

# Comparison and Classification of Protein Structures

Akira R. KINJO  
金城 玲

Institute for Protein Research, Osaka University  
&  
Protein Data Bank Japan

# Protein Data Bank (PDB)

## Worldwide Protein Data Bank (wwPDB)

Welcome to the Worldwide Protein Data Bank (wwPDB)!

The Worldwide Protein Data Bank (wwPDB) consists of organizations that act as deposition, data processing and distribution centers for PDB data.<sup>1</sup> Members are: RCSB PDB (USA), PDBe (Europe) and PDBj (Japan), and BMRB (USA). The wwPDB's mission is to maintain a single PDB archive of macromolecular structural data that is freely and publicly available to the global community.

13-May-2014  
PDB Reaches a New Milestone: 100,000+ Entries

In the weeks leading up to this historic event, wwPDB has looked back at other PDB milestones. (Previously: Building a Community Resource, The Early Structures, Launching Tools for the Next Generation)

I DEPOSITED  
at the 100,000+  
PDB Structures

Depositors: Download this image and write the number of structures deposited.

With this week's update, the PDB archive contains a record 100,147 entries.

Established in 1971, this central, public archive has reached this critical milestone thanks to the efforts of structural biologists throughout the world who contribute their experimentally-determined protein and nucleic acid structure data.

Four wwPDB data centers support online access to three-dimensional structures of biological macromolecules that help researchers understand many facets of biomedicine, agriculture and environmental systems to health and disease to biological energy. The archive is quite large, containing more than 1,000,000 files related to these PDB entries that require more than 249 GBbytes of storage.

more

FULL NEWS

Questions? Info@wwpdb.org  
K. H.M. Bernas, K. Henrick, H. Nakamura (2001) Announcing the worldwide Protein Data Bank. *Nature Structural Biology* 10 (12), p. 990.

wwPDB PARTNER ORGANIZATIONS

PDB PROTEIN DATA BANK PDBe PROTEIN DATA BANK Europe PDBj PROTEIN DATA BANK Japan BMRB BIOLOGICAL MAGNETIC RESONANCE BANK

## Protein Data Bank Japan (PDBj)

www.pdbj.org

100693  
最終更新日: 2014-06-04  
00:00 UTC / 09:00 JST

PDBj  
Protein Data Bank Japan

English 日本語 繁體中文 简體中文 韩国어

ABOUT RCSB PDB BMRB PDBe PDBej LEGACY

ホーム トップページ 相談相談 ヘルプ FAQ お問い合わせ リンク集 PDBアーカイブ

データ登録 ヘルプ ADIT: PDBへの登録 ADIT-NMR データ登録について

新フォーマット PDBx/MLCPTについて

検索 ヘルプ PDB検索 (PDBj) Mine! PDB検索ツール 大阪医工エントリー BMRB検索 Sequence-Navigator Structure-Navigator EM Navigator PDBe Navigator SeSAW Legend Binding Sites (LBMS) 酶の活性部位ナビゲーター 未知のエンタリーのステータス

サービスを探す サービスを表示する リセット

サービスを表示する リセット

最新情報 ニュース (2014年5月20日)  
今年で100万件を記念する企画「オンラインフォラムストリーミング」が、6月17日(火)~20日(金)、興味ある方面にて、開催されます。PDBjから各領域の専門家として参加し、講演を行います。  
ニュース (2014年5月16日)  
PDBjのセンターへ新たに10万台を実装しました  
ニュース (2014年5月16日)  
次世代ゲノムの標準規範と10万件への進歩  
ニュース (2014年5月2日)  
PDBj平均の蛋白構造と10万件への進歩  
ニュース (2014年4月26日)  
2014年6月1日(土)~8日(水)に、大阪大学いちょう前に施設・研究紹介を行います。タンパク質に興味のある方などなたでも歓迎です。  
ニュース (2014年4月26日)  
SP170 (土) ~ 28日 (火)、德国洪堡島で開催される第40回 Asia Pacific Protein Association (APPA) Conferenceにて、講演およびポスター発表を行います。  
ニュース (2014年4月26日)  
PDBjの標準と10万件への進歩  
ニュース (2014年4月17日)  
PDBjで提供したデータがオンライン事典に掲載されました。  
ニュース (2014年4月17日)

パートナー

OBCLS Database Center for Life Science NBDC National Bio-Information Center

The primary database of biological macromolecular structures

# An example of PDB entries

1goF - 検索 - PDBj Mine

100693

2014年03月25日 (2014-03-25)  
00:00 UTC / 09:00 JST

PDBj  
Protein Data Bank Japan

English 日本語 繁体中文 简體中文 한글

search.pdbj.org

watDB RCSB\_PDB BMRB PDBx Lesacie

ホーム 検索情報 実験情報 増強情報 ダウンロード

**1GOF**

NOVEL THIOETHER BOND REVEALED BY A 1.7 ANGSTROMS CRYSTAL STRUCTURE OF GALACTOSE OXIDASE

**1GOFの概要**

分子名稱 GALACTOSE OXIDASE (E.C.1.1.3.9) (PH-4.5)

構造のキーワード OXIDOREDUCTASE(OXYGEN(A))

出来する生物種 Hypomyces rosellus

ポリマー鎖の合計数 1

分子量の合計数 69785.82

著者 Ito, N., Phillips, S.E., Stevens, G., Ogle, Z.B., Mifflin, M.J., Keen, J.N., Yadav, K.D., Knowles, P.R.

引用文献 Novel thioether bond revealed by a 1.7 Å crystal structure of galactose oxidase.  
*Nature*, 350:87-90, 1991  
DOI: 10.1038/350087a0

X-RAY DIFFRACTION (1.7 Å)

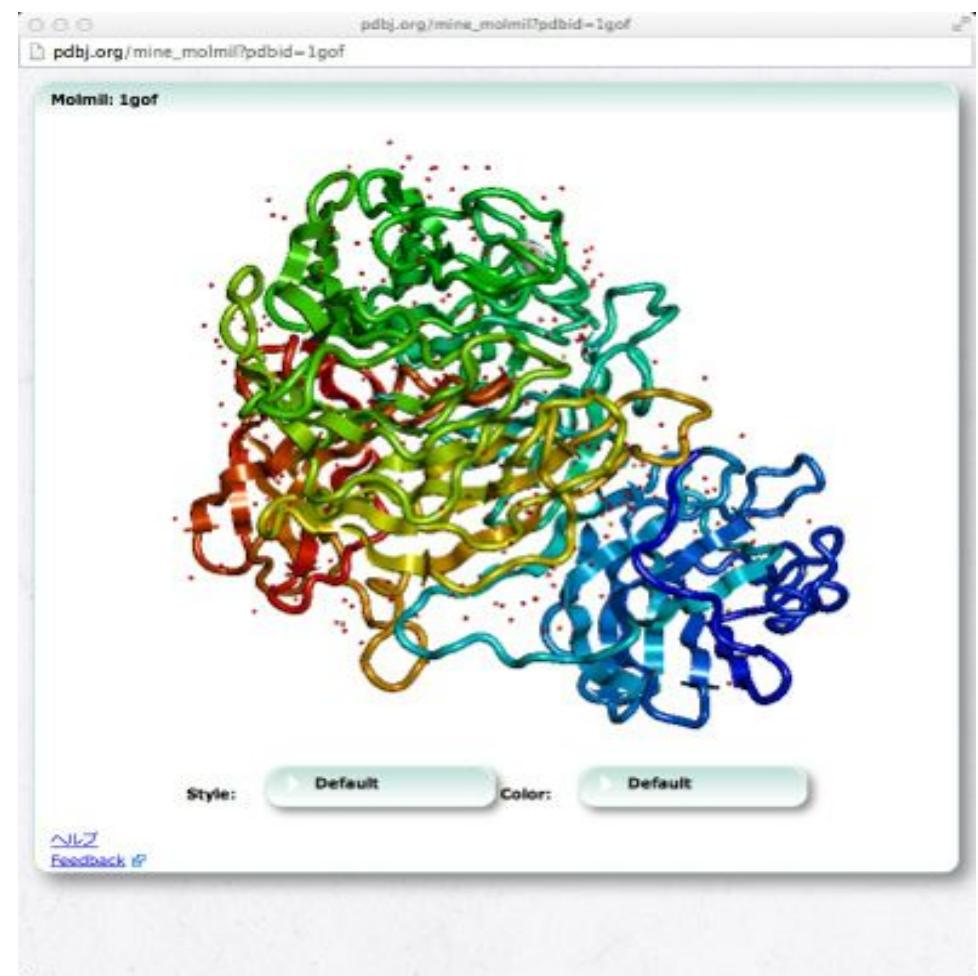
他のデータベース連携

RCSB\_PDB  
PDBx  
Yarobium  
CATH  
Pfam  
SCOP  
VAST  
dPPI  
PISA  
wwPDB/RDF  
電子密度マップ (EDM)  
Uniprot  
Q01745  
KEGG  
1.1.3.9  
Pfam  
PF00754  
PF01344  
PF09118  
ExPacDB (T00005)  
PDB

Copyright © 2013-2014 日本蛋白質資源データバンク

サービス&ソフトウェア

ヘルプ  
[V] 3次元表示ピュア  
万葉 (Yarobium)

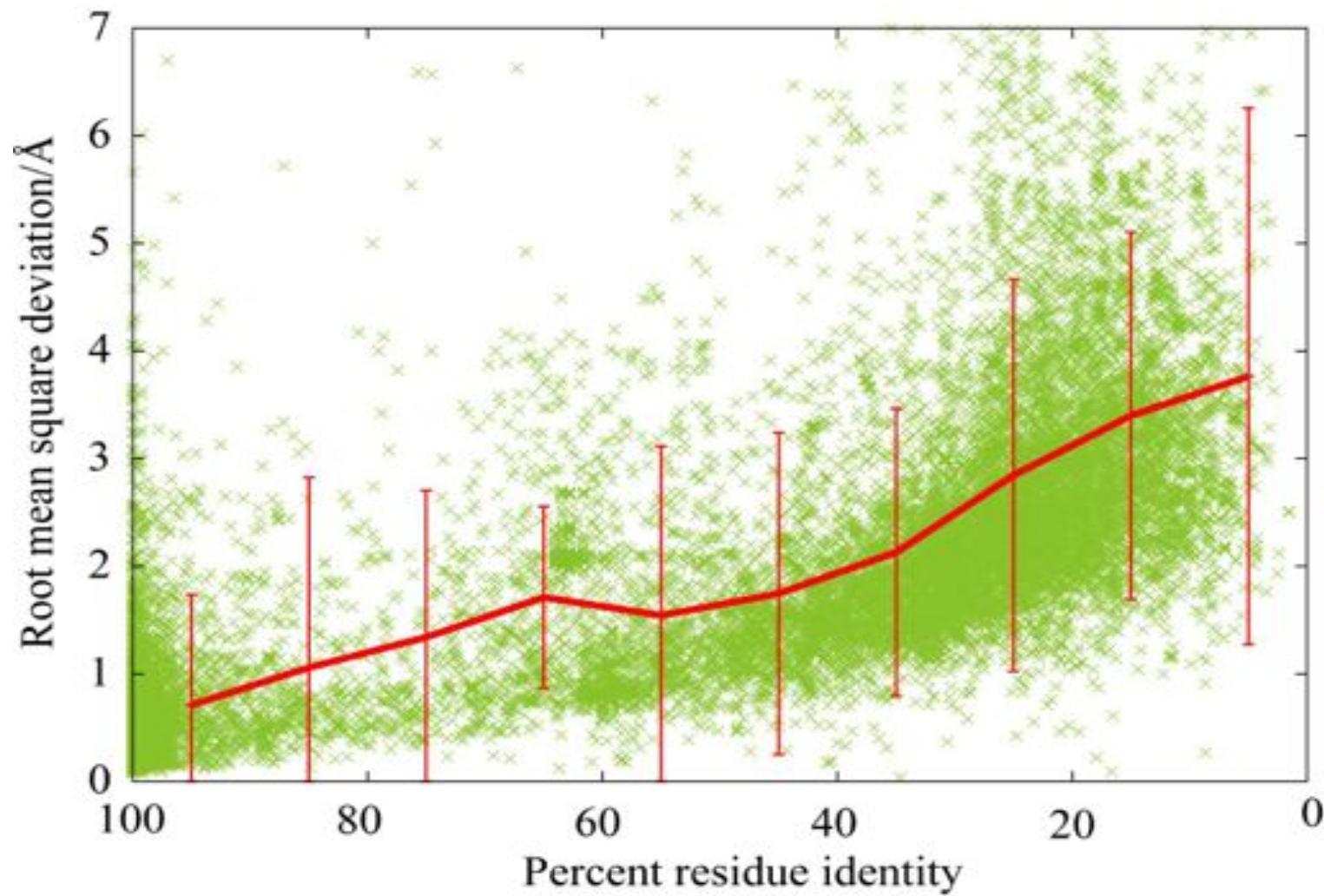


# Protein Structure Comparison

```
[BEGIN ALIGNMENT]
:
E1 H1 E2 H2 E3
SecA : EEEEEEE S HHHHHHHHHHHHHHH S SS EEEEEEE GGGHHHHHHH TT EEEE
3 : RTFFVGGNFKLNGSKQSIKEIVERLNTASIPENVEVICPPATYLDYSVLKKPQVTVG: 62
* * * * * * * * * * * * * * * * * * * * * * * * * * * * *
4 : RKFFVGGNWKMNGDKKSLGELIHTLNGAKLSADTEVVCGAPSIYLDFAKQL-DAKIGVA: 62
SecB : EEEEEEE S HHHHHHHHHHHHHHH S TT EEEEEEE GGGHHHHHHH S - TTS EEE
:
E1 H1 E2 H2 - E3
:
H3 E4 H4 H5 E
SecA : ES SSSSSS TT HHHHHHTT EEEES HHHHHHS HHHHHHHHHHHHHHHHTT E
63 : AQNAYLKASGAFTGENSDQIKDVGAKWILGHSERFSYFHEDDKFIADKTFALGQGVG: 122
*** * * * * * * * * * * * * * * * * * * * * * * * * * * * * *
63 : AQNCYKVPKGAFTEISPAMIKDIGAAWVILGNPERRHVFGESDELIGQKVAHALAEGLG: 122
SecB : ES SSSSSS TT HHHHHHTT EEEES HHHHHHTS HHHHHHHHHHHHHHHHTT E
:
H3 E4 H4 H5 E
:
5 H6 H7 E6 H8
SecA : EEEEE HHHHHHTT HHHHHHHHHHHHHHH S TT EEEEE GGGTTTS HHHHH
123 : VILCIGETLEEKKAGKTLDVVERQLNAVLEEVKDWNTNVVAYEPVWAIGTGLAATPEDAQ: 182
** * * * * * * * * * * * * * * * * * * * * * * * * * * * * * *
123 : VIACIGEKLDEREAGITEKVVFEOTKAIADNVKDWSKVVLAYEPVWAIGTGKTATPQQAQ: 182
SecB : EEEEE HHHHHHTT HHHHHHHHHHHHHHHHTT S GGEEEEEE GGGTTTS HHHHH
:
5 H6 H7 E6 H8
:
H9 E7 E8 H10
SecA : HHHHHHHHHHHHHHH HHHHHH EEEESS TTTGGGGTT TT EEEESGGGGSTTHHH
183 : DIHASIRKFLASKLGDKAASELRILYGGSANGSNAVTFKDKADVDGFLVGGASLKPEFVD: 242
* * * * * * * * * * * * * * * * * * * * * * * * * * * * * * *
183 : EVHEKLRWLKTHVSADAQSTRIIYGGSVTGGNCNELASQHDVDGFLVGGASLKPEFVD: 242
SecB : HHHHHHHHHHHHHHT HHHHHH EEEESS TTT HHHHHHTSTT EEEESGGGGSTTHHH
:
H9 E7 H10 E8 H11
:
SecA : HHHHTT
243 : IINSRN: 248
*** 
243 : IINAKH: 248
SecB : HHTTT
:
```



# Sequence & structure similarities

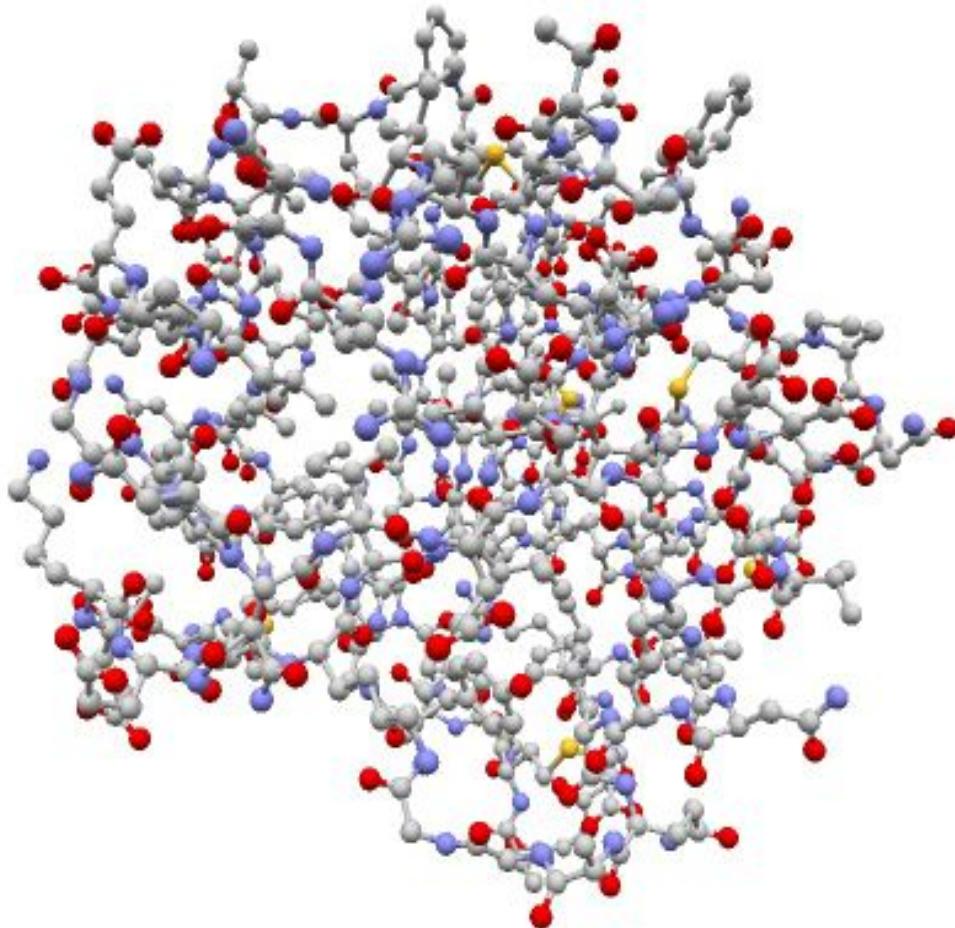


(『タンパク質の立体構造入門』図 4.1 )

# Methods of structure comparison

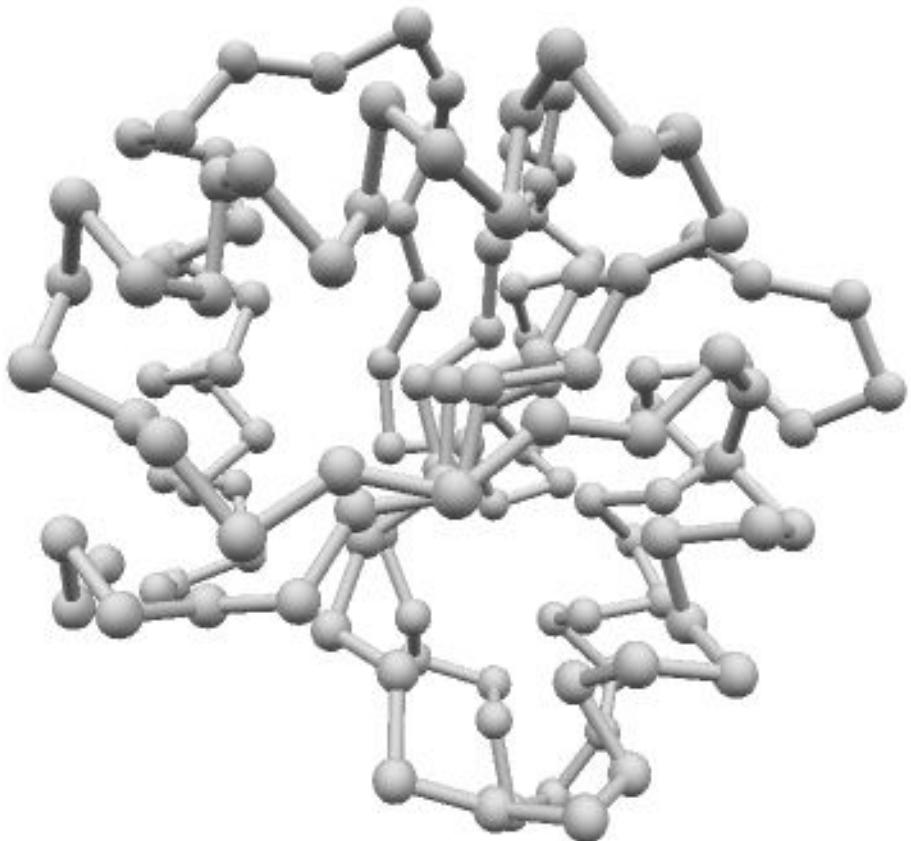
- Visual inspection(!)
  - The “best” method if you are well-trained.
- Algorithms
  - It's an NP-hard problem, so there are a number of approximate methods based on various representations:
    - secondary structure elements (SSE)
    - Amino acid residues
    - Atoms
    - Molecular surface

# Representation: all atoms



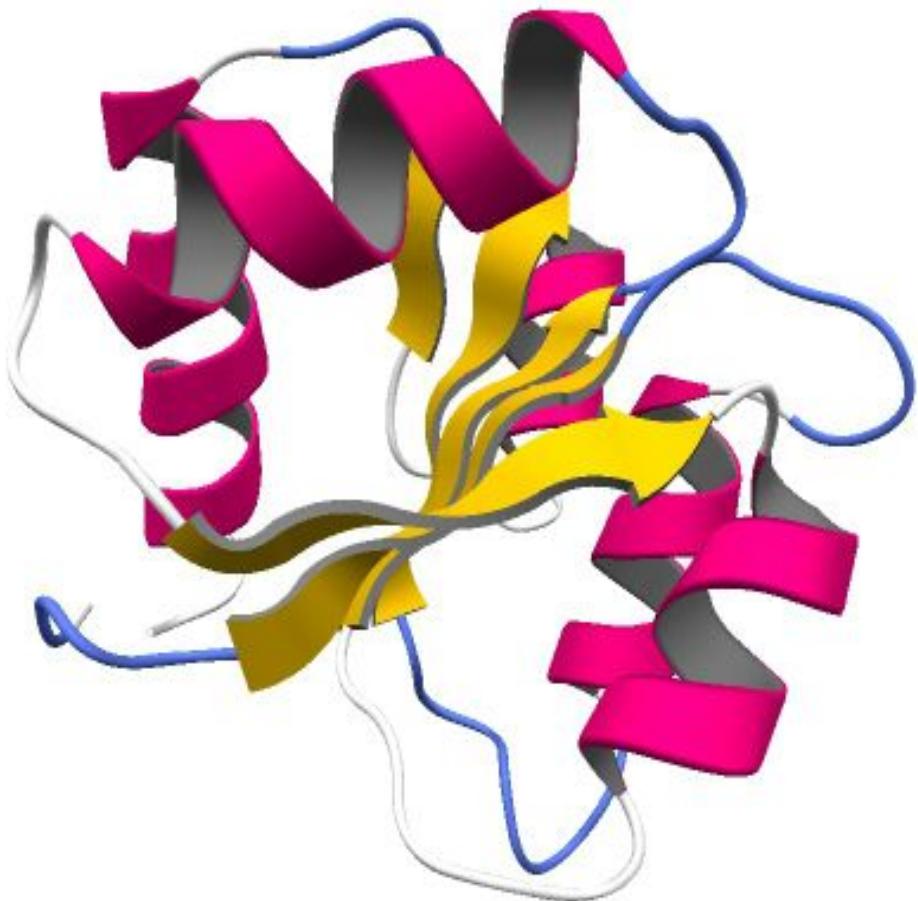
- Dealing with all atoms...
- is difficult. So usually only substructures are treated in this representation.

# Representation: Backbone



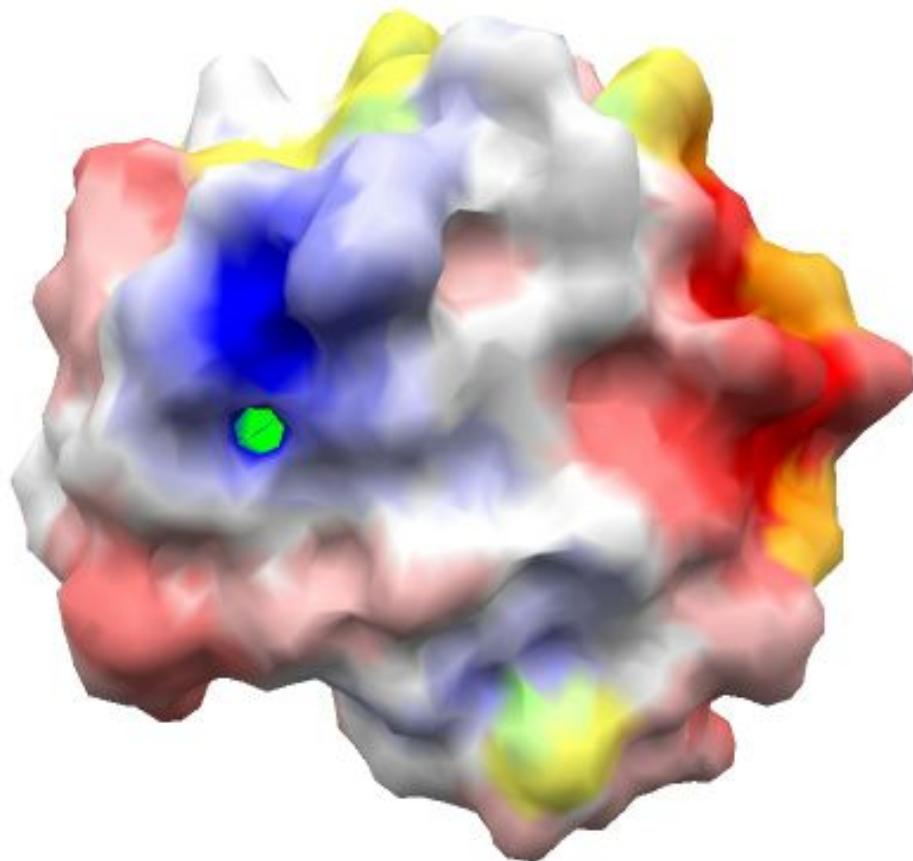
- Using only  $C\alpha$  or  $C\beta$  atoms to reduce computational costs.
- Also compatible with sequence alignment (1 atom / residue)
- Still computationally demanding.

# Representation: 2ndary structures



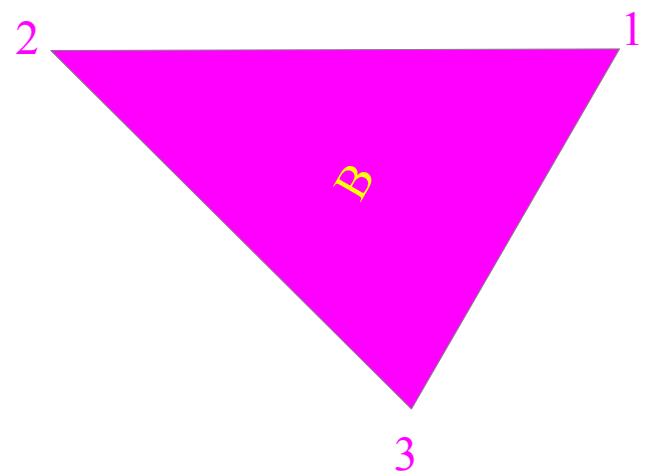
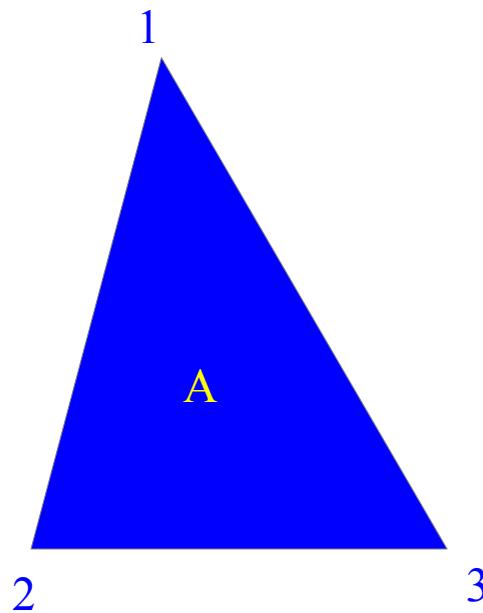
- $\alpha$  helices and  $\beta$  strands as vectors.
- Suitable for finding topological similarities.
- Less cost.

# Representation: Molecular surface



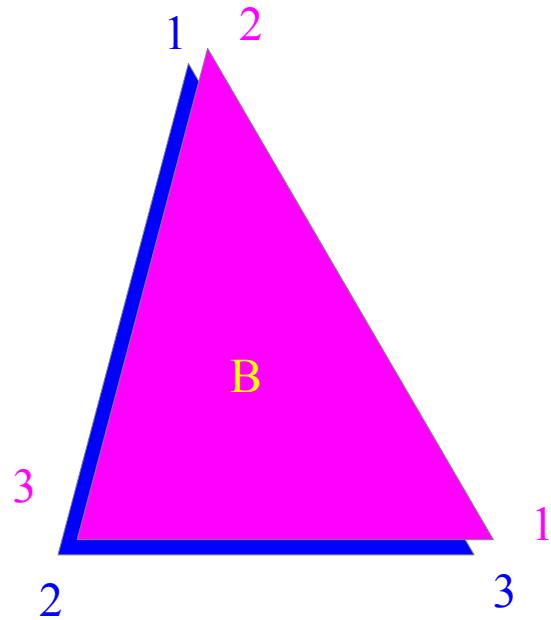
- Protein structure from the view point of a water molecule(?)
- Often used for mapping electrostatic potentials & hydropathy on the structure.

# Basic ideas



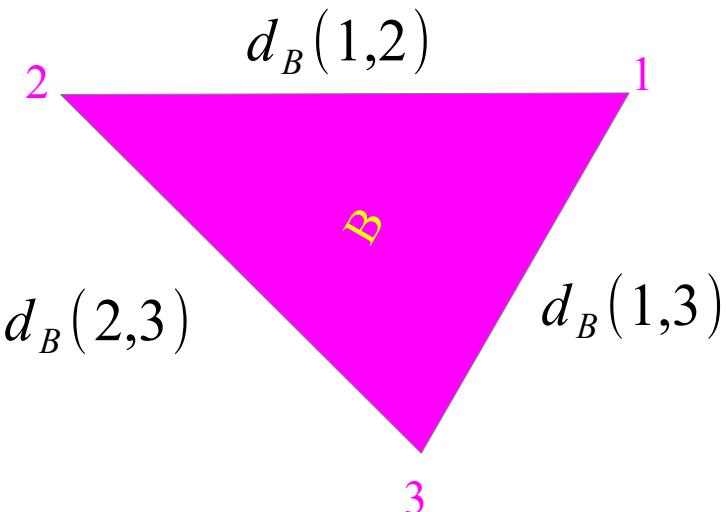
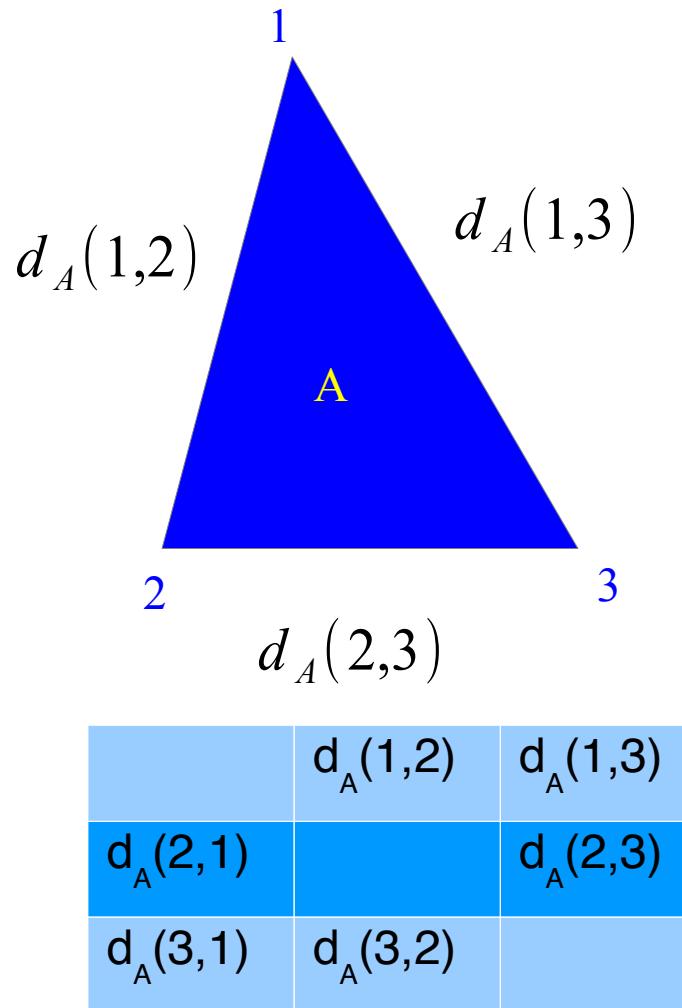
How do you tell the congruence of two triangles?  
(Vertex numbers do not match!)

# Method 1: Coordinate-based



- Actually try to superimpose them!
- Infinite combinations of “translation” & “rotation”.

# Method 2: Distance-based

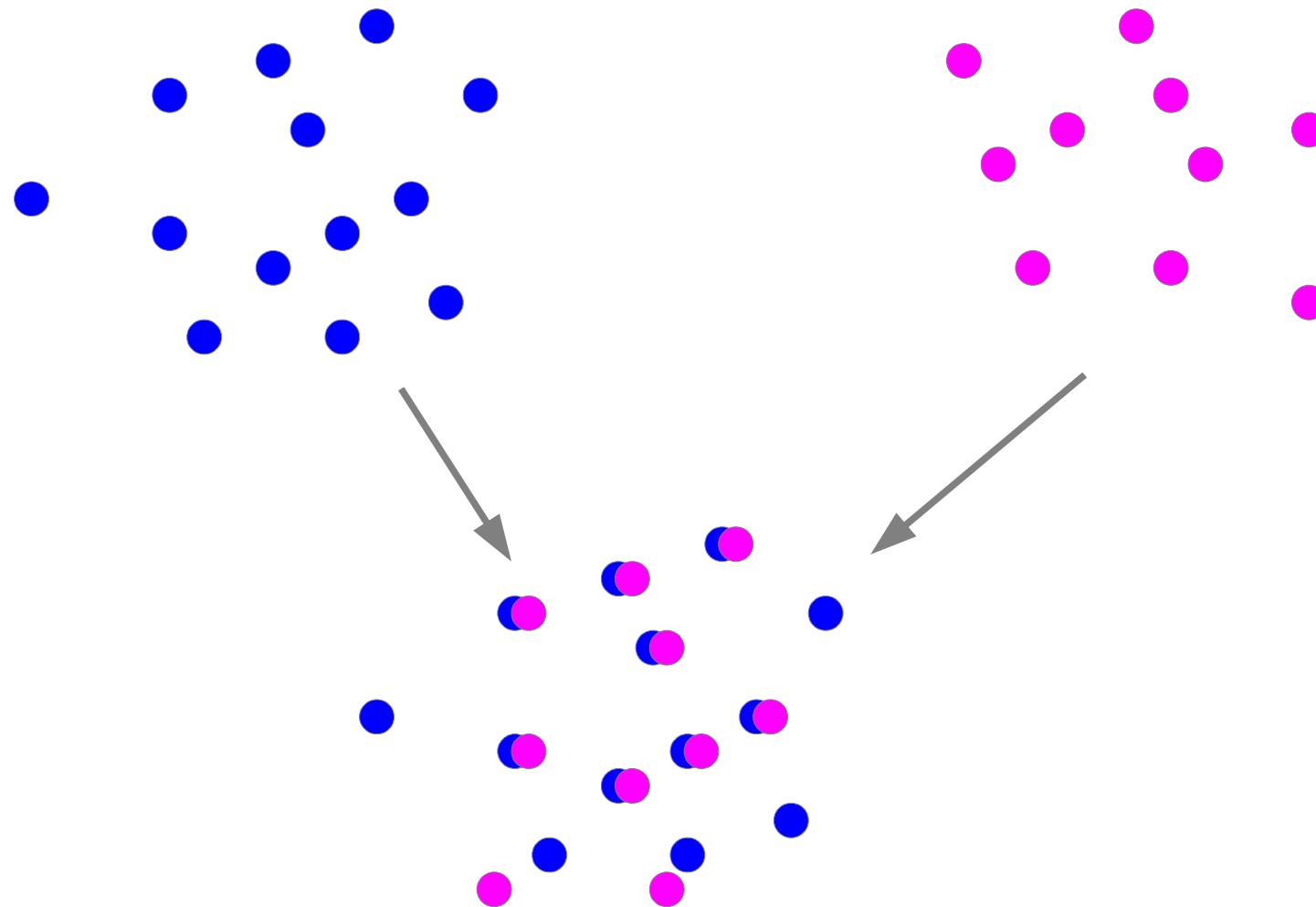


	$d_B(1,2)$	$d_B(1,3)$
$d_B(2,1)$		$d_B(2,3)$
$d_B(3,1)$	$d_B(3,2)$	

Find pairs  $(i,j),(k,l)$  that satisfy  $|d_A(i,j) - d_B(k,l)| = 0$

How many possibilities are there?

# A little more complicated objects



# Summary of comparison methods

- Translation & rotation
  - “Coordinate-based method”
  - Infinite possibilities.
- Comparing the distances between vertices
  - “Distance-based method”
  - Exponentially increasing possibilities.
- In any case, it's a tough problem!

# Coordinate-based method, theory

Let  $A$ ,  $B$  and  $C$  be metric spaces:

$$A = \left( x_1^A, x_2^A, \dots, x_M^A \right) \quad B = \left( x_1^B, x_2^B, \dots, x_N^B \right)$$

$$\begin{aligned} A &\xrightarrow{f} C \\ B &\xrightarrow{g} C \end{aligned}$$

The points in  $A$  and  $B$  are transformed into  $C$ , so the distance between two points, one in  $A$  and the other in  $B$  can be measured in  $C$ .

$$D(A, B) = \sum_{(i, j)} d_C(f(x_i^A), g(x_j^B))$$

The Problem: Find the set of combinations of  $(i, j)$  that minimizes this distance.  
But how do we define  $f$  and  $g$ ?

# Best-fitting problem

Easy case first. Assume the alignment is already known.

$$A = \begin{pmatrix} x_1^A, x_2^A, \dots, x_M^A \end{pmatrix} \quad B = \begin{pmatrix} x_1^B, x_2^B, \dots, x_M^B \end{pmatrix} \quad (\text{the same number of points})$$

$$\text{For all } i=1, \dots, M, (x_i^A, x_i^B) \quad (\text{The } i\text{-th atom in A} \Leftrightarrow \text{The } i\text{-th atom in B})$$

$$\sum_{i=1}^M x_i^A = 0, \sum_{i=1}^M x_i^B = 0 \quad (\text{Both centers of mass are at the origin})$$

$$D(A, B) = \sqrt{\frac{1}{M} \sum_{i=1}^M |x_i^A - R x_i^B|^2} \quad (\text{Rotate B by the rotation matrix R})$$

Now the problem is finding the matrix R (least-square fitting).

This can be solved analytically (Euler angles, singular value decomposition, quaternions)

# Coordinate-based method in practice

- Impossible to try infinite number of transformations
- 3 linearly independent points define a frame.
  - $N!/(N-3)! = N(N-1)(N-2)$
- Consider all combination from two structures
  - $M(M-1)(M-2) \times N(N-1)(N-2)$
  - $M=N=100 \Rightarrow 941,288,040,000$  combinations
- It's finite, but huge!

# Coordinate frame based on 3 points

$$A = \begin{pmatrix} x_1^A & x_2^A & \dots & x_M^A \end{pmatrix}$$

$$\begin{pmatrix} x_i^A & x_j^A & x_k^A \end{pmatrix} \quad \text{3 points}$$

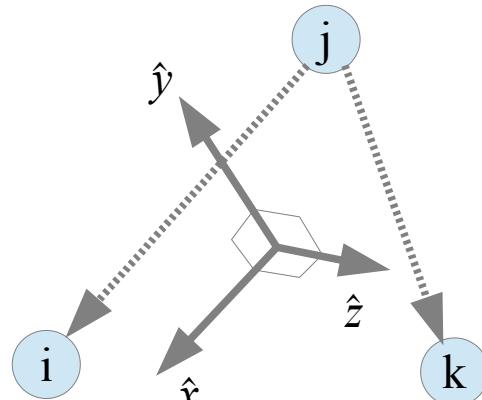
$$\hat{x} = \frac{1}{\|x_i^A - x_j^A\|} (x_i^A - x_j^A) \quad \text{X axis}$$

$$\hat{y} = \frac{1}{\|x_k^A - x_j^A\|} \hat{x} \times (x_k^A - x_j^A) \quad \text{Y axis}$$

$$\hat{z} = \hat{x} \times \hat{y} \quad \text{Z axis}$$

$$O = \frac{1}{3} (x_i^A + x_j^A + x_k^A) \quad \text{Origin}$$

$$\hat{x}_a^A = (\hat{x} \cdot (x_a^A - O), \hat{y} \cdot (x_a^A - O), \hat{z} \cdot (x_a^A - O)), a = 1, \dots, M \quad \text{Transformation}$$



# Simple superposition algorithm

Input: Structure A=x(1)..x(M); Structure B=y(1)..y(N)

Output: Best alignment Ali

Ali := {} --- 初期アライメント(空集合)

**for** (i,j,k) in {1..M} **do** --- Select 3 points from A

basisA := make\_basis(x(i),x(j),x(k)) --- Make a basis

**for** a = 1..M **do**

x'(a) := transform(x(a),basisA)

**for** (l,m,n) in {1..N} **do** --- Select 3 points from B

basisB := make\_basis(y(l),y(m),y(n)) --- Make a basis

S := {} --- Initial (empty) alignment

**for** b = 1..N **do**

y'(b) := transform(y(b),basisB)

(\* After transformation, count neighboring A,B points \*)

**for** a = 1..M **do**

**for** b = 1..N **do**

**if** |x'(a) - y'(b)| < delta

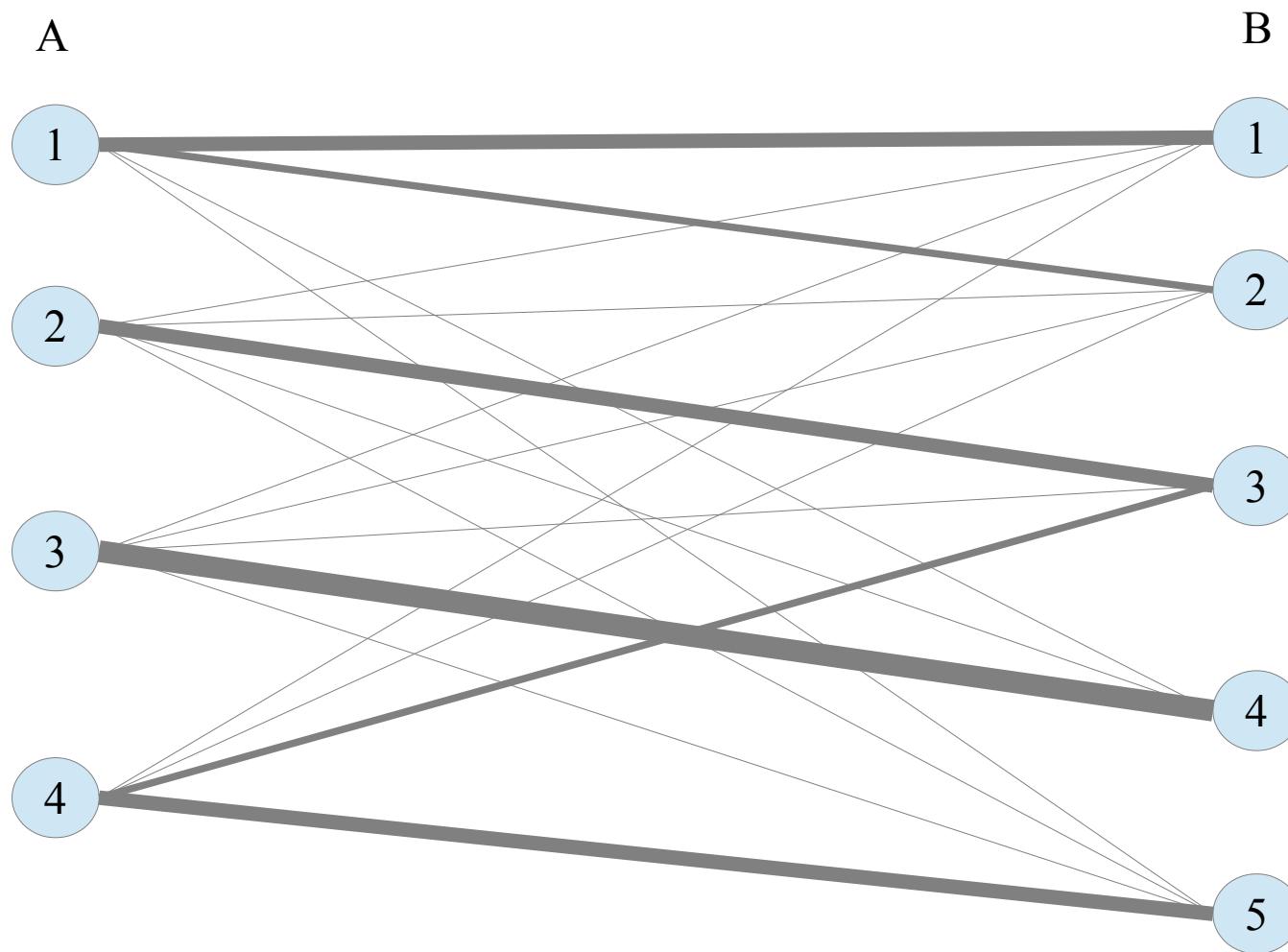
**then** S := S {(a,b)}

--- Add pair to alignment

**if** |S| > |Ali| **then** Ali := S

--- Save the best one!

# A possible result

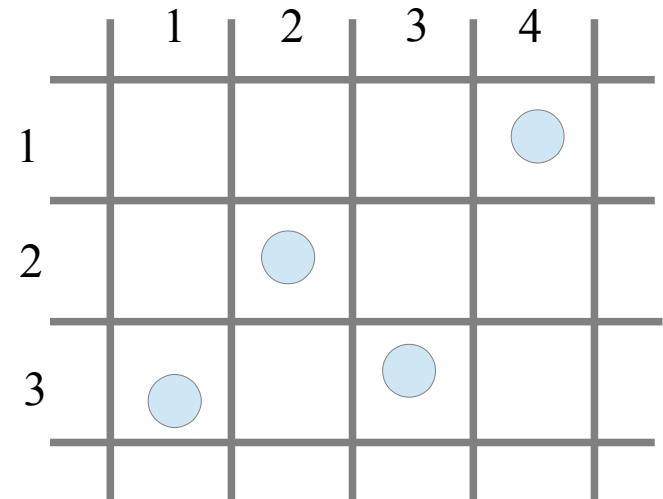


# Geometric Hashing (GH)

- The simple approach is simply too slow.
- Make a dictionary (hash table)  $x' \rightarrow \text{basis}$
- Looking up the dictionary is fast:  $O(1)$ , no loop.

The coordinate after transformed by basisB.

$$(x'(l), y'(l), z'(l)) \rightarrow \text{basisB}$$



# Creating a hash table

Input: Structure B     $y(1) \dots y(N)$

Output: Hash table HB

```
for (l,m,n) in 1..N do
    basisB := make_basis(y(l),y(m),y(n))
    for b = 1..N do
        y'(b) := transform(y(b),basisB)
        HB := HB      (y'(b) => y'(b),basisB)
```

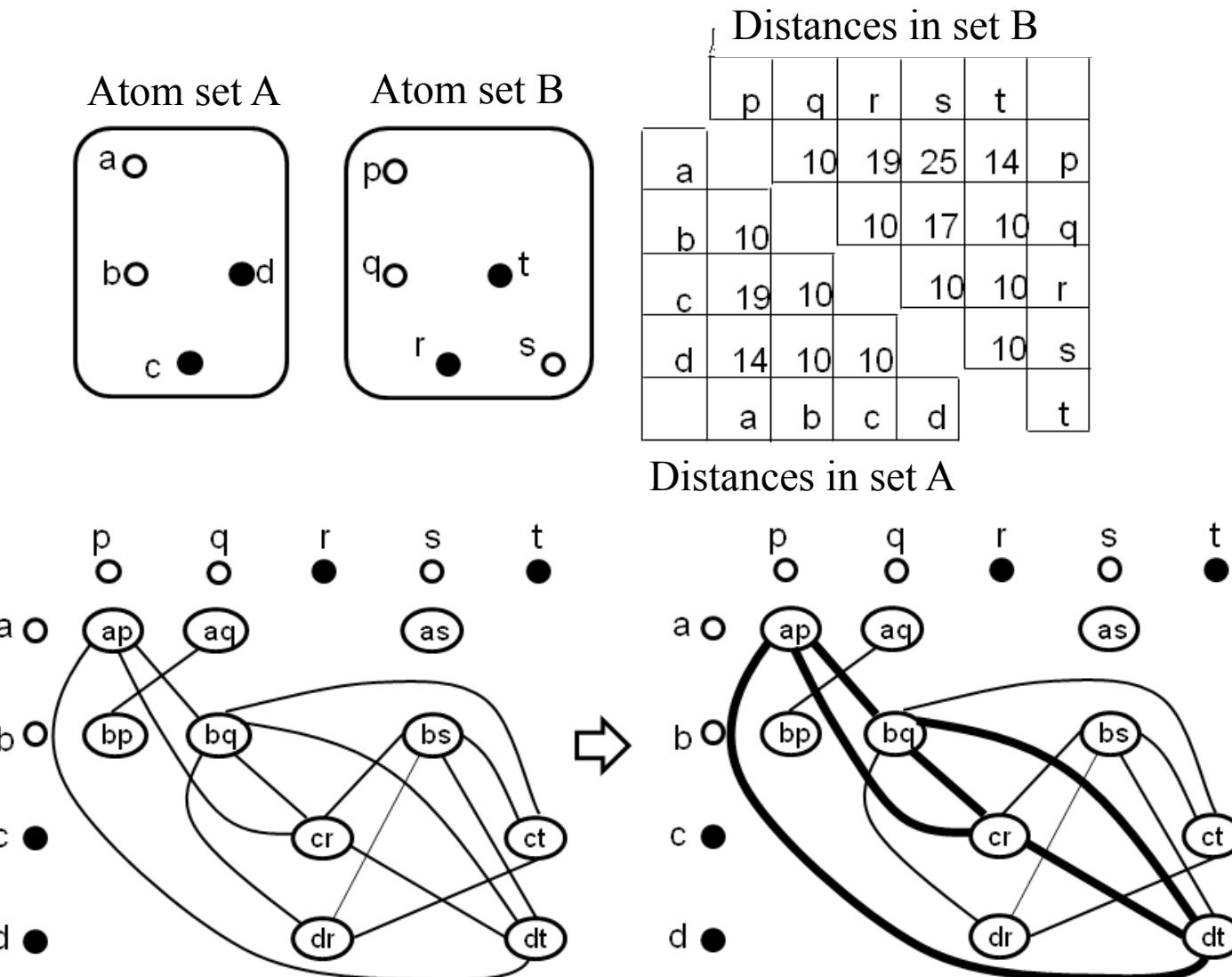
This requires  $N^2(N-1)(N-2)$  steps.

# Structure comparison by GH

```
Input: Structure A=x(1)..x(M); Structure B=y(1)..y(N)
Output: Best alignment Ali
HB := make_hashtable(B) --- Create hash table
for (i,j,k) in {1..M} do --- Select 3 points from A
    basisA := make_basis(x(i),x(j),x(k)) --- Make a basis
    for a = 1..M do
        x'(a) := transform(x(a),basisA)
        for y'(b),basisB in find_hash(x'(a)) --- Find a B-basis
            P(basisA,basisB) := {(a,b)}      P(basisA,basisB)
            --- Add the atom pair
Ali := Max|P(basisA,basisB)|      ---(*) Be careful!
```

The last step (\*) requires a smart data structure!  
Otherwise, this method is as slow as the previous one.

# Distance comparison method



Courtesy of Dr. Takeshi Kawabata

# Basic idea of distance-based method

$$A = (x_1^A, \dots, x_M^A)$$

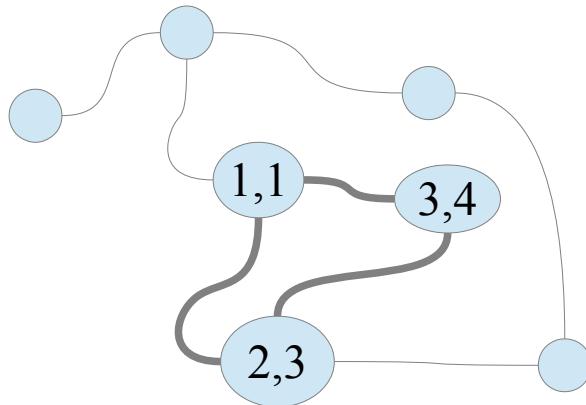
$$B = (x_1^B, \dots, x_N^B)$$

Given A and B, consider all the pairs of A and B points:  $P = \{(x_i^A, x_j^B)\}$

For the pair of pairs (i,j) and (k,l), the two distances (i,k) & (j,l) are similar, draw an edge between the nodes (i,j) & (k,l).

$$\| \|x_i^A - x_k^A\| - \|x_j^B - x_l^B\| \| < \delta$$

Find the subgraph of thus created graph, that is complete and maximum:  
The maximum clique problem



# Algorithm

Bron-Kerbosch (1973)

```
R := empty
P := set of vertices
X := empty
```

```
BronKerbosch1(R,P,X):
    if P and X are both empty:
        report R as a maximal clique
    for each vertex v in P:
        BronKerbosch1(R - {v}, P - N(v), X + N(v))
        P := P \ {v}
        X := X + {v}
```

Where  $N(v)$  is the set of vertices connected with "v".

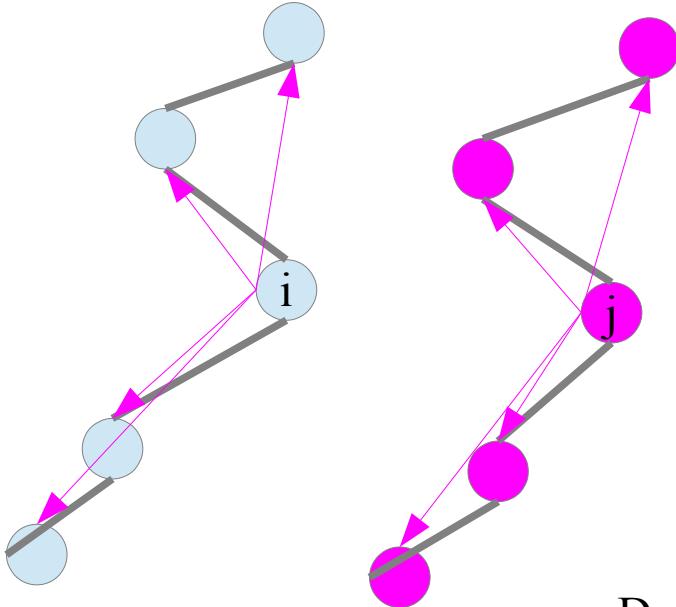
From [http://en.wikipedia.org/wiki/Bron–Kerbosch\\_algorithm](http://en.wikipedia.org/wiki/Bron–Kerbosch_algorithm)

This is an exact algorithm, and may not terminate.

# Double Dynamic Programming

- Distance-based methods are also computationally demanding.
- DDP is a hybrid of coordinate- & distance-based methods
- Applying DP (just as in sequence comparison) in two layers.
- This requires the point set to be ordered.

# DDP: idea



$$A = (x_1^A, \dots, x_M^A)$$

$$B = (x_1^B, \dots, x_N^B)$$

Assume  $(x_i^A, x_j^B)$  is a matching pair of points.

If  $(i, j)$  is really a matching pair, the “scene of A from i” and the “scene of B from j” should look similar.

Define the similarity measure for the “scenes” based on  $(i, j)$ :

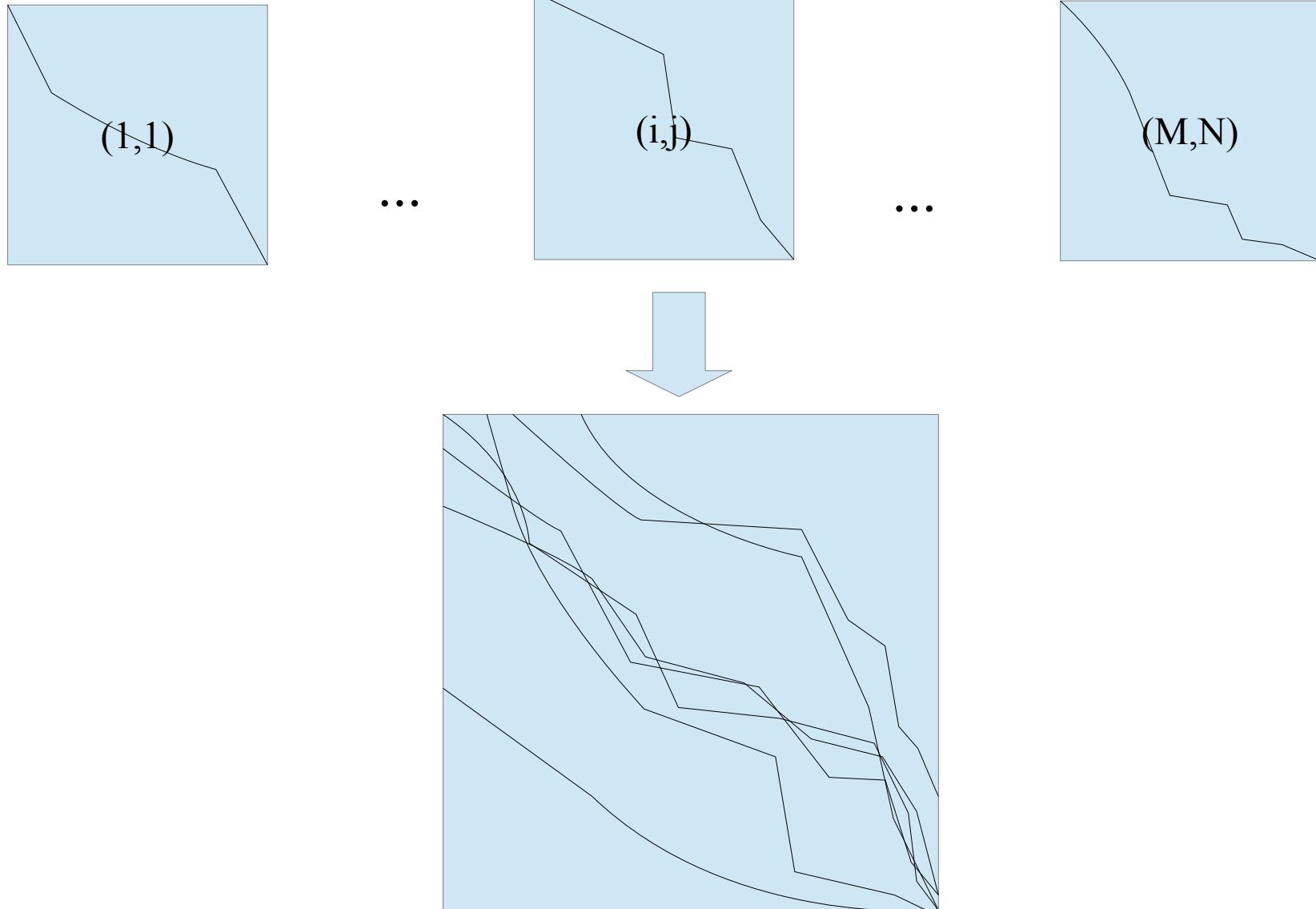
$$s(k, l; i, j) = \frac{1}{|d_A(i, k) - d_B(j, l)| + c}$$

Apply DP by regarding this as a score matrix  $s(k, l)$ , you get the “best” alignment under the assumption that  $(i, j)$  is a matching pair. The score is, say:  $S_1(i, j)$   
( Do this for all possible  $(i, j)$  pairs. )

Then using  $S_1(i, j)$  as a score matrix, apply another DP.

This will yield an approximation to the “best” alignment

# DDP の概念図



# DDP Algorithm

```
# lower level DP
for i=1..M do
    for j=1..N do
        S(i,j) = DP using s(k,l; i,j) --- details omitted.

# upper level DP
for i= 1..M do
    for j= 1..N do
        D := T(i-1,j-1) + S(i,j)
        V := T(i-1,j) - g
        H := T(i,j-1) - g
        T(i,j) := max(d,v,h)
        if T(i,j) = d then P(i,j) := 'D'      --- diagonal
        else if T(i,j) = v then P(i,j) := 'V'  --- vertical
        else T(i,j) := 'H'                    --- horizontal
    done
done
Score := T(M,N)
--- omitting the rest...
```

# Why DDP works

- If  $(i,j)$  is a truly matching pair
  - $S_1(i,j)$  is a large positive value.
- If  $(i,j)$  is not a truly matching pair
  - $S_1(i,j)$  is a small value.
- The scores of truly matching pairs are amplified along the (sub)optimal alignment.

# Summary

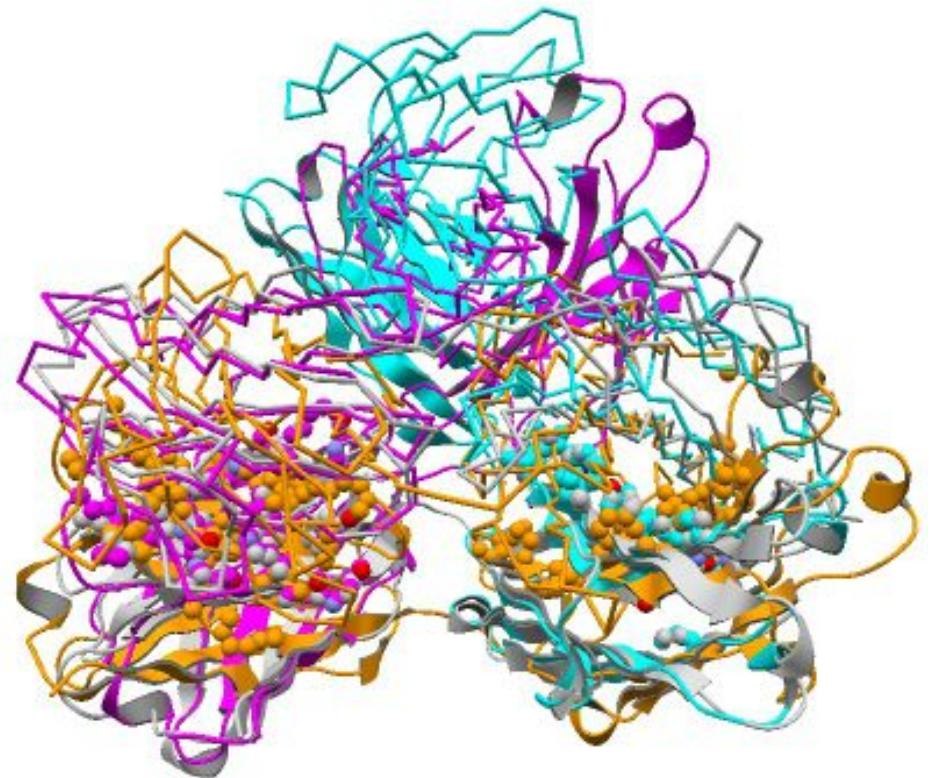
- 2 approaches for structure comparison
  - Coordinate-based
  - Distance-based
- In special cases, dynamic programming can be also used.

## Application

Fast search and flexible alignment of interaction site structures in proteins

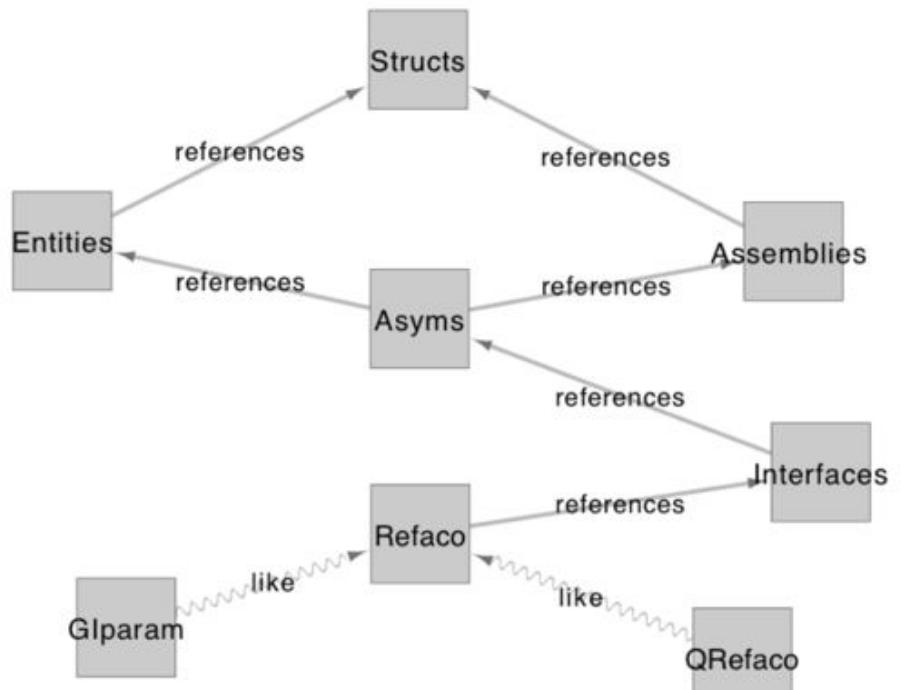
# Objectives

- Search a large database of interaction site structures *quickly*.
- Align flexible structures



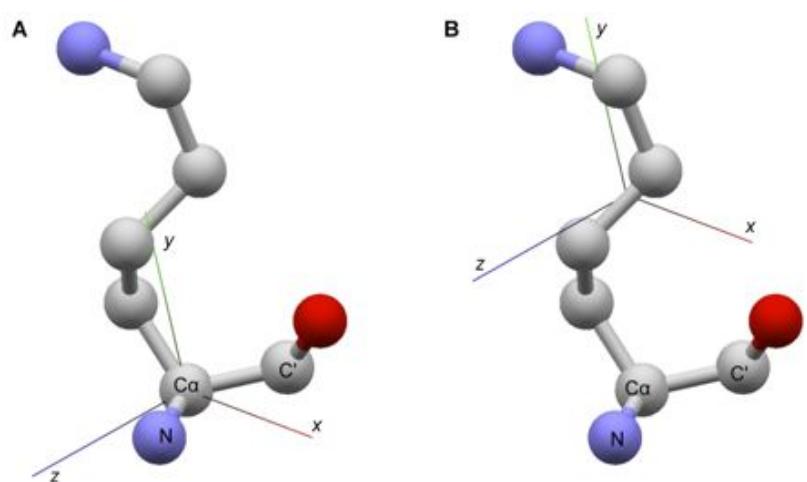
# Database design

- From PDBML files, extract all interaction interfaces (small molecules, proteins, DNA/RNA)
- Biological assemblies are considered.
- Preprocess the local coordinates.



# Local coordinate system

- origin: center of mass of side chain heavy atoms.
- axes: based on backbone N, C $\alpha$ , C' atoms.
- Defined for each residue in an interface.

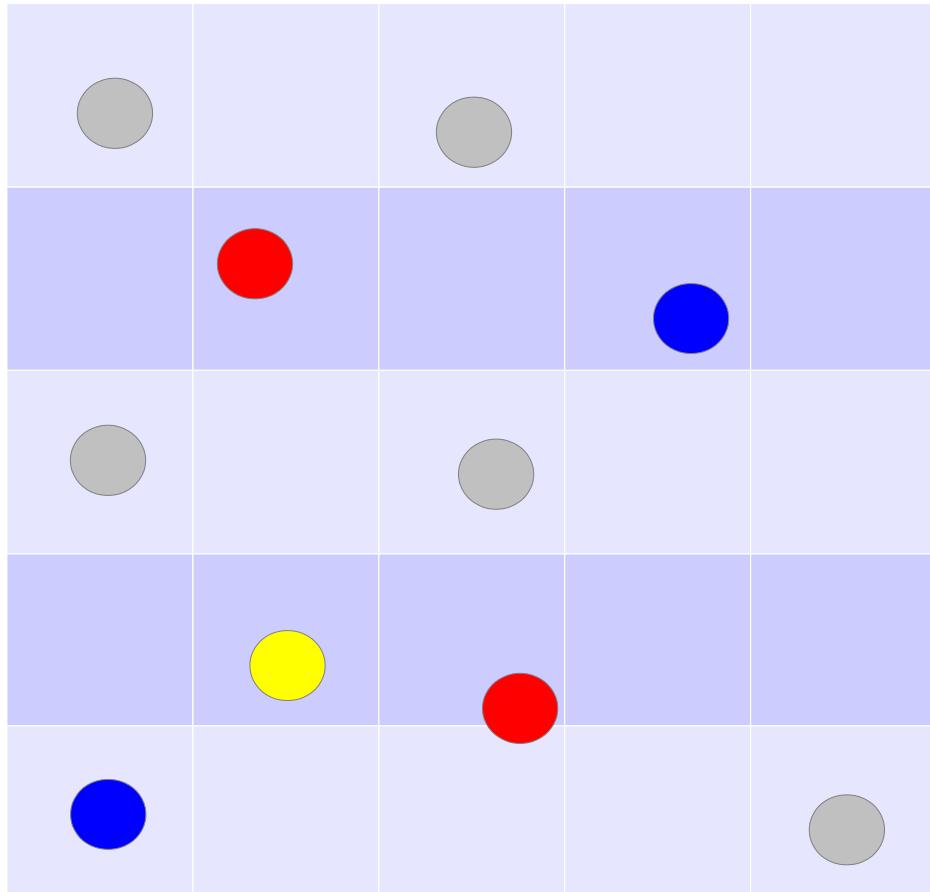


# Structure search by SQL

```
CREATE TABLE Refaco (
    if_id TEXT          /* interface ID */
  , rs_id TEXT          /* affine frame ID */
  , type TEXT           /* interface type */
  , frame BYTEA         /* affine frame */
  , ft01 FLOAT          /* structural features 1~44 */
  , ft02 FLOAT
  , . . . . .
  , ft44 FLOAT
  , lattice BIGINT[ ] /* array of lattice points */
  , natoms INT          /* number of interface atoms */
);
```

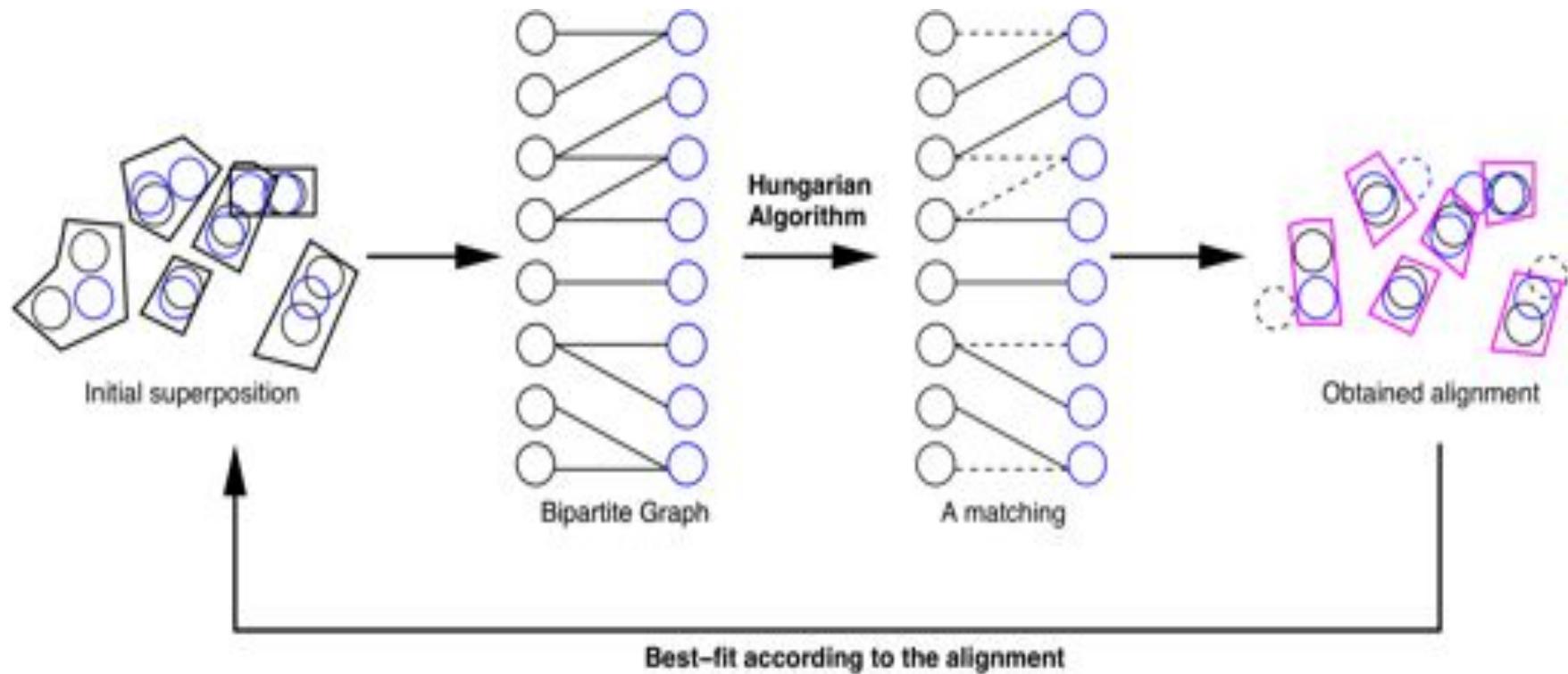
```
1: SELECT t.if_id, t.rs_id, t.type, t.frame, q.rs_id
2: FROM Refaco t, Qrefaco q
3: WHERE t.ft01 BETWEEN q.ft01 - D01 AND q.ft01 + D01
4: AND   t.ft02 BETWEEN q.ft02 - D02 AND q.ft02 + D02
      ...
5: AND   t.ft44 BETWEEN s.ft44 - D44 AND q.ft44 + D44
6: AND   (SELECT COUNT(*)
7:           FROM (SELECT UNNEST(t.lattice)
8:                     INTERSECT
9:                     SELECT UNNEST(q.lattice)) AS x)
10:      > Smin * LEAST(t.natoms, q.natoms)
```

# Discretized local atomic coordinates



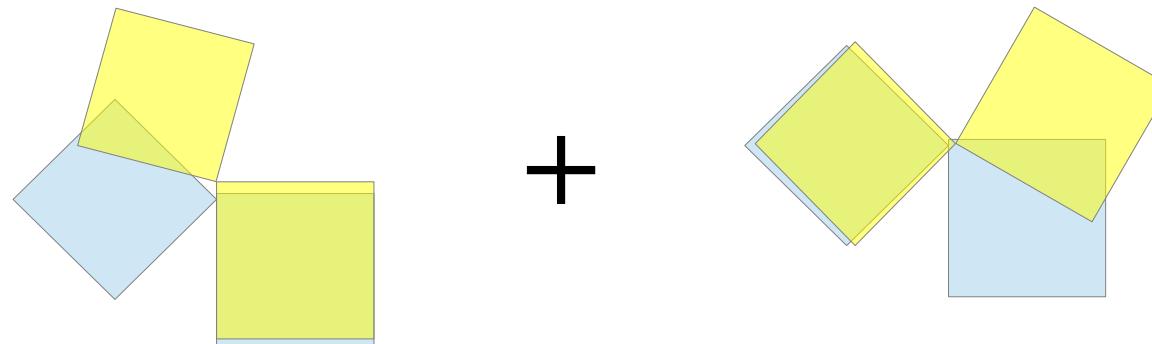
- Local coordinates of interface atoms are discretized and saved in an *array-type* column of the Refaco table.

# Alignment by Hungarian method

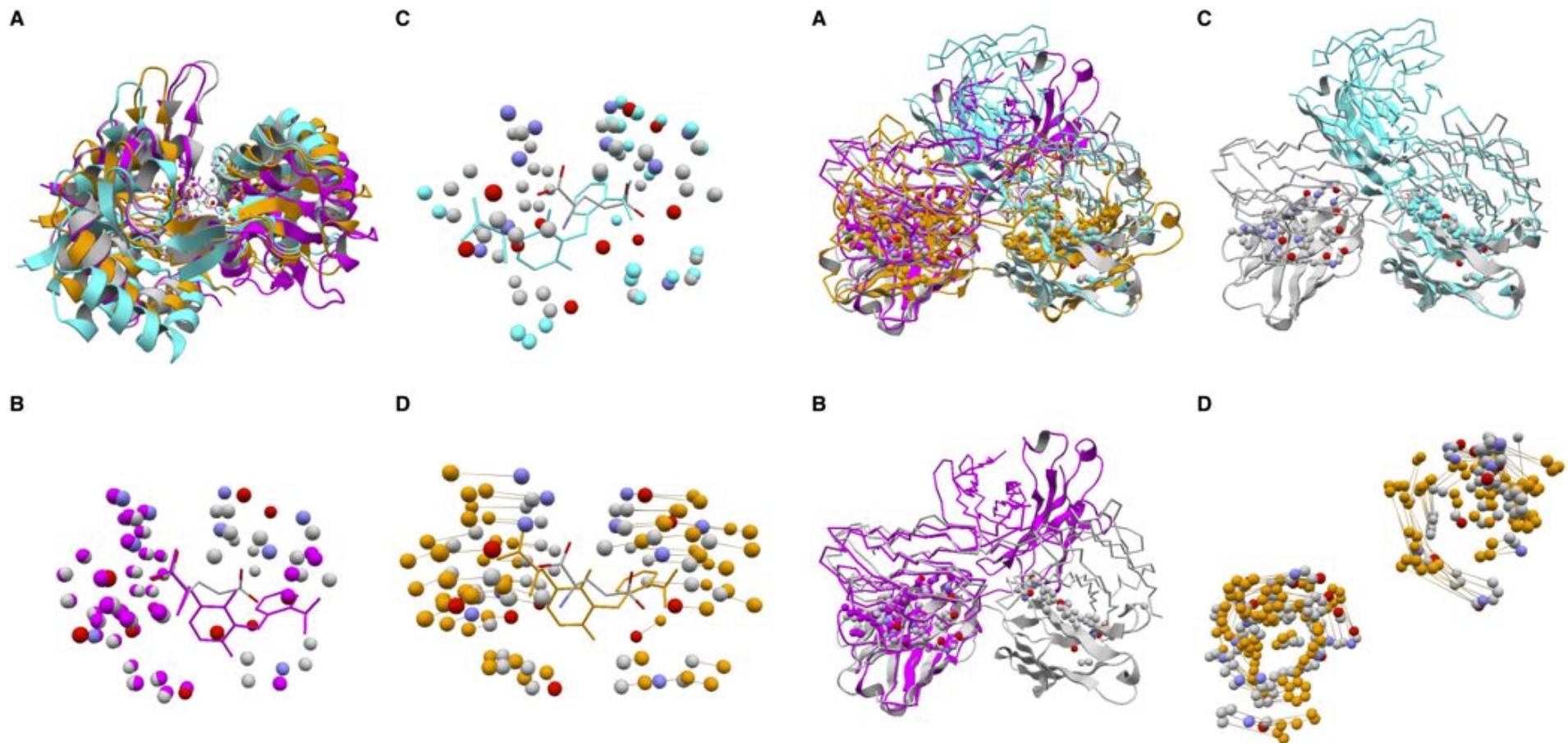


# Flexible alignment

- Rigid alignments based on different affine frames.
- They are merged as long as mutually consistent.
  - Bad in 3D-Euclidean space, Good in patches of locally Euclidean spaces.



# Flexible alignment: examples



# **Applications**

Structure 17:234-246 (2009)

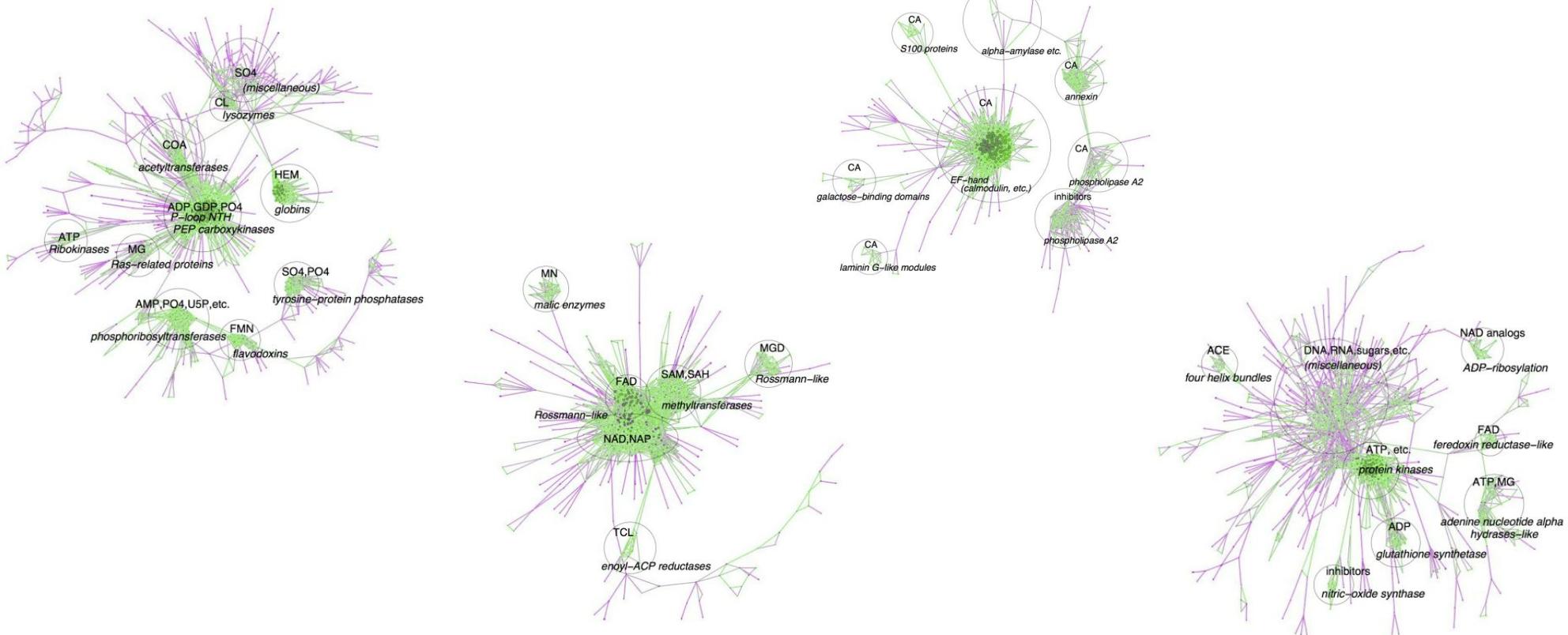
# Comprehensive Structural Classification of Ligand-Binding Motifs in Proteins

Akira R. Kinjo<sup>1,\*</sup> and Haruki Nakamura<sup>1</sup>

<sup>1</sup>Institute for Protein Research, Osaka University, 3-2 Yamadaoka, Suita, Osaka 565-0871, Japan

\*Correspondence: akinjo@protein.osaka-u.ac.jp

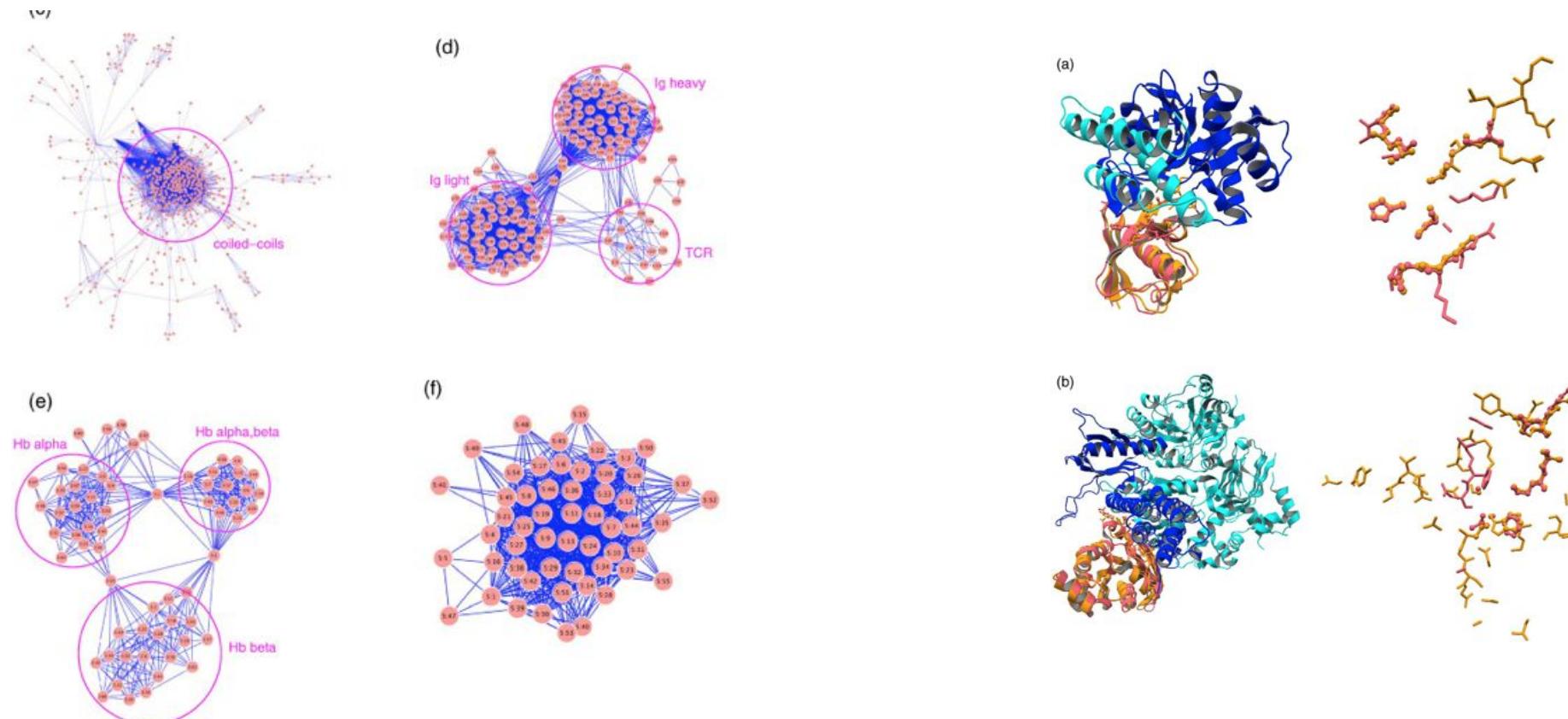
DOI 10.1016/j.str.2008.11.009





# Geometric Similarities of Protein–Protein Interfaces at Atomic Resolution Are Only Observed within Homologous Families: An Exhaustive Structural Classification Study

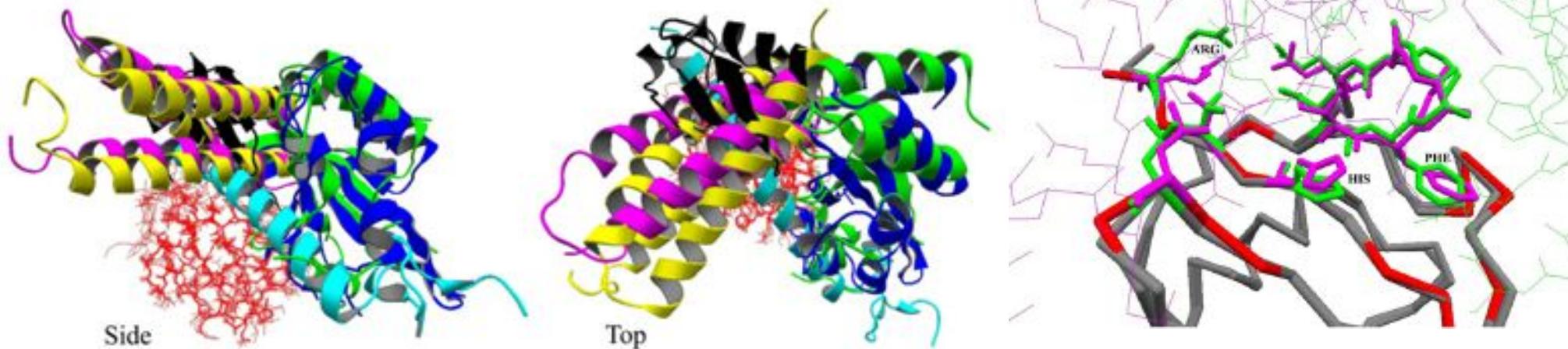
Akira R. Kinjo\* and Haruki Nakamura





# Distinct Roles of Overlapping and Non-overlapping Regions of Hub Protein Interfaces in Recognition of Multiple Partners

Bhaskar Dasgupta<sup>1,2\*</sup>, Haruki Nakamura<sup>1</sup> and Akira R. Kinjo<sup>1</sup>



OPEN  ACCESS Freely available online



# Composite Structural Motifs of Binding Sites for Delineating Biological Functions of Proteins

Akira R. Kinjo\*, Haruki Nakamura

PLoS ONE 7(2): e31437 (2012)

# What is “protein function”?

- Biochemical function
  - Catalytic activity (enzymes), ligand binding...
  - Attributes of proteins themselves.
- Biological function
  - “Development”, “aging”, “memory”,...
  - Behavior of a network of proteins.

# How do you know a function?

- Adenylate kinase 1 (human) in UniProt:

General annotation (Comments)	Hide   Top
Function	Catalyzes the reversible transfer of the terminal phosphate group between ATP and AMP. Small ubiquitous enzyme involved in energy metabolism and nucleotide synthesis that is essential for maintenance and cell growth.
Catalytic activity	$\text{ATP} + \text{AMP} = 2 \text{ ADP}$ .
Subunit structure	Monomer.
Subcellular location	Cytoplasm.
Polymorphism	This enzyme represents the most common of at least five alleles.
Involvement in disease	Defects in AK1 are the cause of hemolytic anemia due to adenylate kinase deficiency [MIM:612631].
Sequence similarities	Belongs to the adenylate kinase family.

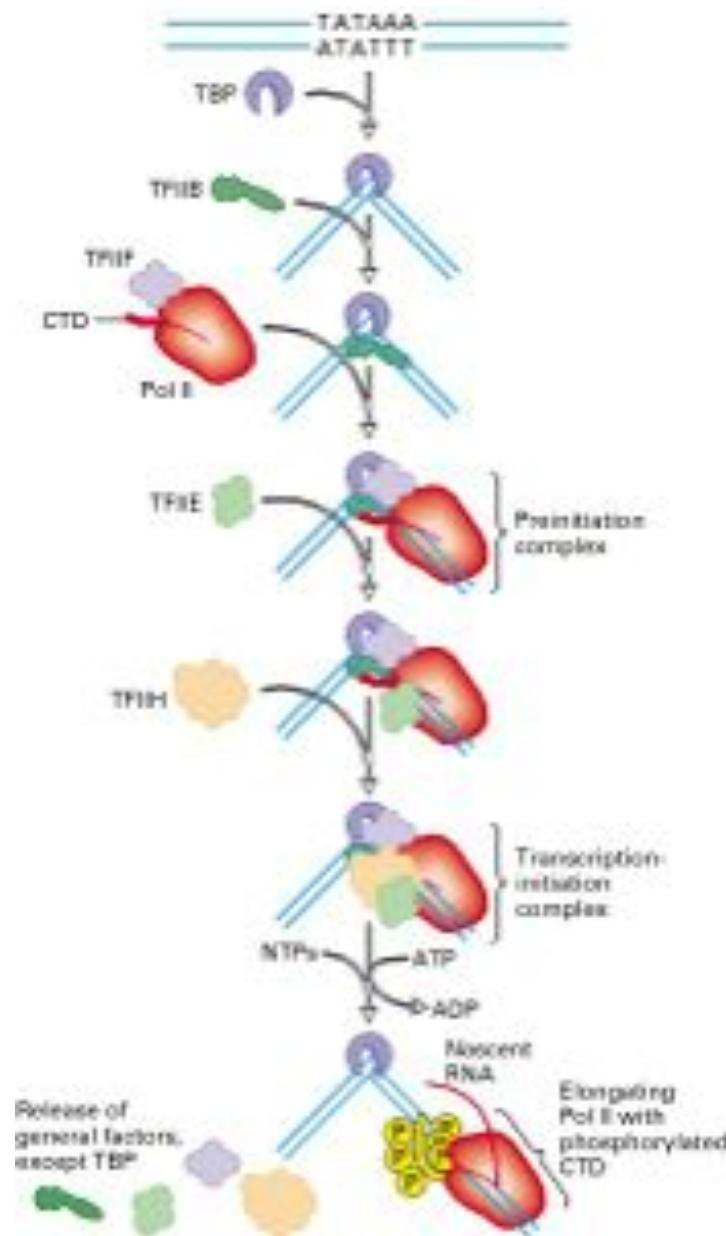
# Another way to annotate adenylate kinase in *Gene Ontology*

GO:0004017 adenylate kinase activity

Catalysis of the reaction: ATP + AMP = 2 ADP.

Term Information	Ancestor Chart	Ancestor Table	Child
Terms	Protein Annotation	Co-occurring Terms	
<a href="#">ID</a> : GO:0004017			
<a href="#">Name</a> : adenylate kinase activity			
<a href="#">Ontology</a> : Molecular Function			
<a href="#">Definition</a> : Catalysis of the reaction: ATP + AMP = 2 ADP.			
<a href="#">Comment</a>			
<a href="#">Secondary IDs</a>			
<a href="#">GONUTS Wiki Page</a>			
<a href="#">Synonyms</a>			
Type	Synonym		
exact	ATP:AMP phosphotransferase activity		
narrow	myokinase activity		
exact	adenylokinase activity		
exact	adenylic kinase activity		
exact	5'-AMP-kinase activity		

# What is protein function?



From *Molecular Cell Biology* 4/e (Fig. 10-50)

# Protein function is ...

A series of  
Interaction States,  
which is ...

An ordered set of  
*binding patterns*

Connecting  
*atomic structures*  
to  
*higher-order biological  
functions*

# Plan

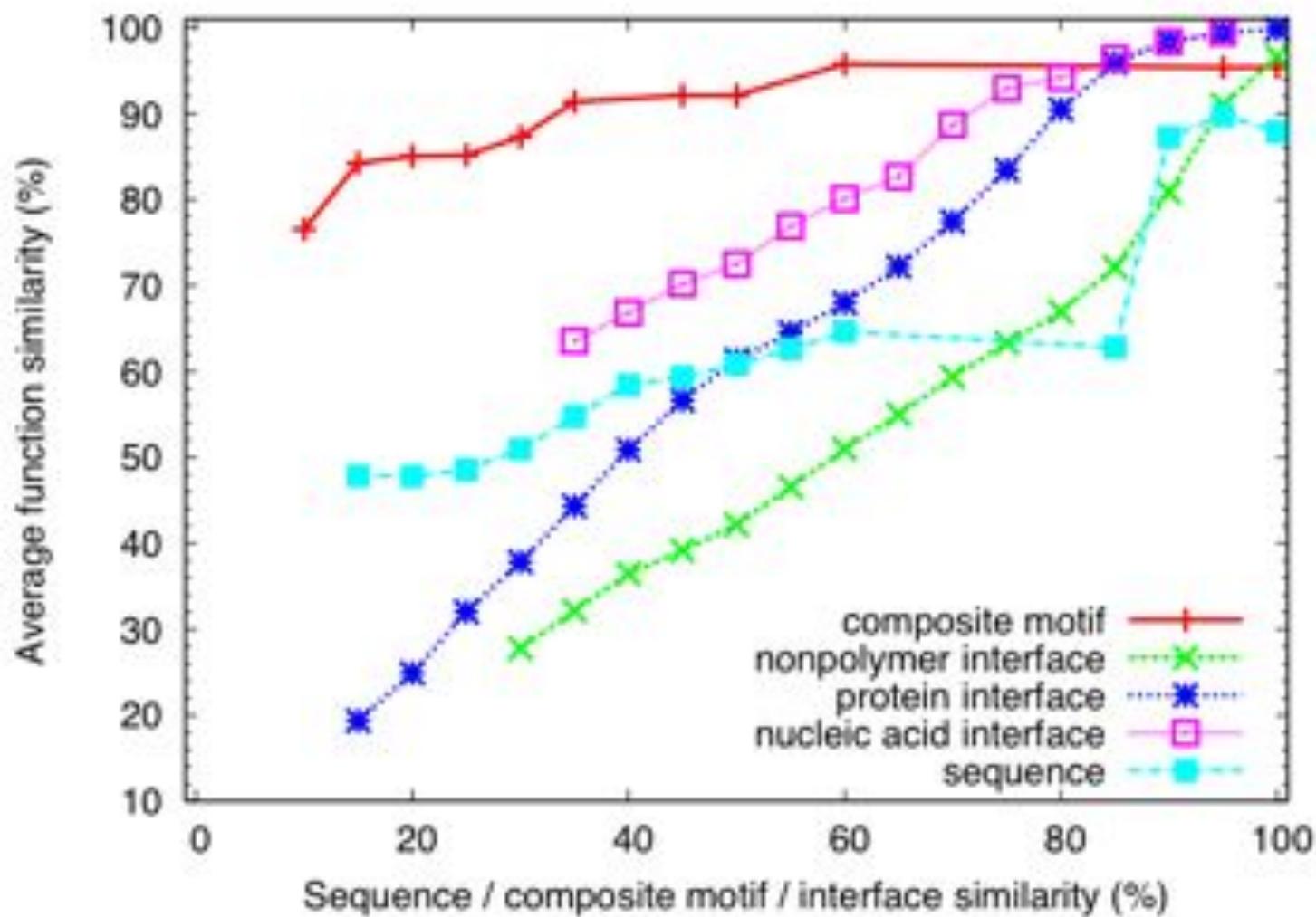
- Identify all patterns of binding sites
  - elementary motifs
- Identify all combinations of elementary motifs in each subunit
  - composite motifs
- Identify all composite motifs associated with a particular protein function (defined by UniProt).
  - meta-composite motifs
  - ~ a network of interaction patterns.

# Materials

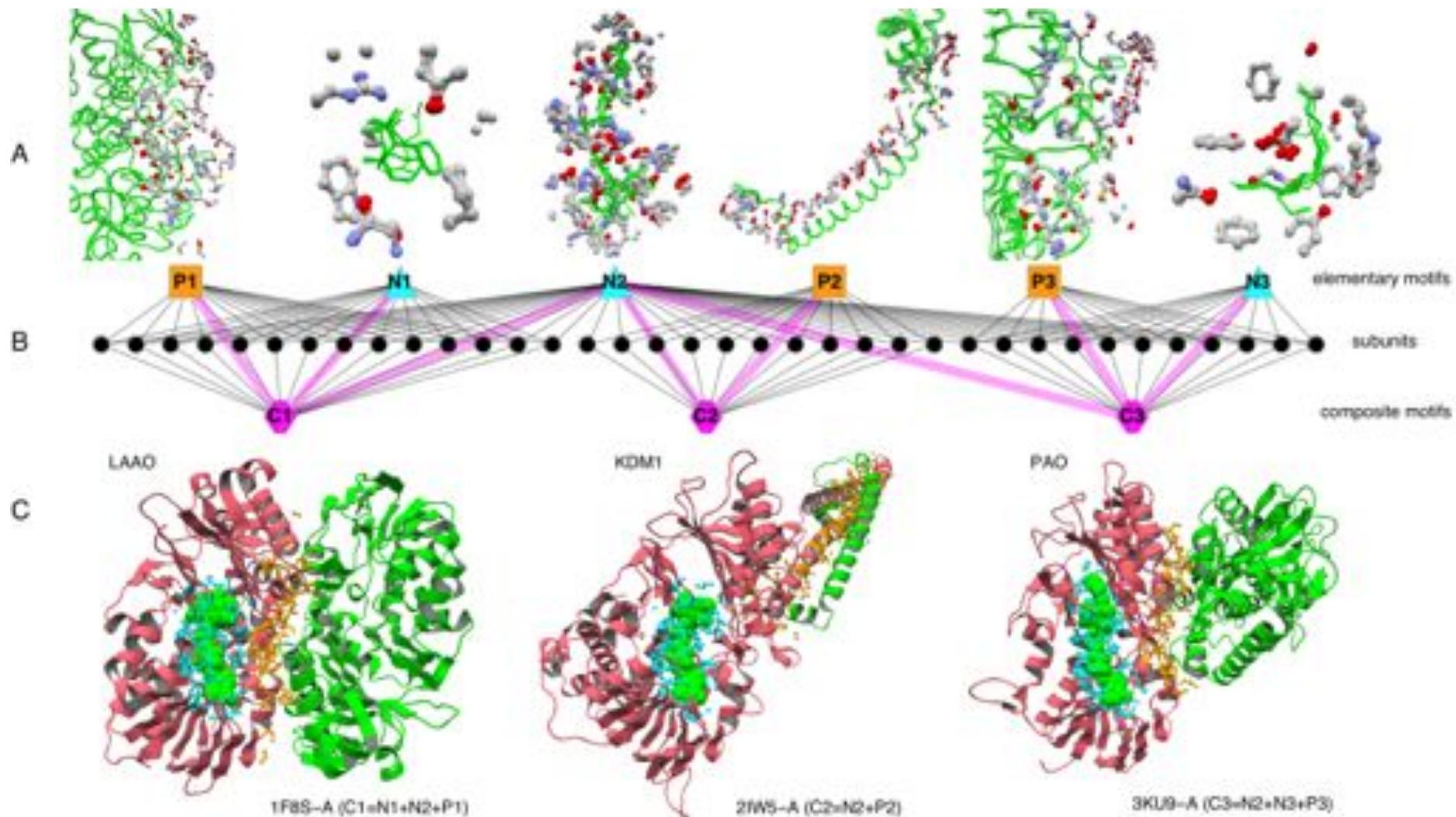
PDB entry	~70,000	
Biological Unit	~80,000	
Subunits	~200,000	~6,000 composite motifs
non-polymer interface	~400,000	~6,000 elementary motifs
protein interface	~350,000	~8,000 elementary motifs
DNA/RNA interface	~20,000	~400 elementary motifs

*No sequence representatives are used!*

# c-motifs vs. functions

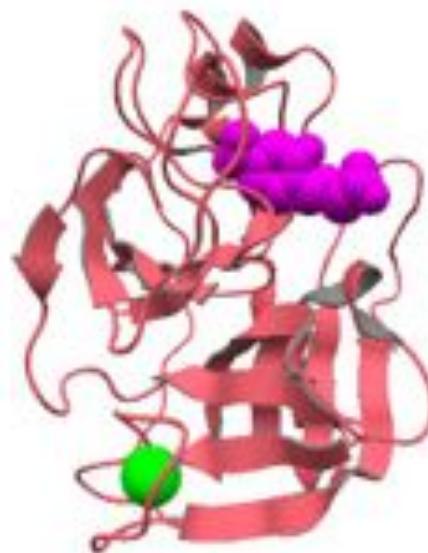


# Composite motifs: example

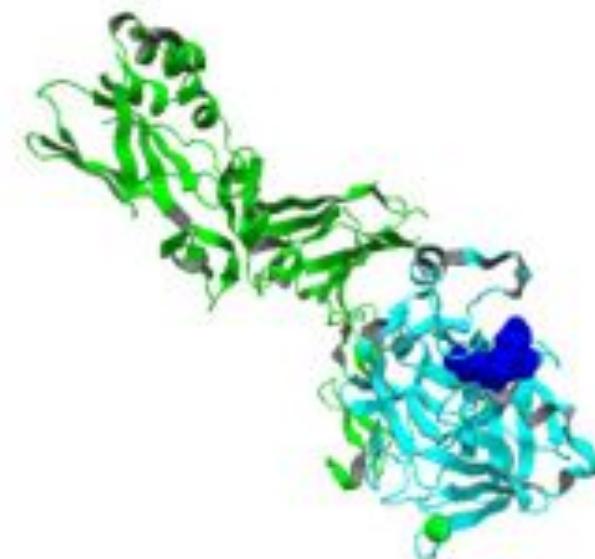


# Examples

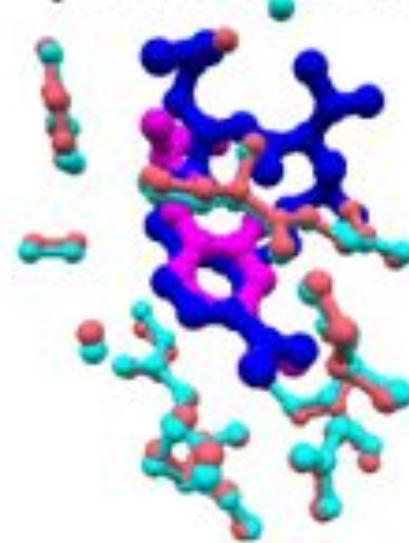
beta-trypsin (EC 3.4.21.4)



Factor VII (EC 3.4.21.21)



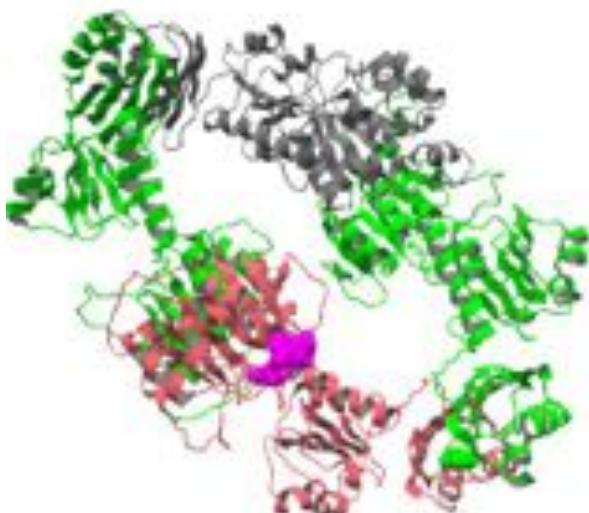
protease inhibitor



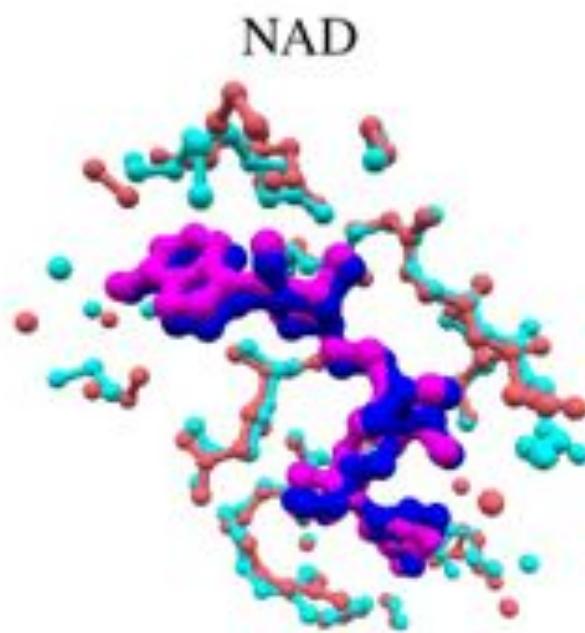
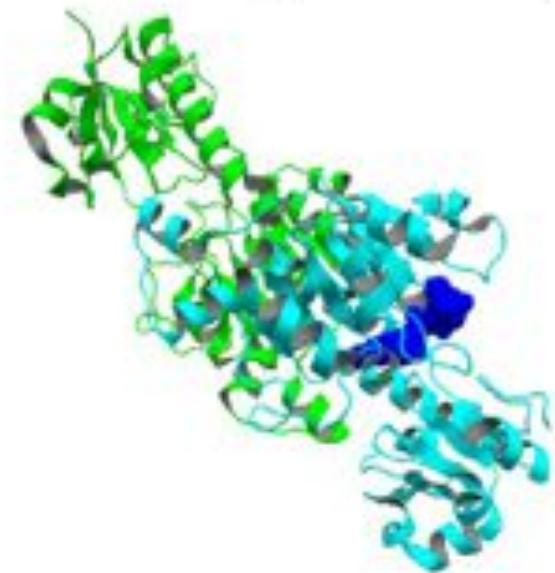
(i.e. catalytic site)

# Same e-motif & fold, different c-motifs & functions

PGDH (EC 1.1.1.95)

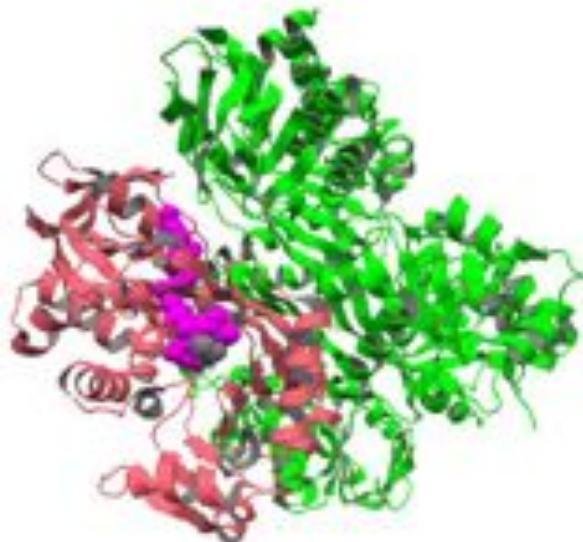


CtBP3 (EC 1.1.1.-)

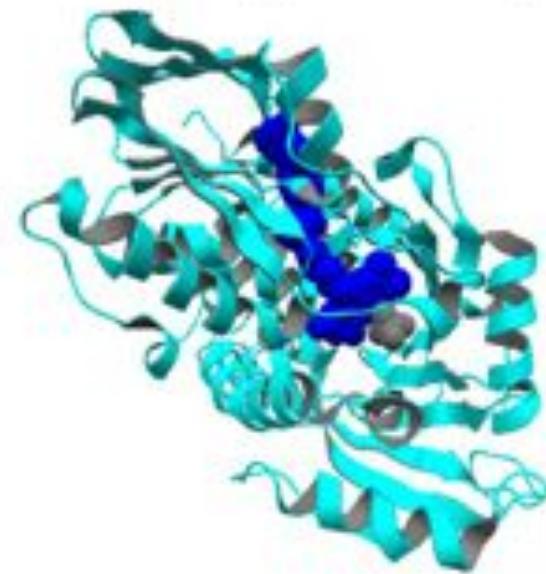


# Same e-motif & fold, different c-motifs & functions

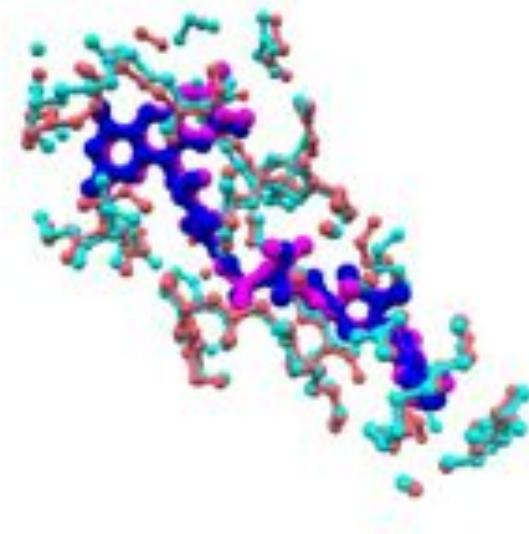
GO (EC 1.4.3.19)



GlpD (EC 1.1.5.3)

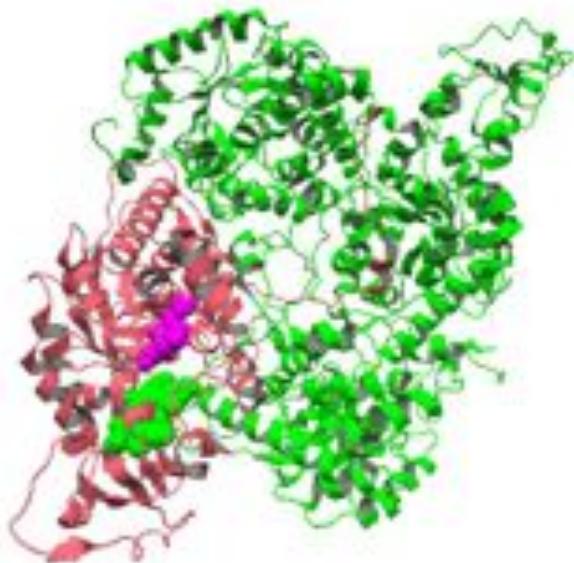


FAD

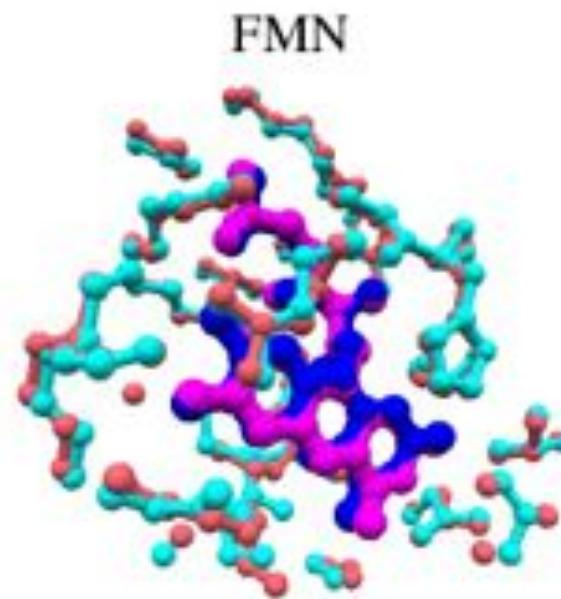
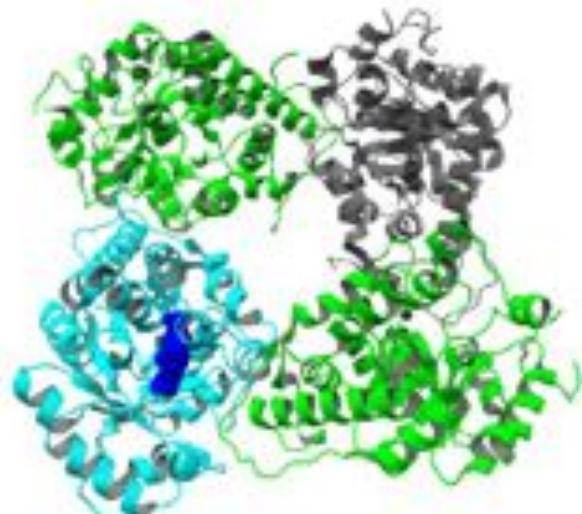


# Same e-motif & fold, different c-motifs & functions

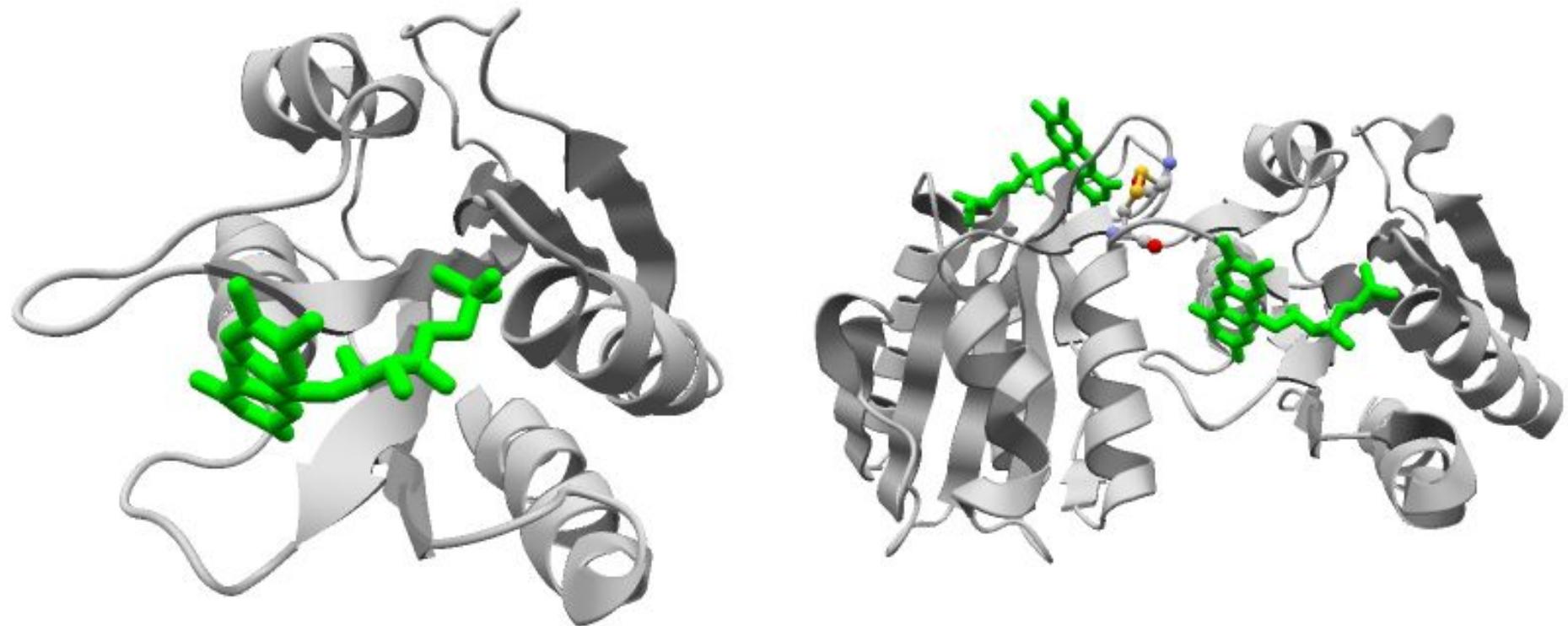
Cyt. b2 (EC 1.1.2.3)



GOX (EC 1.1.3.15)



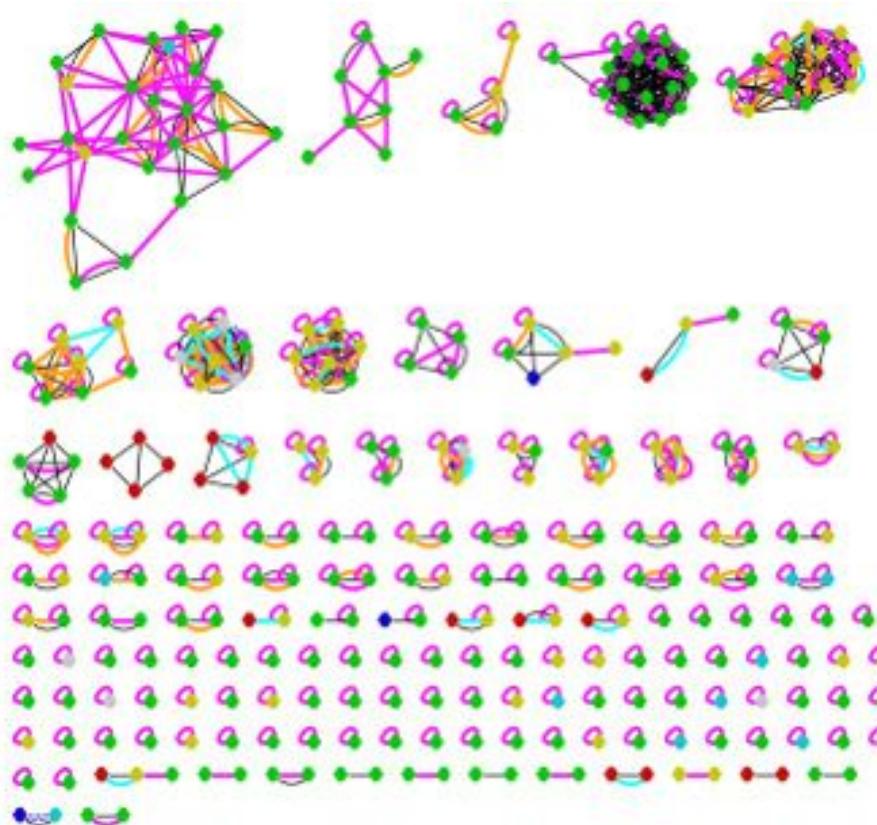
# Flavodoxin: monomer & dimer



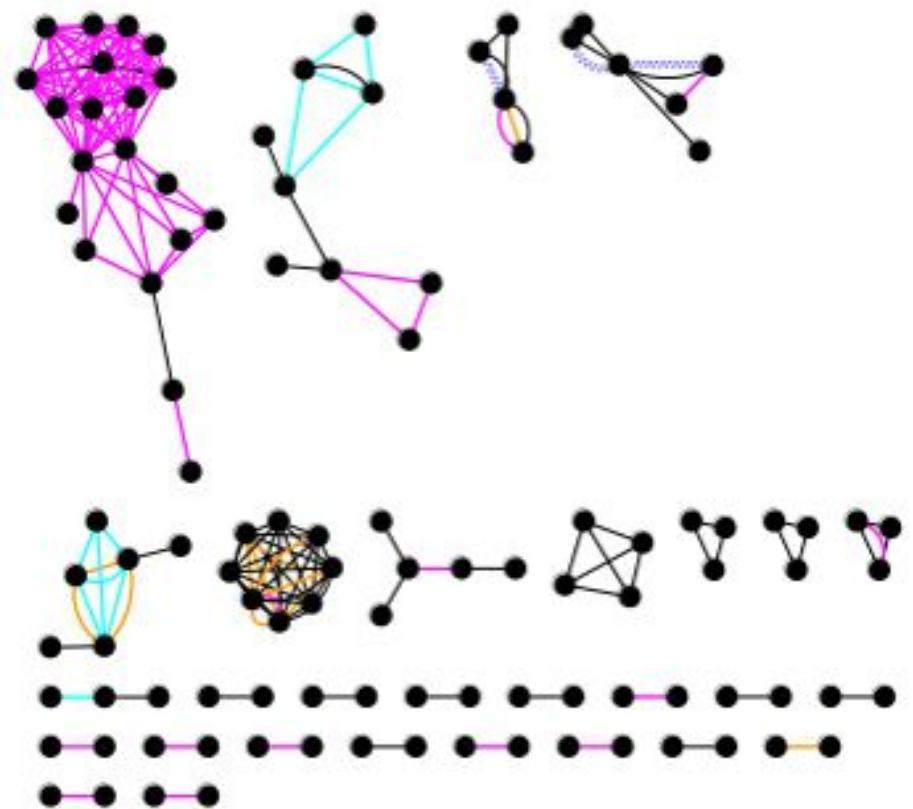
# Meta-motif networks

## “Transcription”

Meta-composite motif

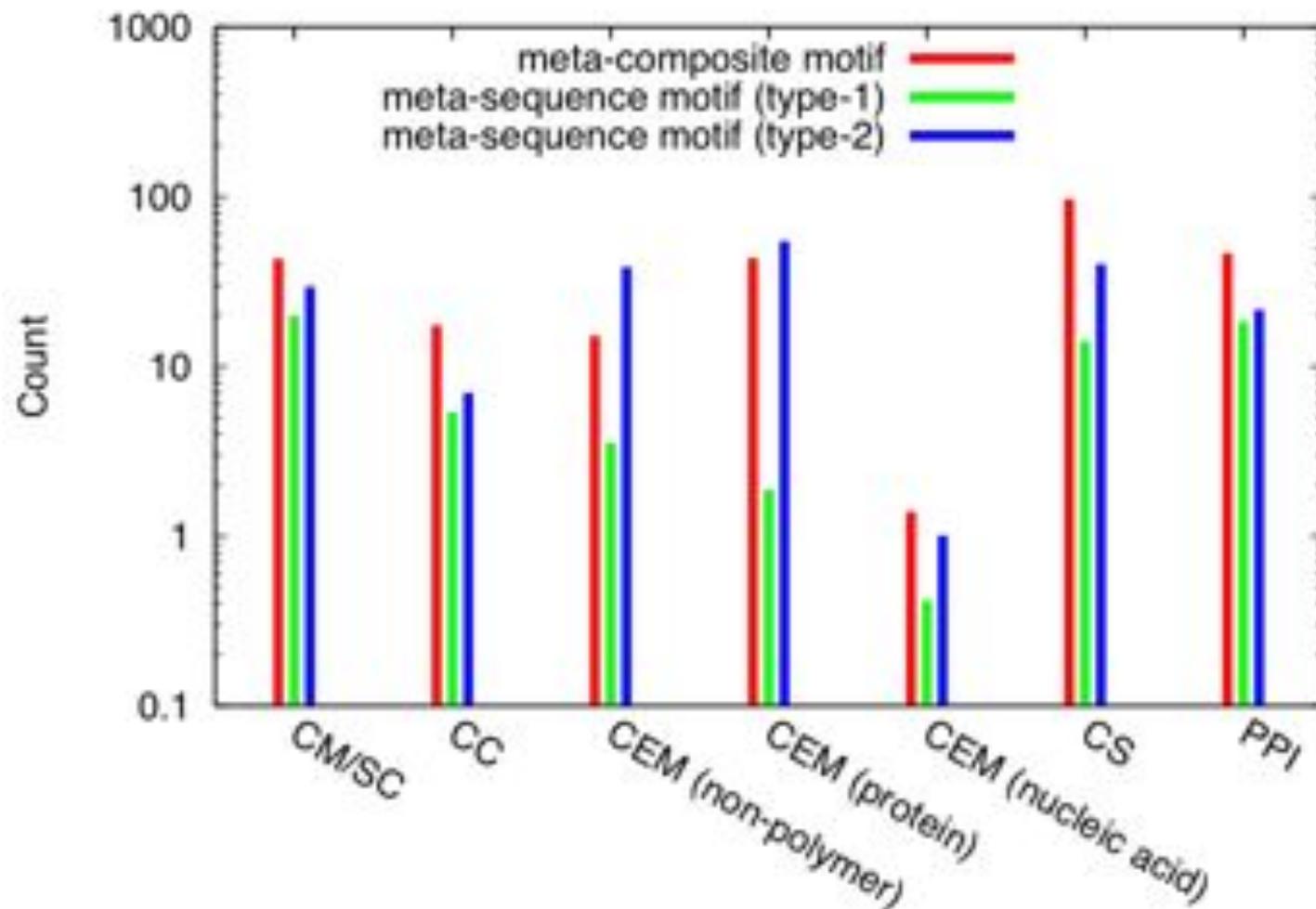


Meta-sequence motif



- node colors indicate interaction states of c-motifs.
- edges colors indicate relations between nodes.

# Network characteristics



type-1 sequence clusters include remote homologs (BLAST e-value < 0.01)  
type-2 sequence clusters consists only of close homologs (100% SID)

# Conclusion

- Composite motifs better distinguish functions
  - different c-motifs  $\Rightarrow$  different functions
  - “Difference” rather than “Similarity”
- Meta-composite motifs provide richer annotations of biological processes.
  - Transitions between interaction/conformational states
  - But time-dependence is still ignored...