蛋白質結構比對搜尋

Protein Structure Comparison and Search

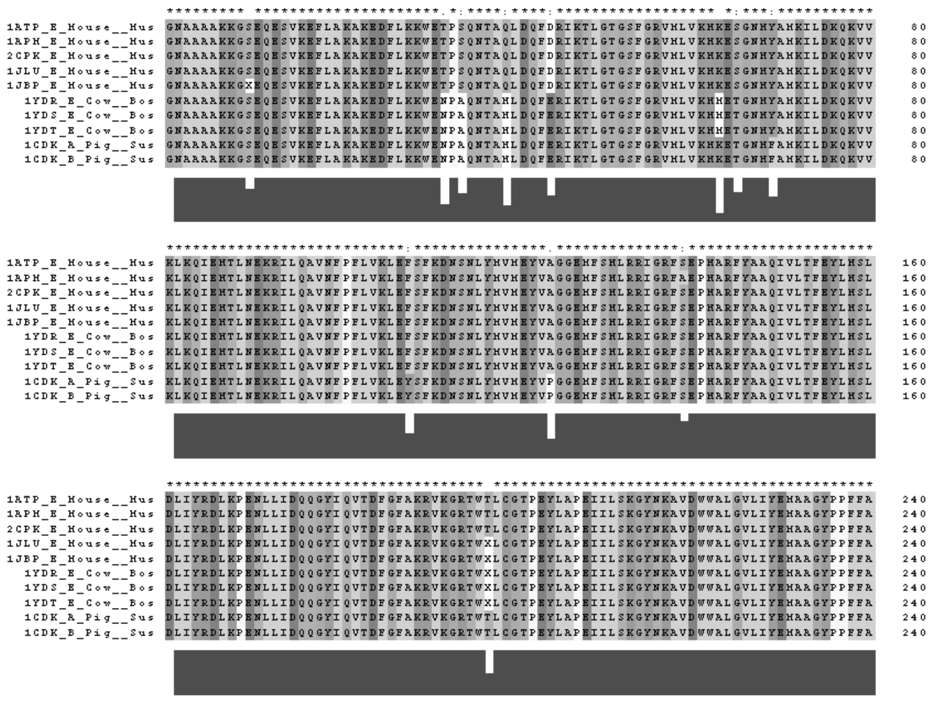
呂平江

國立清華大學

生命科學系/生物資訊與結構生物研究所

2012/06/27

Sequence Comparison



```
                   ********* *********************.*:****:****:****************** *:***:***********
1ATP_E_House__Hus  GNAAAAKKGSEQESVKEFLAKAKEDFLKKWETPSQNTAQLDQFDRIKTLGTGSFGRVMLVKHKESGNHYAMKILDKQKVV   80
1APH_E_House__Hus  GNAAAAKKGSEQESVKEFLAKAKEDFLKKWETPSQNTAQLDQFDRIKTLGTGSFGRVMLVKHKESGNHYAMKILDKQKVV   80
2CPK_E_House__Hus  GNAAAAKKGSEQESVKEFLAKAKEDFLKKWETPSQNTAQLDQFDRIKTLGTGSFGRVMLVKHKESGNHYAMKILDKQKVV   80
1JLU_E_House__Hus  GNAAAAKKGSEQESVKEFLAKAKEDFLKKWETPSQNTAQLDQFDRIKTLGTGSFGRVMLVKHKESGNHYAMKILDKQKVV   80
1JBP_E_House__Hus  GNAAAAKKGXEQESVKEFLAKAKEDFLKKWETPSQNTAQLDQFDRIKTLGTGSFGRVMLVKHKESGNHYAMKILDKQKVV   80
  1YDR_E_Cow__Bos  GNAAAAKKGSEQESVKEFLAKAKEDFLKKWENPAQNTAHLDQFERIKTLGTGSFGRVMLVKHHETGNHYAMKILDKQKVV   80
  1YDS_E_Cow__Bos  GNAAAAKKGSEQESVKEFLAKAKEDFLKKWENPAQNTAHLDQFERIKTLGTGSFGRVMLVKHHETGNHYAMKILDKQKVV   80
  1YDT_E_Cow__Bos  GNAAAAKKGSEQESVKEFLAKAKEDFLKKWENPAQNTAHLDQFERIKTLGTGSFGRVMLVKHHETGNHYAMKILDKQKVV   80
  1CDK_A_Pig__Sus  GNAAAAKKGSEQESVKEFLAKAKEDFLKKWENPAQNTAHLDQFERIKTLGTGSFGRVMLVKHKETGNHFAMKILDKQKVV   80
  1CDK_B_Pig__Sus  GNAAAAKKGSEQESVKEFLAKAKEDFLKKWENPAQNTAHLDQFERIKTLGTGSFGRVMLVKHKETGNHFAMKILDKQKVV   80

                   ***************************:***************.**************:*********************
1ATP_E_House__Hus  KLKQIEHTLNEKRILQAVNFPFLVKLEFSFKDNSNLYMVMEYVAGGEMFSHLRRIGRFSEPHARFYAAQIVLTFEYLHSL  160
1APH_E_House__Hus  KLKQIEHTLNEKRILQAVNFPFLVKLEFSFKDNSNLYMVMEYVAGGEMFSHLRRIGRFAEPHARFYAAQIVLTFEYLHSL  160
2CPK_E_House__Hus  KLKQIEHTLNEKRILQAVNFPFLVKLEFSFKDNSNLYMVMEYVAGGEMFSHLRRIGRFSEPHARFYAAQIVLTFEYLHSL  160
1JLU_E_House__Hus  KLKQIEHTLNEKRILQAVNFPFLVKLEFSFKDNSNLYMVMEYVAGGEMFSHLRRIGRFSEPHARFYAAQIVLTFEYLHSL  160
1JBP_E_House__Hus  KLKQIEHTLNEKRILQAVNFPFLVKLEFSFKDNSNLYMVMEYVAGGEMFSHLRRIGRFSEPHARFYAAQIVLTFEYLHSL  160
  1YDR_E_Cow__Bos  KLKQIEHTLNEKRILQAVNFPFLVKLEFSFKDNSNLYMVMEYVAGGEMFSHLRRIGRFSEPHARFYAAQIVLTFEYLHSL  160
  1YDS_E_Cow__Bos  KLKQIEHTLNEKRILQAVNFPFLVKLEFSFKDNSNLYMVMEYVAGGEMFSHLRRIGRFSEPHARFYAAQIVLTFEYLHSL  160
  1YDT_E_Cow__Bos  KLKQIEHTLNEKRILQAVNFPFLVKLEFSFKDNSNLYMVMEYVAGGEMFSHLRRIGRFSEPHARFYAAQIVLTFEYLHSL  160
  1CDK_A_Pig__Sus  KLKQIEHTLNEKRILQAVNFPFLVKLEYSFKDNSNLYMVMEYVPGGEMFSHLRRIGRFSEPHARFYAAQIVLTFEYLHSL  160
  1CDK_B_Pig__Sus  KLKQIEHTLNEKRILQAVNFPFLVKLEYSFKDNSNLYMVMEYVPGGEMFSHLRRIGRFSEPHARFYAAQIVLTFEYLHSL  160

                   ************************************ *******************************************
1ATP_E_House__Hus  DLIYRDLKPENLLIDQQGYIQVTDFGFAKRVKGRTWTLCGTPEYLAPEIILSKGYNKAVDWWALGVLIYEMAAGYPPFFA  240
1APH_E_House__Hus  DLIYRDLKPENLLIDQQGYIQVTDFGFAKRVKGRTWTLCGTPEYLAPEIILSKGYNKAVDWWALGVLIYEMAAGYPPFFA  240
2CPK_E_House__Hus  DLIYRDLKPENLLIDQQGYIQVTDFGFAKRVKGRTWTLCGTPEYLAPEIILSKGYNKAVDWWALGVLIYEMAAGYPPFFA  240
1JLU_E_House__Hus  DLIYRDLKPENLLIDQQGYIQVTDFGFAKRVKGRTWXLCGTPEYLAPEIILSKGYNKAVDWWALGVLIYEMAAGYPPFFA  240
1JBP_E_House__Hus  DLIYRDLKPENLLIDQQGYIQVTDFGFAKRVKGRTWXLCGTPEYLAPEIILSKGYNKAVDWWALGVLIYEMAAGYPPFFA  240
  1YDR_E_Cow__Bos  DLIYRDLKPENLLIDQQGYIQVTDFGFAKRVKGRTWXLCGTPEYLAPEIILSKGYNKAVDWWALGVLIYEMAAGYPPFFA  240
  1YDS_E_Cow__Bos  DLIYRDLKPENLLIDQQGYIQVTDFGFAKRVKGRTWXLCGTPEYLAPEIILSKGYNKAVDWWALGVLIYEMAAGYPPFFA  240
  1YDT_E_Cow__Bos  DLIYRDLKPENLLIDQQGYIQVTDFGFAKRVKGRTWXLCGTPEYLAPEIILSKGYNKAVDWWALGVLIYEMAAGYPPFFA  240
  1CDK_A_Pig__Sus  DLIYRDLKPENLLIDQQGYIQVTDFGFAKRVKGRTWTLCGTPEYLAPEIILSKGYNKAVDWWALGVLIYEMAAGYPPFFA  240
  1CDK_B_Pig__Sus  DLIYRDLKPENLLIDQQGYIQVTDFGFAKRVKGRTWTLCGTPEYLAPEIILSKGYNKAVDWWALGVLIYEMAAGYPPFFA  240
```

% of Identity　　　**% of similarity**

Introduction to Structure Comparison

- Sander & Schneider (1990)：

  Total available structures：597

  Protein sequences (PDB) → Pairwise alignment

  sequence identity > 30％ 100% have similar 3D-structrure

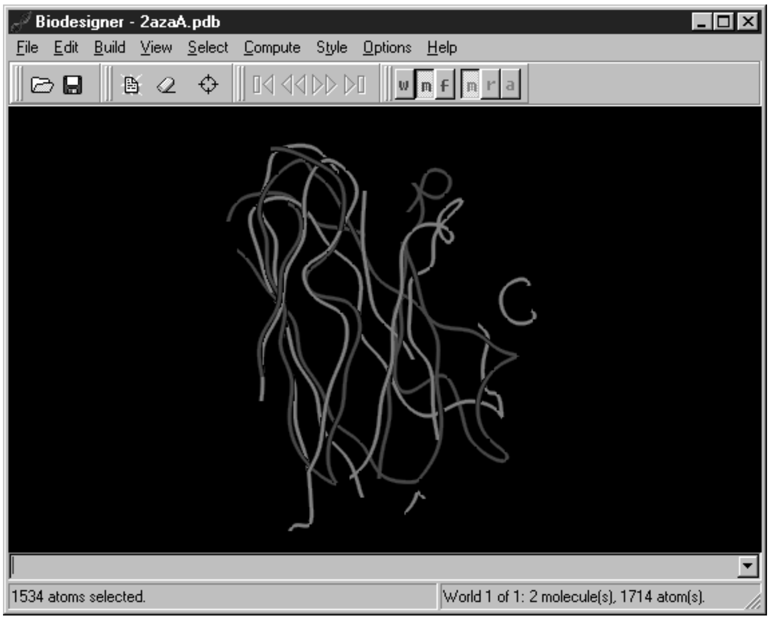
- Burkhard Rost (1999)：

  Total available structures：11,364

  sequence identity > 30% 90% have similar 3D-structure

  sequence identity < 25% 10% have similar 3D-structure

Structure Comparison



1534 atoms selected.

World 1 of 1: 2 molecule(s), 1714 atom(s).

蛋白質結構比對是用來比較兩個蛋白質結構是否相似，通常是計算兩結構距離的方均根差(RMSD, Root Meaning Square Deviation)，而其單位通常是埃(Å) 。

Structural Comparisons – Why?

- Finding similar structure by sequence comparison alone is not enough
- The number of protein structures is increasing rapidly.
- Structure-Structure relationship is more conserved than sequence during evolution

Structural Comparisons – How?

- Two categories of current methods
  - By amino acid sequence alignments.
  - By 3D structural alignments.

Classical Sequence Alignment Methods

- BLAST
  - – Basic Local Alignment Search Tool
- FASTA
  - – FAST-All, reflecting that it can be used for fast protein comparisons

Performance: Rapid but inaccurate*

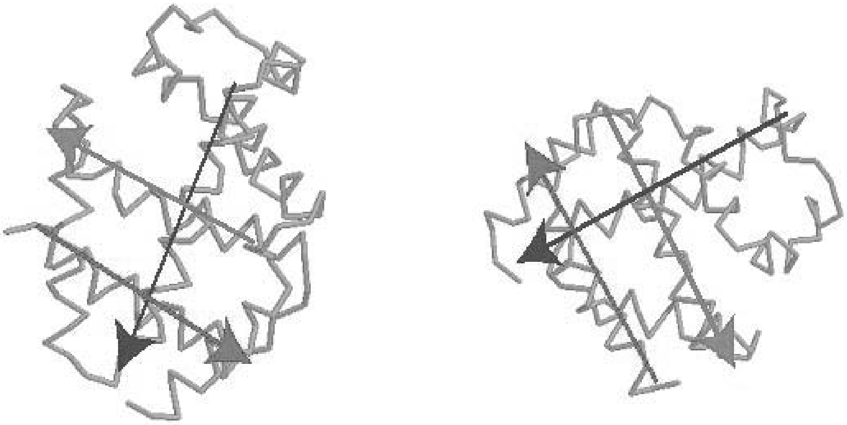
* Kolodny *et al.* (2005) *J Mol Biol.* 346:1173-1188

Conventional Structural Alignment Methods

- Double Dynamic Programming – SSAP
- Distance Alignment Tools – DALI
- Vector Alignment Search Tool – VAST
- Combinatorial Extension – CE
- Fast Alignment Search Tool – FAST
- MAtching Molecular Models Obtained from Theory – MAMMOTH

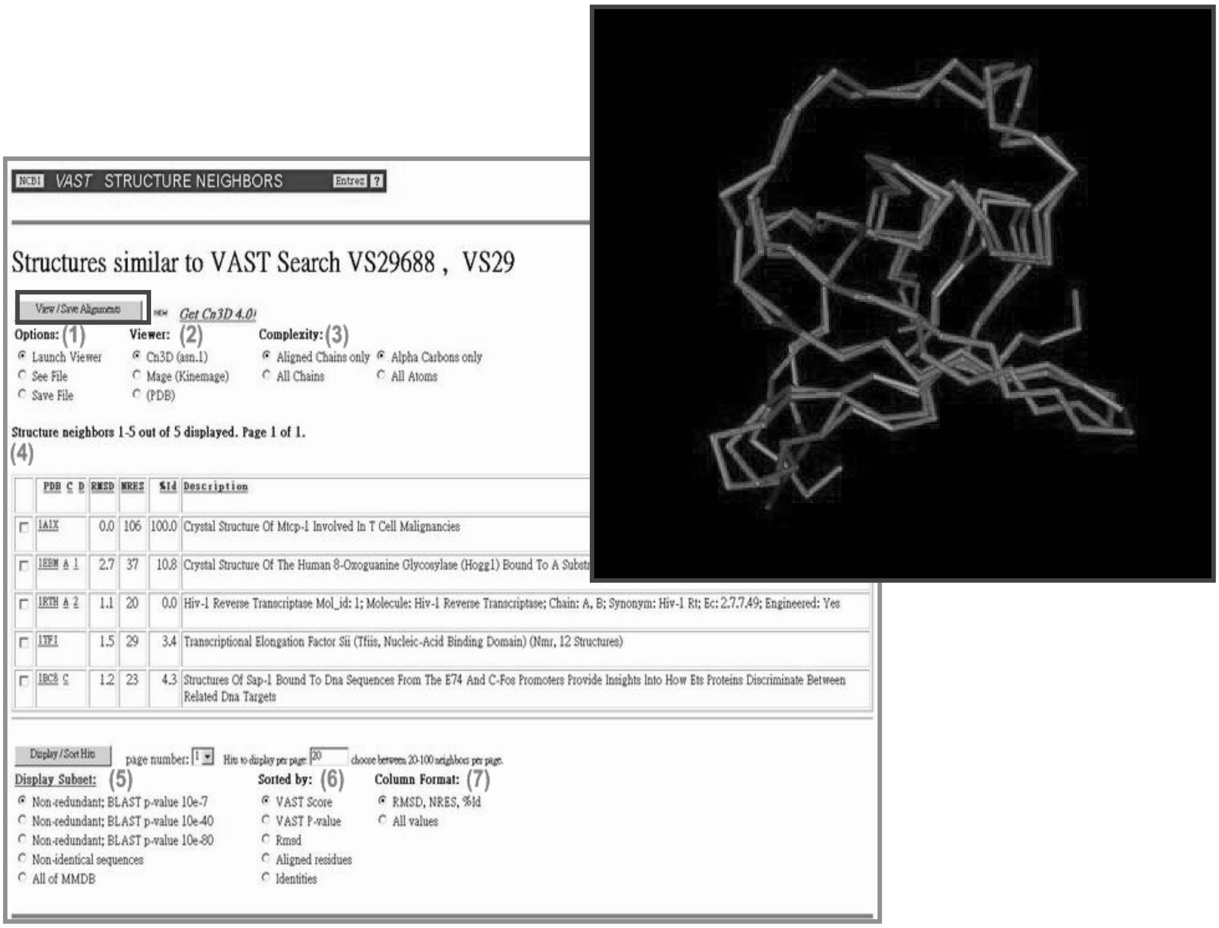
Structure Comparison Methods (VAST)

- VAST (Vector Alignment Search Tool) (Gibrat, Madej,1996)
- Secondary Structure Elements (SSE)



Vector Alignment Search Tool

http://www.ncbi.nlm.nih.gov/Structure/VAST/vast.shtml



NCBI *VAST* STRUCTURE NEIGHBORS Entrez ?

Structures similar to VAST Search VS29688 , VS29

Options: (1)
- ● Launch Viewer
- ○ See File
- ○ Save File

Viewer: (2)
- ● Cn3D (asn.1)
- ○ Mage (Kinemage)
- ○ (PDB)

Complexity: (3)
- ● Aligned Chains only
- ○ All Chains
- ● Alpha Carbons only
- ○ All Atoms

Structure neighbors 1-5 out of 5 displayed. Page 1 of 1.

(4)

	PDB	C	D	RMSD	NRES	%Id	Description
☐	1AIX			0.0	106	100.0	Crystal Structure Of Mtcp-1 Involved In T Cell Malignancies
☐	1EBM	A	1	2.7	37	10.8	Crystal Structure Of The Human 8-Oxoguanine Glycosylase (Hogg1) Bound To A Subst
☐	1RTH	A	2	1.1	20	0.0	Hiv-1 Reverse Transcriptase Mol_id: 1; Molecule: Hiv-1 Reverse Transcriptase; Chain: A, B; Synonym: Hiv-1 Rt; Ec: 2.7.7.49; Engineered: Yes
☐	1TFI			1.5	29	3.4	Transcriptional Elongation Factor Sii (Tfiis, Nucleic-Acid Binding Domain) (Nmr, 12 Structures)
☐	1BC8	C		1.2	23	4.3	Structures Of Sap-1 Bound To Dna Sequences From The E74 And C-Fos Promoters Provide Insights Into How Ets Proteins Discriminate Between Related Dna Targets

page number: 1 Hits to display per page: 20 choose between 20-100 neighbors per page.

Display Subset: (5)
- ● Non-redundant; BLAST p-value 10e-7
- ○ Non-redundant; BLAST p-value 10e-40
- ○ Non-redundant; BLAST p-value 10e-80
- ○ Non-identical sequences
- ○ All of MMDB

Sorted by: (6)
- ● VAST Score
- ○ VAST P-value
- ○ Rmsd
- ○ Aligned residues
- ○ Identities

Column Format: (7)
- ● RMSD, NRES, %Id
- ○ All values

Structure Comparison Methods (DALI)

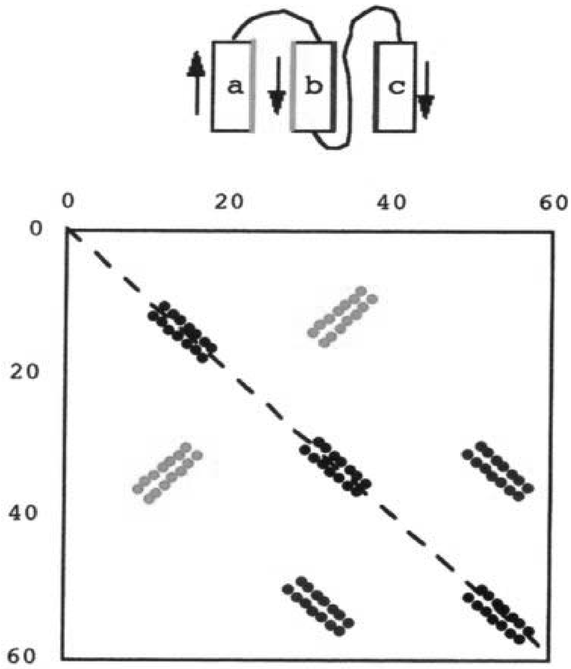
- DALI (distance alignment tools)

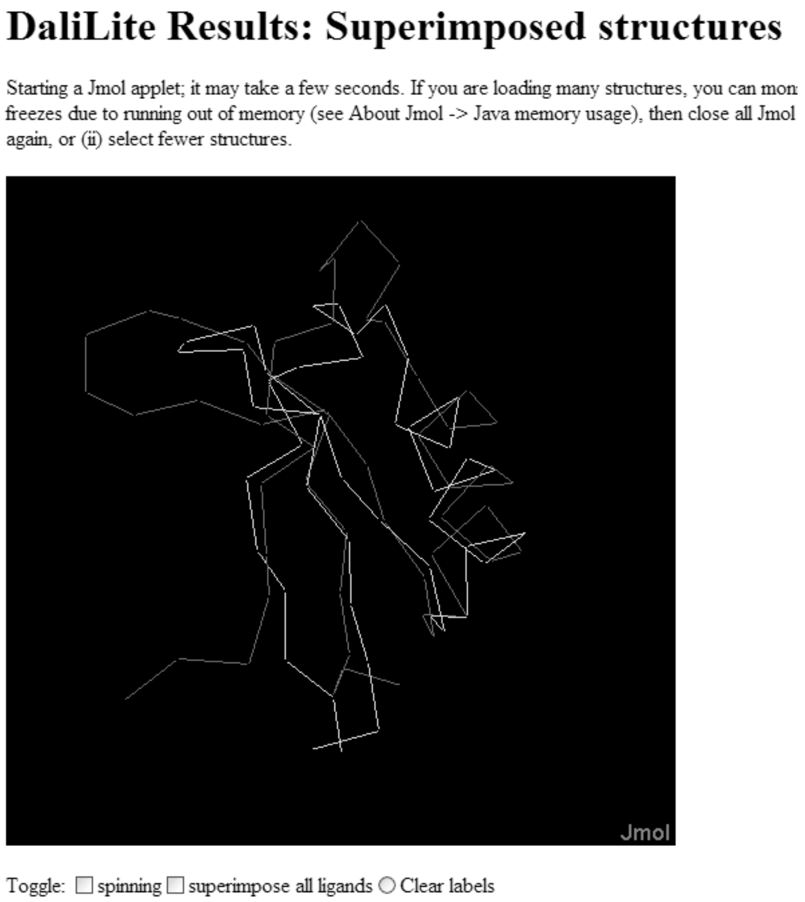
➢ Transfer all the Cα - Cα distance in a protein to distance dot matrix

➢ If Cα-Cαdistance > 12 Å

`delete the dot`

➢ 3D→2D





DaliLite Results: Superimposed structures

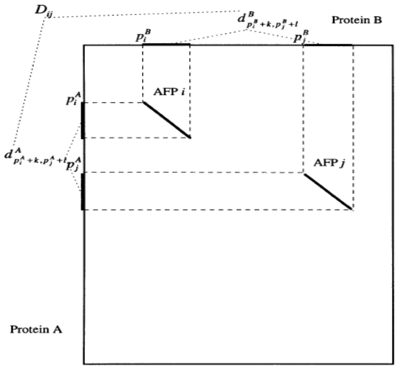
Starting a Jmol applet; it may take a few seconds. If you are loading many structures, you can mon freezes due to running out of memory (see About Jmol -> Java memory usage), then close all Jmol again, or (ii) select fewer structures.

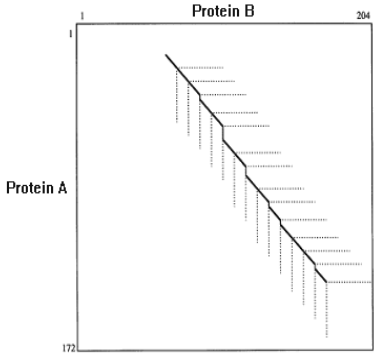
Toggle: □ spinning □ superimpose all ligands ○ Clear labels

Structure Comparison Methods (CE)

- combinatorial extension

$$D_{ij} = \frac{1}{m^2}\left(\sum_{k=0}^{m-1} \sum_{l=0}^{m-1} \left| \; d^{A}_{p^A_i+k,\,p^A_j+l} - d^{B}_{p^B_i+k,\,p^B_j+l} \; \right| \right)$$





CE是一種計算結構比對的方法，是由美國聖地牙哥超級電腦中心所提供的。CE是利用區域的幾何特性(alpha碳原子間的向量)來進行比對，將配對到的片段稱為AFPs (aligned fragment pairs)，再利用演算法來得到最好的RMSD值。

Protein structure classification databases

- SCOP (structure classification of proteins)
  - ✓ based on expert definition of structural similarities
  - ✓ http://scop.mrc-lmb.cam.ac.uk/scop/
- CATH (classification by class, architecture, topology and homology)
  - ✓ based on SSAP (secondary structure alignment program)
  - ✓ University College, London (Orengo 1997)
  - ✓ http://www.biochem.ucl.ac.uk/bsm/cath/
- FSSP (fold classification based on structure-structure alignment of proteins)
  - ✓ based on pairwise structure alignment of PDB by DALI
  - ✓ http://www2.embl-ebi.ac.uk/dali/fssp/fssp.html
- MMDB (molecular modelling database)
  - ✓ PDB has been categorized into structurally related groups in MMDB by VAST (vector alignment search tool)
  - ✓ http://www.ncbi.nlm.nih.gov/Structure/MMDB/mmdb.shtml
- CE (combinatorial extension)
  - ✓ based on pairwise structure alignment of PDB by CE
  - ✓ http://cl.sdsc.edu/ce.html

國立清華大學
National Tsing Hua University

Searching similar protein structures of the specified protein in PDB

SARST – Structural similarity search Aided by Ramachandran Sequential Transformation

Introduction to SARST

- SARST transforms 3D protein structures into 1D text sequences and recruit blast to perform protein structural alignment searches
- Features
  - high speed
  - reasonable compromise of the accuracy
  - giving statistically meaningful results

Speed vs. Accuracy: Incompatible?

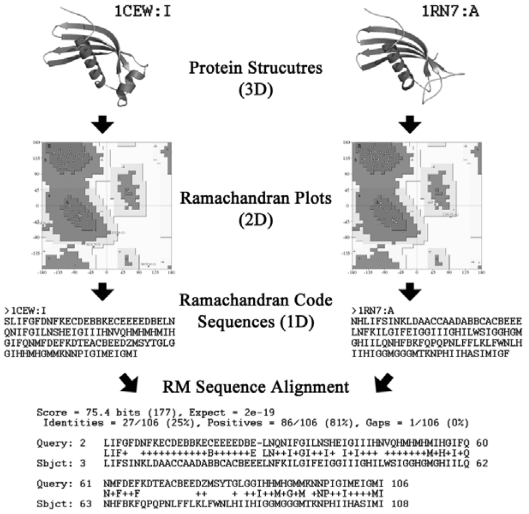
- Possible solution: the linear encoding method

  3D structure ➔ 1D text sequence

- Example:

  SARST

  – <u>S</u>tructure <u>A</u>lignment by <u>R</u>amachandran <u>S</u>earch <u>T</u>ool



```
>1CEW:I
SLIFGFDNFKECDEBBKECEEEEDBELN
QNIFGILNSHEIGIIIHNVQHMHMHMIH
GIFQNMFDEFKDTEACBEEDZMSYTGLG
GIHHMHGMMKNNPIGIMEIGMI
```

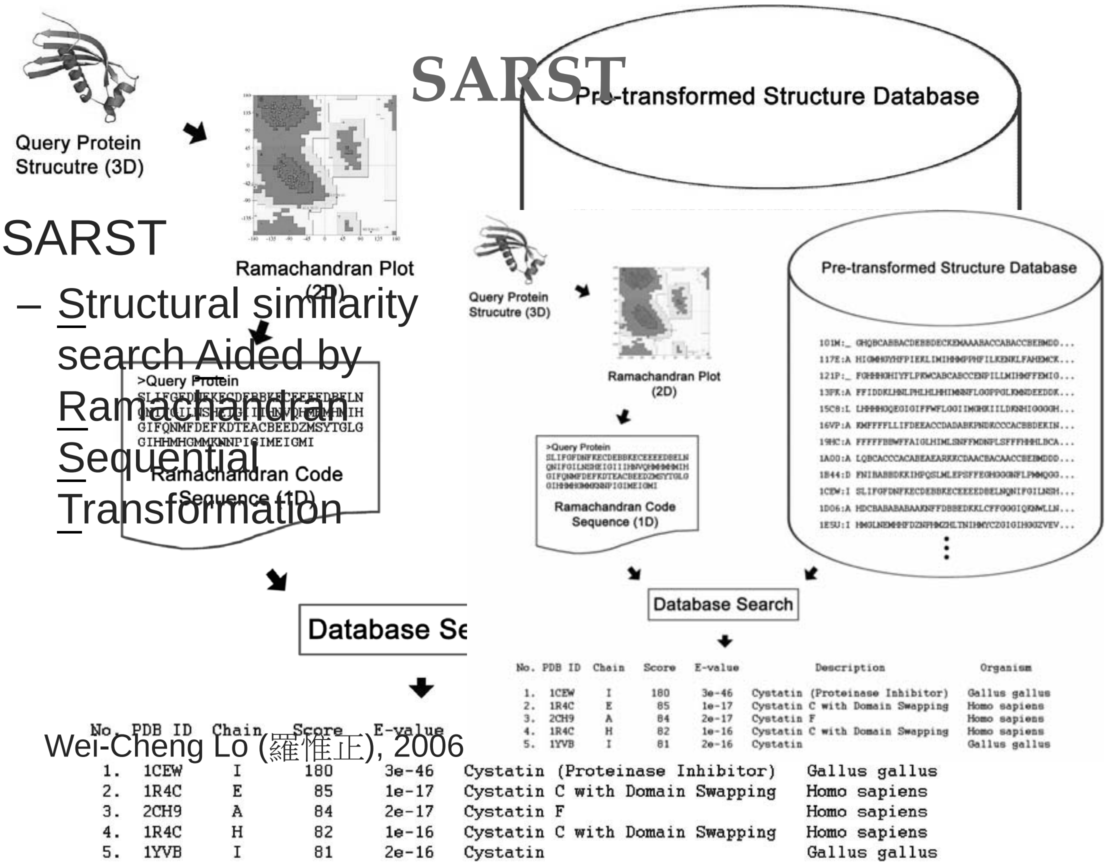
```
>1RN7:A
NHLIFSINKLDAACCAADABBCACBEEE
LNFKILGIFEIGGIIIGHILWSIGGHGM
GHIILQNHFBKFQPQPNLFFLKLFWNLH
IIHIGGMGGGMTKNPHIIHASIMIGF
```

```
Score = 75.4 bits (177), Expect = 2e-19
 Identities = 27/106 (25%), Positives = 86/106 (81%), Gaps = 1/106 (0%)

Query: 2   LIFGFDNFKECDEBBKECEEEEDBE-LNQNIFGILNSHEIGIIIHNVQHMHMHMIHGIFQ 60
            LIF+  +++++++++++B++++++E LN++I+GI++I+ I+I+++ +++++++M+H+I+Q
Sbjct: 3   LIFSINKLDAACCAADABBCACBEEELNFKILGIFEIGGIIIGHILWSIGGHGMGHIILQ 62

Query: 61  NMFDEFKDTEACBEEDZMSYTGLGGIHHMHGMMKNNPIGIMEIGMI 106
            N+F++F         ++    + ++I++M+G+M +NP++I++++MI
Sbjct: 63  NHFBKFQPQPNLFFLKLFWNLHIIHIGGMGGGMTKNPHIIHASIMI 108
```

SARST

- SARST

  – <u>S</u>tructural similarity search <u>A</u>ided by <u>R</u>amachandran <u>S</u>equential <u>T</u>ransformation



```
>Query Protein
SLIFGFDNFKECDEBBKECEEEEDBELN
QNIFGILNSHEIGIIIHNVQHMHMHMIH
GIFQNMFDEFKDTEACBEEDZMSYTGLG
GIHHMHGMMKNNPIGIMEIGMI
```

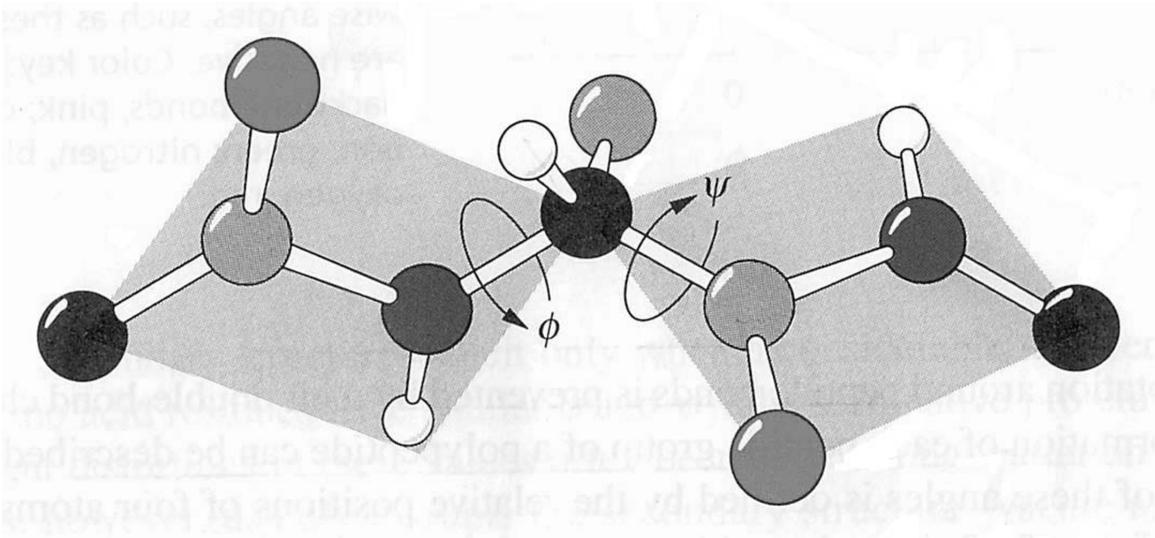
```
101M:_ GHQBCABBACDEBBDECKEMAAABACCABACCBEBMDD...
117E:A HIGMHGYHFPIEKLIMIHHMPPHFILKENKLFAHEMCK...
121P:_ FGHHHGHIYFLPKWCABCABCCENPILLMIHMFFEMIG...
13PK:A FFIDDKLHNLPHLHLHHIMNNFLGGPPGLKMNDEEDDK...
15C8:L LHHHHGQEGIGIFFWFLGGIIMGHKIIILDKNHIGGGGH...
16VP:A KMFFFFLLIFDEEACCDADABKPNDKCCCACBBDEKIN...
19HC:A FFFFFBBWFFAIGLHIMLSNFFMDNPLSFFFHHHLBCA...
1A00:A LQBCACCCACABEAEARKKCDAACBACAACCBEBMDDD...
1B44:D FNIBABBDKKIHPQSLMLEPSFFEGHGGGNFLPMMQGG...
1CEW:I SLIFGFDNFKECDEBBKECEEEEDBELNQNIFGILNSH...
1D06:A HDCBABABABAAKNFFDBBEDKKLCFFGGGIQKNWLLN...
1ESU:I HMGLNEMHHFDZNPHMZHLTNIHMYCZGIGIHGGZVEV...
   ⋮
```

```
No. PDB ID  Chain  Score  E-value  Description                        Organism

 1.  1CEW     I     180    3e-46   Cystatin (Proteinase Inhibitor)    Gallus gallus
 2.  1R4C     E      85    1e-17   Cystatin C with Domain Swapping    Homo sapiens
 3.  2CH9     A      84    2e-17   Cystatin F                         Homo sapiens
 4.  1R4C     H      82    1e-16   Cystatin C with Domain Swapping    Homo sapiens
 5.  1YVB     I      81    2e-16   Cystatin                           Gallus gallus
```

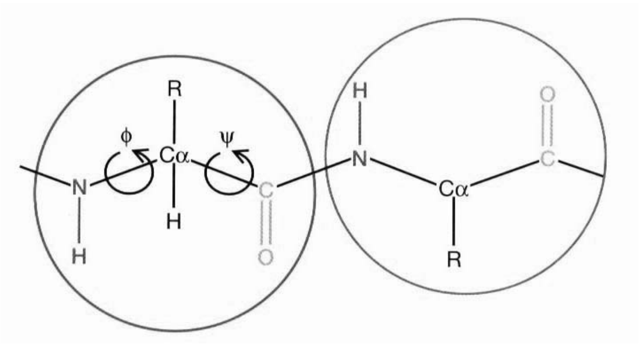
```
No. PDB ID  Chain  Score  E-value  Description                        Organism

 1.  1CEW     I     180    3e-46   Cystatin (Proteinase Inhibitor)    Gallus gallus
 2.  1R4C     E      85    1e-17   Cystatin C with Domain Swapping    Homo sapiens
 3.  2CH9     A      84    2e-17   Cystatin F                         Homo sapiens
 4.  1R4C     H      82    1e-16   Cystatin C with Domain Swapping    Homo sapiens
 5.  1YVB     I      81    2e-16   Cystatin                           Gallus gallus
```

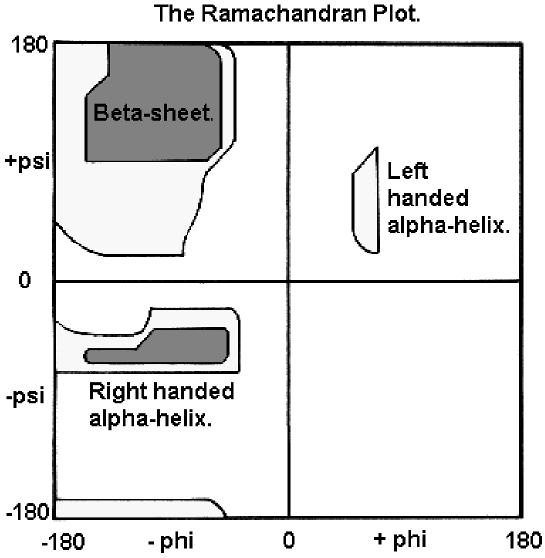
Wei-Cheng Lo (羅惟正), 2006

Phi (φ) and Psi (ψ) Angles

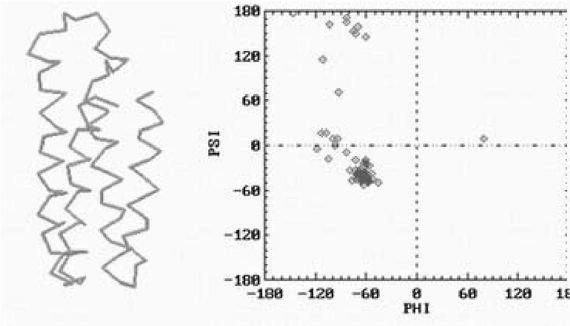


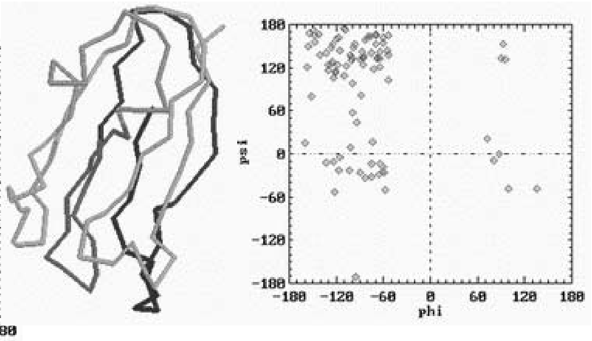
Ramachandran Plot



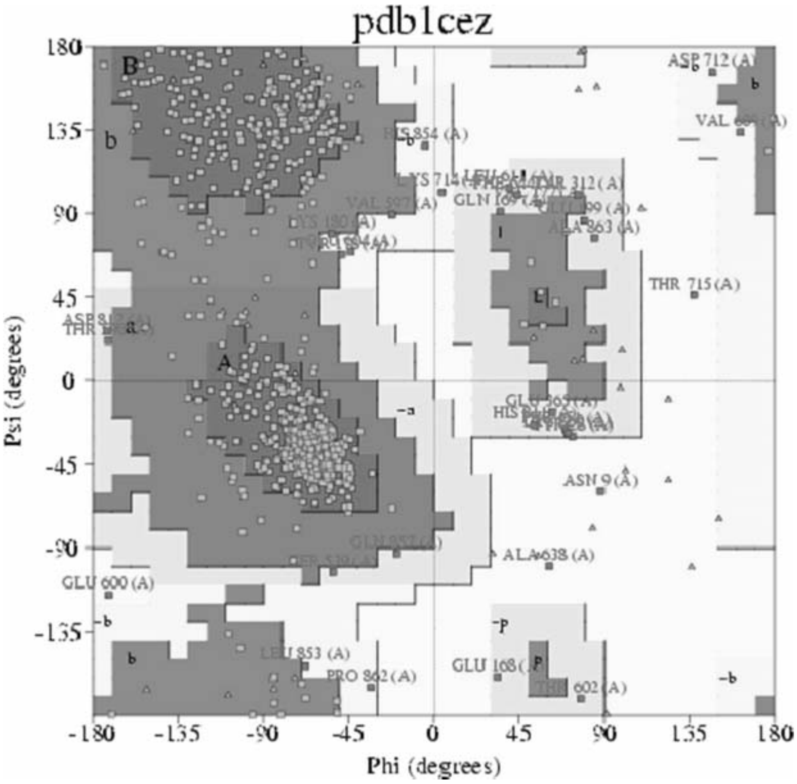


all helix all beta

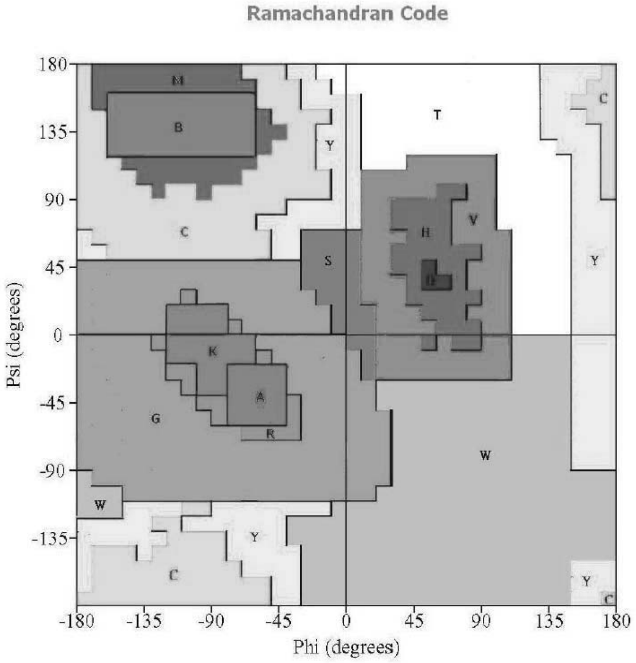




The Ramachandran Map

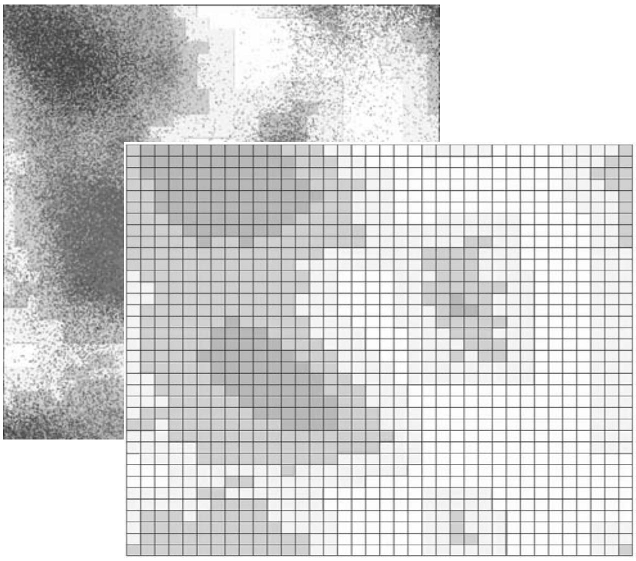


An Example of Organized Ramachandran Map



SARST

Distance Determination of the Cells

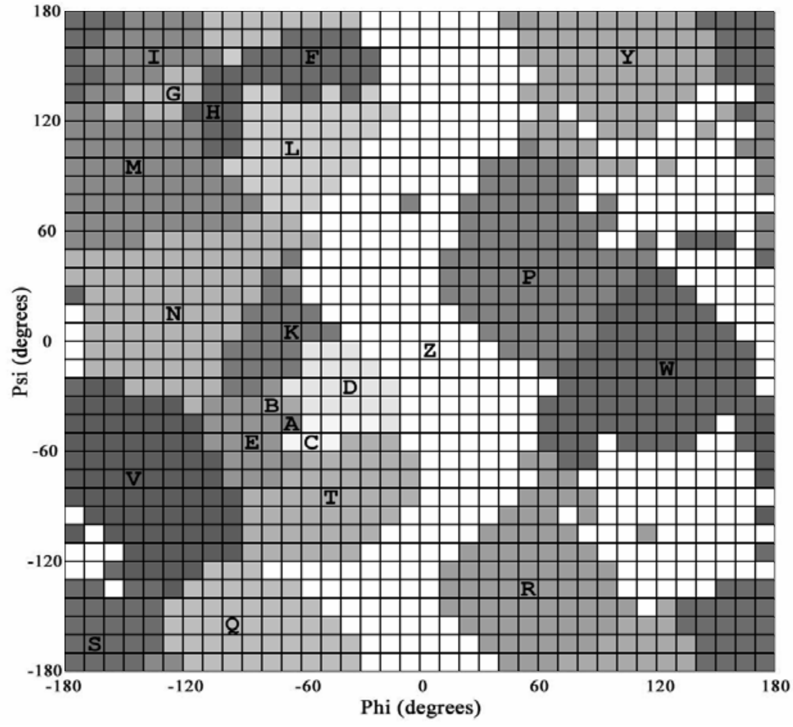


Root-Square Angular Distance

$$RSAD = \sqrt{(\Delta\varphi)^2 + (\Delta\psi)^2}$$

SARST

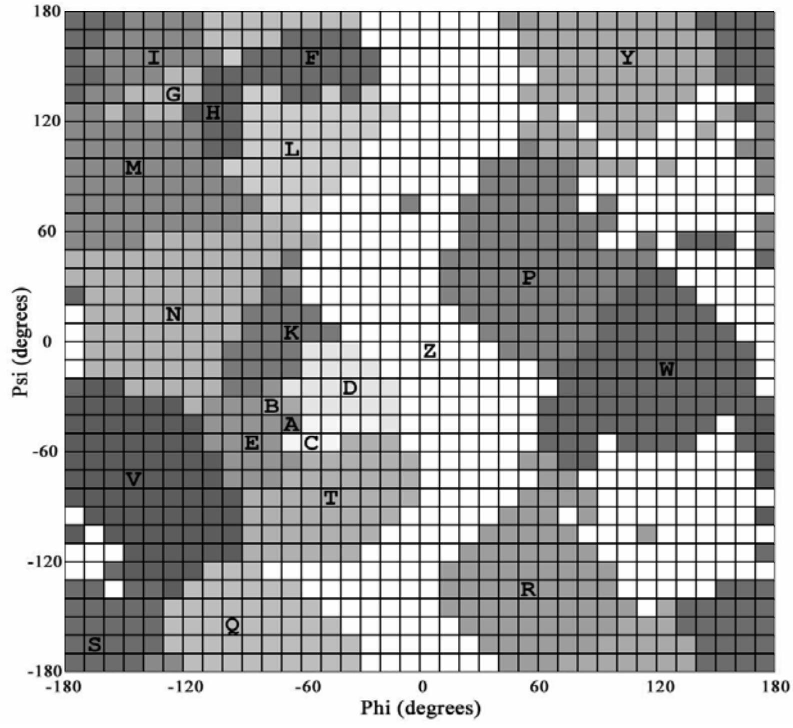
Ramachandran Sequential Transformation



- Nearest-neighbor clustering
- 1,296 cells were clustered into 22 groups
- Each group was assigned with a symbol, i.e. Ramachandran code

SARST

Ramachandran Sequential Transformation



○ ○ ○ ○ ○

RM Seq: I L L P C

How to Evaluate Similarities?

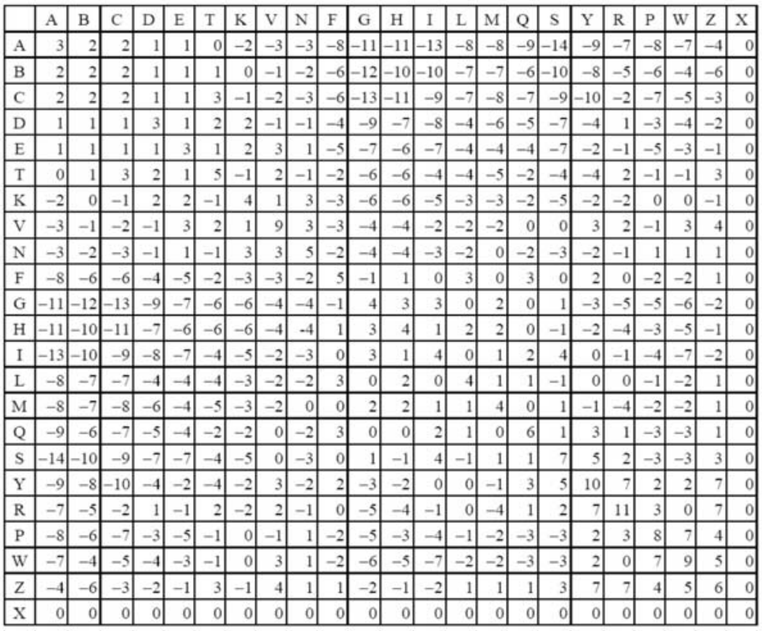
AAAAWWW
AAAAAWW

WWWWAAA
WWWWWAA

Are they equally similar?
Score A:A = ?
Score W:W = ?
Score A:W and W:A = ?

SARST

BLOSUM-like Scoring Matrix for SARST



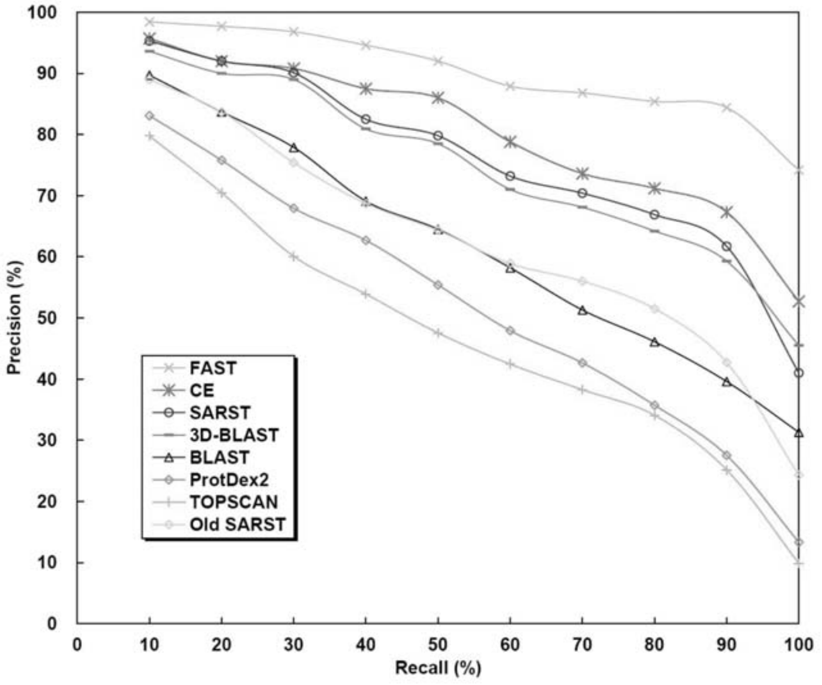
	A	B	C	D	E	T	K	V	N	F	G	H	I	L	M	Q	S	Y	R	P	W	Z	X
A	3	2	2	1	1	0	−2	−3	−3	−8	−11	−11	−13	−8	−8	−9	−14	−9	−7	−8	−7	−4	0
B	2	2	2	1	1	1	0	−1	−2	−6	−12	−10	−10	−7	−7	−6	−10	−8	−5	−6	−4	−6	0
C	2	2	2	1	1	3	−1	−2	−3	−6	−13	−11	−9	−7	−8	−7	−9	−10	−2	−7	−5	−3	0
D	1	1	1	3	1	2	2	−1	−1	−4	−9	−7	−8	−4	−6	−5	−7	−4	1	−3	−4	−2	0
E	1	1	1	1	3	1	2	3	1	−5	−7	−6	−7	−4	−4	−4	−7	−2	−1	−5	−3	−1	0
T	0	1	3	2	1	5	−1	2	−1	−2	−6	−6	−4	−4	−5	−2	−4	−4	2	−1	−1	3	0
K	−2	0	−1	2	2	−1	4	1	3	−3	−6	−6	−5	−3	−3	−2	−5	−2	−2	0	0	−1	0
V	−3	−1	−2	−1	3	2	1	9	3	−3	−4	−4	−2	−2	−2	0	0	3	2	−1	3	4	0
N	−3	−2	−3	−1	1	−1	3	3	5	−2	−4	−4	−3	−2	0	−2	−3	−2	−1	1	1	1	0
F	−8	−6	−6	−4	−5	−2	−3	−3	−2	5	−1	1	0	3	0	3	0	2	0	−2	−2	1	0
G	−11	−12	−13	−9	−7	−6	−6	−4	−4	−1	4	3	3	0	2	0	1	−3	−5	−5	−6	−2	0
H	−11	−10	−11	−7	−6	−6	−6	−4	-4	1	3	4	1	2	2	0	−1	−2	−4	−3	−5	−1	0
I	−13	−10	−9	−8	−7	−4	−5	−2	−3	0	3	1	4	0	1	2	4	0	−1	−4	−7	−2	0
L	−8	−7	−7	−4	−4	−4	−3	−2	−2	3	0	2	0	4	1	1	−1	0	0	−1	−2	1	0
M	−8	−7	−8	−6	−4	−5	−3	−2	0	0	2	2	1	1	4	0	1	−1	−4	−2	−2	1	0
Q	−9	−6	−7	−5	−4	−2	−2	0	−2	3	0	0	2	1	0	6	1	3	1	−3	−3	1	0
S	−14	−10	−9	−7	−7	−4	−5	0	−3	0	1	−1	4	−1	1	1	7	5	2	−3	−3	3	0
Y	−9	−8	−10	−4	−2	−4	−2	3	−2	2	−3	−2	0	0	−1	3	5	10	7	2	2	7	0
R	−7	−5	−2	1	−1	2	−2	2	−1	0	−5	−4	−1	0	−4	1	2	7	11	3	0	7	0
P	−8	−6	−7	−3	−5	−1	0	−1	1	−2	−5	−3	−4	−1	−2	−3	−3	2	3	8	7	4	0
W	−7	−4	−5	−4	−3	−1	0	3	1	−2	−6	−5	−7	−2	−2	−3	−3	2	0	7	9	5	0
Z	−4	−6	−3	−2	−1	3	−1	4	1	1	−2	−1	−2	1	1	1	3	7	7	4	5	6	0
X	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

BLOSUM algorithm:
Henikoff and Henikoff. (1992) *Proc Natl Acad Sci USA.* **89:**10915-10919

Speed Evaluation

Method	Average time per query (sec)	Average time per comparison (sec)	Relative to SARST
CE	82,789.20	2.43E+00	243,497.65
FAST	6,241.57	1.83E–01	18,357.56
TOPSCAN	85.08	2.50E–03	250.24
YAKUSA	35.6	1.05E–03	104.71
3D-BLAST	9.07	2.66E–04	26.68
ProtDex2	0.76	2.23E–05	2.24
BLAST	0.30	8.76E–06	0.88
SARST	0.34	9.98E–06	1.00
SARST (2 CPUs)	0.16	4.70E–06	0.47

Accuracy Evaluation



Information retrieval

- **Recall**

  the ability to extract answers

- **Precision**

  the ability to give correct answers

Next...

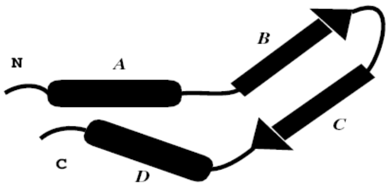
http://sarst.life.nthu.edu.tw/iSARST

CPSARST - Circular **P**ermutation **S**earch **A**ided by **R**amachandran **S**equential **T**ransformation

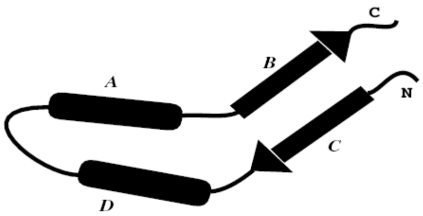
Lo WC, Lyu PC: ***CPSARST: an efficient circular permutation search tool applied to the detection of novel protein structural relationships.*** Genome Biology 2008,9:R11.

Circular Permutation (CP)

- Circular permutation of a protein can be visualized as if the original N- and C-termini were linked and new ones created elsewhere[1].
- In most of the cases, naturally occurring CPs have similar 3D structures and conserved biological functions[2].
- Efficient CP search tool is not available yet.



The sequence: ..A..B..C..D..



The sequence ..C..D..A..B..

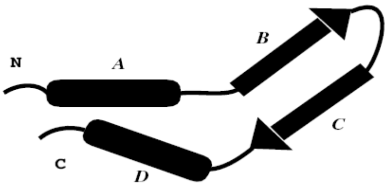
1. Uliel S et al.: **A simple algorithm for detecting circular permutations in proteins.** *Bioinformatics* 1999,**15**:930-936.
2. Lindqvist Y, Schneider G: **Circular permutations of natural protein sequences: structural evidence.** *Curr Opin Struct Biol* 1997,**7**:422-427.

Natural Circular Permutants

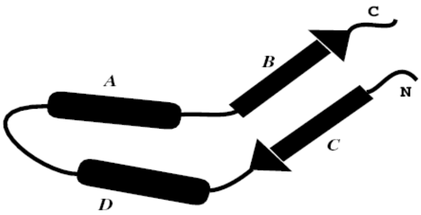
- Plant lectins
- Transaldolases
- DNA and other methyltransferases
- Ferredoxins
- Proteinase inhibitors
- Bacterial β-glucanases
- Swaposins
- Glucosyltransferases
- β-glucosidases
- SLH domains
- C2 domains
- FMN-binding proteins
- Double-φβ-barrels
- Glutathione synthetases

Circular Permutation (CP)

- Circular permutation of a protein can be visualized as if the original N- and C-termini were linked and new ones created elsewhere[1].
- In most of the cases, CPs have similar 3D structures and conserved biological functions[2].
- Efficient CP search tool is not available yet.



The sequence: ..A..B..C..D..



The sequence ..C..D..A..B..

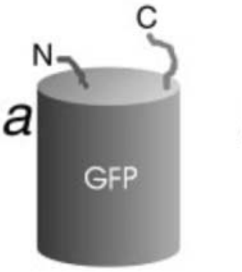
1. Uliel S et al.: **A simple algorithm for detecting circular permutations in proteins.** *Bioinformatics* 1999,**15**:930-936.
2. Lindqvist Y, Schneider G: **Circular permutations of natural protein sequences: structural evidence.** *Curr Opin Struct Biol* 1997,**7**:422-427.

Applications of Circular Permutation

- Folding researches.
- Determination of structurally and functionally important segments[1,2].
- Modification (enhancement) of the activity and/or stability[3-5].
- Creation of novel fusion proteins, the tethered sites of which are not confined to the native termini[5,6].

1. Anand.B. et al. Nucleic Acid Res 2006;34:2196-2205.
2. Gebhard.LG. et al. J Mol Biol 2006;358:280-288.
3. Qian.Z., Lutz.S. J Am Chem Soc 2005;127:13466-13467.
4. Schwartz.TU. et al. Protein Sc 2004;13:2814-2818.
5. Kojima.M. et al. J Biosci Bioeng 2005;100:197-202
6. Baird.GS. et al. Proc Natl Acad Sci USA 1999;96:11241-11246.

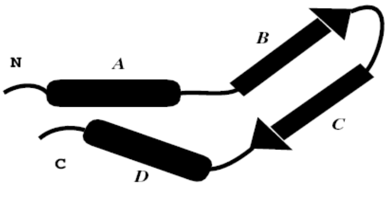
Fluorescent Calcium Sensor with CP



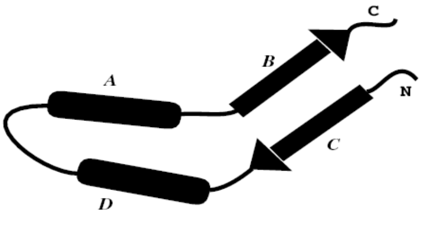
G.S. Baird, et al. **Circular permutation and receptor insertion within green fluorescent proteins.** *PNAS* 1999;96:11241-11246

Circular Permutation (CP)

- Circular permutation of a protein can be visualized as if the original N- and C-termini were linked and new ones created elsewhere[1].
- In most of the cases, naturally occurring CPs have similar 3D structures and conserved biological functions[2].
- **Efficient CP search tool is not available yet.**



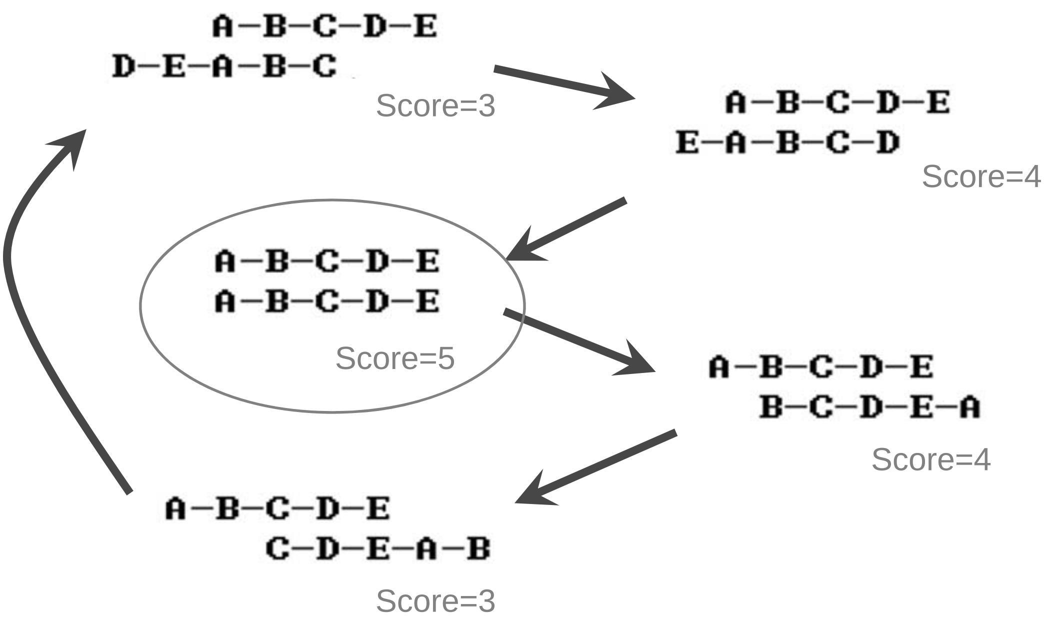
The sequence: ..A..B..C..D..



The sequence ..C..D..A..B..

1. Uliel S et al.: **A simple algorithm for detecting circular permutations in proteins.** *Bioinformatics* 1999,**15**:930-936.
2. Lindqvist Y, Schneider G: **Circular permutations of natural protein sequences: structural evidence.** *Curr Opin Struct Biol* 1997,**7**:422-427.

Basic Approach to the Detection of CP



A-B-C-D-E
D-E-A-B-C
Score=3

A-B-C-D-E
E-A-B-C-D
Score=4

A-B-C-D-E
A-B-C-D-E
Score=5

A-B-C-D-E
B-C-D-E-A
Score=4

A-B-C-D-E
C-D-E-A-B
Score=3

The Basic Idea of CPSARST

Target: A-B-C-D-E

Query: D-E-A-B-C

A-B-C-D-E

Answer 1: A-B-C-D-E
D-E-A-B-C

D-E-A-B-C-D-E-A-B-C

Score = 2

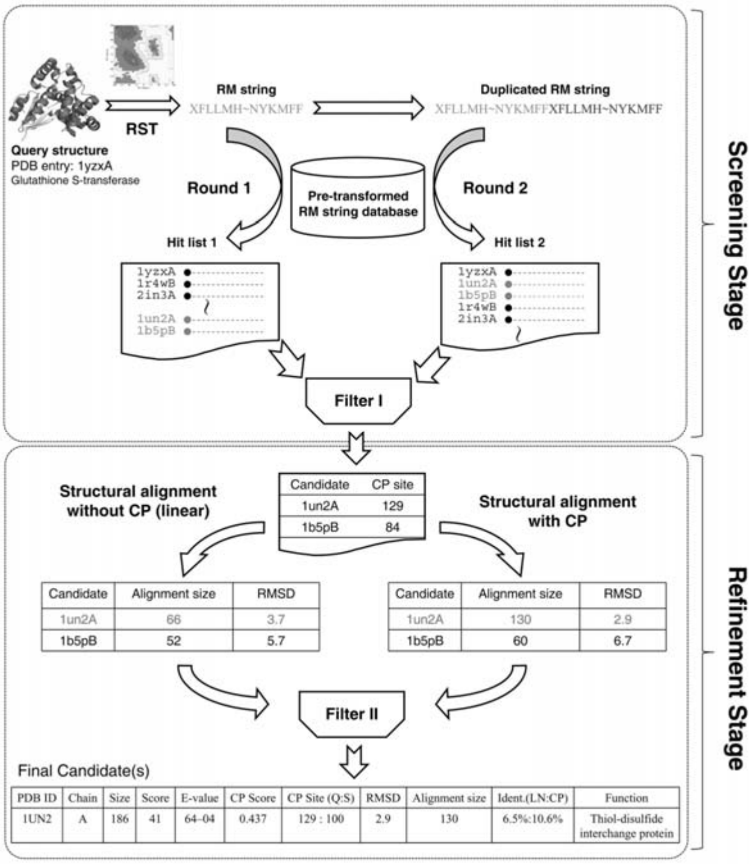
CP site

Answer 2: A-B-C-D-E
D-E-A-B-C

Score = 3

Score = 5

The Double Filter-and-Refine Strategy



Candidate	CP site
1un2A	129
1b5pB	84

Candidate	Alignment size	RMSD
1un2A	66	3.7
1b5pB	52	5.7

Candidate	Alignment size	RMSD
1un2A	130	2.9
1b5pB	60	6.7

Final Candidate(s)

PDB ID	Chain	Size	Score	E-value	CP Score	CP Site (Q:S)	RMSD	Alignment size	Ident.(LN:CP)	Function
1UN2	A	186	41	64–04	0.437	129 : 100	2.9	130	6.5%:10.6%	Thiol-disulfide interchange protein

Statistics of protein structural database searches by CPSARST

Database			nrPDB-90	nrSCOP-90
No. of proteins			14,422	11,688
No. of candidate pairs	1. Detected by amino acid sequence		5,020	1,802
	2. Detected only by Ramachandran string		252,287	196,533
	3. Confirmed after the refinement stage	Total	2,911	4,228
		Symmetric CP	682	1,161
Total No. of protein pairs			208.0×10^6	136.6×10^6
Total running time (minutes)			3,942	1,974
No. of protein pairs scanned per minute			52,764	69,204

Speed Advantage of CPSARST

- 4 times faster than UFAU (sequence-based)
  - Uliel S et al.: **A simple algorithm for detecting circular permutations in proteins.** *Bioinformatics* 1999,**15**:930-936.
- 8,824 times faster than SAMO (structure-based)
  - Chen L et al.: **Revealing divergent evolution, identifying circular permutations and detecting active-sites by protein structure comparison**. *BMC Struct Biol* 2006, **6**:18.
- CPSARST requires only 1.7 minute to scan the current PDB (~90,000 polypeptides).

Performance of pair-wise comparisons for natural candidate CP pairs over various sequence identities

$$\frac{\text{RMSD}}{\left(\frac{\text{Alignment size}}{\text{Average protein size}}\right)^{1.5}}$$

Identity (%)	No. of candidate CP pairs	Structural diversity		
		CPSARST	SHEBA	SAMO
≤ 10	823	6.309	11.180	4.396
10 ~20	152	5.864	13.881	4.994
20 ~ 30	11	3.581	4.506	3.363
30 ~ 40	33	1.868	3.284	2.210
40 ~ 50	40	1.755	3.096	1.544
> 50	9	1.385	2.247	1.520

Lu G: **Top: A new method for protein structure comparisons and similarity searches.** *J Appl Cryst* 2000,**33**:176-183.

Top 20 homologs retrieved from nrPDB by DALI for hypothetical protein YlqF

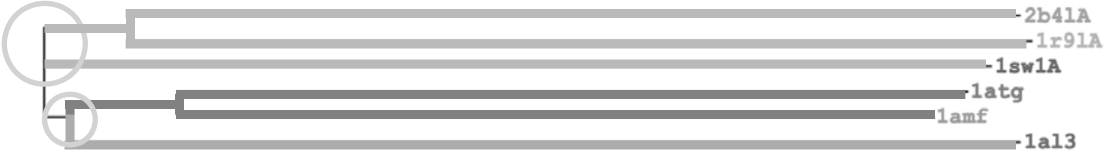
No.	PDB entry / Size	Function
1	1pujA / 261	Conserved hypothetical protein YlqF
2	1u0lA / 278	Probable GTPase
3	1ctqA / 166	p21h-Ras-1 fragment
4	1ejjA / 508	Phosphoglycerate mutase (isomerase)
5	1gpmA / 501	Amidotransferase, GMP synthetase
6	1efcA / 386	Elongation factor Eftu (RNA binding)
7	1hrkA / 359	Ferrochelatase fragment (lyase)
8	1ni5A / 428	Putative cell cycle protein Mesj
9	1dpgA / 485	Glucose 6-phosphate reductase
10	2hjgA / 390	GTP-binding protein engA
11	1veeA / 134	Unknown function proline-rich protein
12	1cqxA / 403	Flavohemoprotein (lipid binding)
13	2p8zT / 813	Elongation factor 2
14	1mkyA / 400	Probable GTP-binding protein
15	1dar / 615	Elongation factor G (translational GTPase)
16	1kk1A / 397	Eif2gamma mutant
17	1hurA / 180	Human ADP-ribosylation factor 1
18	1fdr / 244	Flavodoxin reductase
19	2clsA / 179	Rho-related GTP-binding protein
20	1wcwA / 254	Uroporphyrinogen III synthase
21	1ak1 / 308	Ferrochelatase

Top 20 circular permutants detected from nrPDB by CPSARST for hypothetical protein YlqF

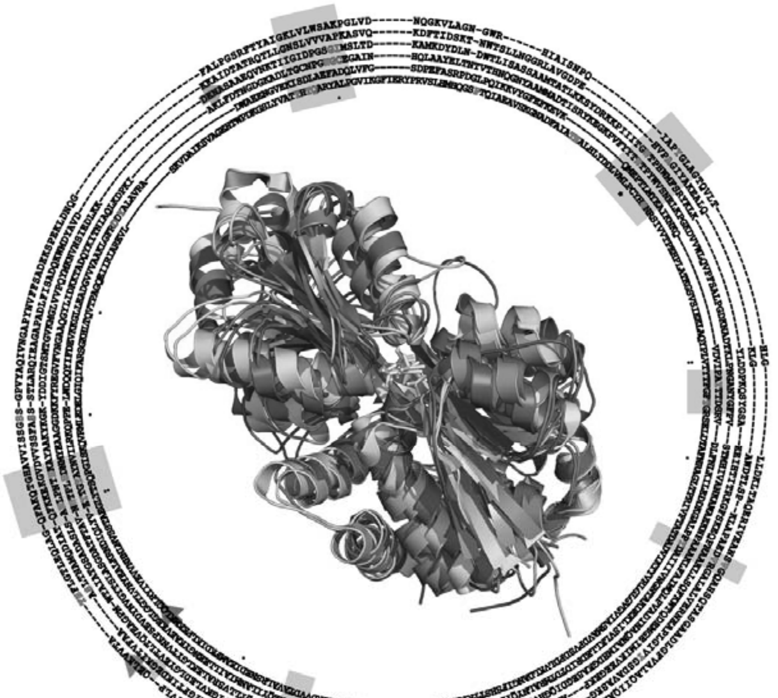
No.	PDB entry / Size	Function
1	1ZBD / 203	Rabphilin-3A
2	1KY2 / 182	GTP-binding
3	2F7S / 217	Ras-related protein Rab-27B protein YPT7P
4	2NZJ / 175	GTP-binding protein REM 1
5	1T91 / 207	Ras-related protein Rab-7
6	1X3S / 195	Ras-related protein Rab-18
7	1YU9 / 175	GTP-binding protein, GTPase domain
8	2EW1 / 201	Ras-related protein Rab-30
9	2GF9 / 189	Ras-related protein Rab-3D
10	1YVD / 169	Ras-related protein Rab-22A
11	1PUI / 210	Probable GTP-binding protein engB
12	2O52 / 200	Ras-related protein Rab-4B
13	1U8Y / 168	Ras-related protein Ral-A
14	1HUQ / 164	Rab5C, GTPase domain
15	2HUP / 201	Ras-related protein Rab-43
16	1FZQ / 181	ADP-ribosylation factor-like protein 3
17	2OCB / 180	Ras-related protein Rab-9B
18	1OIV / 191	Ras-related protein Rab-11A
19	2FN4 / 181	Ras-related protein R-Ras
20	1Z0F / 179	Rab14, member Ras oncogene family

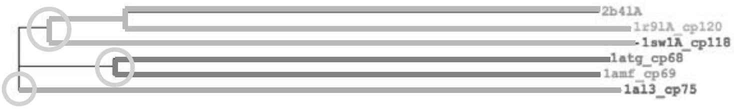
Multiple Alignment of Raw Sequences

```
1atg      --------------------------------------------ELKVVTATNFLGTLEQLAGQFAKQTGHAVVISSGSSGPVYAQIVNGAPYNVFFSADEKSPEKLDN----QGFALPG 72
1amf      -----------------------------------------DEGKITVFAAASLTNAMQDIATQFKKEKGVDVVSSFASSSTLARQIEAGAPADLFISADQKWMDYAVD----KKAIDTA 75
1al3      MKLQQLRYIVEVVNHNLNVSSTAEGLYTSQPGISKQVRMLEDELGIQIFARSGKHLTQVTPAGQEIIRIAREVLSKVDAIKSVAGEHTWPDKGSLYVATTHTQARYALPGV-IKGFIERY 119
1sw1A     ------------GSQSSERVVIGSKPFNEQYILANMIAILLEENGYKAEVKEGLGGTLVNYEALKRNDIQLYVEYTGTAYNVILRKQPPELWDQQYIFDEVKKGLLEADG--VVVAAKLG 106
2b41A     ----------------------DENASAAEQVNKTIIGIDPGSGIMSLTDKAMKDYDLNDWTLISASSAAMTATLKKSYDRKKPIIITGWTPHWMFSRYKLKYLDDPKQSYGSAEEIHTI 98
1r91A     ---------------ADLPGKGITVNPVQSTITEETFQTLLVSRALEKLGYTVNKPSEVDYNVGYTSLASGDATFTAVNWTPLHDNMYEAAGGDKKFYREGVFVNGAAQGYLIDKKTADQ 105
                                                                                    .  .

1atg      SRFTYAIGKLVLWSAKPGLVDNQGKVLAGNGWR--------------HIAISNPQIAPYGLAGTQVLTHLGLLD---------------------KLTAQERIVEANSVGQAHSQTASGA 157
1amf      TRQTLLGNSLVVVAPKASVQKDFT-IDSKTNWTSLLN--------GGRLAVGDPEHVPAGIYAKEALQKLGAWD---------------------TLSP--KLAPAEDVRGALALVERNE 163
1al3      PRVSLHMHQGSPTQIAEAVSKGNADFAIATEALHLYDDLVMLPCYHWNRSIVVTPEHPLATKGSVSIEELAQYP---------------------LVTYTFGFTGRSELDTAFNRAGLTP 218
1sw1A     FRDDYALAVRADWAEENGVEKISDLAEFADQLVFGSD---------PEFASRPDGLPQIKKVYGFEFKEVKQME---------------------PTLMYEAIKNKQVDVIPAYTTDSRV 196
2b41A     TRKGFSKEQPNAAKLLSQFKWTQDEMGEIMIKVEEGE---------KPAKVAAEYVNKHKDQIAEWTKGVQKVK---------------------GDKINLAYVAWDSEIASTNVIGKVL 188
1r91A     YKITN-IAQLKDPKIAKLFDTNGDGKADLTGCNPGWGCEGAINHQLAAYELTNTVTHNQGNYAAMMADTISRYKEGKPVFYYTWTPYWVSNELKPGKDVVWLQVPFSALPGDKNADTKLP 224
           :                .                                                  :                                    .

1atg      ADLGFVALAQIIQAAAKIPGSHWFPPANYYEPIVQQAVITKST--------------AEKANAEQFMSWMK--GPKAVAIIKAAGYVLPQ---------------- 231
1amf      APLGIVYGSDAVASKG-VKVVATFPEDSHKK--VEYPVAVVEG--------------HNNATVKAFYDYLK--GPQAAEIFKRYGFTIK----------------- 233
1al3      RIVFTATDADVIKTYVRLGLGVGVIASMAVDPVSDPDLVKLDANGIFSHSTTKIGFRRSTFLRSYMYDFIQRFAPHLTRDVVDTAVALRSNEDIEAMFKDIKLPEK 324
1sw1A     DLFNLKILEDDKGALPPYDAIIIVNGNTAKDEKLISVLKLLEDR----------IDTDTMRALNYQYDVEKKDAREIAMSFLKEQGLVK----------------- 275
2b41A     EDLGYEVTLTQVEAGPMWTAIATGSADASLSAWLPNTHKAYAAKYKG---------KYDDIGTSMTGVKMGLVVPQYMKNVNSIEDLKK----------------- 268
1r91A     NGANYGFPVSTMHIVANKAWAEKNPAAAKLFAIMQLPVADINAQNAIMHDG----KASEGDIQGHVDGWIKAHQQQFDGWVNEALAAQK----------------- 309
                                                                                    .   .
```



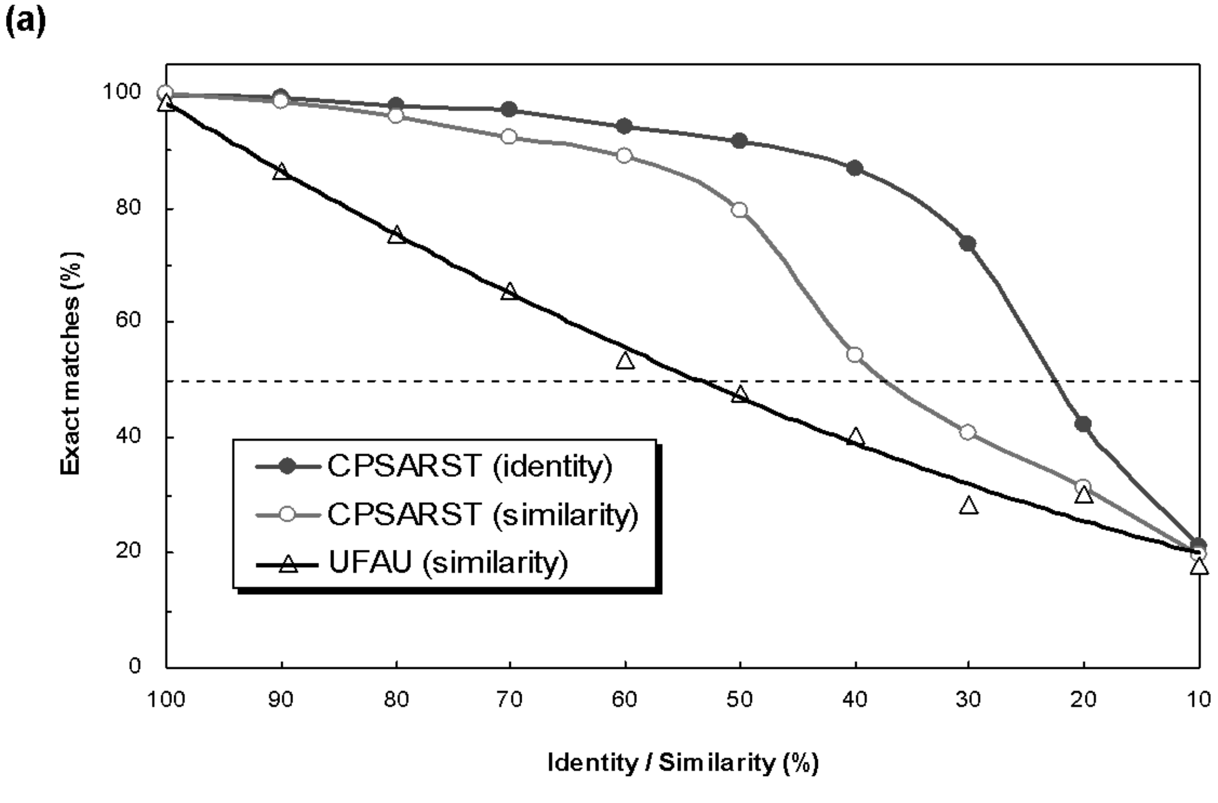
Multiple Alignment of Circularly-Permutated Sequences





Possible Applications of CPSARST

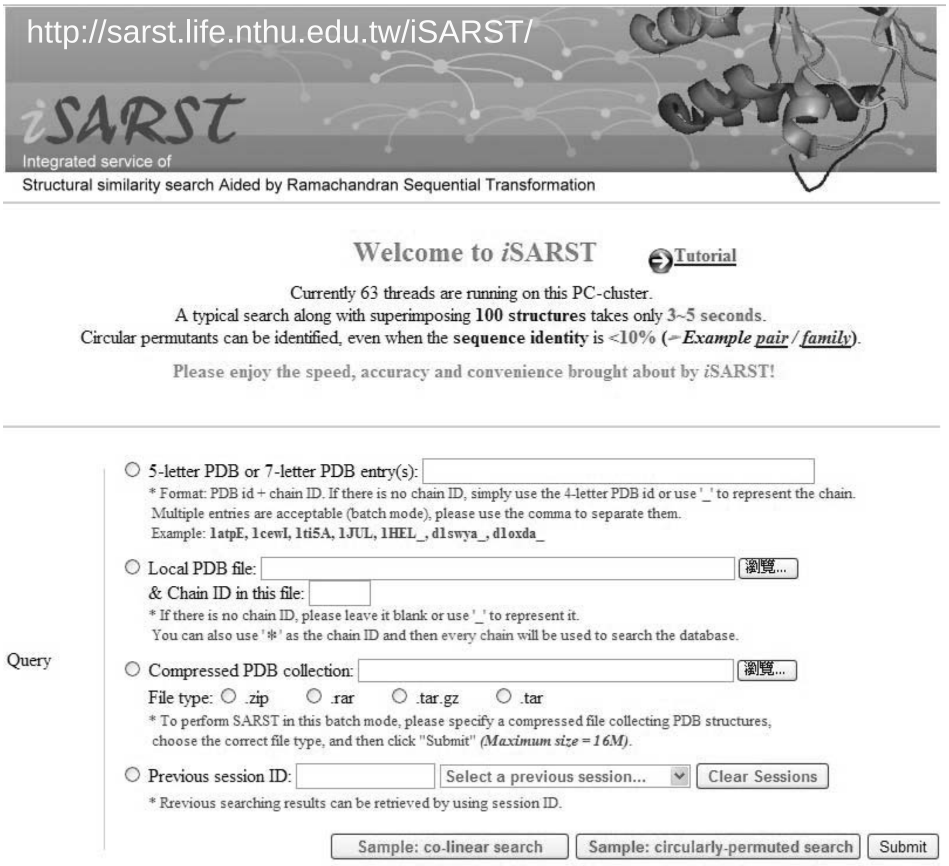
- Bank-against-bank searches are achievable.
- Develop automated procedures such as the functional assignment system for novel hypothetical proteins
- Construct CP database



Next…

iSARST
Integrated service of
Structural similarity search Aided by Ramachandran Sequential Transformation

http://sarst.life.nthu.edu.tw/iSARST


http://sarst.life.nthu.edu.tw/iSARST/

iSARST
Integrated service of
Structural similarity search Aided by Ramachandran Sequential Transformation

Welcome to *i*SARST   Tutorial

Currently 63 threads are running on this PC-cluster.
A typical search along with superimposing **100 structures** takes only **3~5 seconds**.
Circular permutants can be identified, even when the **sequence identity** is **<10%** (***Example pair / family***).

Please enjoy the speed, accuracy and convenience brought about by *i*SARST!

Query

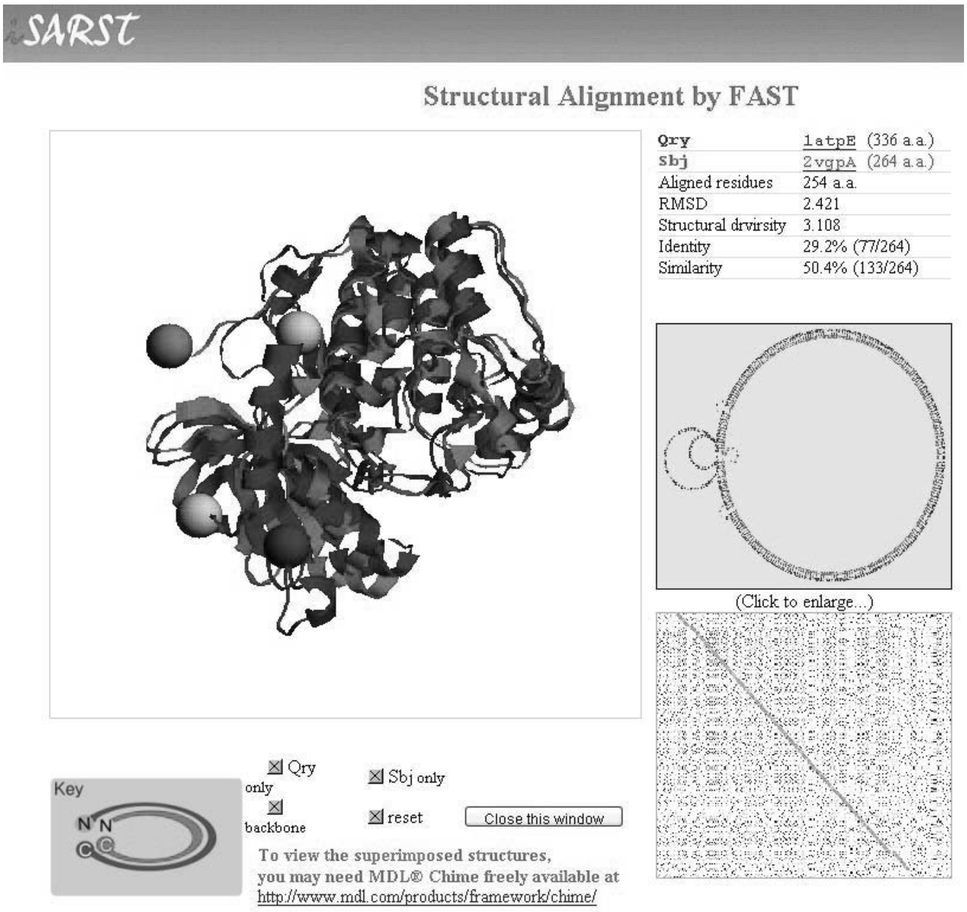
○ 5-letter PDB or 7-letter PDB entry(s):
* Format: PDB id + chain ID. If there is no chain ID, simply use the 4-letter PDB id or use '_' to represent the chain.
Multiple entries are acceptable (batch mode), please use the comma to separate them.
Example: **1atpE, 1cewI, 1ti5A, 1JUL, 1HEL_, d1swya_, d1oxda_**

○ Local PDB file:
& Chain ID in this file:
* If there is no chain ID, please leave it blank or use '_' to represent it.
You can also use '*' as the chain ID and then every chain will be used to search the database.

○ Compressed PDB collection:
File type: ○ .zip ○ .rar ○ .tar.gz ○ .tar
* To perform SARST in this batch mode, please specify a compressed file collecting PDB structures,
choose the correct file type, and then click "Submit" ***(Maximum size = 16M)***.

○ Previous session ID:
* Rrevious searching results can be retrieved by using session ID.

瀏覽...
瀏覽...
Select a previous session...
Clear Sessions
Sample: co-linear search
Sample: circularly-permuted search
Submit



iSARST

Structural Alignment by FAST

Qry	1atpE (336 a.a.)
Sbj	2vgpA (264 a.a.)
Aligned residues	254 a.a.
RMSD	2.421
Structural drvirsity	3.108
Identity	29.2% (77/264)
Similarity	50.4% (133/264)

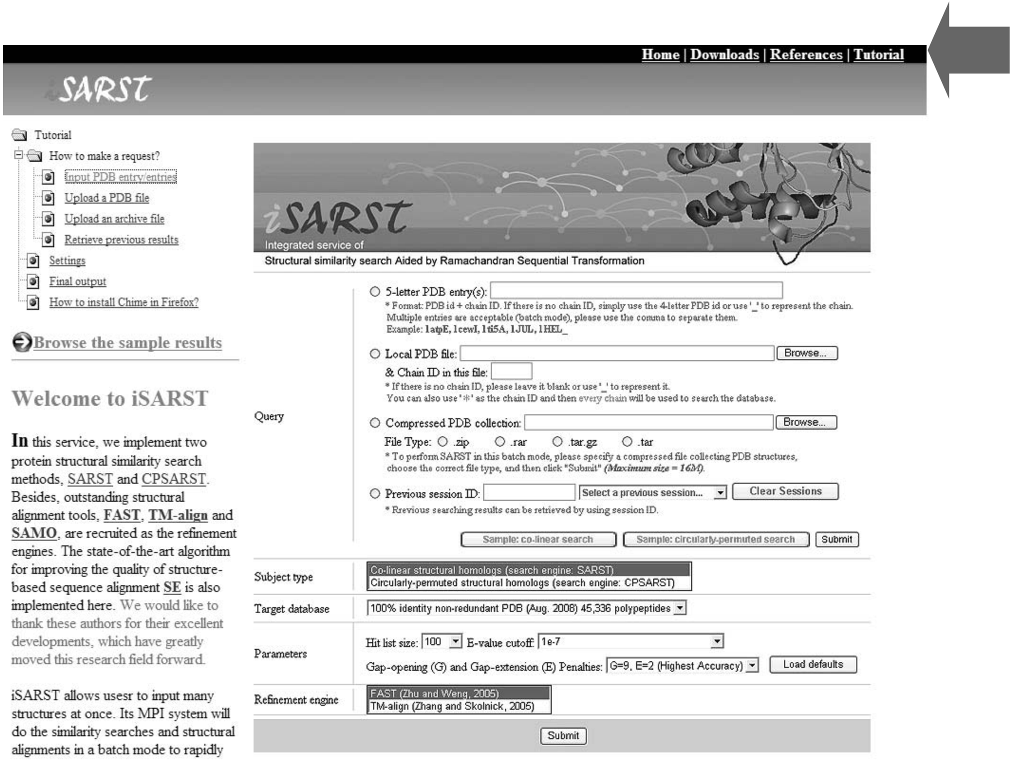
(Click to enlarge...)

☒ Qry only ☒ Sbj only

☒ backbone ☒ reset

To view the superimposed structures, you may need MDL® Chime freely available at http://www.mdl.com/products/framework/chime/

Tutorial of *i*SARST



Home | Downloads | References | Tutorial
iSARST
Tutorial
How to make a request?
Input PDB entry/entries
Upload a PDB file
Upload an archive file
Retrieve previous results
Settings
Final output
How to install Chime in Firefox?
Browse the sample results

Welcome to iSARST

In this service, we implement two protein structural similarity search methods, SARST and CPSARST. Besides, outstanding structural alignment tools, **FAST**, **TM-align** and **SAMO**, are recruited as the refinement engines. The state-of-the-art algorithm for improving the quality of structure-based sequence alignment **SE** is also implemented here. We would like to thank these authors for their excellent developments, which have greatly moved this research field forward.

iSARST allows usesr to input many structures at once. Its MPI system will do the similarity searches and structural alignments in a batch mode to rapidly

iSARST

Integrated service of

Structural similarity search Aided by Ramachandran Sequential Transformation

Query

○ 5-letter PDB entry(s):

* Format: PDB id + chain ID. If there is no chain ID, simply use the 4-letter PDB id or use '_' to represent the chain. Multiple entries are acceptable (batch mode), please use the comma to separate them.
Example: **1atpE, 1cewI, 1ti5A, 1JUL, 1HEL_**

○ Local PDB file: Browse...

& Chain ID in this file:

* If there is no chain ID, please leave it blank or use '_' to represent it.
You can also use '*' as the chain ID and then every chain will be used to search the database.

○ Compressed PDB collection: Browse...

File Type: ○ .zip ○ .rar ○ .tar.gz ○ .tar

* To perform SARST in this batch mode, please specify a compressed file collecting PDB structures, choose the correct file type, and then click "Submit" ***(Maximum size = 16M)***.

○ Previous session ID: Select a previous session... Clear Sessions

* Rrevious searching results can be retrieved by using session ID.

Subject type: Co-linear structural homologs (search engine: SARST) / Circularly-permuted structural homologs (search engine: CPSARST)

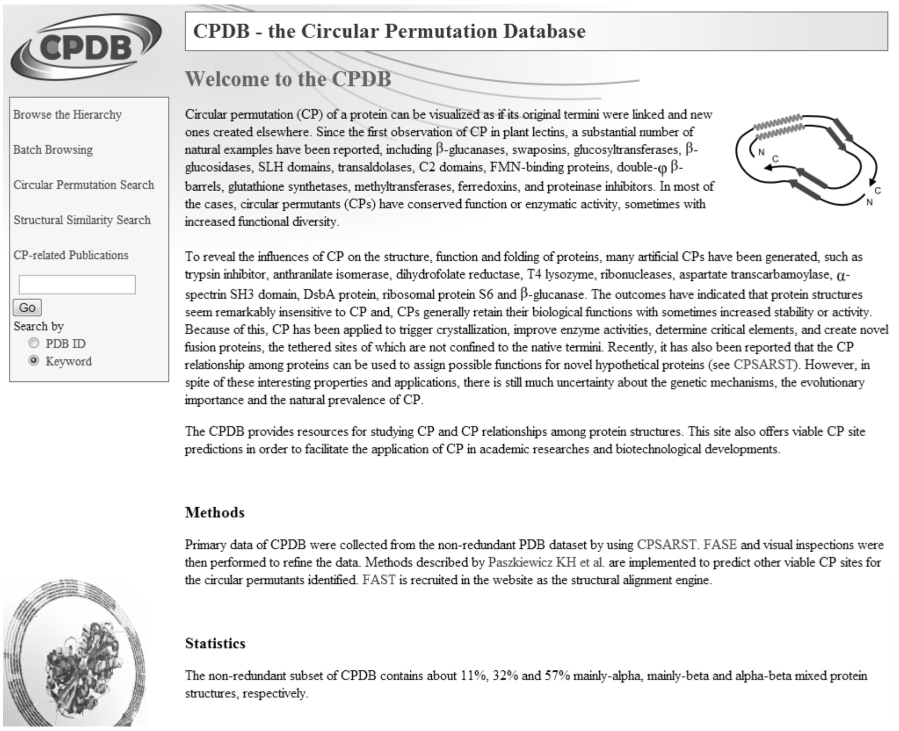
Target database: 100% identity non-redundant PDB (Aug. 2008) 45,336 polypeptides

Parameters: Hit list size: 100 E-value cutoff: 1e-7
Gap-opening (G) and Gap-extension (E) Penalties: G=9, E=2 (Highest Accuracy) Load defaults

Refinement engine: FAST (Zhu and Weng, 2005) / TM-align (Zhang and Skolnick, 2005)

http://sarst.life.nthu.edu.tw/iSARST/hlp/tutorial.php

CPDB:
a database of circular permutation in proteins



CPDB - the Circular Permutation Database

Welcome to the CPDB

Circular permutation (CP) of a protein can be visualized as if its original termini were linked and new ones created elsewhere. Since the first observation of CP in plant lectins, a substantial number of natural examples have been reported, including β-glucanases, swaposins, glucosyltransferases, β-glucosidases, SLH domains, transaldolases, C2 domains, FMN-binding proteins, double-φ β-barrels, glutathione synthetases, methyltransferases, ferredoxins, and proteinase inhibitors. In most of the cases, circular permutants (CPs) have conserved function or enzymatic activity, sometimes with increased functional diversity.

To reveal the influences of CP on the structure, function and folding of proteins, many artificial CPs have been generated, such as trypsin inhibitor, anthranilate isomerase, dihydrofolate reductase, T4 lysozyme, ribonucleases, aspartate transcarbamoylase, α-spectrin SH3 domain, DsbA protein, ribosomal protein S6 and β-glucanase. The outcomes have indicated that protein structures seem remarkably insensitive to CP and, CPs generally retain their biological functions with sometimes increased stability or activity. Because of this, CP has been applied to trigger crystallization, improve enzyme activities, determine critical elements, and create novel fusion proteins, the tethered sites of which are not confined to the native termini. Recently, it has also been reported that the CP relationship among proteins can be used to assign possible functions for novel hypothetical proteins (see CPSARST). However, in spite of these interesting properties and applications, there is still much uncertainty about the genetic mechanisms, the evolutionary importance and the natural prevalence of CP.

The CPDB provides resources for studying CP and CP relationships among protein structures. This site also offers viable CP site predictions in order to facilitate the application of CP in academic researches and biotechnological developments.

Methods

Primary data of CPDB were collected from the non-redundant PDB dataset by using CPSARST. FASE and visual inspections were then performed to refine the data. Methods described by Paszkiewicz KH et al. are implemented to predict other viable CP sites for the circular permutants identified. FAST is recruited in the website as the structural alignment engine.

Statistics

The non-redundant subset of CPDB contains about 11%, 32% and 57% mainly-alpha, mainly-beta and alpha-beta mixed protein structures, respectively.

http://sarst.life.nthu.edu.tw/cpdb/

Thanks for your attentions.