave) in seven different families with known structures (three families each have two blocks in the set, corresponding to different regions of their structures). These regions fit well with each other and also perfectly superimpose the β -barrels (Figure 4(b), Table 3). The β -barrel strands in the regions are from different positions in the protein sequence. Thus, aligning the sequences according to the structural superposition requires circular permutations of the primary sequence (Table 3). Structural superimposition of β -barrels by randomly chosen strands showed that these results are not due to the radial symmetry of the barrels (not shown). Our finding defines the TIM barrels phosphate-binding pocket²⁰ as a Rossmann-type fold ligand binding motif (see above). Recent structural evidence for evolution of TIM barrels by a half-barrel duplication supports our finding of pairs of similar motifs within TIM barrel

families.²⁵ We demonstrate how a local sequence similarity identifies a global structural fold.

The structural similarity between the segments and full folds of the TIM barrel CYRCA sets could not be identified by comparing the sequences from each family with other sequences or with blocks. PSI-BLAST identified sequence similarity between only seven of the 17 families in Table 3 (not finding three of the families with known TIM barrel structures). Sequence to blocks searches identified sequence similarity between just two families. Thus, CYRCA analysis predicts local and global protein structures for ten families where no such experimental data are currently available (Table 3), while other advanced sequence search methods can give only very partial forecasts. Most of these families are well connected with Z-scores ≥ 5.6 , while three are connected with lower scores (Figure 4(a) and Table 3). The prediction for these last families is less certain. Nevertheless, we predict (with either high or low certainty levels) the presence of a TIM barrel fold in all members of each of these ten families.

ClpP protease and crotonase folds

CYRCA analysis clustered three blocks from the Enoyl-CoA hydratase/isomerase, ClpP protease, and U2 peptidase protein families. Enoyl-

CoA hydratases and isomerases are eubacterial and eukaryotic enzymes that function in fatty acid metabolism.²⁶ They form homotrimers or homo-hexamers composed of two trimer rings. This fold is termed the crotonase fold after one of its members. The active sites are in clefts on the outer surface of each monomer.^{27,28} ClpP is the catalytic subunit of the Clp ATP-dependent protease, found in eubacteria, and eukaryotes, where they are encoded in nuclear and chloroplast genomes and expressed in the chloroplast and mitochondria.²⁹ Clp proteolytic core is a ClpP homo-oligomer made up of two heptameric rings forming a central cavity where proteins are cleaved in a catalytic triad-type active site.³⁰ Less is known about the third family of eubacterial and archaeal U2 peptidases, apart from the activity of some members, and no structural information is available.

The sequence similarity between the blocks is across 14-21 amino acid residues. The corresponding segments from ClpP and enoyl-CoA hydratase proteins have the same structure (0.96 Å RMSD across 17 residues) and fitting them also superimposes the core fold of the two proteins (3.2 Å RMSD across 151 aligned residues). Nevertheless, the active site in each protein monomer is at a different position (Figure 5). Moreover, the common segments are positioned differently in the pro-

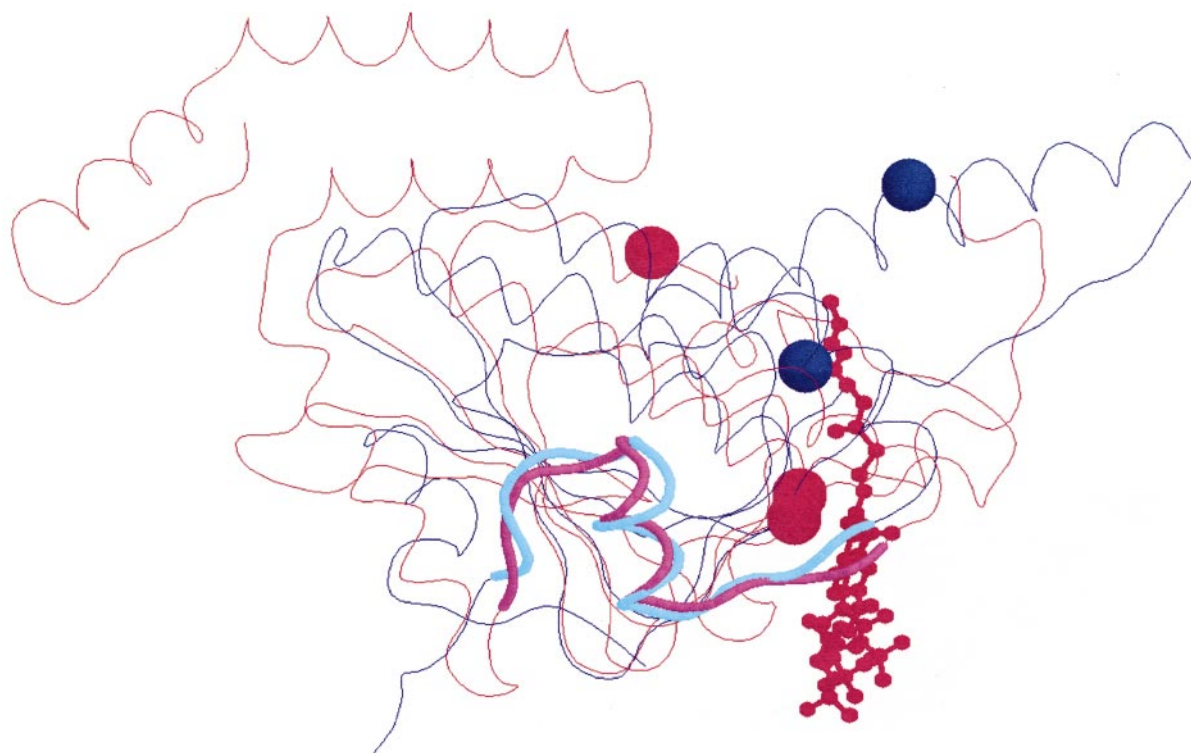


Figure 5. ClpP and crotonase folds. *Escherichia coli* ClpP (PDB structure 1TYF) is shown in blue and rat Enoyl-CoA hydratase (2DUB) is shown in red. Two monomers superimposed by the CYRCA identified common segments: 1TYF 172-188 in cyan and 2DUB 196-212 in magenta. The protein's α -carbon atoms are traced, active-site residues (1TYF 144, 164 and 2DUB 97, 122, 161) are shown as spheres and the Enoyl-CoA hydratase octanoyl-CoA ligand is shown in a "ball-and-stick" representation.

tein oligomers and in each oligomer the monomers interact with each other differently. In the enoyl-CoA hydratases, the segments are at the center of the rings, while in the ClpP proteins the segments are on the periphery of the rings. Superimposing monomers from the two structures by the segments, we found the rings perpendicular to each other (not shown).

These two protein families have the same monomer structural fold and similar sequence motifs but their function, position of the active sites and overall structure are totally different. While their monomer folds are the same, most probably originating from a common ancestor, they are utilized in dissimilar ways for different functions. This fold similarity was previously identified in comparing the structures³¹ but cannot be identified by comparing the sequences to other sequences or to blocks.

Discussion

This work extends the approach of multiple alignment-to-multiple alignment comparison, identifying similar sets of protein sequence motifs appearing in different contexts. Our method essentially creates multiple alignment of multiple-alignments. We believe this is a significant advance in sequence analysis, enhancing our ability to practically use protein motif information.

There are several advantages in our choice of protein motifs as the basic unit for analysis. First, the modular nature of proteins naturally lends itself to description by motifs; i.e. locally conserved sequence regions. While groups of motifs can describe specific families, each motif is not necessarily associated with a single family. A motif can appear in separate families and repeated within one family.^{32,33} Next, a large number of sequence motifs are available from several extensive databases.^{14,34}

The short local nature of blocks is also the basis for the success of the consistency approach we took. Since each block usually depicts a single feature (protein motif), it is easier to identify consistent sets. In contrast, when whole sequences or global multiple-alignments are compared the similarity usually consists of several ungapped segments. Not all these segments will necessarily be consistent, even in sets of genuinely related sequences or multiple alignments. This greatly complicates the identification of consistent relations described here.

CYRCA analysis can identify sequence relations beyond the range of advanced sequence-to-sequence, sequence-to-blocks and block-blocks analyses. We used very permissive criteria for identifying BLAST/PSI-BLAST and BLIMPS hits, and our basic parameters are those generally used in these methods (i.e. Grishin, 2000). Nevertheless, most of the CYRCA relations could not be identified by these methods. Realistically, even hits found by

these criteria would not serve as convincing evidence. Users would probably question results where more than half the sequences did *not* identify the target family. In contrast, the built-in consistency analysis of CYRCA allowed confidence even in sets that are not fully connected.

Our approach could be improved in two areas: multiple-alignment databases and multiple-alignment comparison method. Currently, only part of all known sequence families is present in multiple-alignment databases and not all motifs are identified in families that are present. Progress is continuously made in these two areas as more raw data become available and bioinformatics and motif-finding methods are refined. For example, when we performed our analysis, only 55% of all the sequences in the SwissProt database (44,163) were assigned to any block in the Blocks database. Of these sequences, 20% (8680) were in the 61 CYRCA sets we found. In the next release of the Blocks+ database (June 2000), 61% of the SwissProt sequences were assigned to blocks and CYRCA-identified sets now included 30% of those assigned sequences (14,485). We expect this trend to continue, leading to increased coverage. Next, we believe that unidentified genuine relations are present in the data we analyzed. The number of possibly related block pairs not found in CYRCA sets is about 1500 more than that predicted in a comparison of shuffled blocks (Table 1). We are working on improving the LAMA program to refine the discrimination between false and genuine block pair relations.

An important caveat regarding enhanced sensitivity is demonstrated in our results. As shown in the relation of the ClpP and crotonase folds, homologous sequences, and even proteins with the same structure, may have different functions.³¹ A principal aim of sequence analysis is function determination. Structure identification is often a means to infer function and mechanism rather than an end in itself. Thus, correctly predicting the fold for a sequence will be misleading if its function is different from that assigned to this fold. Such cases were previously recognized after protein structure determination^{31,36} when the function of the studied proteins was known. As sequence analysis methods become more sensitive and most analyzed sequences have no functional annotations, special care must be taken to avoid wrong functional assignments derived from correct structural predictions.

This work identifies many block sets whose similarity could be proven by structures or be supported by description for some blocks. Function and/or structure is then predicted for the other blocks in the set. Furthermore, the information predicted for each block is bequeathed to all its constituent sequences. This approach can be used to automatically annotate databases of sequences and of multiple-alignments. Blocks WWW sites (<http://blocks.fhcrc.org>) now include links to CYRCA sets, allowing simple exploration and

cross-referencing of related blocks. LAMA searches of the Blocks database are followed by a CYRCA analysis and links are provided to blocks sets to which the queries can be joined.

One remarkable aspect of our results is the identification of global structure correspondence from local sequence similarity. Sequence similarity between sequence regions as short as eight amino acid residues predicted not only the region's structure similarity but also the similarity of the folds in which they are found (Figures 3(b) and 5). Protein sequence motifs can thus serve as markers for structure folds, identifying the structure similarity of regions whose sequences diverged beyond recognition.

Methods and Data Sources

Sequence and structure comparison methods

To evaluate CYRCA, we used other sensitive sequence analysis methods. Each CYRCA set includes three or more blocks, and each block includes motifs from different sequences. We examined whether these sequences are found similar to sequences in other blocks using BLAST 2.0 and PSI-BLAST 2.0 sequence-to-sequence comparison programs⁴ or to other blocks using the BLIMPS and BLKPROB sequence-to-multiple-alignment programs.^{37,38} PSI-BLAST implicitly uses multiple-alignment-to-sequence comparison after its first round of comparison and can thus also be viewed as a multiple-alignment-to-sequence comparison method.

Sequences were compared to SwissProt database release 38 with the BLAST programs and to the Blocks+ database of January 2000 with the BLIMPS program. *E*-value thresholds for all programs were 0.001. PSI-BLAST was executed with three iterations. Given a CYRCA set, we examined each sequence in all blocks. Outputs were checked for the presence of sequences or blocks participating in the examined set. An identification required a detection of a block (for BLIMPS) or any of its sequences (for BLAST and PSI-BLAST) by at least a quarter of the sequences of the query block. Filtering for low-complexity sequence regions of search queries or search queries and database targets did not improve PSI-BLAST results.

Some blocks include sequences with determined structure in their motif regions. These could be used in structure-to-structure comparisons. These were done with the CE program³⁹ for identifying similar region/s across long regions in proteins and with the PROFIT 1.8 program for superimposing two specified protein regions (<http://www.biochem.ucl.ac.uk/~martin/programs>).

Data sources

The Blocks+ database (<http://blocks.fhcrc.org>) of January 2000 was used in the comparison. Protein structures were retrieved from the RCSB protein database (<http://www.rcsb.org/pdb>).

Acknowledgements

We thank Steven and Jorja Henikoff for critically reading the manuscript and providing helpful suggestions. This research was supported by The Israel Science Foundation, founded by The Israel Academy of Sciences and Humanities. This work was supported, in part, by a grant from the NIH to Steven Henikoff (GM29009).

References

- Gribskov, M., McLachlan, A. D. & Eisenberg, D. (1987). Profile analysis: detection of distantly related proteins. *Proc. Natl Acad. Sci. USA*, **84**, 4355-4358.
- Tatusov, R. L., Altschul, S. F. & Koonin, E. V. (1994). Detection of conserved segments in proteins: iterative scanning of sequence databases with alignment blocks. *Proc. Natl Acad. Sci. USA*, **91**, 12091-12095.
- Henikoff, S. & Henikoff, J. G. (1997). Embedding strategies for effective use of information from multiple sequence alignments. *Protein Sci.* **6**, 698-705.
- Altschul, S. F., Madden, T. L., Schäffer, A. A., Zhang, J., Zhang, Z., Miller, W. & Lipman, D. J. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucl. Acids Res.* **25**, 3389-3402.
- Park, J., Karplus, K., Barrett, C., Hughey, R., Haussler, D., Hubbard, T. & Chothia, C. (1998). Sequence comparisons using multiple sequences detect three times as many remote homologues as pairwise methods. *J. Mol. Biol.* **284**, 1201-1210.
- Henikoff, S., Pietrokovski, S. & Henikoff, J. G. (1998). Superior performance in protein homology detection with the Blocks Database servers. *Nucl. Acids Res.* **26**, 309-312.
- Pietrokovski, S. (1996). Searching databases of conserved sequence regions by aligning protein multiple-alignments. *Nucl. Acids Res.* **24**, 3836-3845.
- Rychlewski, L., Jaroszewski, L., Li, W. & Godzik, A. (2000). Comparison of sequence profiles. Strategies for structural predictions using sequence information. *Protein Sci.* **9**, 232-241.
- Altschul, S. F. & Lipman, D. J. (1990). Protein database searches for multiple alignments. *Proc. Natl Acad. Sci. USA*, **87**, 5509-5513.
- Pietrokovski, S. & Henikoff, S. (1997). A helix-turn-helix DNA-binding motif predicted for transposases of DNA transposons. *Mol. Gen. Genet.* **254**, 689-695.
- Herbert, A., Alfken, J., Kim, Y. G., Mian, I. S., Nishikura, K. & Rich, A. (1997). A Z-DNA binding domain present in the human editing enzyme, double-stranded RNA adenosine deaminase. *Proc. Natl Acad. Sci. USA*, **94**, 8421-8426.
- van Pouderooyen, G., Ketting, R. F., Perrakis, A., Plasterk, R. H. A. & Sixma, T. K. (1997). Crystal structure of the specific DNA-binding domain of Tc3 transposase of *C. elegans* in complex with transposon DNA. *EMBO J.* **16**, 6044-6054.
- Schwartz, T., Rould, M. A., Lowenhaupt, K., Herbert, A. & Rich, A. (1999). Crystal structure of the Zalpha domain of the human editing enzyme ADAR1 bound to left-handed Z-DNA. *Science*, **284**, 1841-1845.

14. Henikoff, S., Henikoff, J. G. & Pietrokovski, S. (1999). Blocks + : a non-redundant database of protein alignment blocks derived from multiple compilations. *Bioinformatics*, **15**, 471-479.
15. Rost, B. (1996). PHD: predicting one-dimensional protein structure by profile-based neural networks. *Methods Enzymol.* **266**, 525-539.
16. Rao, S. T. & Rossmann, M. G. (1973). Comparison of super-secondary structures in proteins. *J. Mol. Biol.* **76**, 241-256.
17. Bellamacina, C. R. (1996). The nicotinamide dinucleotide binding motif: a comparison of nucleotide binding proteins. *FASEB J.* **10**, 1257-1269.
18. Sigrell, J. A., Cameron, A. D., Jones, T. A. & Mowbray, S. L. (1997). Purification, characterization, and crystallization of *Escherichia coli* ribokinase. *Protein Sci.* **6**, 2474-2476.
19. Djordjevic, S. & Stock, A. M. (1997). Crystal structure of the chemotaxis receptor methyltransferase CheR suggests a conserved structural motif for binding S-adenosylmethionine. *Structure*, **5**, 545-558.
20. Bork, P., Gellerich, J., Groth, H., Hooft, R. & Martin, F. (1995). Divergent evolution of a beta/alpha-barrel subclass: detection of numerous phosphate-binding sites by motif search. *Protein Sci.* **4**, 268-274.
21. Lo Conte, L., Ailey, B., Hubbard, T., Brenner, S., Murzin, A. & Chothia, C. (2000). SCOP: a structural classification of proteins database. *Nucl. Acids Res.* **28**, 257-259.
22. Brody, S. S., Gough, S. P. & Kannangara, C. G. (1999). Predicted structure and fold recognition for the glutamyl tRNA reductase family of proteins. *Proteins: Struct. Funct. Genet.* **37**, 485-493.
23. Schlosser, A., Hamann, A., Bossemeyer, D., Schneider, E. & Bakker, E. P. (1993). NAD⁺ binding to the *Escherichia coli* K(+)-uptake protein TrkA and sequence similarity between TrkA and domains of a family of dehydrogenases suggest a role for NAD⁺ in bacterial transport. *Mol. Microbiol.* **9**, 533-543.
24. Reardon, D. & Farber, G. K. (1995). The structure and evolution of alpha/beta barrel proteins. *FASEB J.* **9**, 497-503.
25. Lang, D., Thoma, R., Henn-Sax, M., Sterner, R. & Wilmanns, M. (2000). Structural evidence for evolution of the beta/alpha barrel scaffold by gene duplication and fusion. *Science*, **289**, 1546-1550.
26. Babbitt, P. C., Kenyon, G. L., Martin, B. M., Charest, H., Slyvestre, M., Scholten, J. D., Chang, K. H., Liang, P. H. & Dunaway-Mariano, D. (1992). Ancestry of the 4-chlorobenzoate dehalogenase: analysis of amino acid sequence identities among families of acyl:adenyl ligases, enoyl-CoA hydratases/isomerases, and acyl-CoA thioesterases. *Biochemistry*, **31**, 5594-5604.
27. Benning, M. M., Taylor, K. L., Liu, R. Q., Yang, G., Xiang, H., Wesenberg, G., Dunaway-Mariano, D. & Holden, H. M. (1996). Structure of 4-chlorobenzoyl coenzyme A dehalogenase determined to 1.8 Å resolution: an enzyme catalyst generated *via* adaptive mutation. *Biochemistry*, **35**, 8103-8109.
28. Engel, C. K., Mathieu, M., Zeelen, J. P., Hiltunen, J. K. & Wierenga, R. K. (1996). Crystal structure of enoyl-coenzyme A (CoA) hydratase at 2.5 angstroms resolution: a spiral fold defines the CoA-binding pocket. *EMBO J.* **15**, 5135-5145.
29. Porankiewicz, J., Wang, J. & Clarke, A. K. (1999). New insights into the ATP-dependent Clp protease: *Escherichia coli* and beyond. *Mol. Microbiol.* **32**, 449-458.
30. Wang, J., Hartling, J. A. & Flanagan, J. M. (1997). The structure of ClpP at 2.3 Å resolution suggests a model for ATP-dependent proteolysis. *Cell*, **91**, 447-456.
31. Murzin, A. G. (1998). How far divergent evolution goes in proteins. *Curr. Opin. Struct. Biol.* **8**, 380-387.
32. Henikoff, S. & Henikoff, J. G. (1994). Protein family classification based on searching a database of blocks. *Genomics*, **19**, 97-107.
33. Henikoff, S., Greene, E. A., Pietrokovski, S., Bork, P., Attwood, T. K. & Hood, L. (1997). Gene families: the taxonomy of protein paralogs and chimeras. *Science*, **278**, 609-614.
34. Apweiler, R., Attwood, T. K., Bairoch, A., Bateman, A., Birney, E., Biswas, M., Bucher, P., Cerutti, L., Corpet, F., Croning, M. D., Durbin, R., Falquet, L., Fleischmann, W., Gouzy, J., Hermjakob, H. & Hulo, N., *et al.* (2001). The InterPro database, an integrated documentation resource for protein families, domains and functional sites. *Nucl. Acids Res.* **29**, 37-40.
35. Grishin, N. V. (2000). C-terminal domains of *Escherichia coli* topoisomerase I belong to the zinc-ribbon superfamily. *J. Mol. Biol.* **299**, 1165-1177.
36. Piatigorsky, J. & Wistow, G. (1991). The recruitment of crystallins: new functions precede gene duplication. *Science*, **252**, 1078-1079.
37. Henikoff, S., Henikoff, J. G., Alford, W. J. & Pietrokovski, S. (1995). Automated construction and graphical presentation of protein blocks from unaligned sequences. *Gene*, **163**, 17-26.
38. Henikoff, J. G., Greene, E. A., Pietrokovski, S. & Henikoff, S. (2000). Increased coverage of protein families with the blocks database servers. *Nucl. Acids Res.* **28**, 228-230.
39. Shindyalov, I. N. & Bourne, P. E. (1998). Protein structure alignment by incremental combinatorial extension (CE) of the optimal path. *Protein Eng.* **11**, 739-747.

Edited by B. Holland

(Received 6 October 2000; received in revised form 11 January 2001; accepted 12 January 2001)