Comparison and Classification of Protein Structures

Akira R. KINJO
金城　玲

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

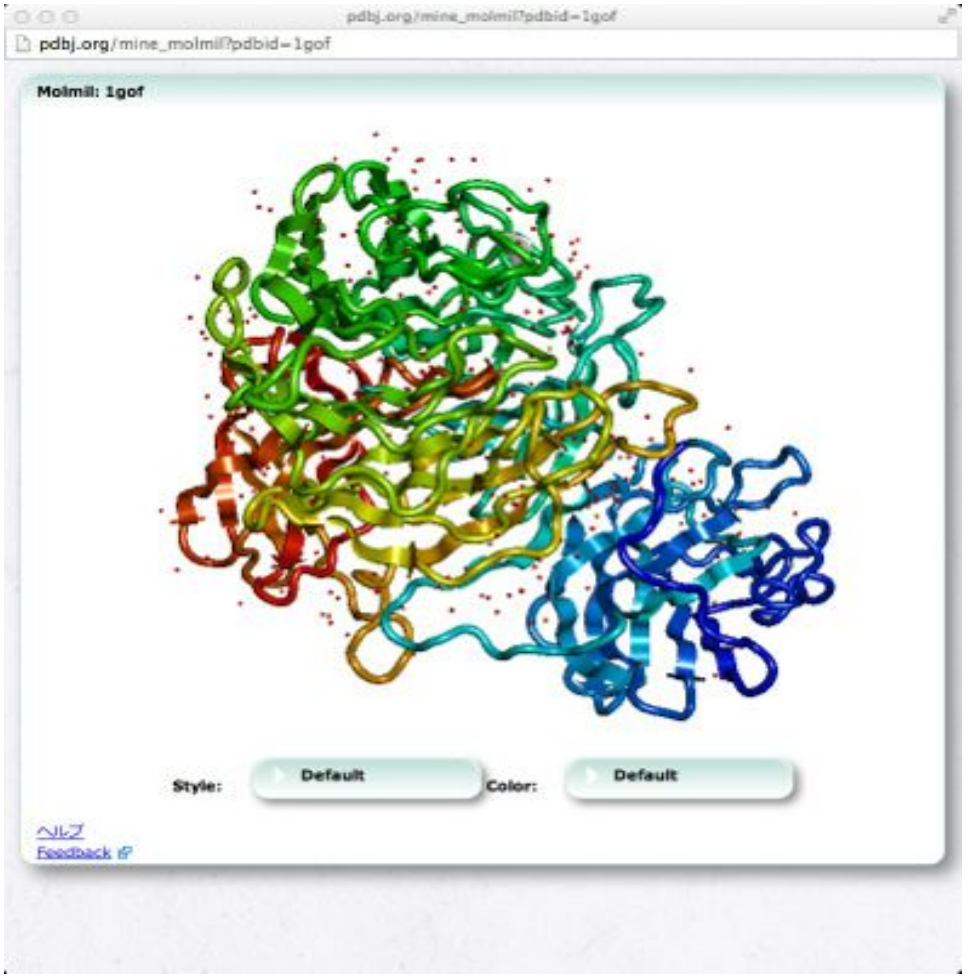
Protein Data Bank (PDB)

Worldwide Protein Data Bank (wwPDB)

Protein Data Bank Japan (PDBj)

The primary database of biological macromolecular structures

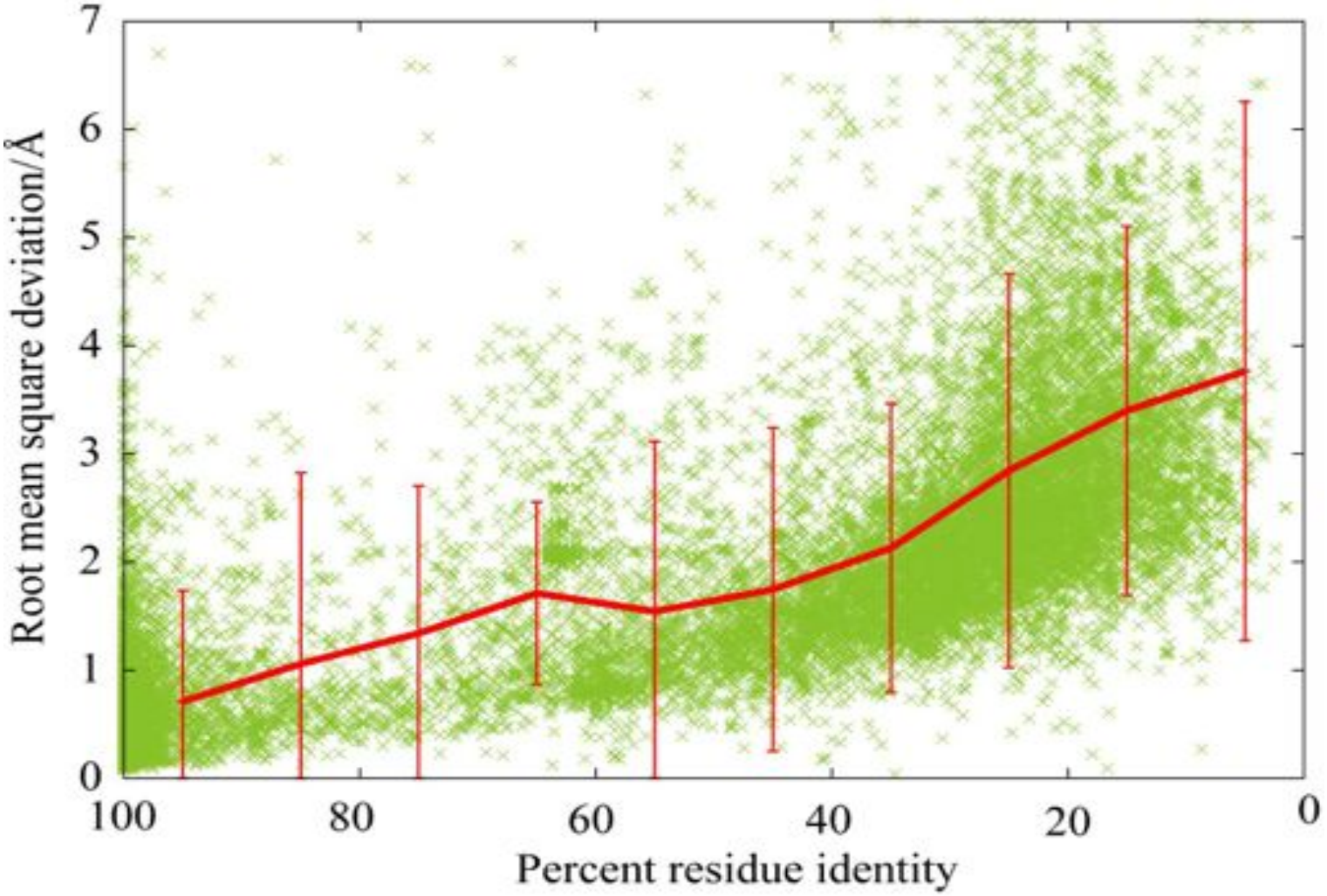
An example of PDB entries



Protein Structure Comparison

```
[BEGIN ALIGNMENT]
     :  E1          H1                  E2         H2          E3
SecA :  EEEEEE  S   HHHHHHHHHHHHHS   SSEEEEEE  GGGHHHHHHH   TTEEEE
   3 :RTFFVGGNFKLNGSKQSIKEIVERLNTASIPENVEVVICPPATYLDYSVSLVKKPQVTVG:  62
      * ****** * ** * *  *    ** *      ***   *  ***            *
   4 :RKFFVGGNWKMNGDKKSLGELIHTLNGAKLSADTEVVCGAPSIYLDFARQKL-DAKIGVA:  62
SecB :  EEEEEE  S   HHHHHHHHHHHHS   TTEEEEEEE GGGHHHHHHHS- TTSEEE
     :  E1          H1                  E2         H2      -    E3

     :                H3        E4    H4        H5              E
SecA :ES   SSSSSS TT   HHHHHHTT  EEEES HHHHHHS   HHHHHHHHHHHHHTT E
  63 :AQNAYLKASGAFTGENSVDQIKDVGAKWVILGHSERRSYFHEDDKFIADKTKFALGQGVG: 122
      *** *    ****** *   *** ** *****  ***  * * *  *  *   **  * *
  63 :AQNCYKVPKGAFTGEISPAMIKDIGAAWVILGNPERRHVFGESDELIGQKVAHALAEGLG: 122
SecB :ES   SSSSSS TT   HHHHHHHT  EEEES HHHHHTS   HHHHHHHHHHHHHTT E
     :                H3        E4    H4        H5              E

     :5       H6      H7                  E6                H8
SecA :EEEEE   HHHHHTT HHHHHHHHHHHHHHH S  TTEEEEE  GGGTTTS    HHHHH
 123 :VILCIGETLEEKKAGKTLDVVERQLNAVLEEVKDWTNVVVAYEPVWAIGTGLAATPEDAQ: 182
      ** **** * *  ** *  **  *  *    ****  ** **********  ***  **
 123 :VIACIGEKLDEREAGITEKVVFEQTKAIADNVKDWSKVVLAYEPVWAIGTGKTATPQQAQ: 182
SecB :EEEEE   HHHHHTT HHHHHHHHHHHHHTT S GGGEEEEE  GGGTTTS    HHHHH
     :5       H6      H7                  E6                H8

     :               H9      E7                   E8          H10
SecA :HHHHHHHHHHHHHH HHHHHH  EEEESS  TTTGGGGTT TT  EEEESGGGGSTTHHH
 183 :DIHASIRKFLASKLGDKAASELRILYGGSANGSNAVTFKDKADVDGFLVGGASLKPEFVD: 242
        *   *  *     *  *   ** ****  * *        *****************
 183 :EVHEKLRGWLKTHVSDAVAQSTRIIYGGSVTGGNCKELASQHDVDGFLVGGASLKPEFVD: 242
SecB :HHHHHHHHHHHHHT HHHHHH  EEEESS  TTTHHHHHTSTT  EEEESGGGGSTTHHH
     :               H9      E7         H10        E8          H11

     :
SecA :HHHTT
 243 :IINSRN: 248
      ***
 243 :IINAKH: 248
SecB :HHTTT
     :
```

Sequence & structure similarities

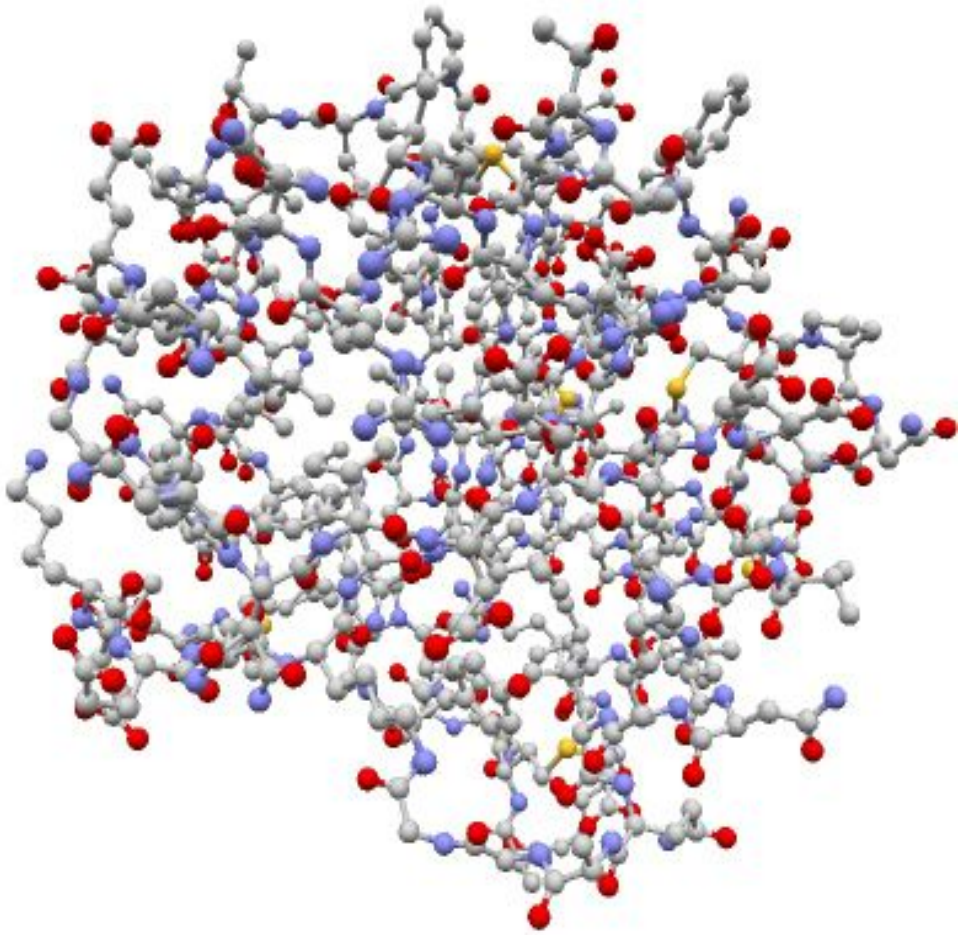


（『タンパク質の立体構造入門』図 4.1 ）

Methods of structure comparison

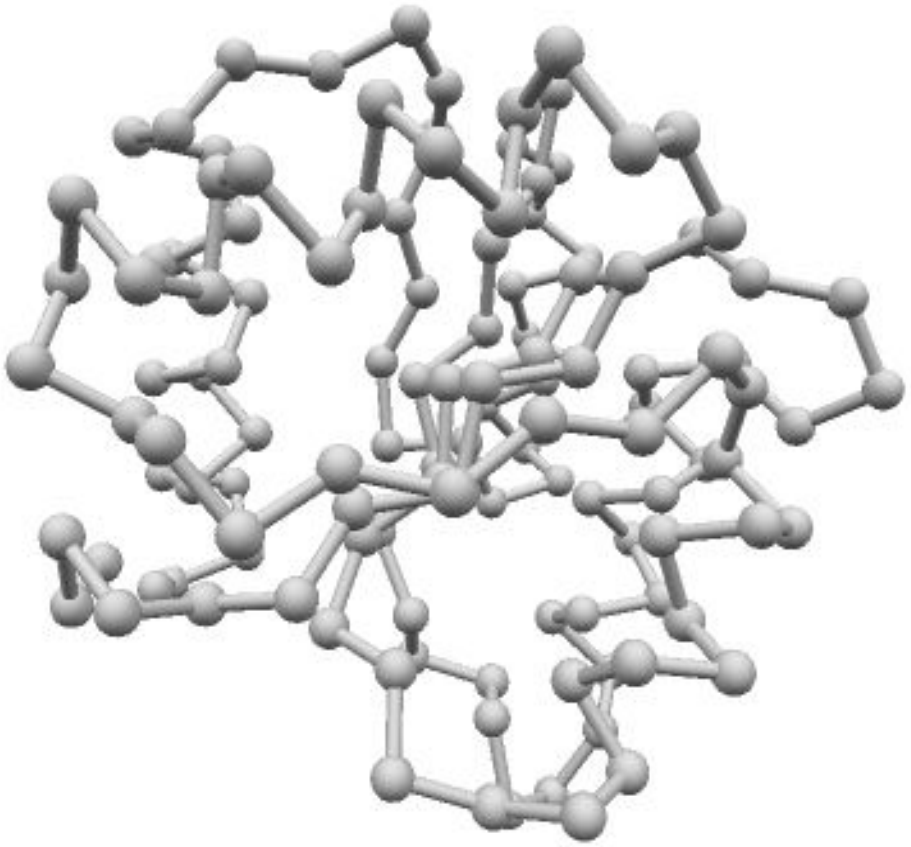
- Visual inspection(!)
  - The “best” method if you are well-trained.
- Algorithms
  - It's an NP-hard problem, so there are a number of approximate methods based on various representations:
    - secondary structure elements (SSE)
    - Amino acid residues
    - Atoms
    - Molecular surface

Representation: all atoms



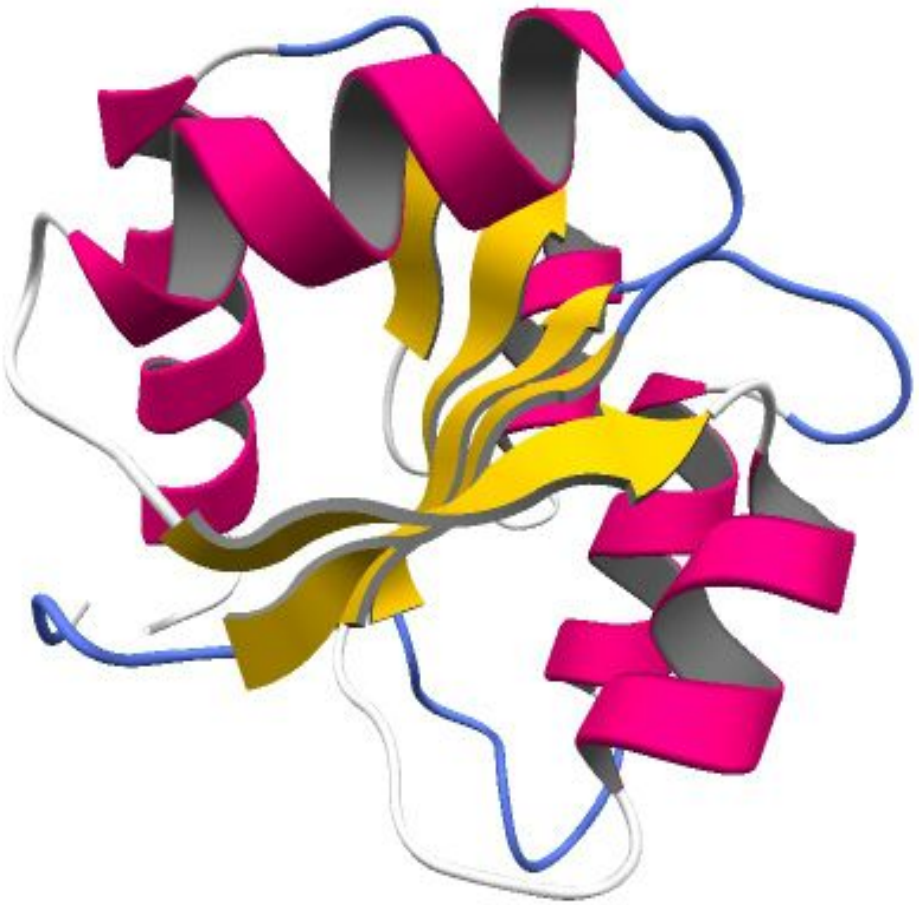
- Dealing with all atoms...
- is difficult. So usually only substructures are treated in this representation.

Representation: Backbone



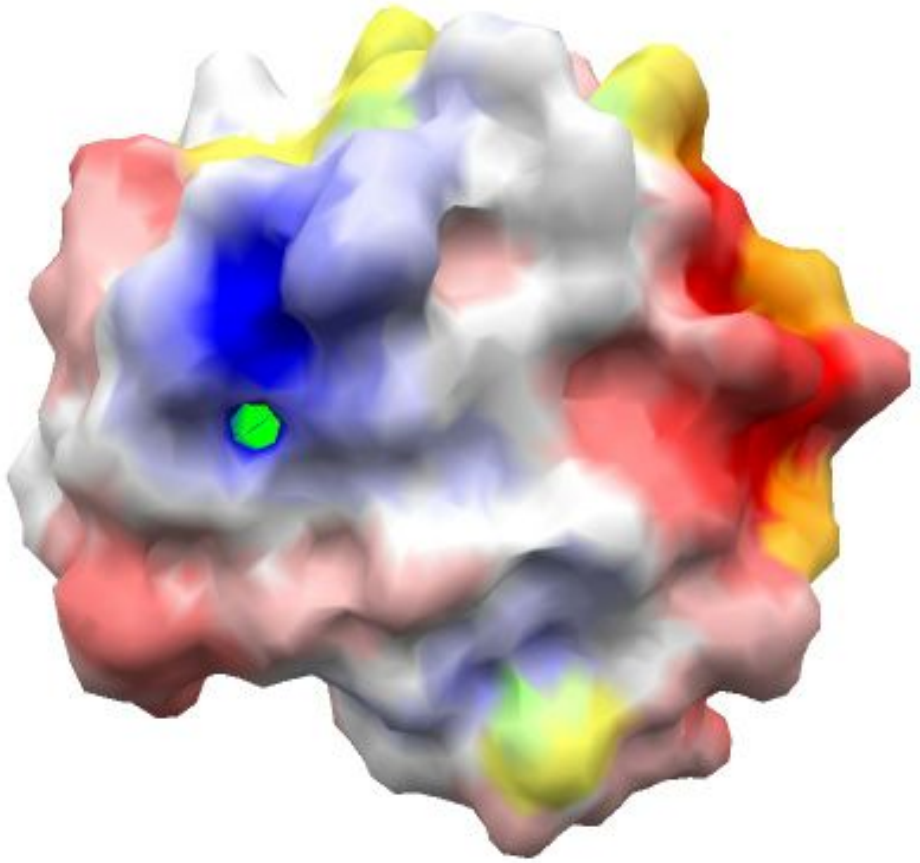
- Using only Cα or Cβ atoms to reduce computational costs.
- Also compatible with sequence alignment (1 atom / residue)
- Still computationally demanding.

Representation: 2ndary structures



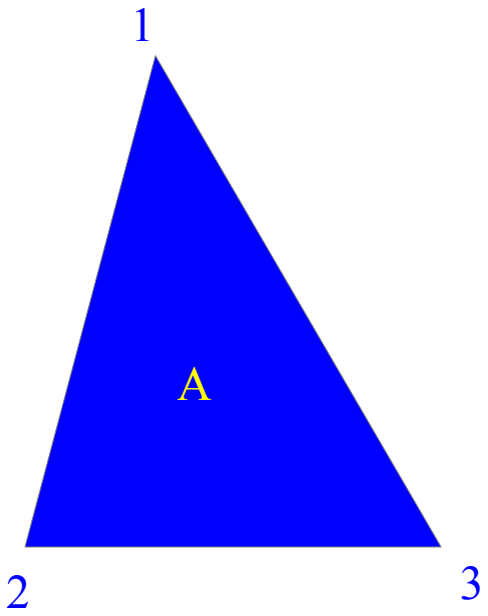
- α helices and β strands as vectors.
- Suitable for finding topological similarities.
- Less cost.

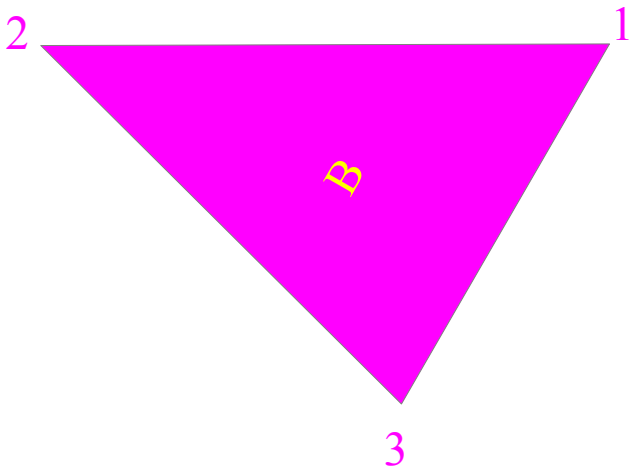
Representation: Molecular surface



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

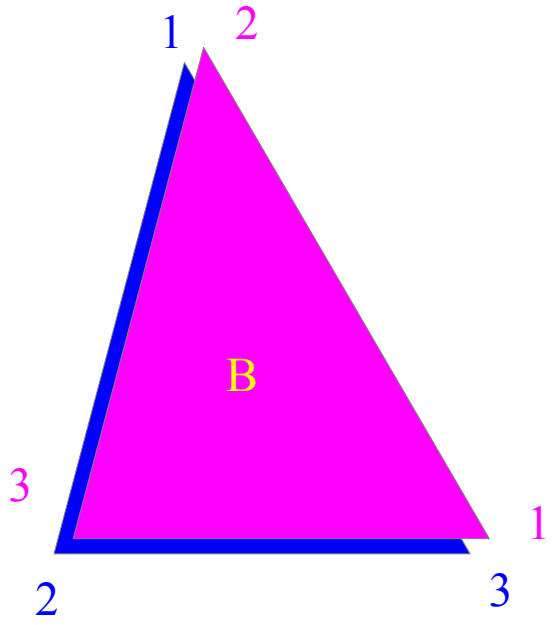
Basic ideas





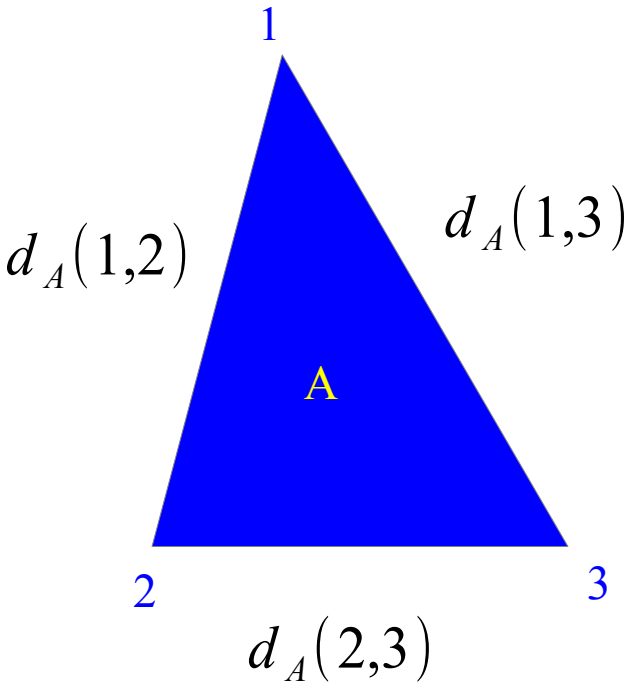
How do you tell the congruence of two triangles?
(Vertex numbers do not match!)

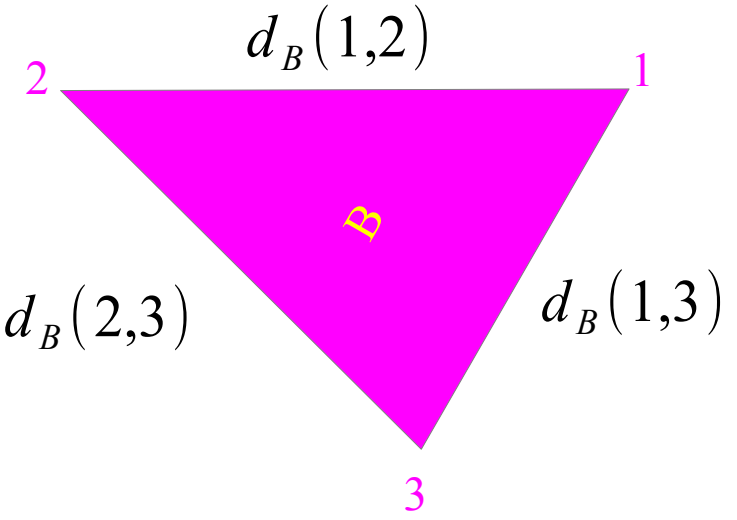
Method 1: Coordinate-based



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

Method 2: Distance-based





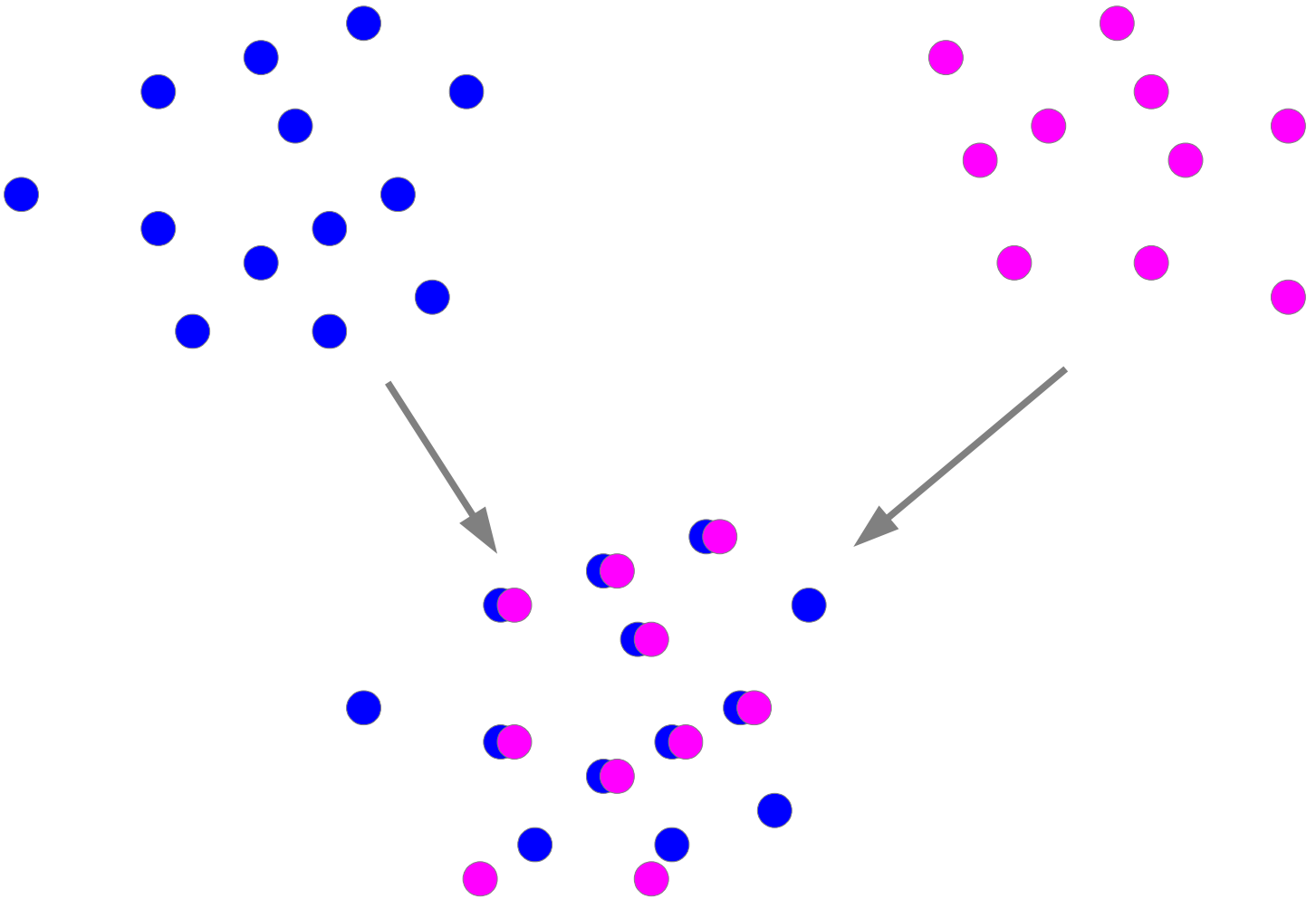
	$d_A(1,2)$	$d_A(1,3)$
$d_A(2,1)$		$d_A(2,3)$
$d_A(3,1)$	$d_A(3,2)$	

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

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:

$$A = \left(x^A_{1,} x^A_{2,} \ldots, x^A_M \right) \qquad B = \left(x^B_{1,} x^B_{2,} \ldots, x^B_N \right)$$

$$A \xrightarrow{f} C$$
$$B \xrightarrow{g} C$$

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.

$$D(A, B) = \sum_{(i,\, j)} d_C\left(f\left(x^A_i \right), g\left(x^B_j \right) \right)$$

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.

$$A=\left(x_{1,}^{A}\, x_{2,}^{A}, \ldots, x_{M}^{A}\right) \qquad B=\left(x_{1,}^{B}\, x_{2,}^{B}, \ldots, x_{M}^{B}\right) \qquad \text{(the same number of points)}$$

$$For\ all\ i=1,\cdots, M, \left(x_{i}^{A}, x_{i}^{B}\right) \qquad \text{(The i-th atom in A <=> The i-th atom in B)}$$

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

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

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

$$A = \left(x^A_{1,} x^A_{2,} \ldots, x^A_M \right)$$

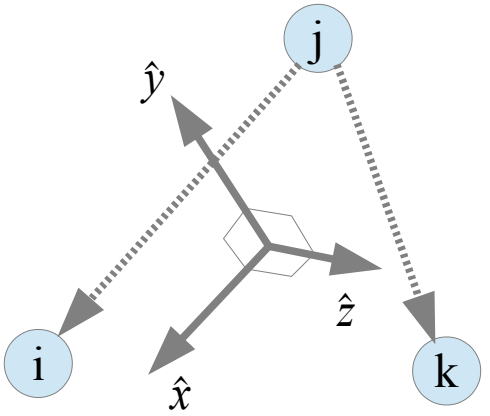
$\left(x^A_i, x^A_j, x^A_k \right)$ 3 points

$$\hat{x} = \frac{1}{\| x^A_i - x^A_j \|} \left(x^A_i - x^A_j \right)$$ X axis

$$\hat{y} = \frac{1}{\| x^A_k - x^A_j \|} \hat{x} \times \left(x^A_k - x^A_j \right)$$ Y axis

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

$$O = \frac{1}{3} \left(x^A_i + x^A_j + x^A_k \right)$$ Origin

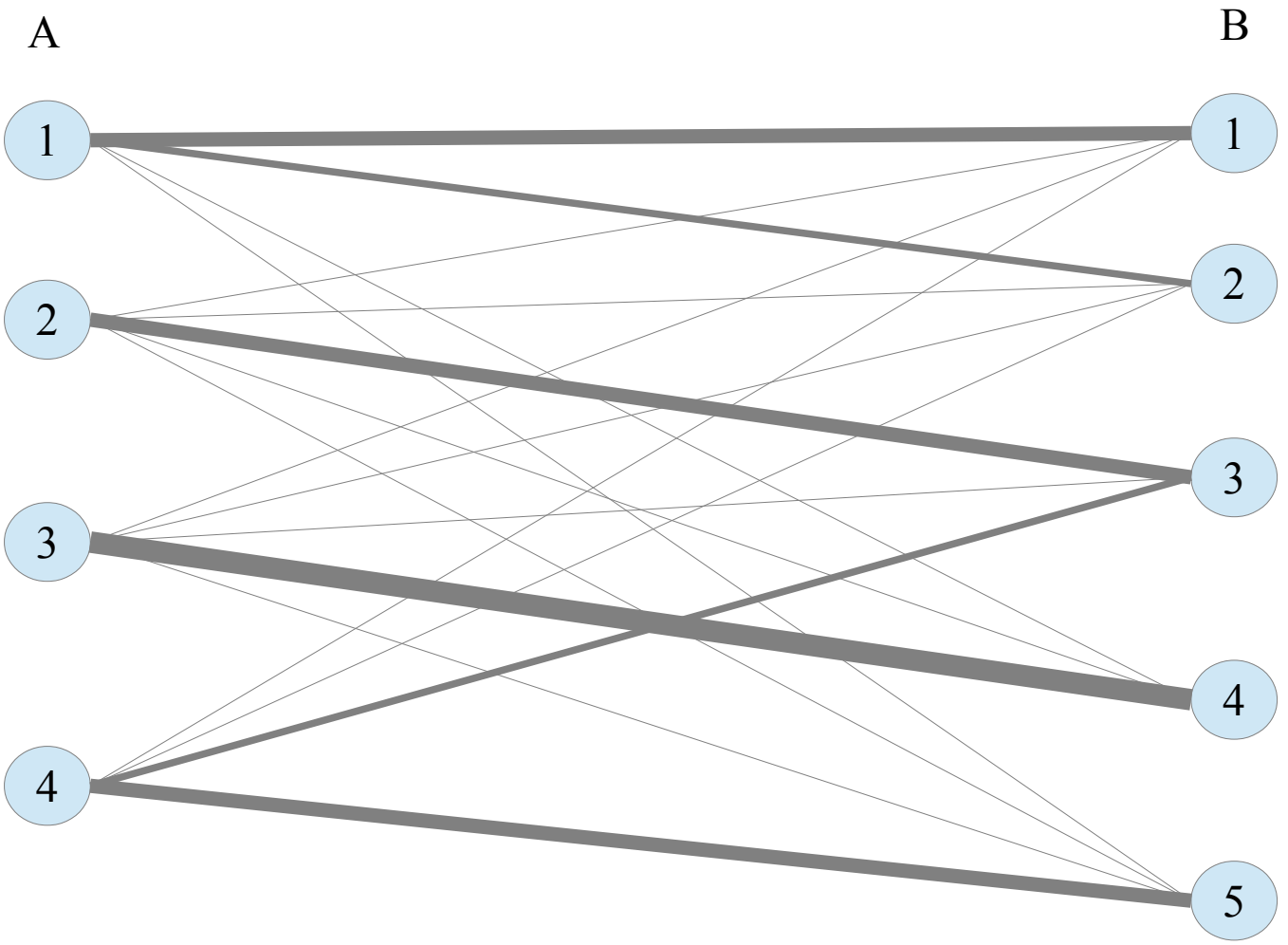


$$\hat{x}^A_a = \left(\hat{x} \cdot \left(x^A_a - O \right), \hat{y} \cdot \left(x^A_a - O \right), \hat{z} \cdot \left(x^A_a - O \right) \right), a = 1, \ldots, M$$ Transformation

Simple superposition algorithm

```
Input: Structure A=x(1)..x(M); Structure B=y(1)..y(N)
Output: Best alignment Ali
Ali := {}   --- 初期アラインメント(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

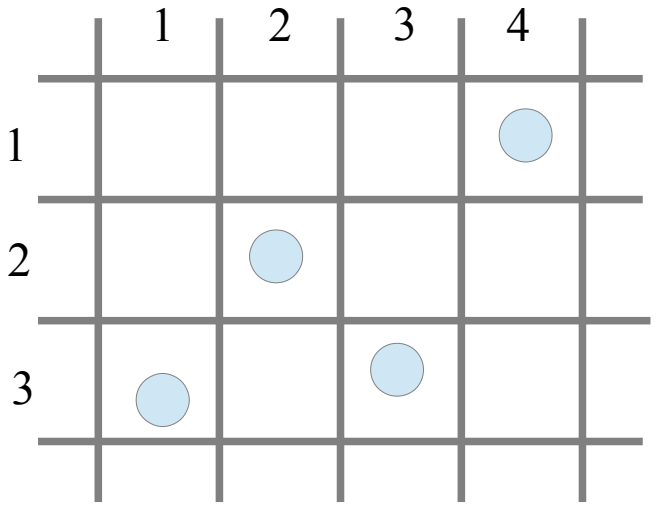


Geometric Hashing (GH)

- The simple approach is simply too slow.
- Make a dictionary (hash table) x' -> 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)..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)
```

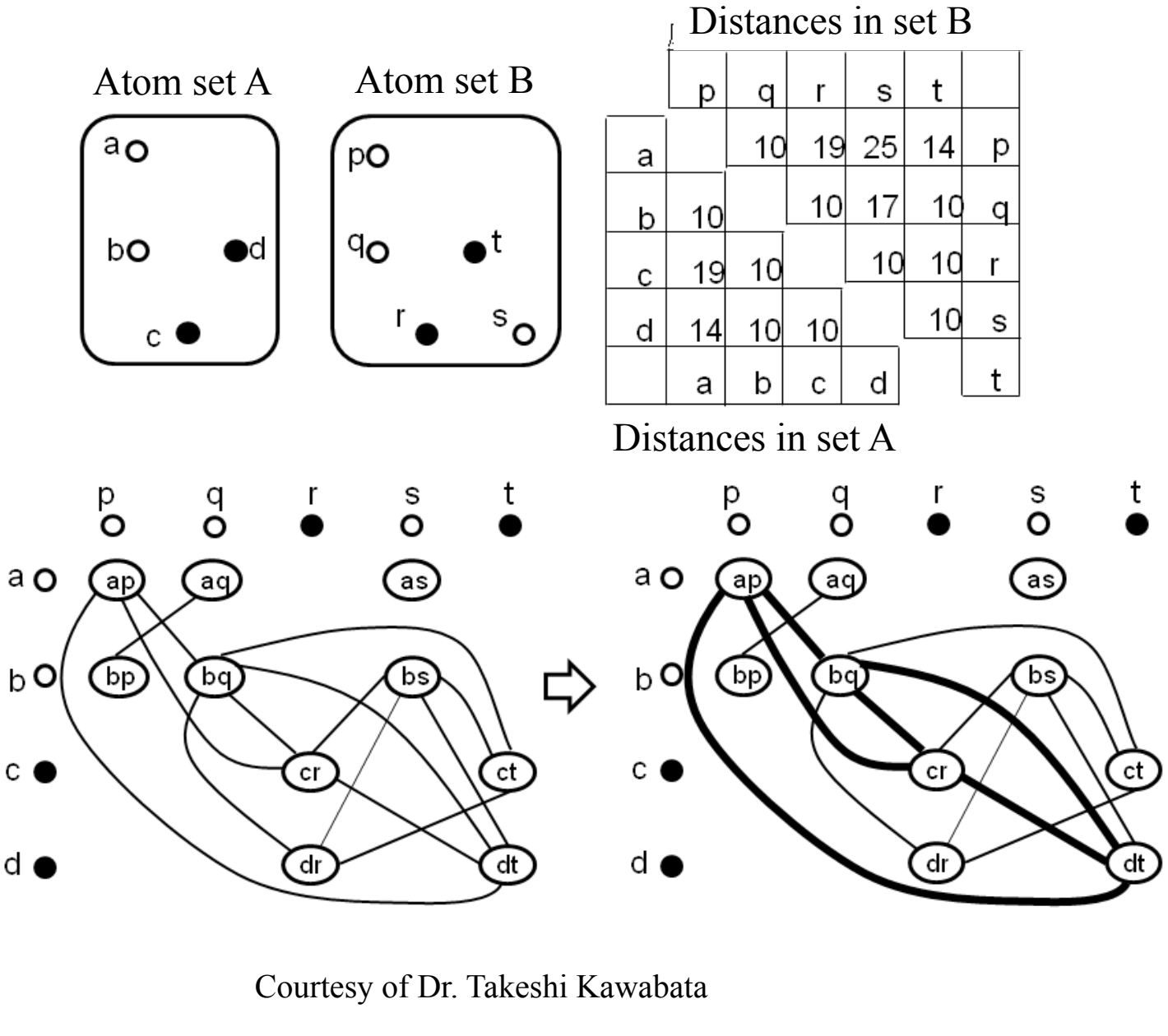
This requires $N^2(N-1)(N-2)$ steps.

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

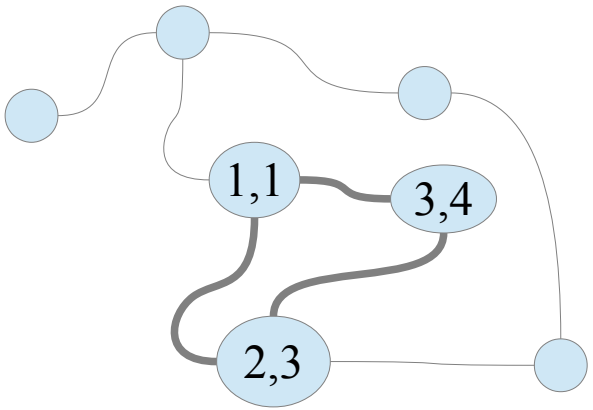
$A = (x^A_1, \ldots, x^A_M)$
$B = (x^B_1, \ldots, x^B_N)$

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

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).

$$\left| \left\| x^A_i - x^A_k \right\| - \left\| x^B_j - x^B_l \right\| \right| < \delta$$

Find the subgraph of thus created graph, that is complete and max imum:
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”.

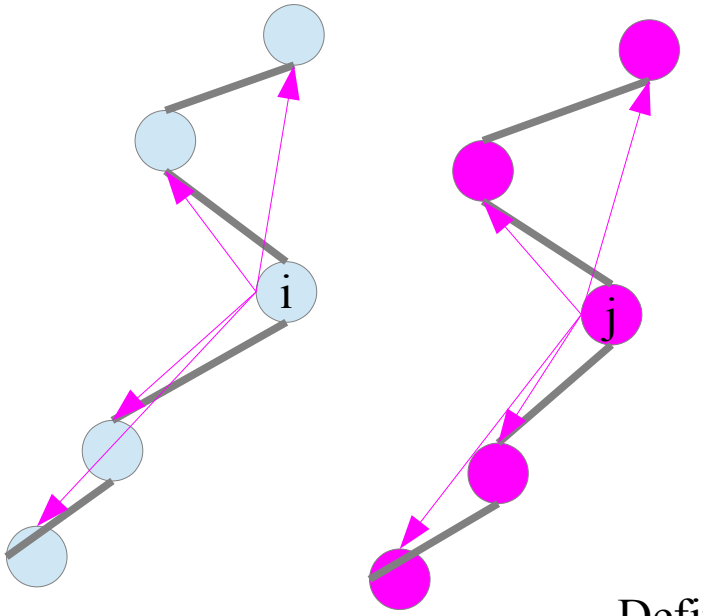
From http://en.wikipedia.org/wiki/Bron–Kerbosch_algorithm

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



$$A=(x^A_{1,}\ldots,x^A_M)$$
$$B=(x^B_{1,}\ldots,x^B_N)$$

Assume (x^A_i, x^B_j) 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):

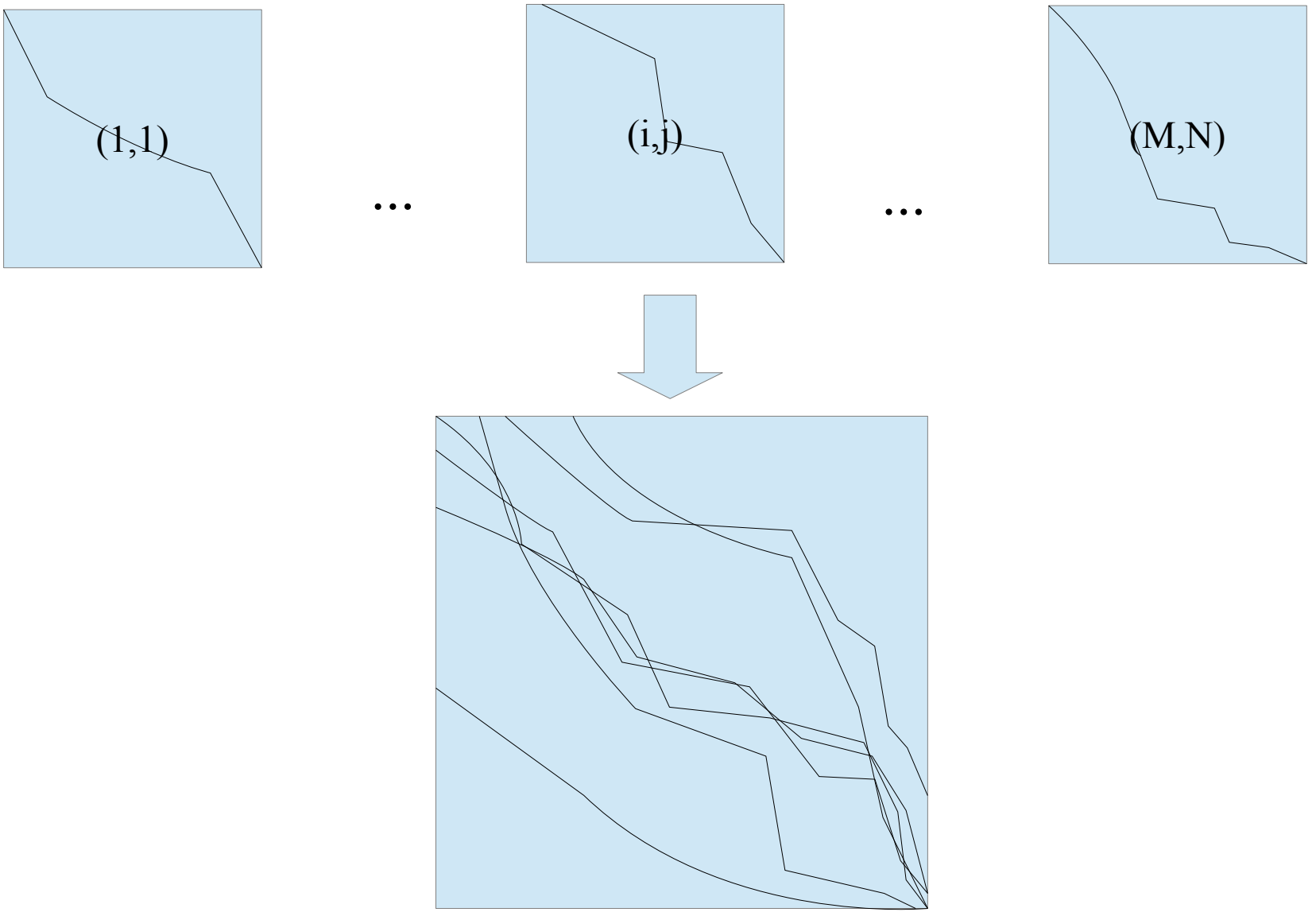
$$s(k,l;i,j)=\frac{1}{|d_A(i,k)-d_B(j,l)|+c}$$

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.

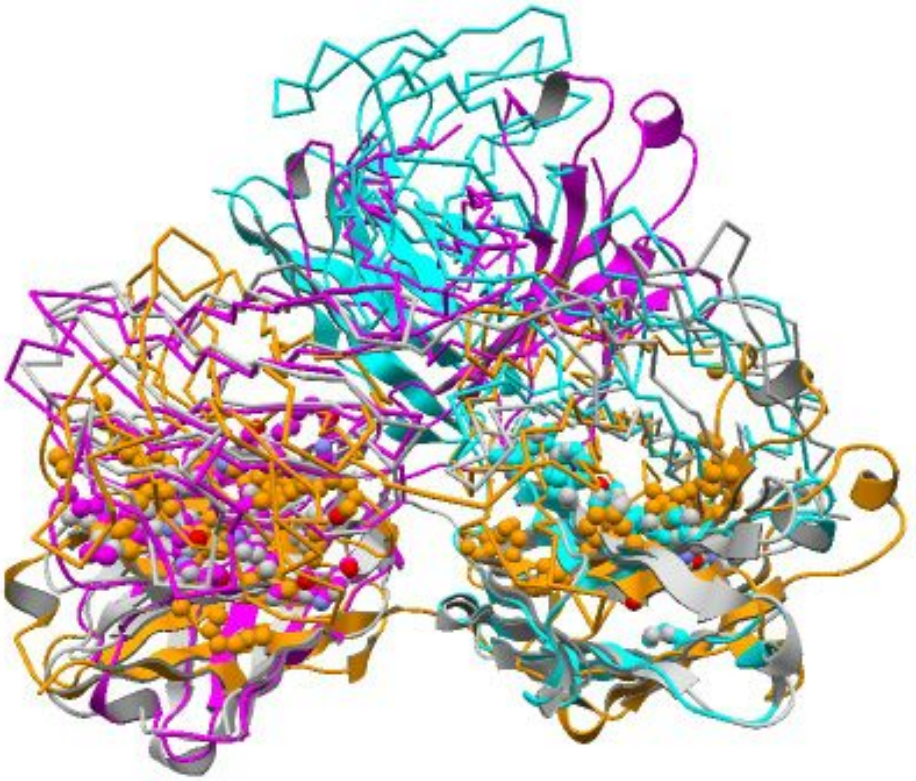
Summary

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

Fast search and flexible alignment of interaction site structures in proteins

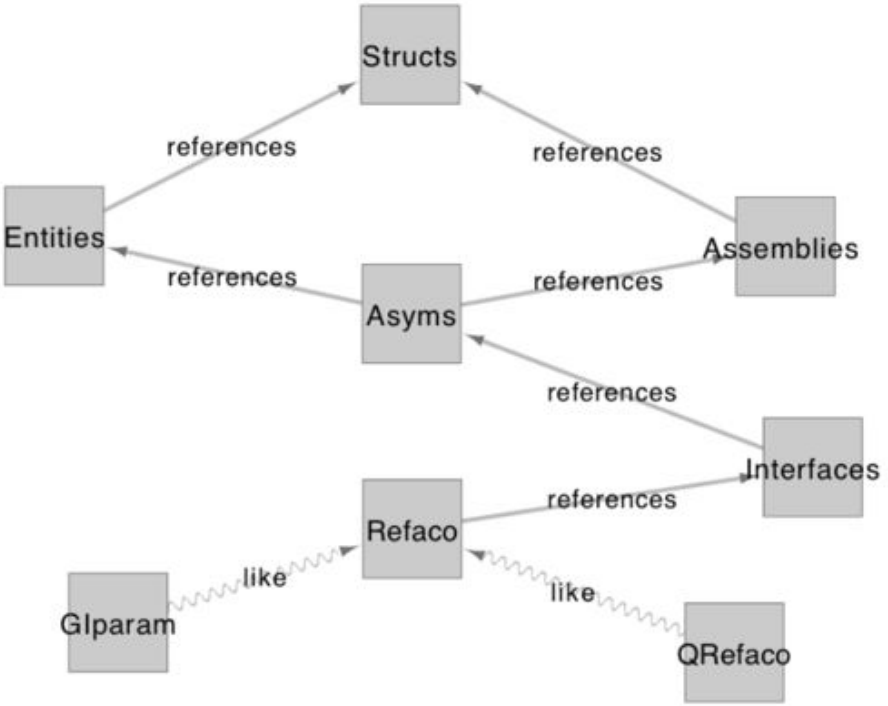
Objectives

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



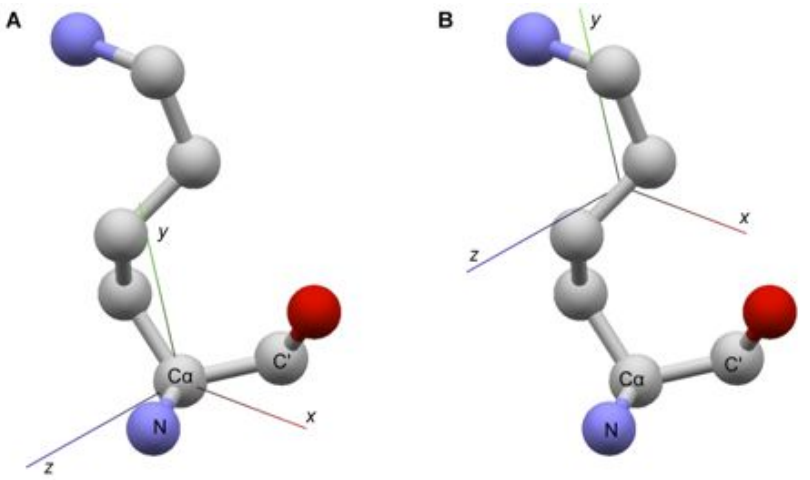
Database design

- 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, Cα, C' atoms.
- Defined for each residue in an interface.

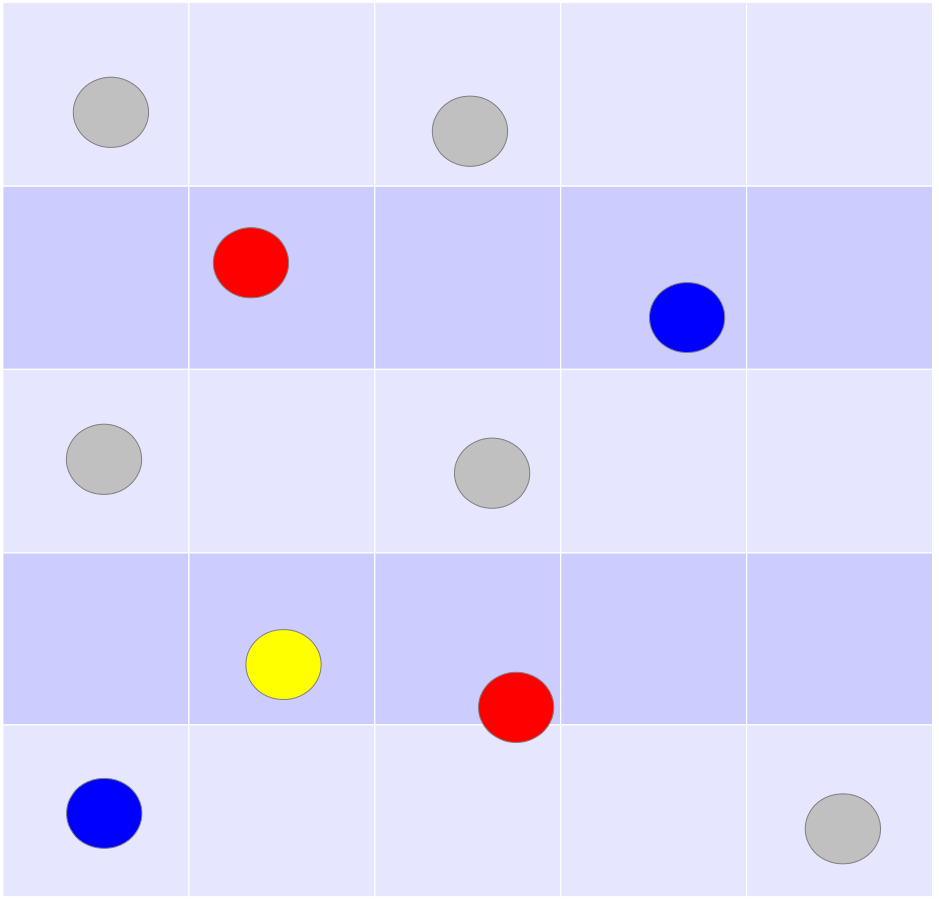


Structure search by SQL

```
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 */
 );
```

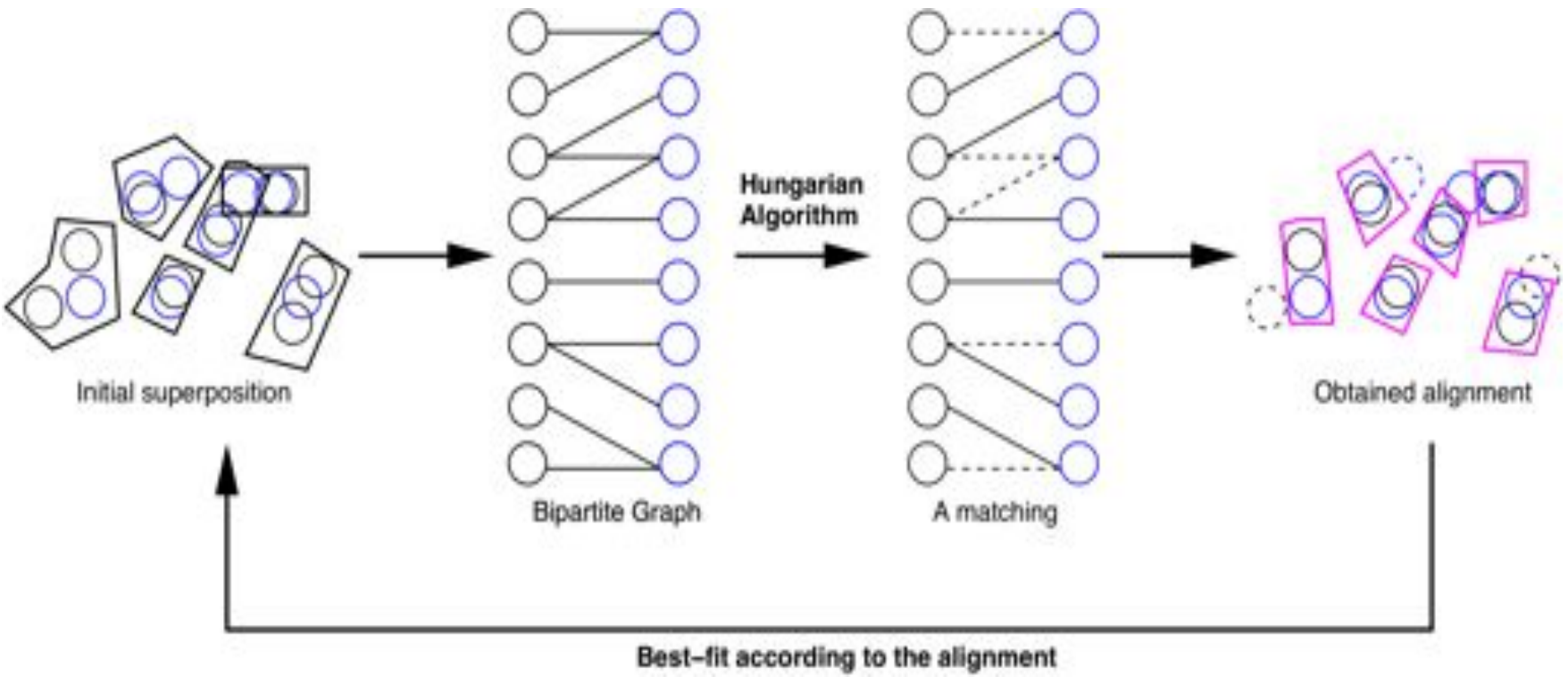
```
 1: SELECT t.if_id, t.rs_id, t.type, t.frame, q.rs_id
 2: FROM  Refaco t, Qrefaco q
 3: WHERE t.ft01 BETWEEN q.ft01 - D01 AND q.ft01 + D01
 4: AND   t.ft02 BETWEEN q.ft02 - D02 AND q.ft02 + D02
    ...
 5: AND   t.ft44 BETWEEN s.ft44 - D44 AND q.ft44 + D44
 6: AND   (SELECT COUNT(*)
 7:         FROM (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.

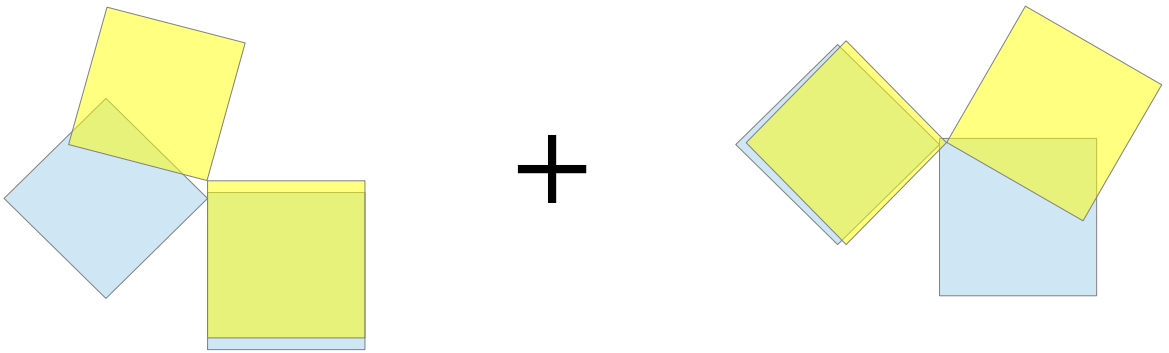
Alignment by Hungarian method



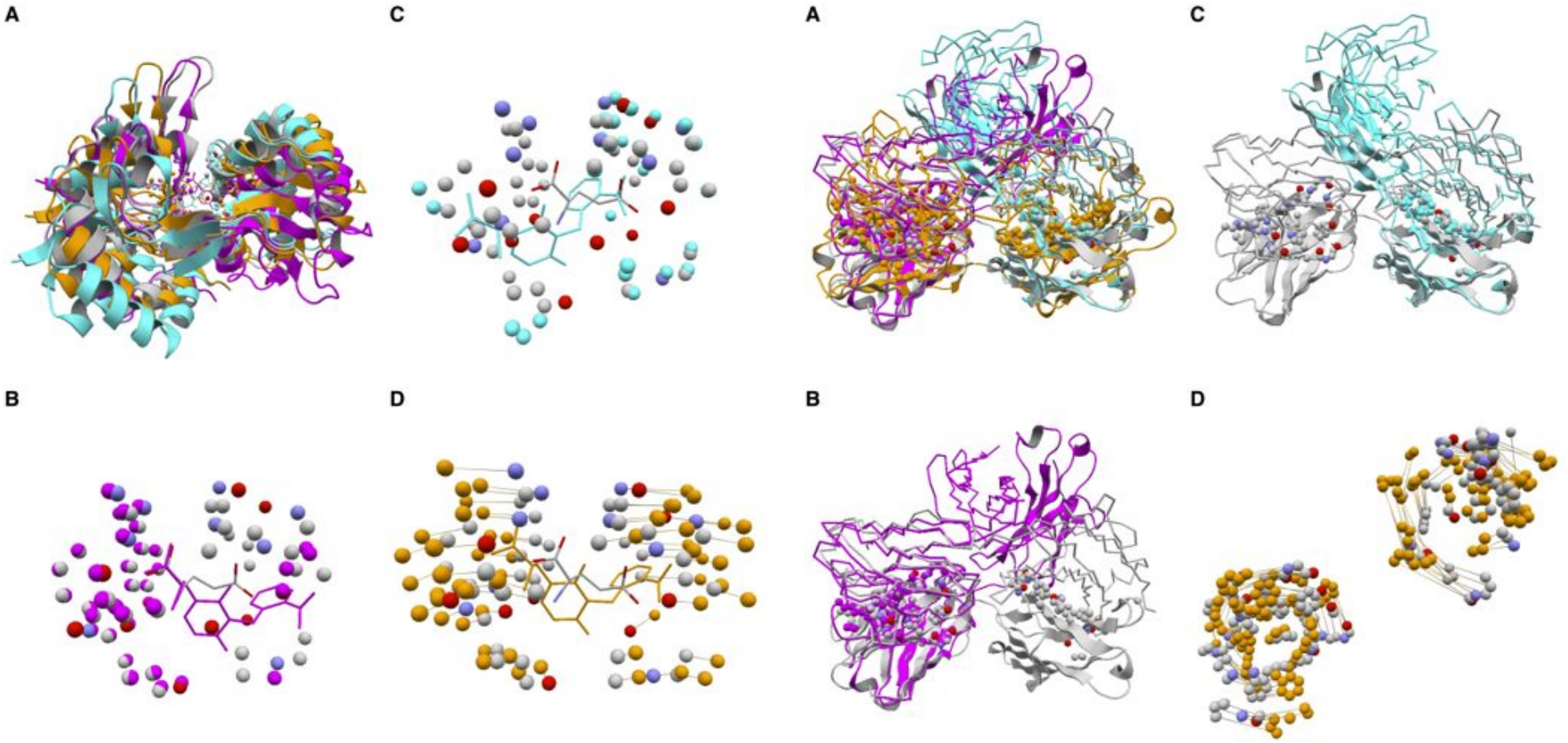
Flexible alignment

- 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

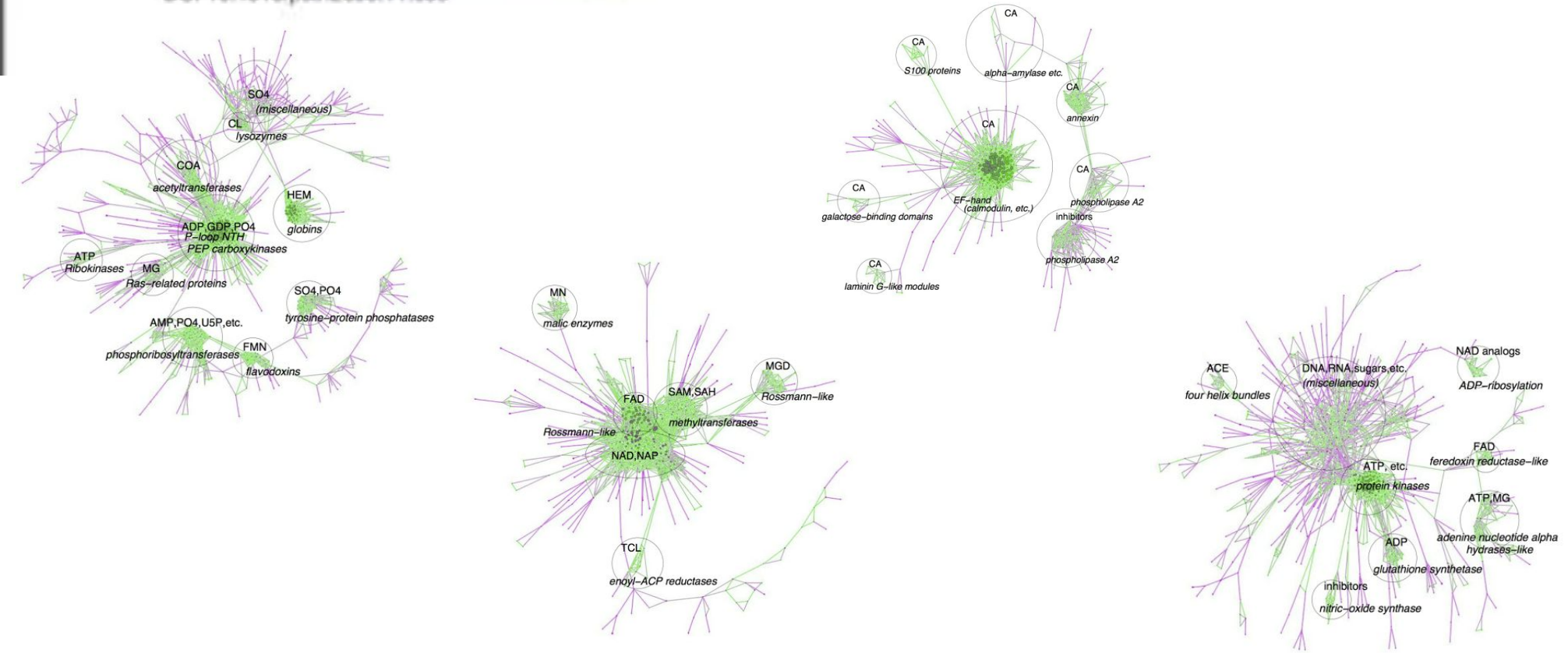
Cell PRESS

Structure
Article

Structure **17**:234-246 (2009)

Comprehensive Structural Classification of Ligand-Binding Motifs in Proteins

Akira R. Kinjo[1,*] and Haruki Nakamura[1]
[1]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



doi:10.1016/j.jmb.2010.04.028 J. Mol. Biol. (2010) 399, 526–540

JMB

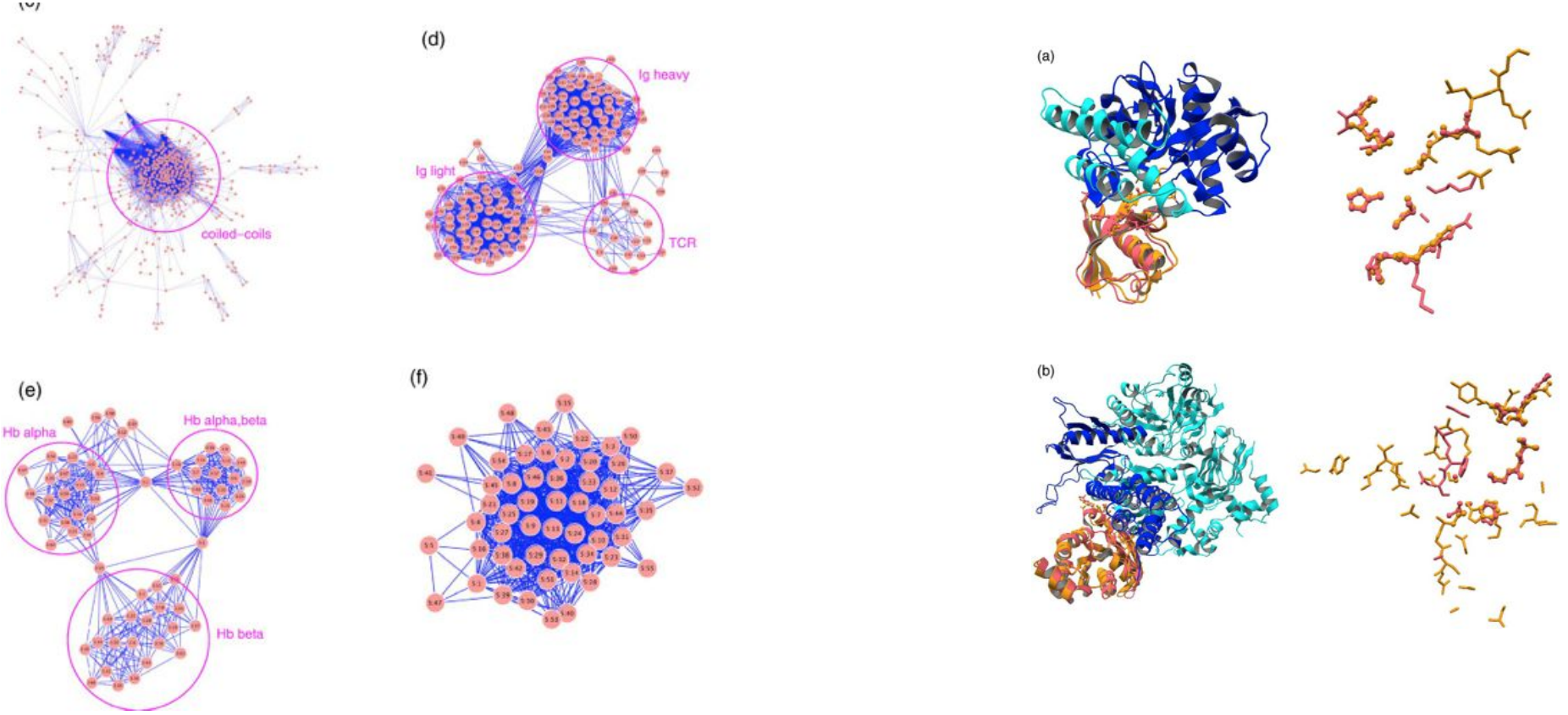
Available online at www.sciencedirect.com

ScienceDirect

ELSEVIER

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



doi:10.1016/j.jmb.2011.06.027 J. Mol. Biol. (2011) 411, 713–727

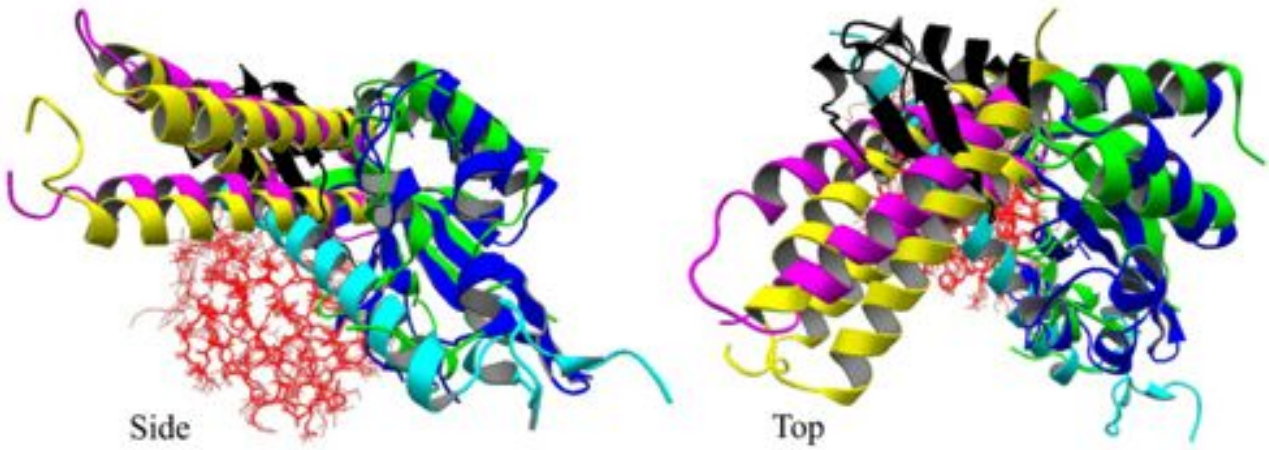
Contents lists available at www.sciencedirect.com

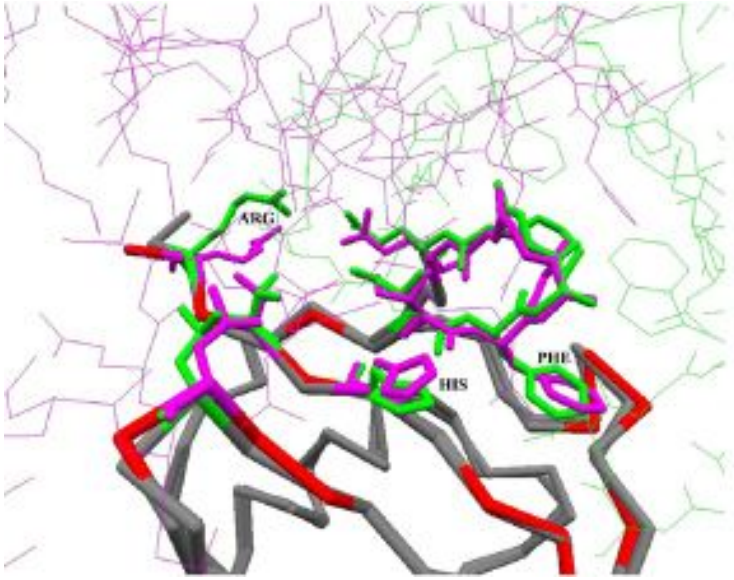
Journal of Molecular Biology

journal homepage: http://ees.elsevier.com.jmb

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

Bhaskar Dasgupta[1,2*], Haruki Nakamura[1] and Akira R. Kinjo[1]





OPEN ACCESS Freely available online

PLoS one

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	ATP + AMP = 2 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

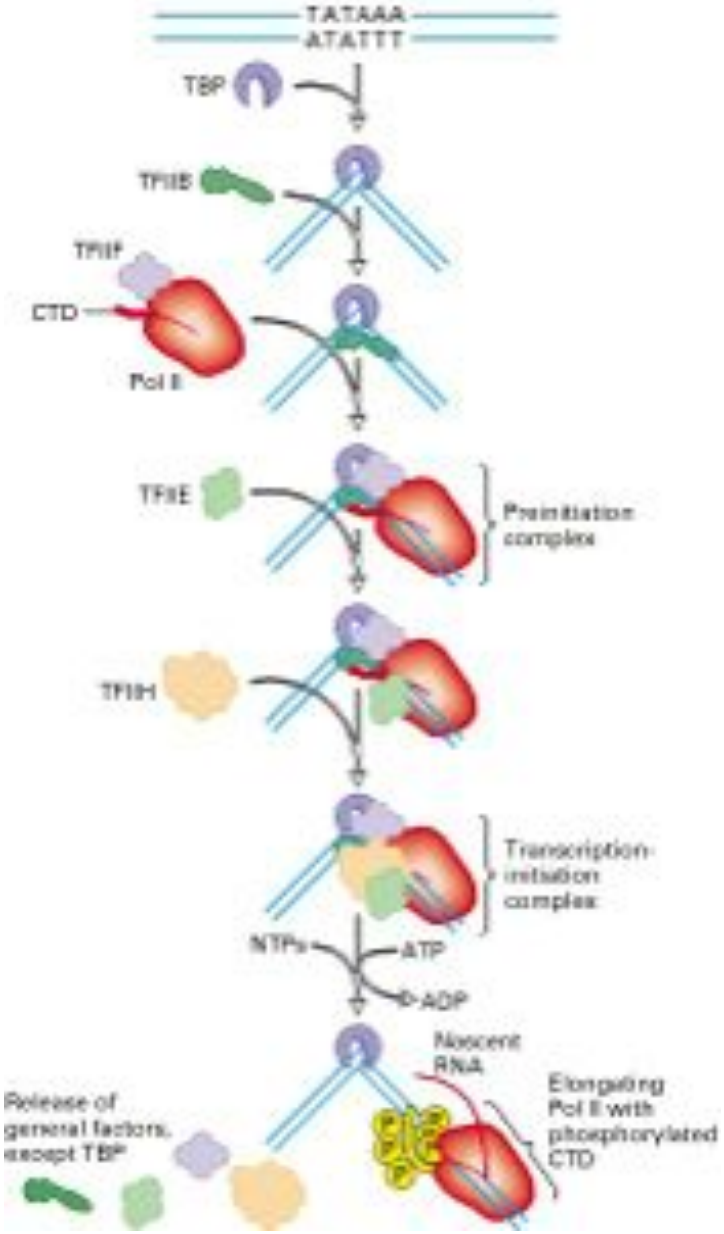
ID	GO:0004017
Name	adenylate kinase activity
Ontology	Molecular Function
Definition	Catalysis of the reaction: ATP + AMP = 2 ADP.
Comment	
Secondary IDs	

GONUTS Wiki Page

Synonyms

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

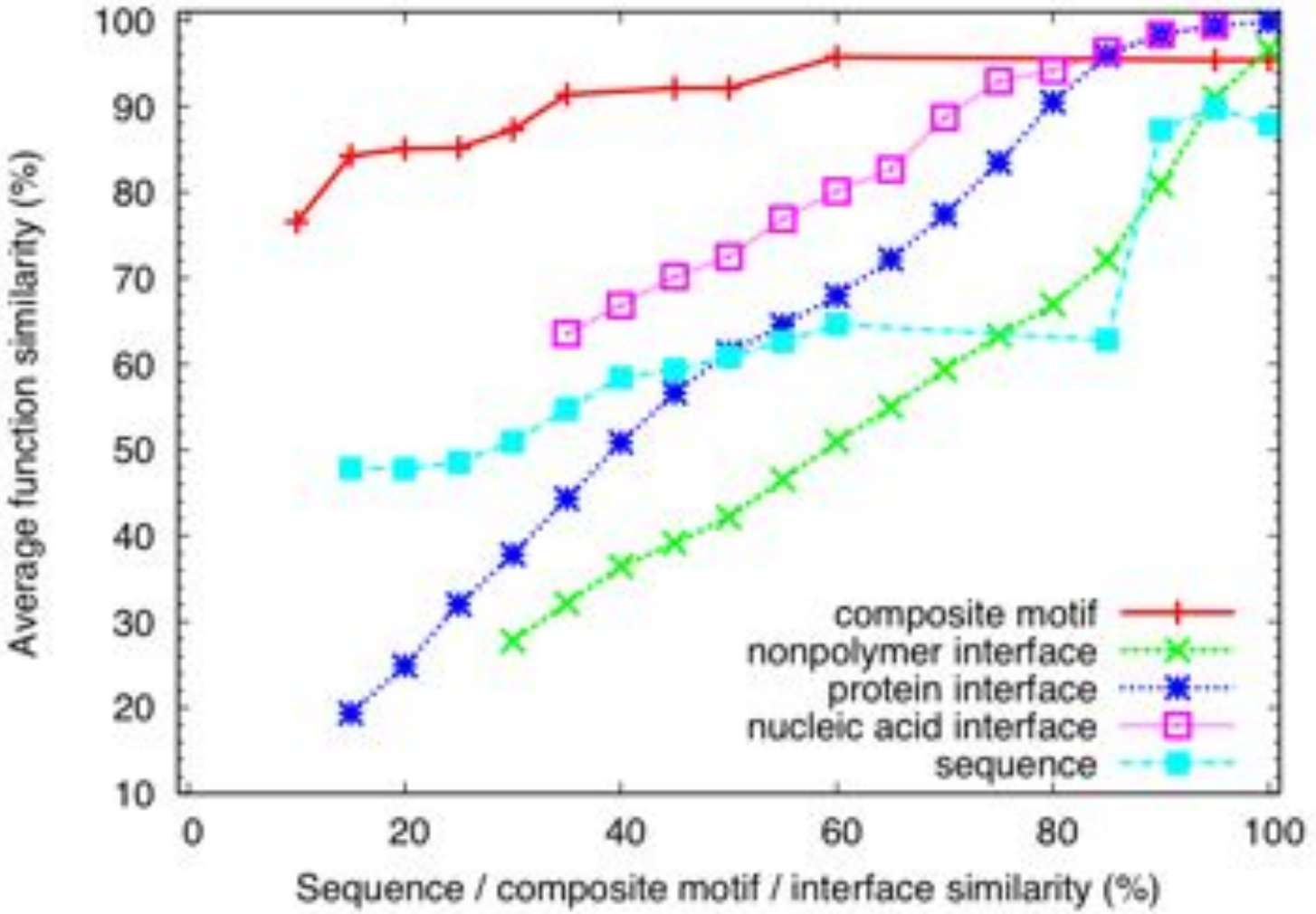
  ~ a network of interaction patterns.

Materials

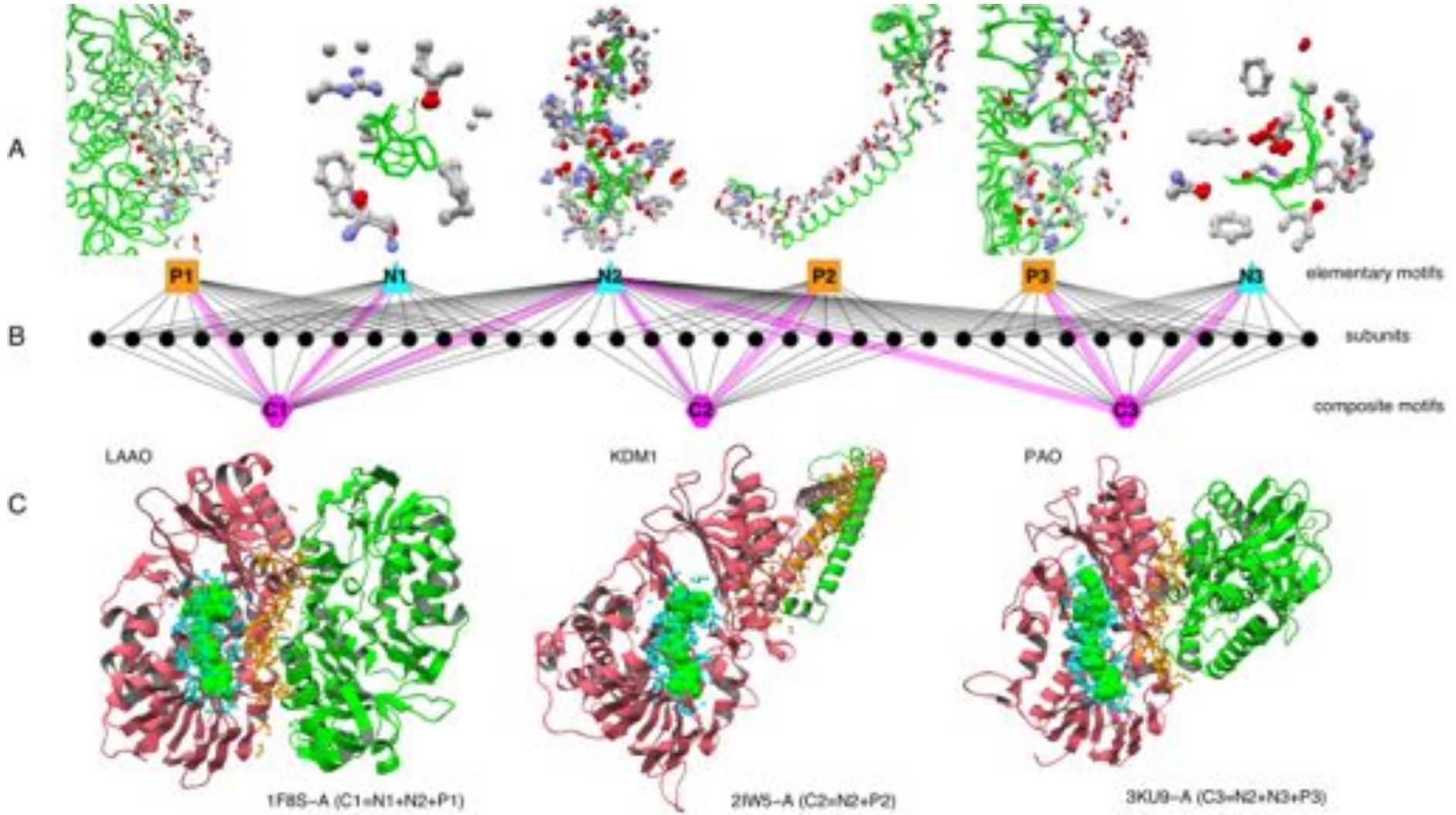
PDB entry	~70,000	
Biological Unit	~80,000	
Subunits	~200,000	~6,000 composite motifs
non-polymer interface	~400,000	~6,000 elementary motifs
protein interface	~350,000	~8,000 elementary motifs
DNA/RNA interface	~20,000	~400 elementary motifs

No sequence representatives are used!

c-motifs vs. functions



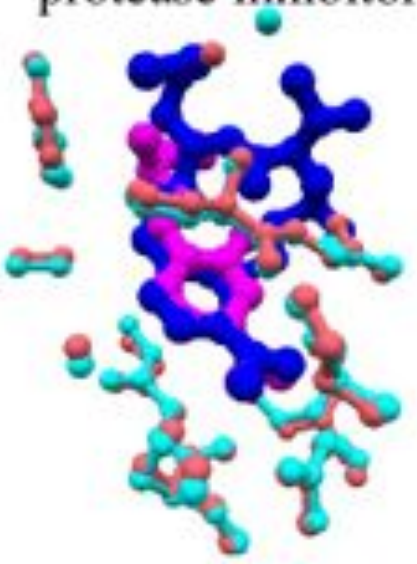
Composite motifs: example



Examples

beta–trypsin (EC 3.4.21.4)

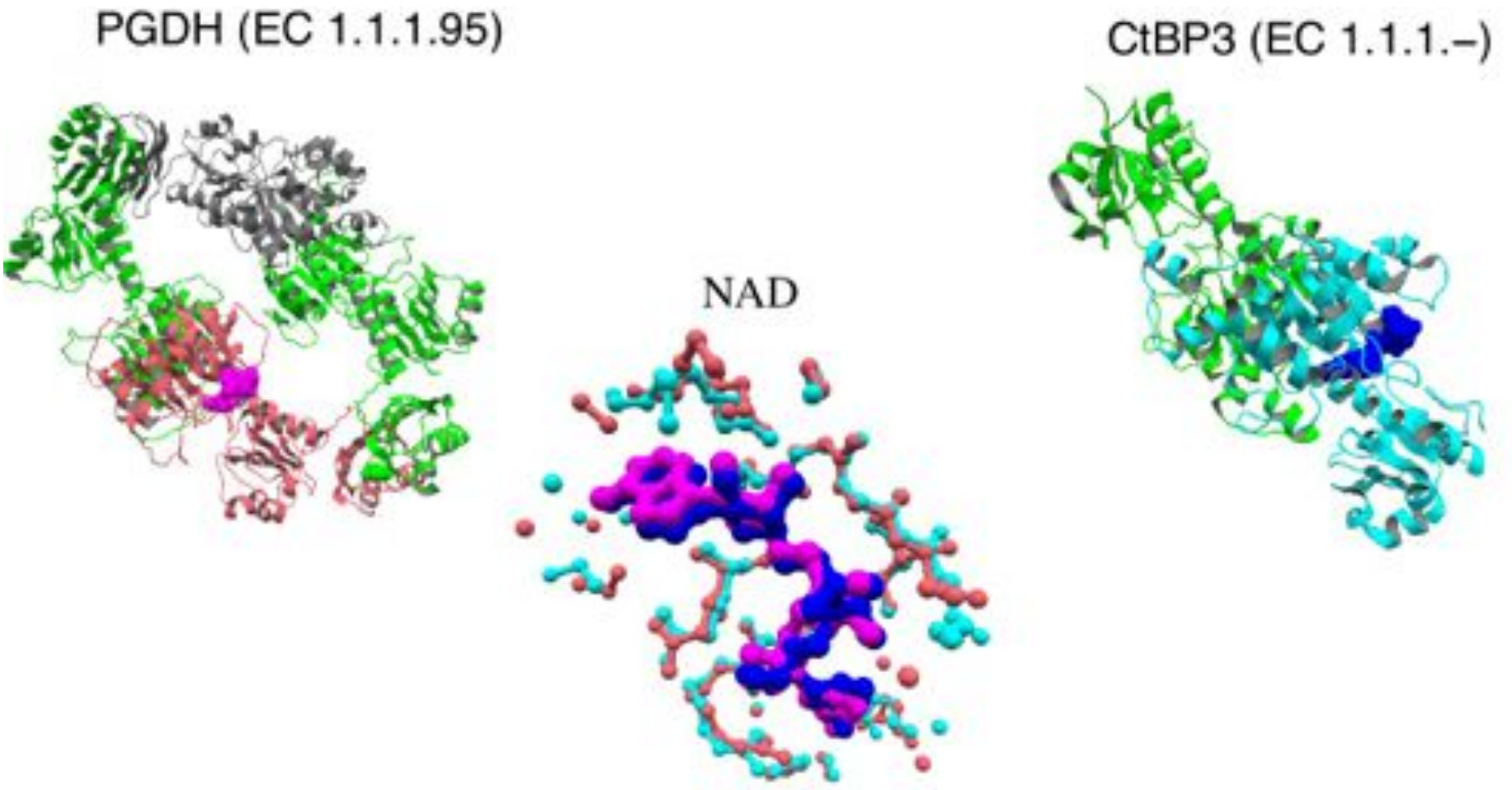
protease inhibitor



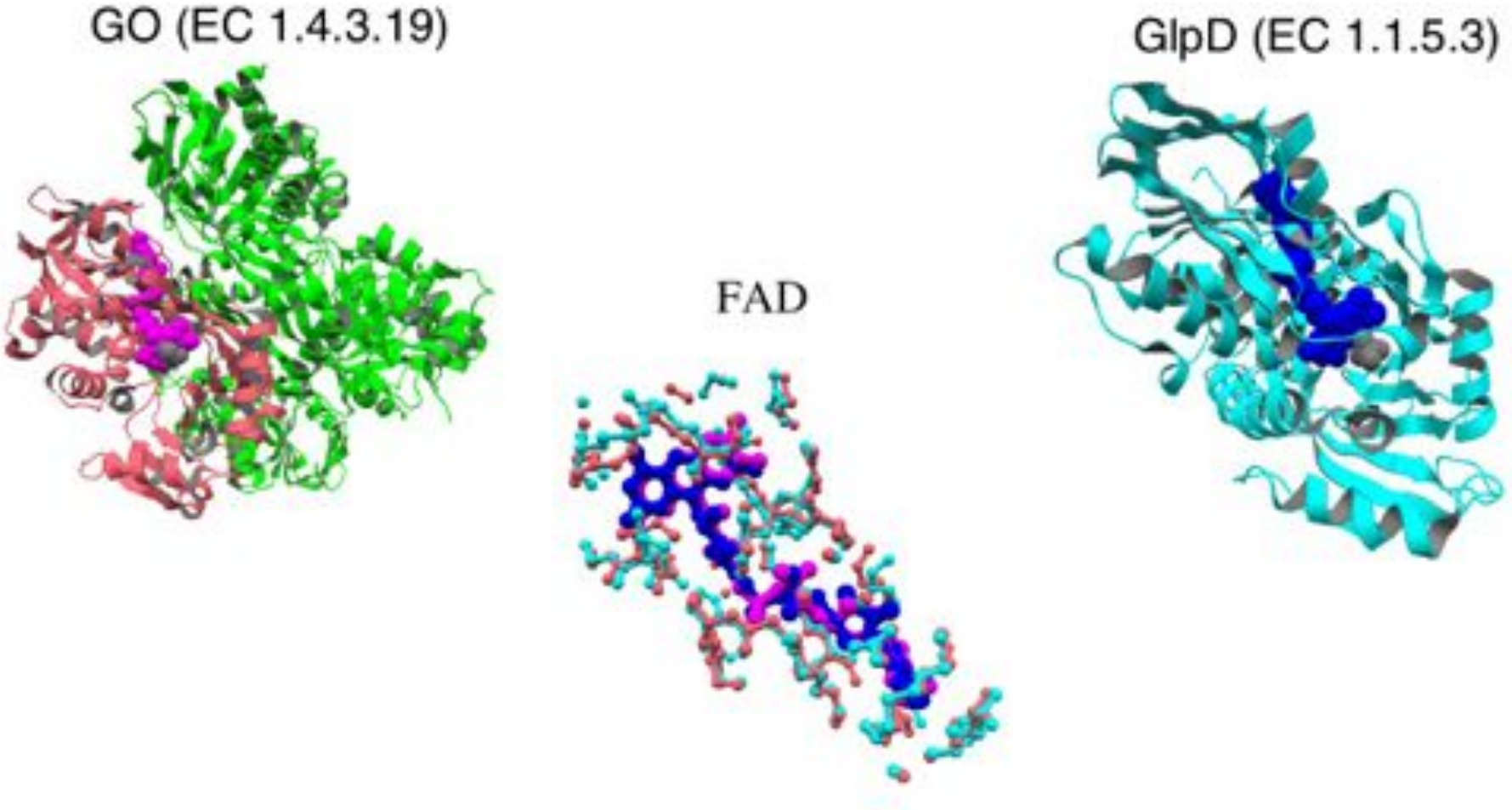
Factor VII (EC 3.4.21.21)

(i.e. catalytic site)

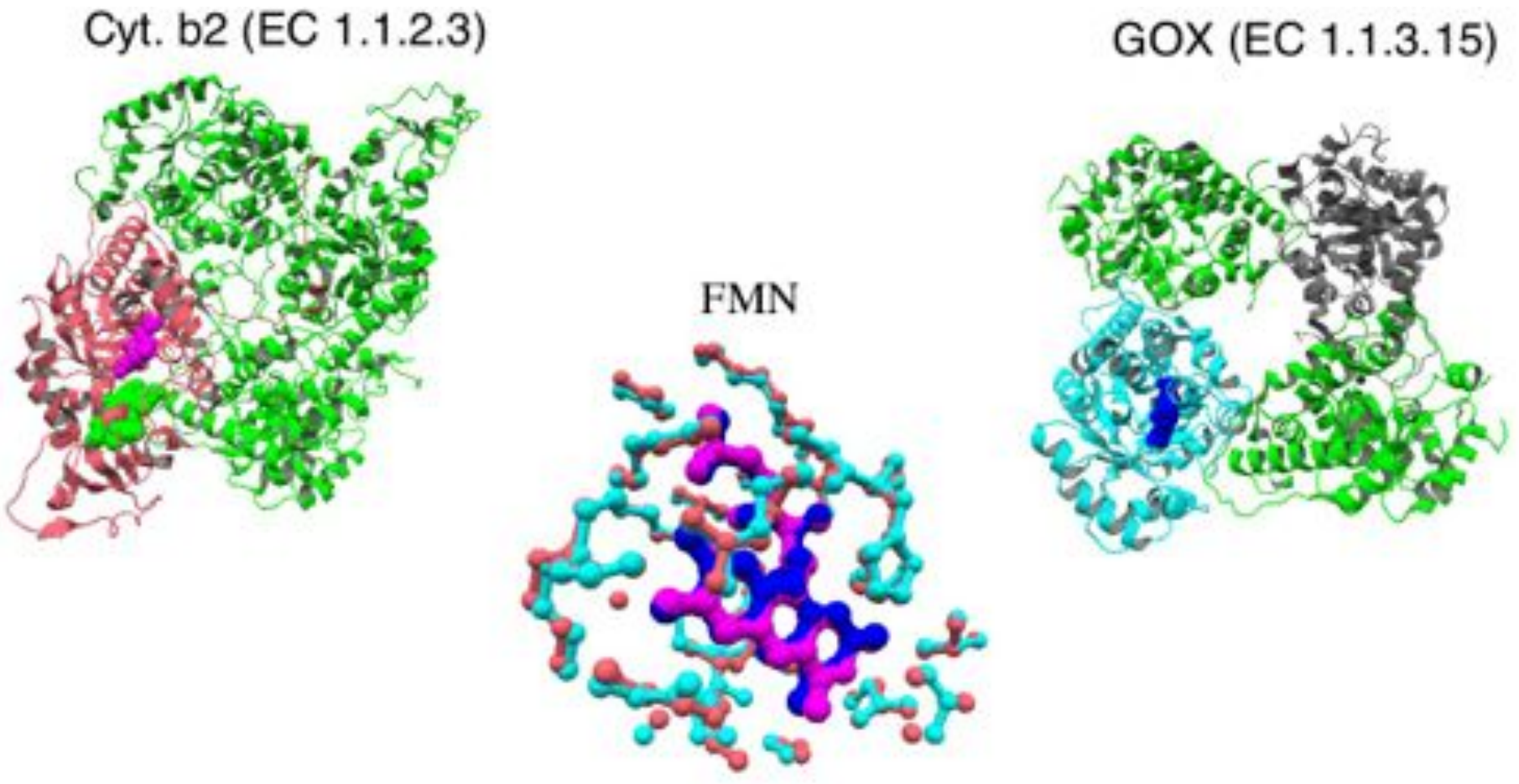
Same e-motif & fold, different c-motifs & functions



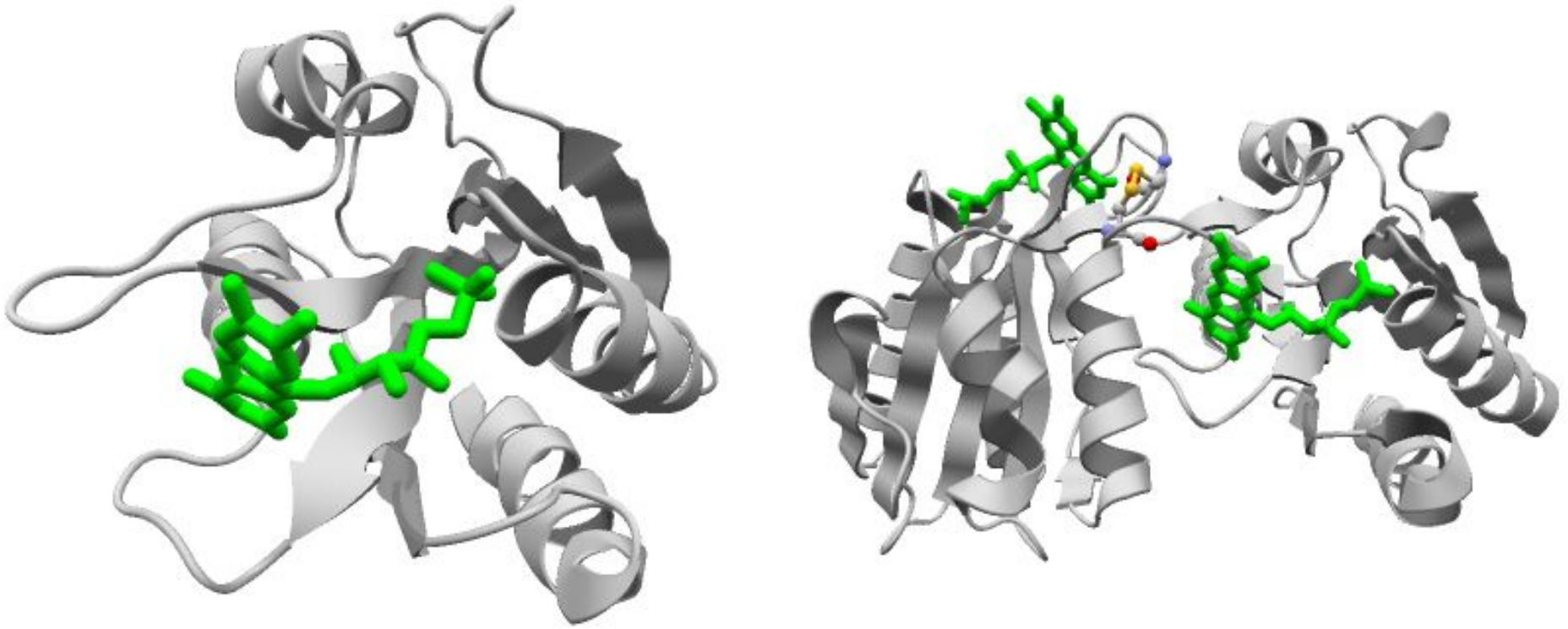
Same e-motif & fold, different c-motifs & functions



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

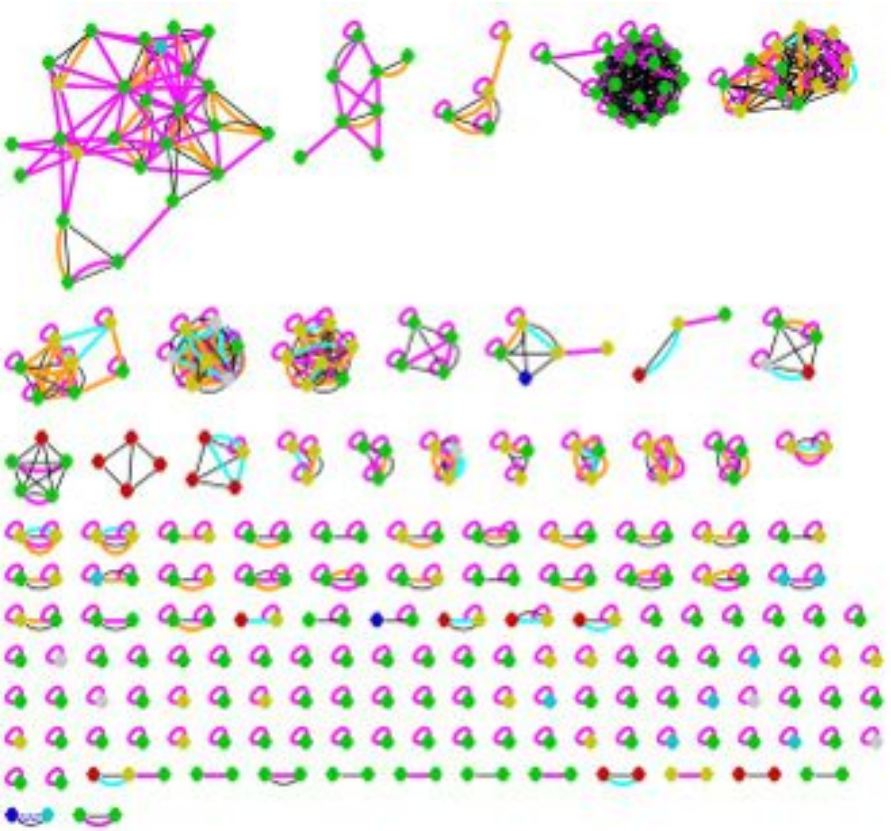


Flavodoxin: monomer & dimer

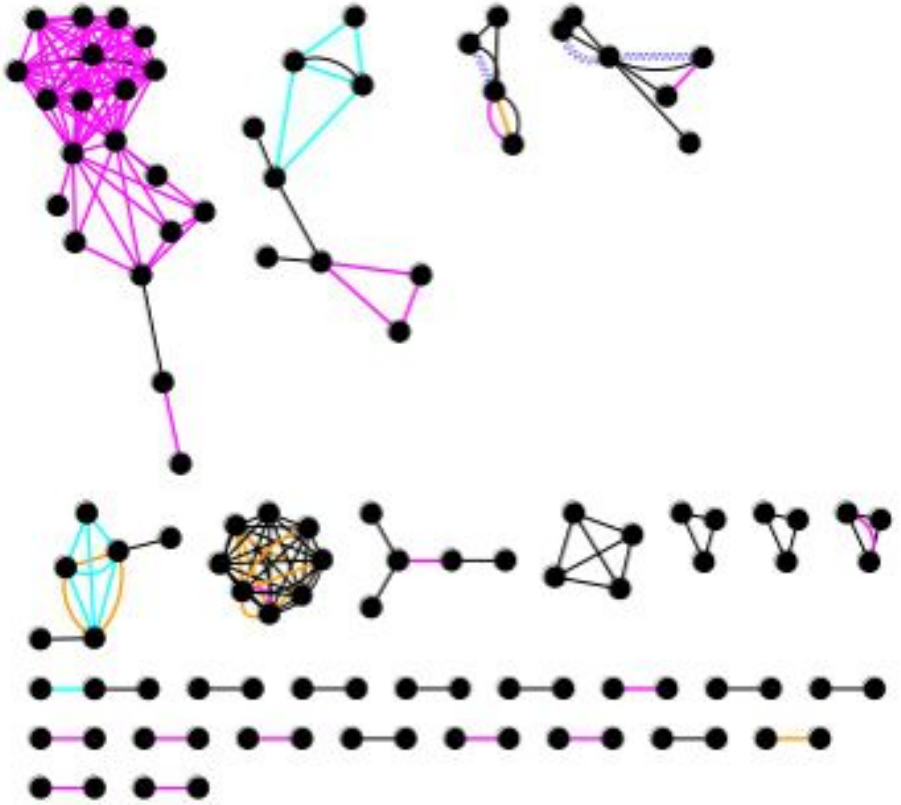


Meta-motif networks "Transcription"

Meta-composite motif

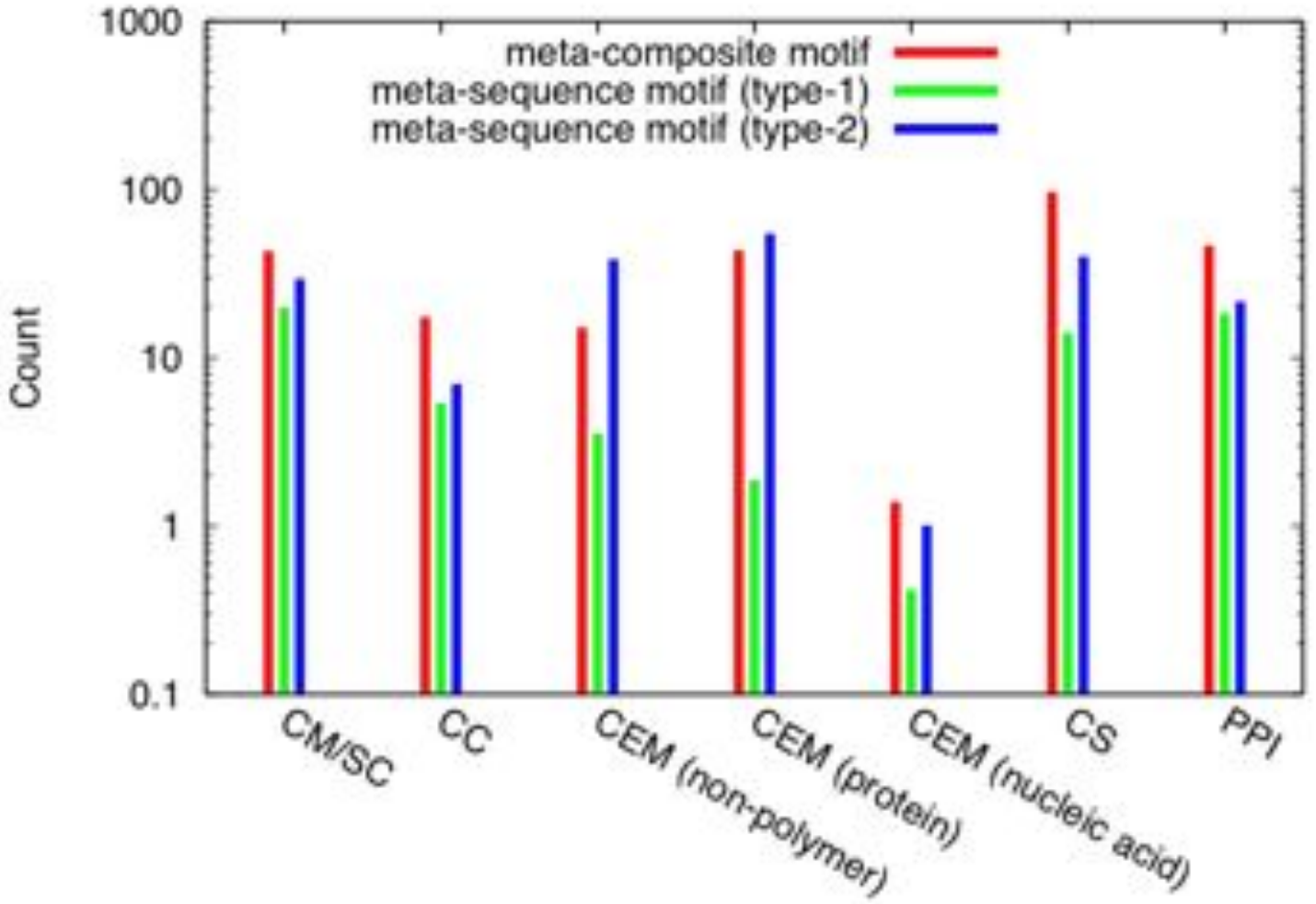


Meta-sequence motif



- 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 => 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...