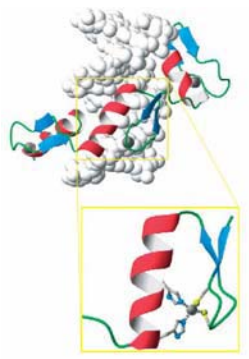
Structural Motif Finding

- Automated discovery of 3D motifs for protein function annotation (Polacco and Babbit, Bioinformatics, January 2006)
- Overall comparison of protein structures may not identify similarities among functionally significant amino acids or atoms involved in a protein function’s mechanism.
  - e.g., amino acids on the surface of the protein vs. buried amino acids

What are motifs?

- Wikipedia: a **structural motif** is a three-dimensional structural element or fold within a protein, which appears also in a variety of other proteins.
- In proteins, structure motifs usually consist of just a few elements, e.g. the 'helix-turn-helix' has just three.

Motif examples

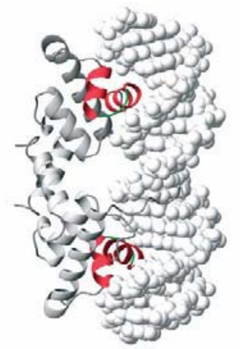


Zinc finger motif

Two beta strands with an alpha helix end folded over to bind a zinc ion. This motif is seen in transcription factors.

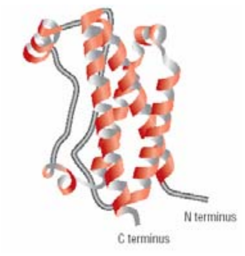
A fragment derived from a mouse gene regulatory protein is shown, with three zinc fingers bound spirally in the major groove of a DNA molecule. The inset shows the coordination of a zinc atom by characteristically spaced cysteine and histidine residues in a single zinc finger motif.
The image is of PDB: 1aay

Motif examples



Helix-turn-helix The DNA-binding domain of the bacterial gene regulatory protein lambda repressor, with the two helix-turn-helix motifs shown in color. The two helices closest to the DNA are the reading or recognition helices, which bind in the major groove and recognize specific gene regulatory sequences in the DNA. (PDB 1lmb)

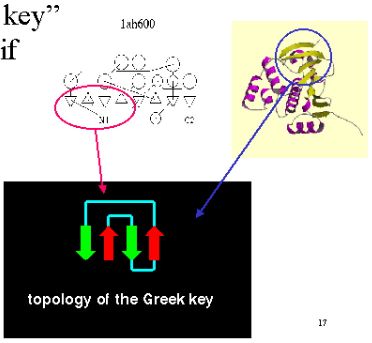
Motif examples



Four-helix bundle motif
The four-helix bundle motif can comprise an entire protein domain, and occurs in proteins with many different biochemical functions. Shown here is human growth hormone, a signaling molecule.

Motif examples

“Greek key” motif



Motifs

- The term motif is used in two different ways in structural biology. The first refers to a particular amino-acid sequence that is characteristic of a specific biochemical function.
  - example: CXX(XX)CXXXXXXXXXXXXHXXXH
- Sequence motifs can be recognized by inspecting the amino-acid sequence. Databases of such motifs exist:
  - e.g., PROSITE (http://www.expasy.ch/prosite/)

Motifs

- The second use of the term motif refers to a set of contiguous secondary structure elements that have a particular functional significance.
  - e.g. helix-turn-helix, Greek-key motif
- Usually, sequence motifs are more indicative of certain function, because a shared structural motif does not always imply similar function. However, detecting functional motifs from sequence alone is difficult due to variable spacing, different ordering of functional residues.

Motif databases

- Which configuration is functionally important?
  - Determined by experts; therefore, accumulation of known motifs is slow
- The catalytic site atlas (CSA) provides 147 non-redundant active site motifs for enzymes.
- Automated methods of identifying repeated patterns in protein structures generates more motifs (450 non-redundant ligand-biding patterns in PINTS database). However, they may not provide specific functional information.

GASPS

- Polacco and Babbit, 2006
- Genetic Algorithm Search for Patterns in Structures
- GASPS goals:
  - for a group of proteins GASPS should find the motif most useful for identifying the group
  - GASPS should rely on known functional residues as little as possible

GASPS

- In this study a motif is a small set of residues (<10) taken from a single chain
- Each residue is modeled with two points: Carbon-alpha and the side-chain geometrical center.
- They use a previous method named SPASM (Kleywegt, 1999) to find matches.
- They first find exact sequence matches and then compute an RMSD between the matches.

GASPS

- Only the match with the best RMSD is considered from each structure.
- They identify motifs from a set of related proteins with this method and then they test the discovered motif’s discriminative power by conducting leave-one-out cross validation experiments.
  - Five groups of proteins of different functions. A motif is extracted from each group. Then, try to predict the function of a new protein.

Another technique

- The fragment transformation method to detect the protein structural motifs (Lu et al., Proteins, 2006)
- Motivation: motifs with conserved residues and constrained geometry are easy to detect. However, it is challenging to detect general structural motifs like ββα-metal binding motif, which have variable conformation and sequence. Such motifs are currently identified by manual procedures using sequence and structure analysis.

Overview

- A structural alignment algorithm that combines both structural and sequence information to identify local structural motifs.
- The methods is tested to detect *ββα-metal binding* motif and the *treble clef* motif.

ββα-metal binding motif

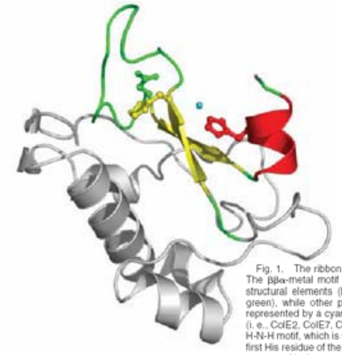


Fig. 1. The ribbon representation of colicin ColE7 (PDB code: 7CEI). The ββα-metal motif is rendered in colors according to the secondary structural elements (helices in red, β-strands in yellow, and loops in green), while other parts of the protein in gray. The Zn metal ion is represented by a cyan sphere. The ββα-metal motif of the colicin family (i. e., ColE2, ColE7, ColE8, and ColE9) contains the so-called conserved H-N-H motif, which is shown in a ball-and-stick model. However, only the first His residue of the sequence motif (i.e., on the first β-strand) is strictly conserved in the general ββα-metal motif of other proteins. All molecular images in this work are created by PyMOL (http://www.pymol.org).

The method

- Let *S* be the structural motif of length *n*, and *T* be the target protein of length *m*. *S* and *T* are represented by their C_α coordinates, ($\mathbf{x}_1$, $\mathbf{x}_2$,..., $\mathbf{x}_n$) and ($\mathbf{y}_1$, $\mathbf{y}_2$,..., $\mathbf{y}_m$) respectively.
- The basic unit of the structural alignment is a triplet composed of three consecutive C_α atoms. The structures *S* and *T* can be expressed in terms of these triplets, that is, $S=\{\sigma_1, \sigma_2,..., \sigma_{n\text{-}2}\}$ and $T=\{\tau_1, \tau_2,..., \tau_{m\text{-}2}\}$

  $\sigma_i=(\mathbf{x}_i,\mathbf{x}_{i+1},\mathbf{x}_{i+2})$ and $\tau_i=(\mathbf{y}_i,\mathbf{y}_{i+1},\mathbf{y}_{i+2})$

The method

- An (*m*-2)x(*n*-2) matrix between σ and τ triplets can be constructed,

$$\mathbf{M} = \left| \begin{array}{cccc} M_{11} & M_{12} & \dots & M_{1,n-2} \\ M_{21} & M_{22} & \dots & M_{2,n-2} \\ \dots & \dots & \dots & \dots \\ M_{m-2,1} & M_{m-2,2} & \dots & M_{m-2,n-2} \end{array} \right|$$

where the element $M_{i,j}$ is a rigid body transformation matrix from σ_i to τ_j, that is, $M_{i,j}\,\sigma_i = \tau_j$

Example transformation

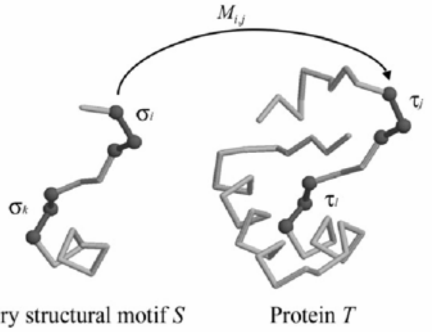


Fig. 2. σ_i and σ_k are two arbitrary triplet units of the query structural motif *S*, and τ_j and τ_l are two arbitrary triplet units of the target protein *T*. The triplet σ_i is transformed to τ_j through the transformation matrix $M_{i,j}$.

Triplet clustering

- Define the distance between two matched pairs of triplets as the Cartesian distance between one pair of triplets when the others' transformation is applied to them.
- Formally:
  - D^{ij}_{kl} = cartesian_distance($M_{i,j}\sigma_k$, τ_l)
  - The smaller the distance the more similar are the transformations $M_{i,j}$ and $M_{k,l}$
  - In other words, the orientation between the triplet pairs (σ_i ,τ_j) and (σ_k ,τ_l) is more similar.

Triplet clustering

- They cluster the triplets with a very simple algorithm called single-linkage algorithm.
- For two triplet pairs (σ_i ,τ_j) and (σ_k ,τ_l), where i≠k and j≠l, if the distance D^{ij}_{kl} is smaller than D_0 (a constant threshold, 3 A° in their experiments) then put them in the same cluster.
- If any two elements in two different clusters satisfy the same criteria, then merge the clusters.

Triplet clustering

- After clustering, each cluster represents a structural alignment between the source motif *S* and the target protein *T*.
- Each cluster is given a score based on some scoring function, which is a combination of 3 ranked scores: RMSD distance, sequence alignment score, and secondary structure alignment score.

Results: *ββα-metal binding* motif

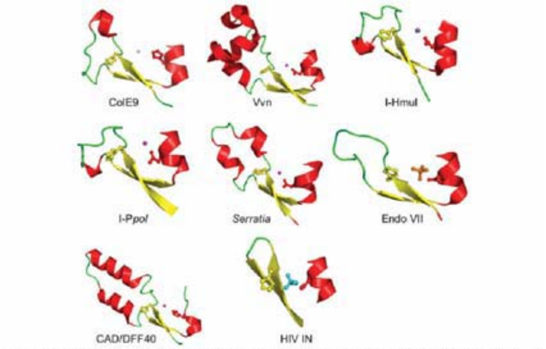


Fig. 3. The identified ββα-metal motifs (using ColE9 (1EMV:B) as the reference) in Vvn (1OUO:A), I-HmuI (1U3E:M), I-*PpoI* (1EVW), *Serratia* nuclease (1QL0), the T4 Endo VII (1E7L:A), CAD/DFF40 (1V0D), and HIV IN (1EX4:A). The conserved His residue, the ligand and the binding residue (His, Asn or Asp), which is the underlined residue in Table II, are shown in the ball-and-stick model.

Results: *treble clef finger* motif

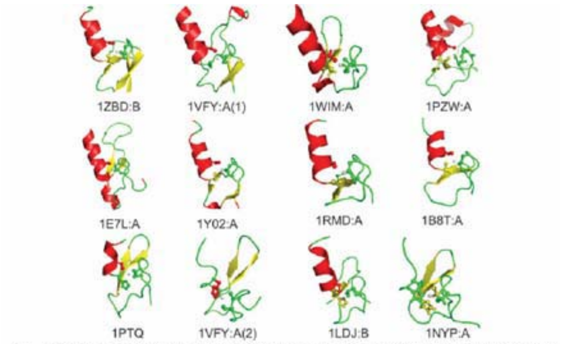


Fig. 4. The identified treble clef finger motifs [using G protein Rab3A (1ZBD:B) as the reference in Vps27p (1VFY:A(1)], UbcM4-interacting protein 4 (1WIM:A), transcription factor Grauzone (1PZW:B), the T4 Endo VII (1E7L:A), CARP2 (1Y02:A), Hrs (1DVP:A), RAG1 dimerization domain (1RMD), CRP1 (1B8T:A), Protein kinase C (1PTQ:A), Vps27p (1VFY:A(2), ring-box protein 1 (1LDJ:B), and LIM4 (1NYP:A).