7.91 April 1, 2004 Amy Keating

Protein Structure

Outline of the next part of the course

4/1	Protein Structure Comparison & Classification
4/6	Principles of Molecular Mechanics
4/8	X-ray crystallography and NMR
4/13	Modeling Mutants and Homologs
4/15	Threading and Ab Initio Structure Prediction
4/22	Computational Protein Design

7.91 April 1, 2004 Amy Keating

Introduction to Protein Structure & Classification

Protein structures

basics
where to find them
how to look at them
what they can tell you
structural and evolutionary
comparisons

PDB ID: 1HCL

Schulze-Gahmen, U., J. Brandsen, H. D. Jones, D. O. Morgan, L. Meijer, J. Vesely, S. H. Kim. "Multiple Modes of Ligand Recognition: Crystal Structures of Cyclin-dependent Protein Kinase 2 in Complex with ATP and Two Inhibitors, Olomoucine and Isopentenyladenine." *Proteins* 22 (1995): 378.

The Protein Data Bank (PDB - http://www.pdb.org/) is the single worldwide repository for the processing and distribution of 3-D biological macromolecular structure data.

Berman, H. M., J. Westbrook, Z. Feng, G. Gilliland, T. N. Bhat, H. Weissig, I. N. Shindyalov, and P. E. Bourne. The Protein Data Bank. *Nucleic Acids Research* 28 (2000): 235-242

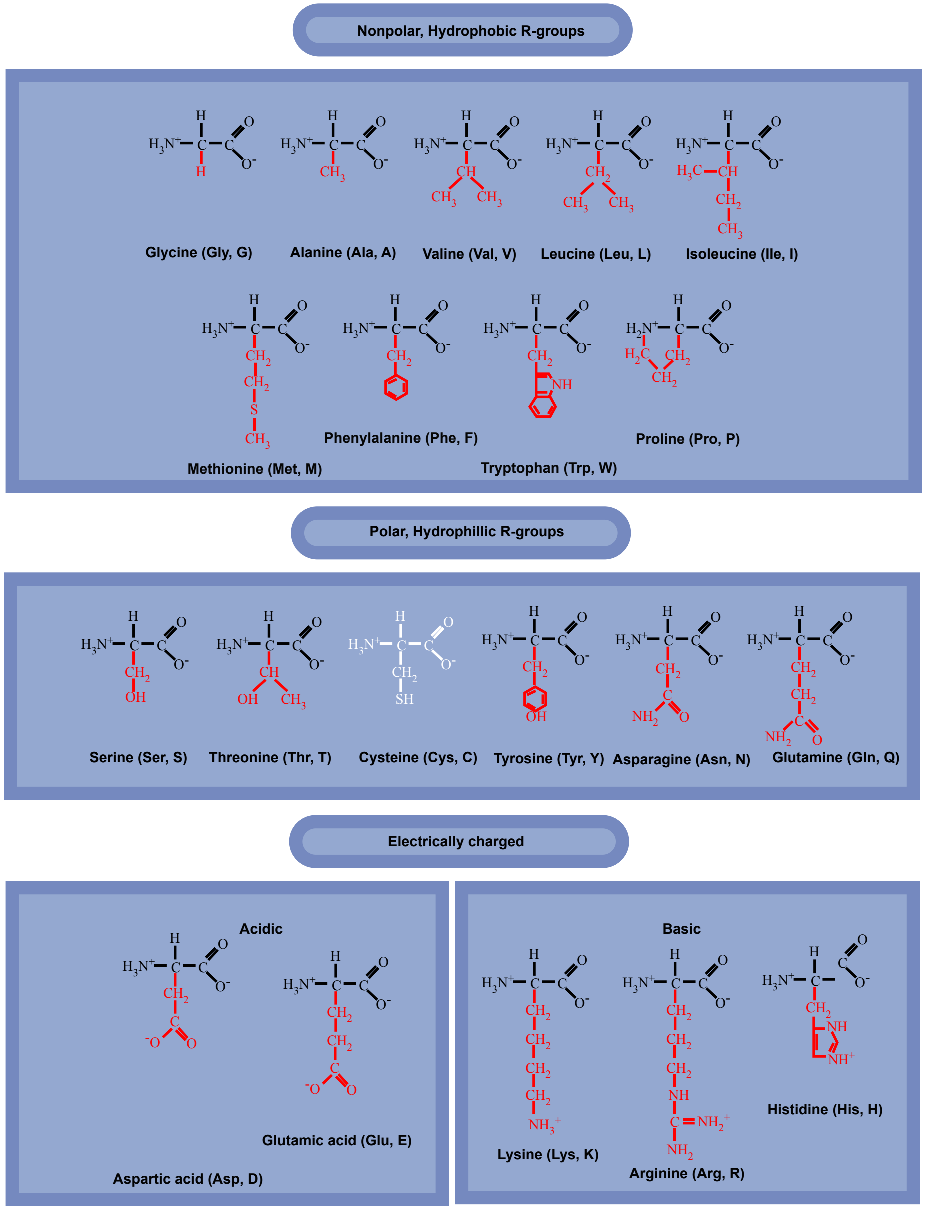
(PDB Advisory Notice on using materials available in the archive: http://www.rcsb.org/pdb/advisory.html)

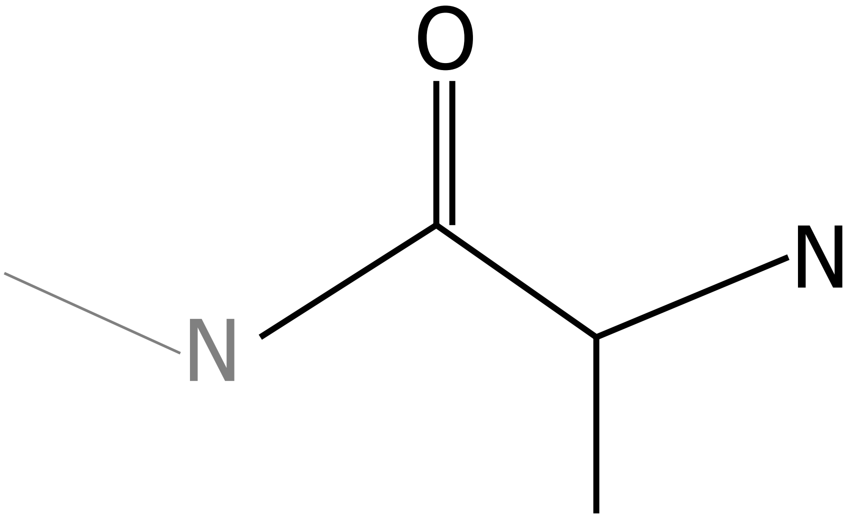
Review of protein structure hierarchy

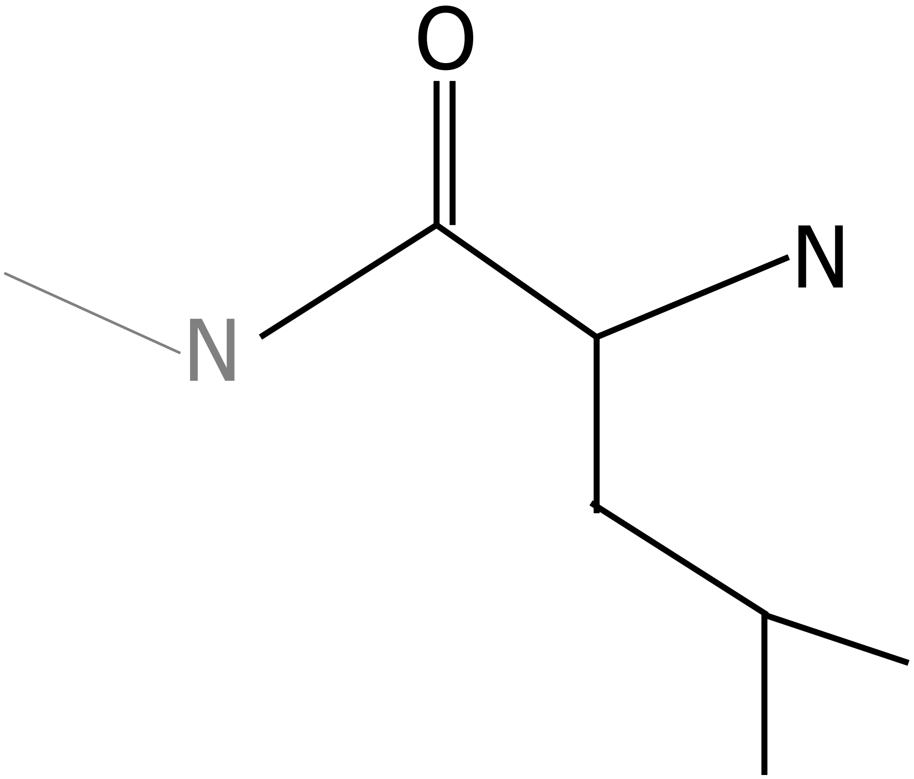
- *Primary structure*

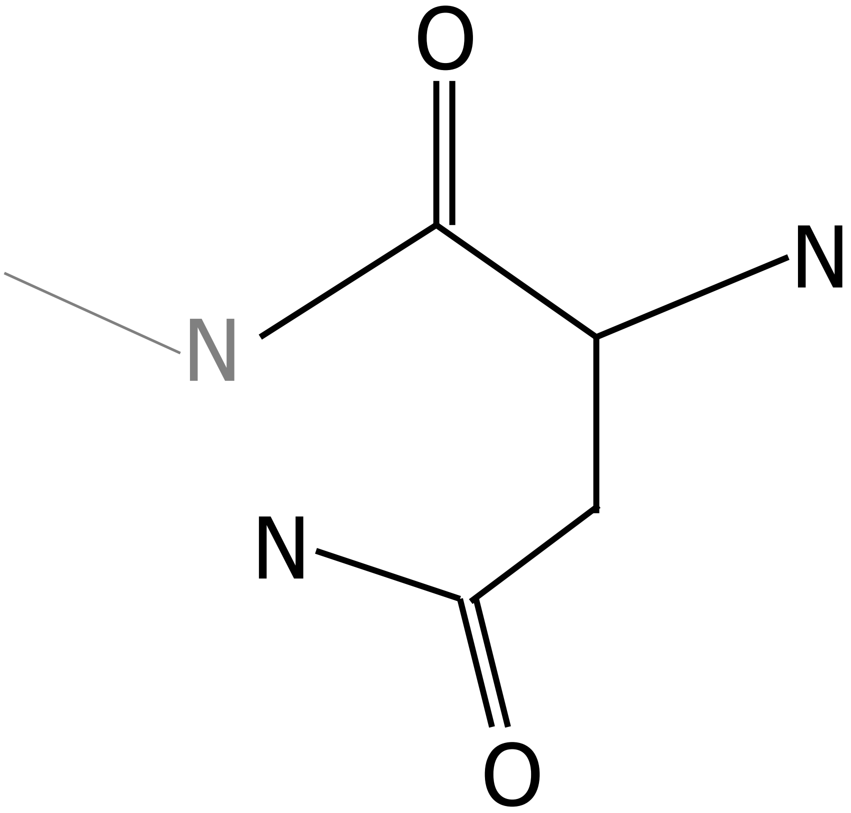
  MAAAAAAGPEMVRGQVF

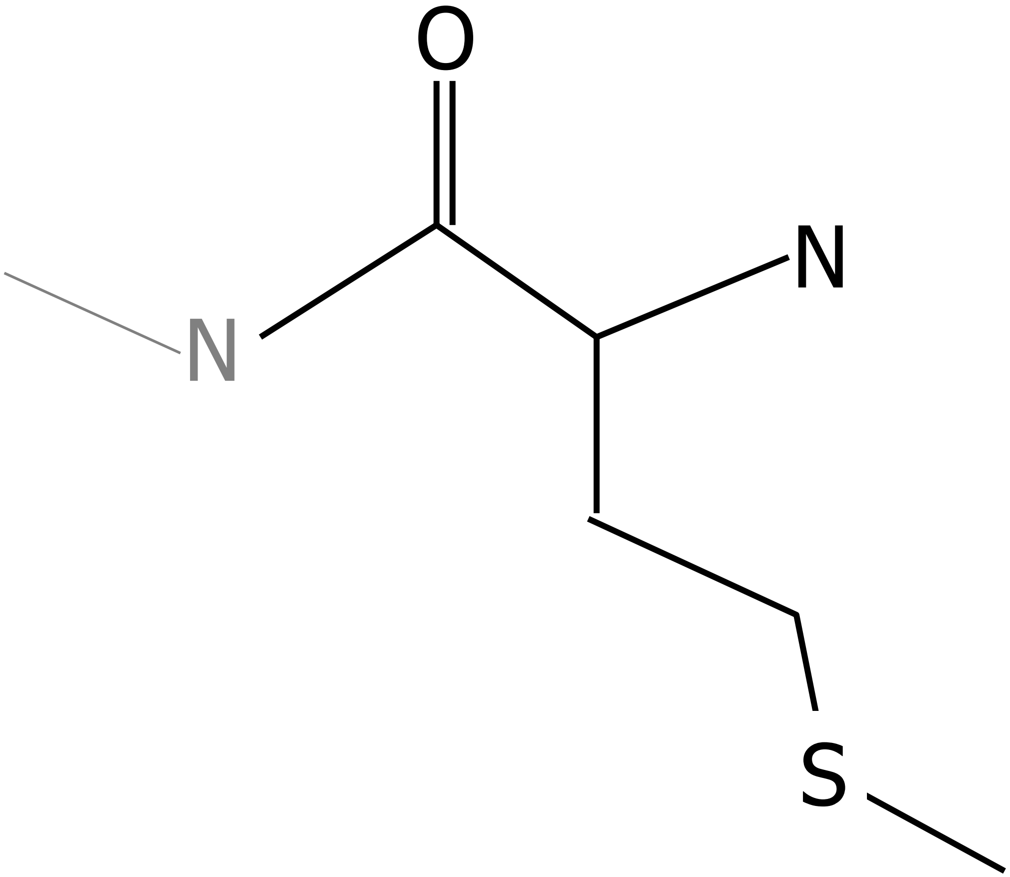
- 20 amino acids
  - hydrophobic/hydrophilic
  - acidic/basic
  - large/small
  - specialized (Gly,Pro,Cys)



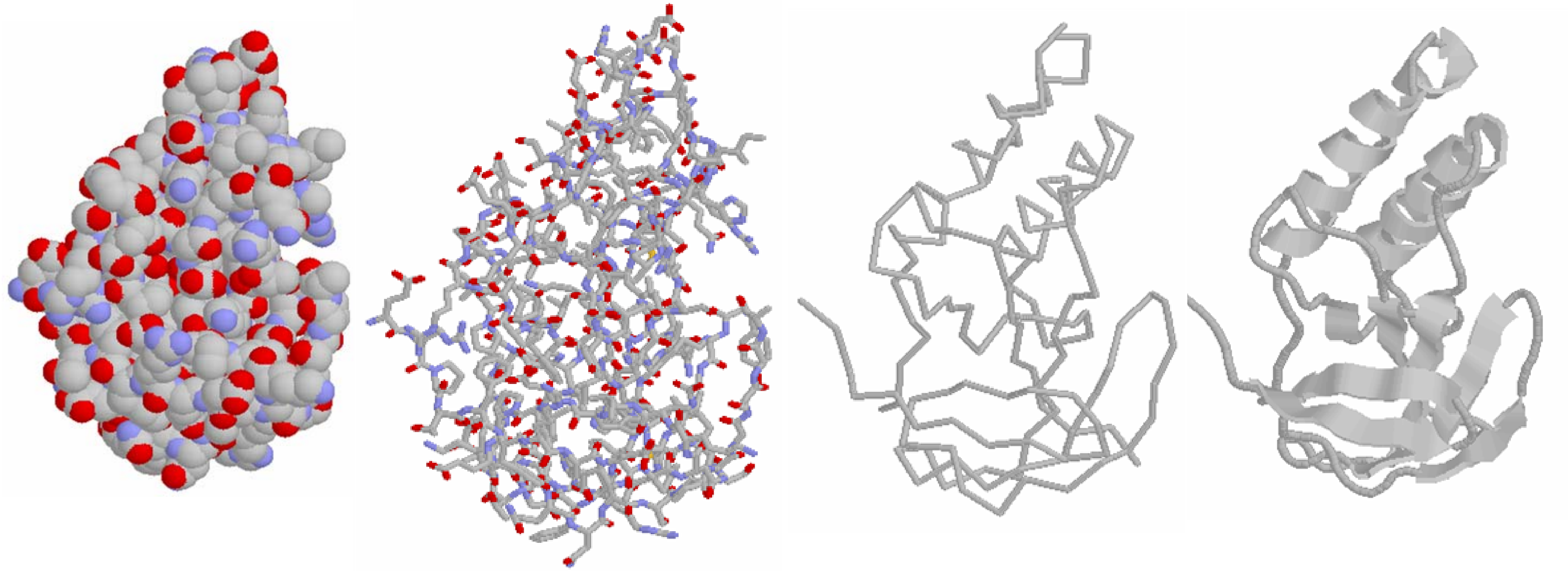

O
N
N


O
N
N


O
N
N
N
O

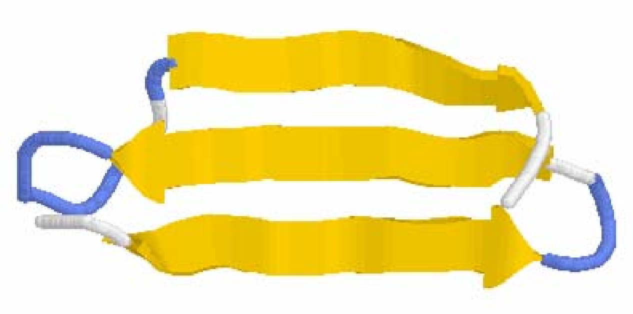

O
N
N
S

Representations of Protein Structure



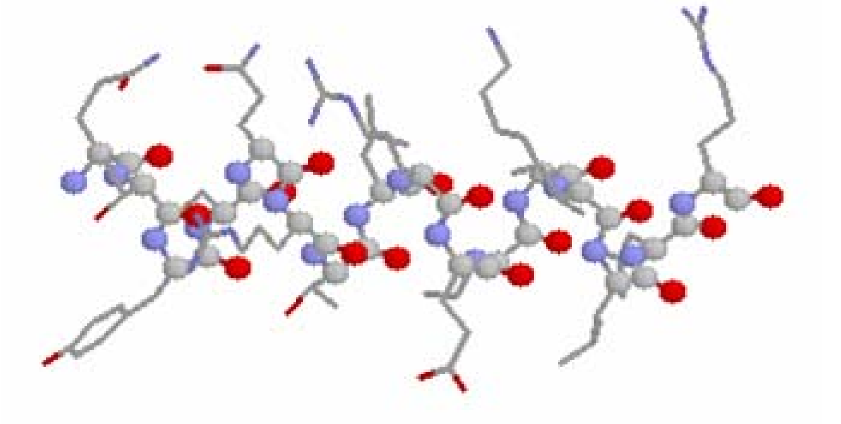
Review of protein structure hierarchy

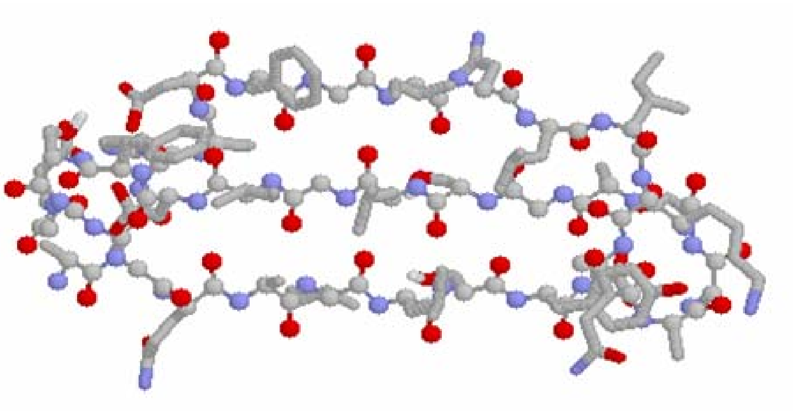
- *Secondary structure - why do you get regular secondary structure?*



α-helices

β-strands



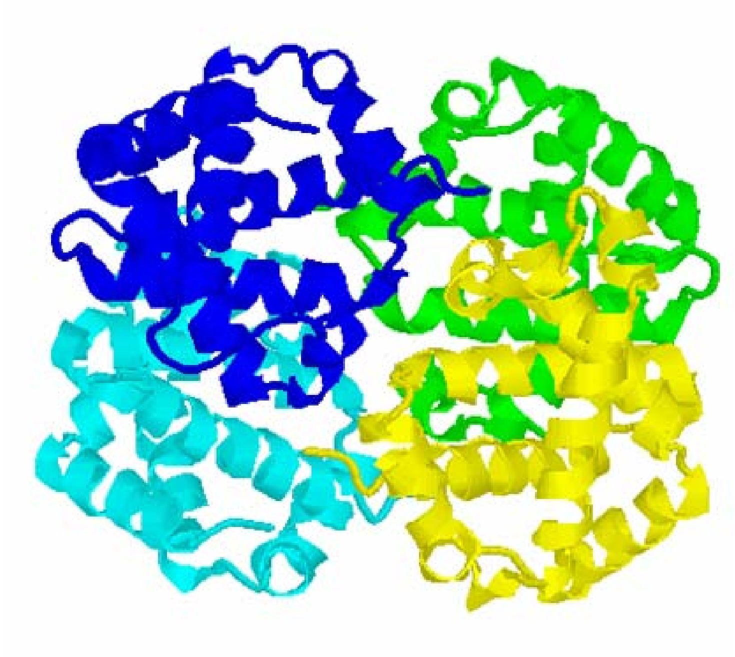


```
SGAYGSVCAA FDTKTGHRVA VKKLSRPFQS IIHAKRTYRE LRLLKHMKHE
    EEEEEE EE     EEE EEEE       HHHHHHHHHH HHHHHH
```

Review of protein structure hierarchy

- *Tertiary structure*
- *Quaternary structure*

N-terminal domain of kinase

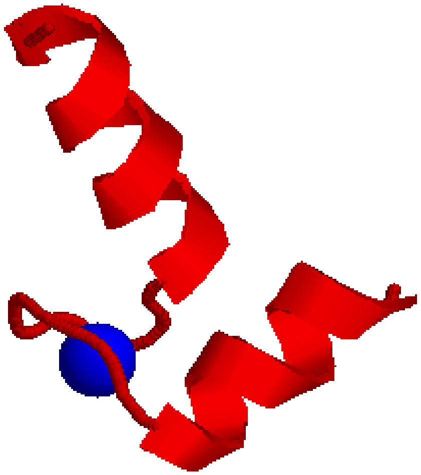


hemoglobin

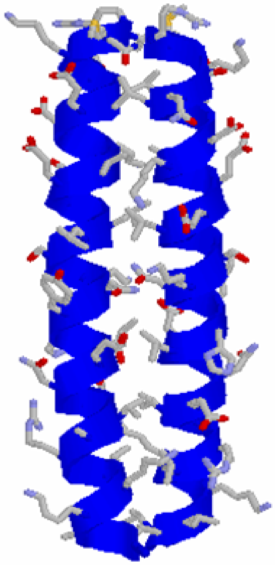
Why do you get compact/globular tertiary structures?

Other units of protein structure

Motifs

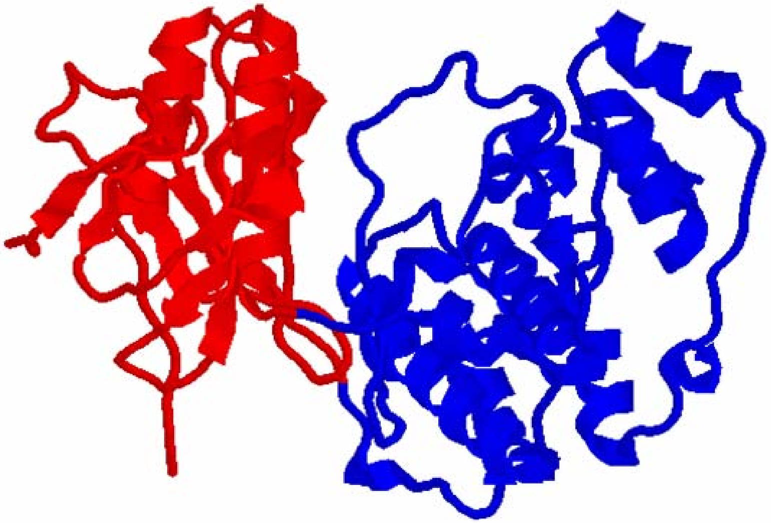


EF hand



coiled coil

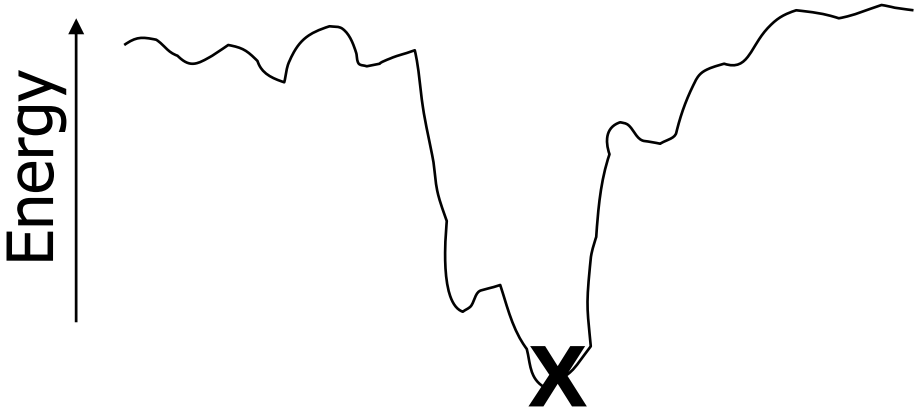
Domains



Sequence determines structure.
How?

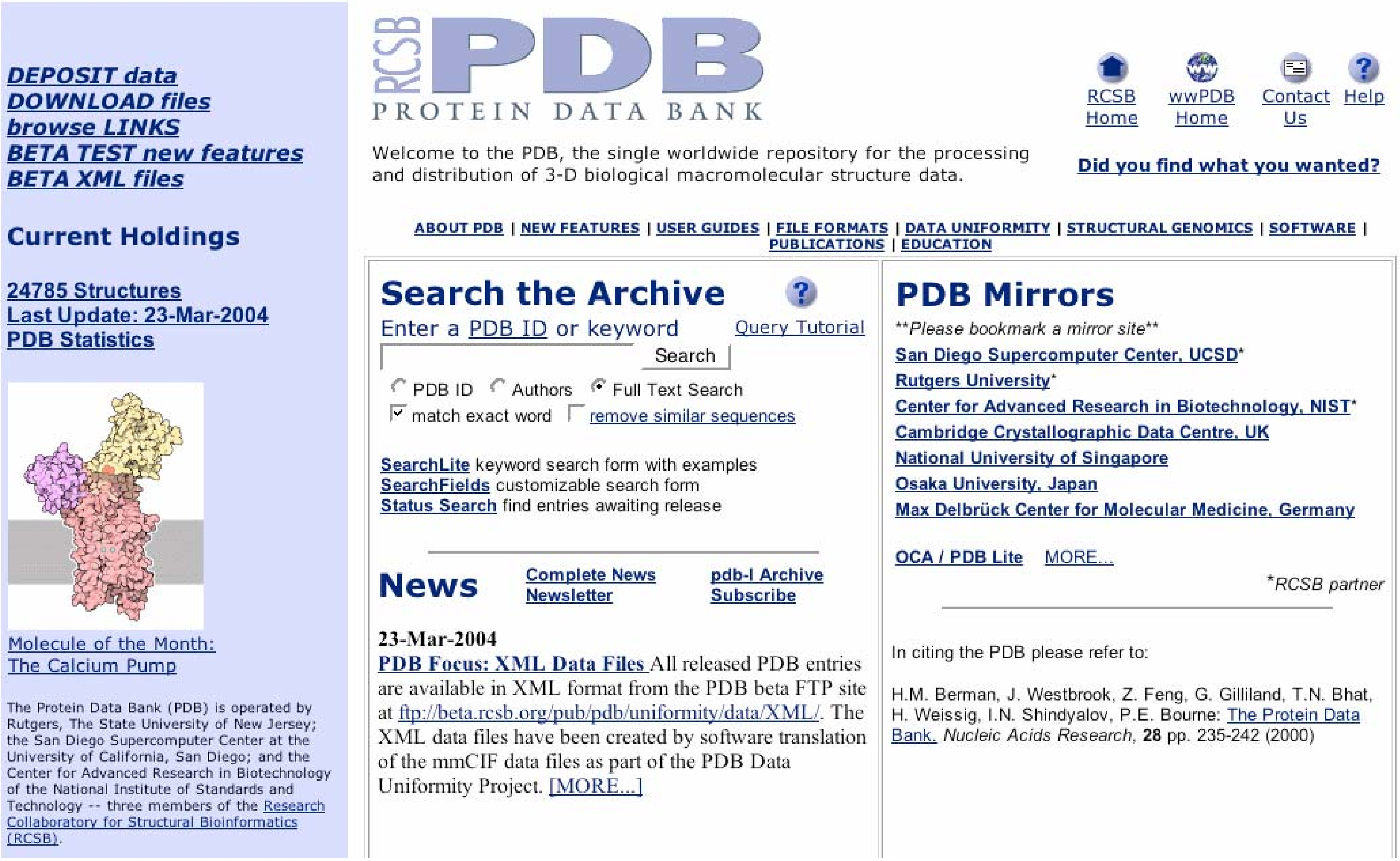
- Secondary structure preferences (satisfy H bonds)
- Hydrophobic/polar patterning
- Steric complementarity
- Electrostatics

Interactions are both LOCAL and NONLOCAL in sequence



Where do protein structures live?

www.rcsb.org/pdb



DEPOSIT data
DOWNLOAD files
browse LINKS
BETA TEST new features
BETA XML files

Current Holdings

24785 Structures
Last Update: 23-Mar-2004
PDB Statistics

Molecule of the Month:
The Calcium Pump

The Protein Data Bank (PDB) is operated by Rutgers, The State University of New Jersey; the San Diego Supercomputer Center at the University of California, San Diego; and the Center for Advanced Research in Biotechnology of the National Institute of Standards and Technology -- three members of the Research Collaboratory for Structural Bioinformatics (RCSB).

RCSB PDB
PROTEIN DATA BANK

RCSB Home | wwPDB Home | Contact Us | Help

Welcome to the PDB, the single worldwide repository for the processing and distribution of 3-D biological macromolecular structure data.

Did you find what you wanted?

ABOUT PDB | NEW FEATURES | USER GUIDES | FILE FORMATS | DATA UNIFORMITY | STRUCTURAL GENOMICS | SOFTWARE | PUBLICATIONS | EDUCATION

Search the Archive

Enter a PDB ID or keyword Query Tutorial

Search

○ PDB ID ○ Authors ● Full Text Search
☑ match exact word ☐ remove similar sequences

SearchLite keyword search form with examples
SearchFields customizable search form
Status Search find entries awaiting release

News

Complete News
Newsletter
pdb-l Archive
Subscribe

23-Mar-2004
PDB Focus: XML Data Files All released PDB entries are available in XML format from the PDB beta FTP site at ftp://beta.rcsb.org/pub/pdb/uniformity/data/XML/. The XML data files have been created by software translation of the mmCIF data files as part of the PDB Data Uniformity Project. [MORE...]

PDB Mirrors

Please bookmark a mirror site

San Diego Supercomputer Center, UCSD*
Rutgers University*
Center for Advanced Research in Biotechnology, NIST*
Cambridge Crystallographic Data Centre, UK
National University of Singapore
Osaka University, Japan
Max Delbrück Center for Molecular Medicine, Germany

OCA / PDB Lite MORE...

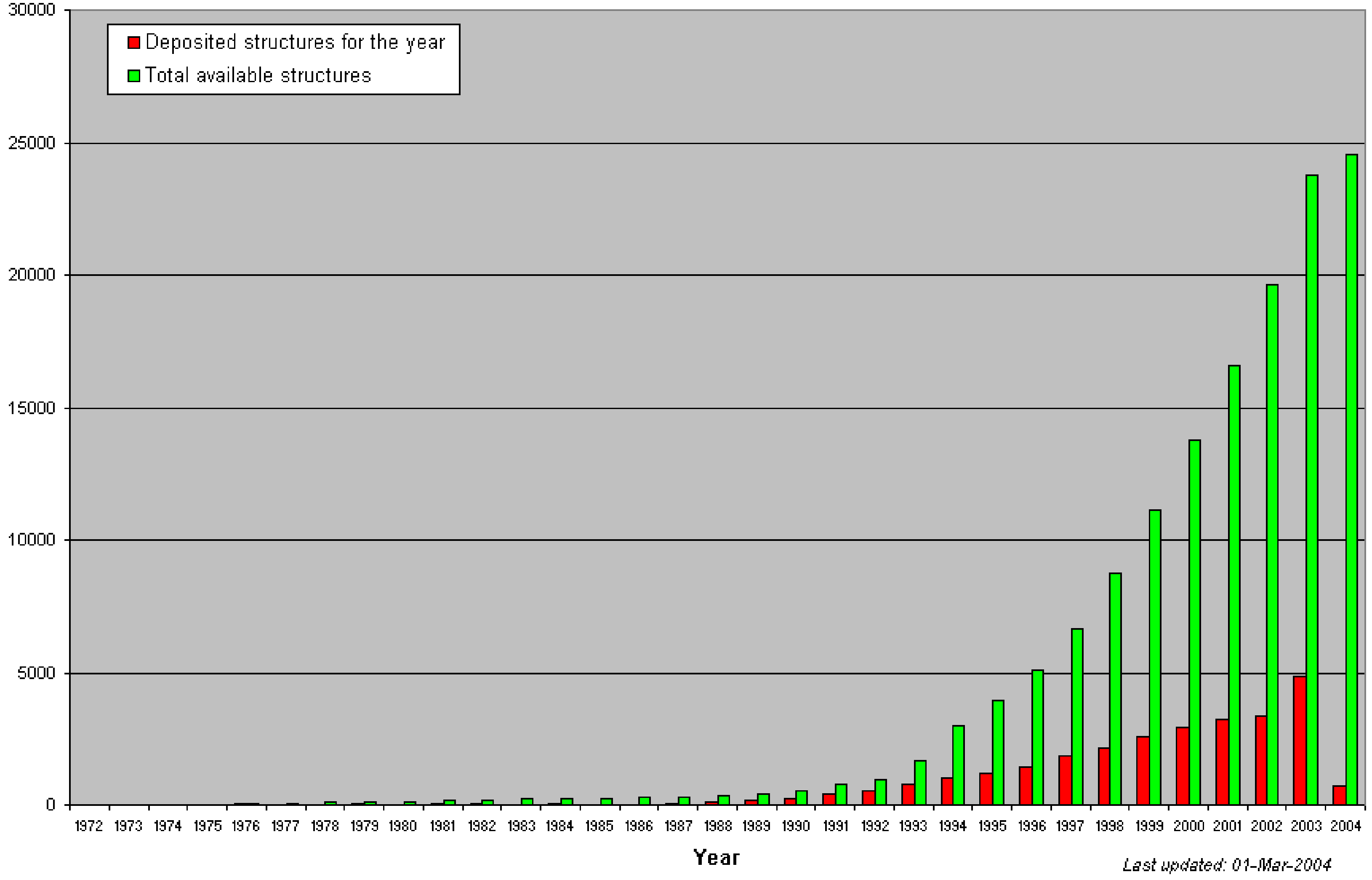
**RCSB partner*

In citing the PDB please refer to:

H.M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T.N. Bhat, H. Weissig, I.N. Shindyalov, P.E. Bourne: The Protein Data Bank. *Nucleic Acids Research*, **28** pp. 235-242 (2000)

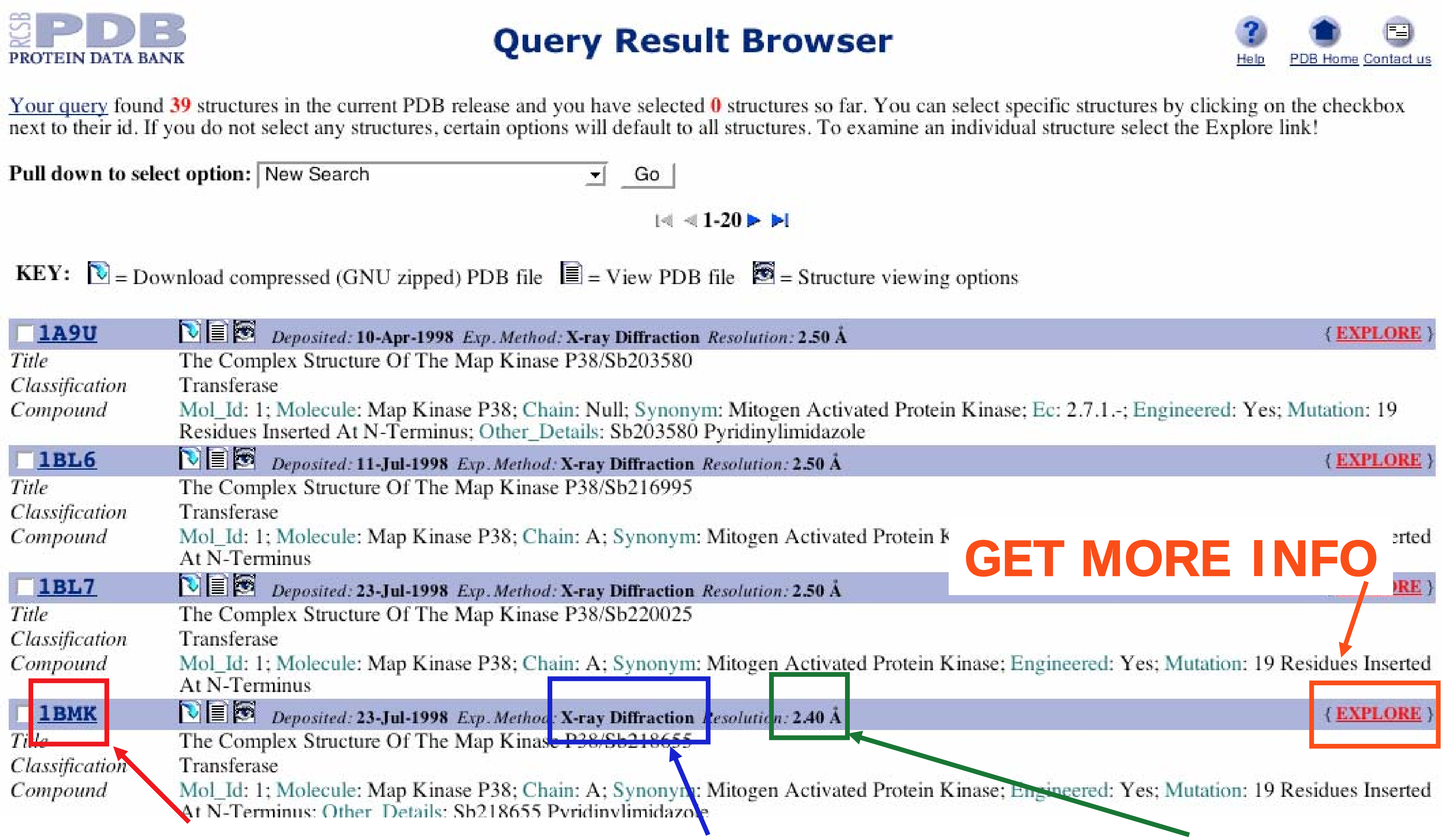
24,785 structures now in the PDB!

Compare: SwissProt 146,193, TrEMBL 1,070,786



Last updated: 01-Mar-2004

Finding structures in the PDB



RCSB PDB PROTEIN DATA BANK

Query Result Browser

Your query found 39 structures in the current PDB release and you have selected 0 structures so far. You can select specific structures by clicking on the checkbox next to their id. If you do not select any structures, certain options will default to all structures. To examine an individual structure select the Explore link!

Pull down to select option: New Search ▾ Go

KEY: = Download compressed (GNU zipped) PDB file = View PDB file = Structure viewing options

□ **1A9U** *Deposited:* **10-Apr-1998** *Exp. Method:* **X-ray Diffraction** *Resolution:* **2.50 Å** { EXPLORE }
Title The Complex Structure Of The Map Kinase P38/Sb203580
Classification Transferase
Compound Mol_Id: 1; Molecule: Map Kinase P38; Chain: Null; Synonym: Mitogen Activated Protein Kinase; Ec: 2.7.1.-; Engineered: Yes; Mutation: 19 Residues Inserted At N-Terminus; Other_Details: Sb203580 Pyridinylimidazole

□ **1BL6** *Deposited:* **11-Jul-1998** *Exp. Method:* **X-ray Diffraction** *Resolution:* **2.50 Å** { EXPLORE }
Title The Complex Structure Of The Map Kinase P38/Sb216995
Classification Transferase
Compound Mol_Id: 1; Molecule: Map Kinase P38; Chain: A; Synonym: Mitogen Activated Protein K erted At N-Terminus

□ **1BL7** *Deposited:* **23-Jul-1998** *Exp. Method:* **X-ray Diffraction** *Resolution:* **2.50 Å** RE }
Title The Complex Structure Of The Map Kinase P38/Sb220025
Classification Transferase
Compound Mol_Id: 1; Molecule: Map Kinase P38; Chain: A; Synonym: Mitogen Activated Protein Kinase; Engineered: Yes; Mutation: 19 Residues Inserted At N-Terminus

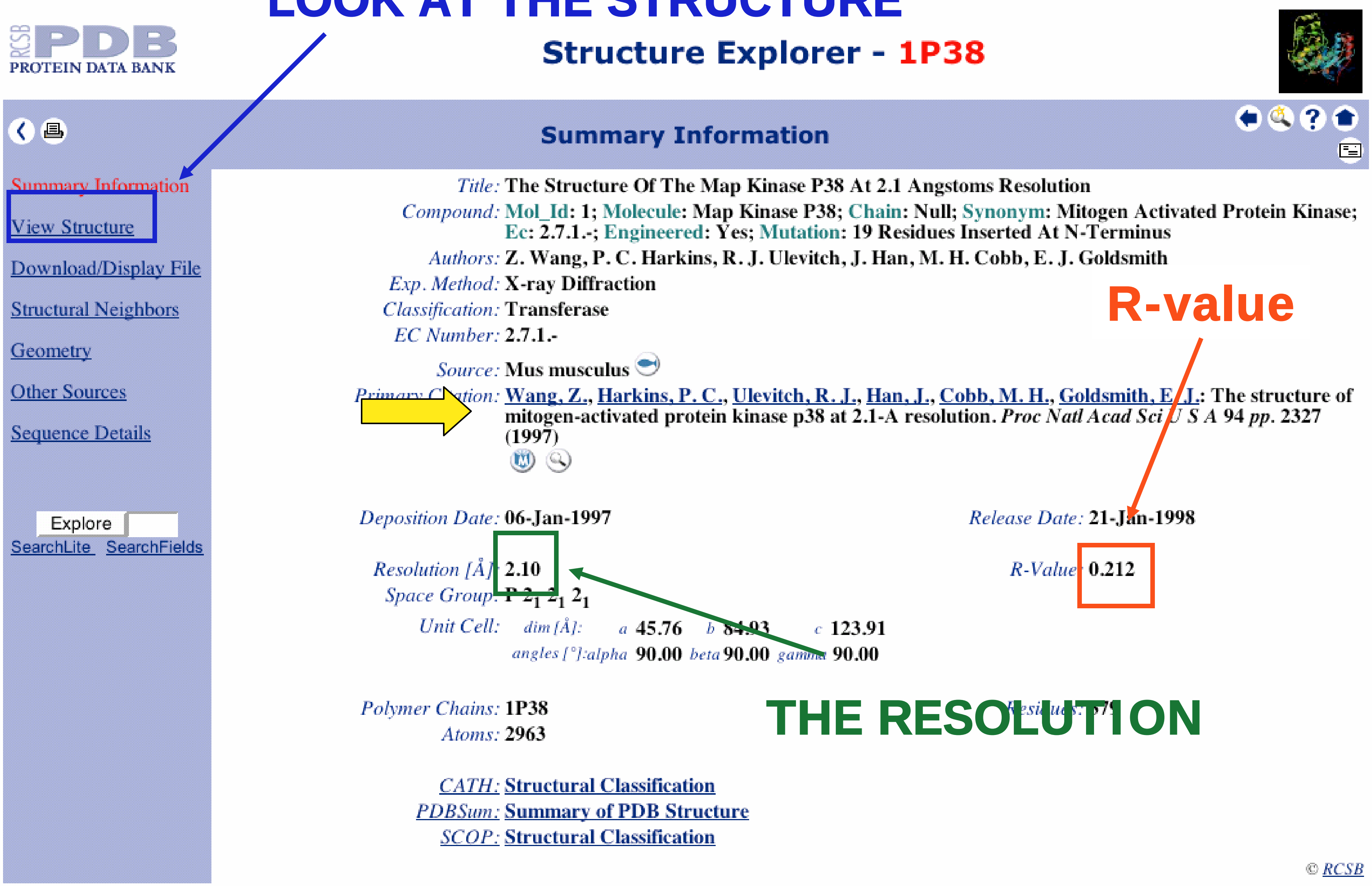
□ **1BMK** *Deposited:* **23-Jul-1998** *Exp. Method:* **X-ray Diffraction** *Resolution:* **2.40 Å** { EXPLORE }
Title The Complex Structure Of The Map Kinase P38/Sb218655
Classification Transferase
Compound Mol_Id: 1; Molecule: Map Kinase P38; Chain: A; Synonym: Mitogen Activated Protein Kinase; Engineered: Yes; Mutation: 19 Residues Inserted At N-Terminus; Other_Details: Sb218655 Pyridinylimidazole

GET MORE INFO

THE PDB CODE THE TECHNIQUE THE RESOLUTION

Exploring structures in the PDB

LOOK AT THE STRUCTURE



RCSB PDB PROTEIN DATA BANK

Structure Explorer - 1P38

Summary Information

Summary Information

View Structure

Download/Display File

Structural Neighbors

Geometry

Other Sources

Sequence Details

Explore

SearchLite SearchFields

Title: The Structure Of The Map Kinase P38 At 2.1 Angstoms Resolution

Compound: Mol_Id: 1; Molecule: Map Kinase P38; Chain: Null; Synonym: Mitogen Activated Protein Kinase; Ec: 2.7.1.-; Engineered: Yes; Mutation: 19 Residues Inserted At N-Terminus

Authors: Z. Wang, P. C. Harkins, R. J. Ulevitch, J. Han, M. H. Cobb, E. J. Goldsmith

Exp. Method: X-ray Diffraction

Classification: Transferase

EC Number: 2.7.1.-

Source: Mus musculus

Primary Citation: Wang, Z., Harkins, P. C., Ulevitch, R. J., Han, J., Cobb, M. H., Goldsmith, E. J.: The structure of mitogen-activated protein kinase p38 at 2.1-A resolution. *Proc Natl Acad Sci U S A* 94 *pp.* 2327 (1997)

Deposition Date: 06-Jan-1997

Release Date: 21-Jan-1998

Resolution [Å]: 2.10

R-Value: 0.212

Space Group: P 2_1 2_1 2_1

Unit Cell: *dim [Å]:* *a* 45.76 *b* 84.93 *c* 123.91

angles [°]:alpha 90.00 *beta* 90.00 *gamma* 90.00

Polymer Chains: 1P38

Residues: 379

Atoms: 2963

CATH: Structural Classification

PDBSum: Summary of PDB Structure

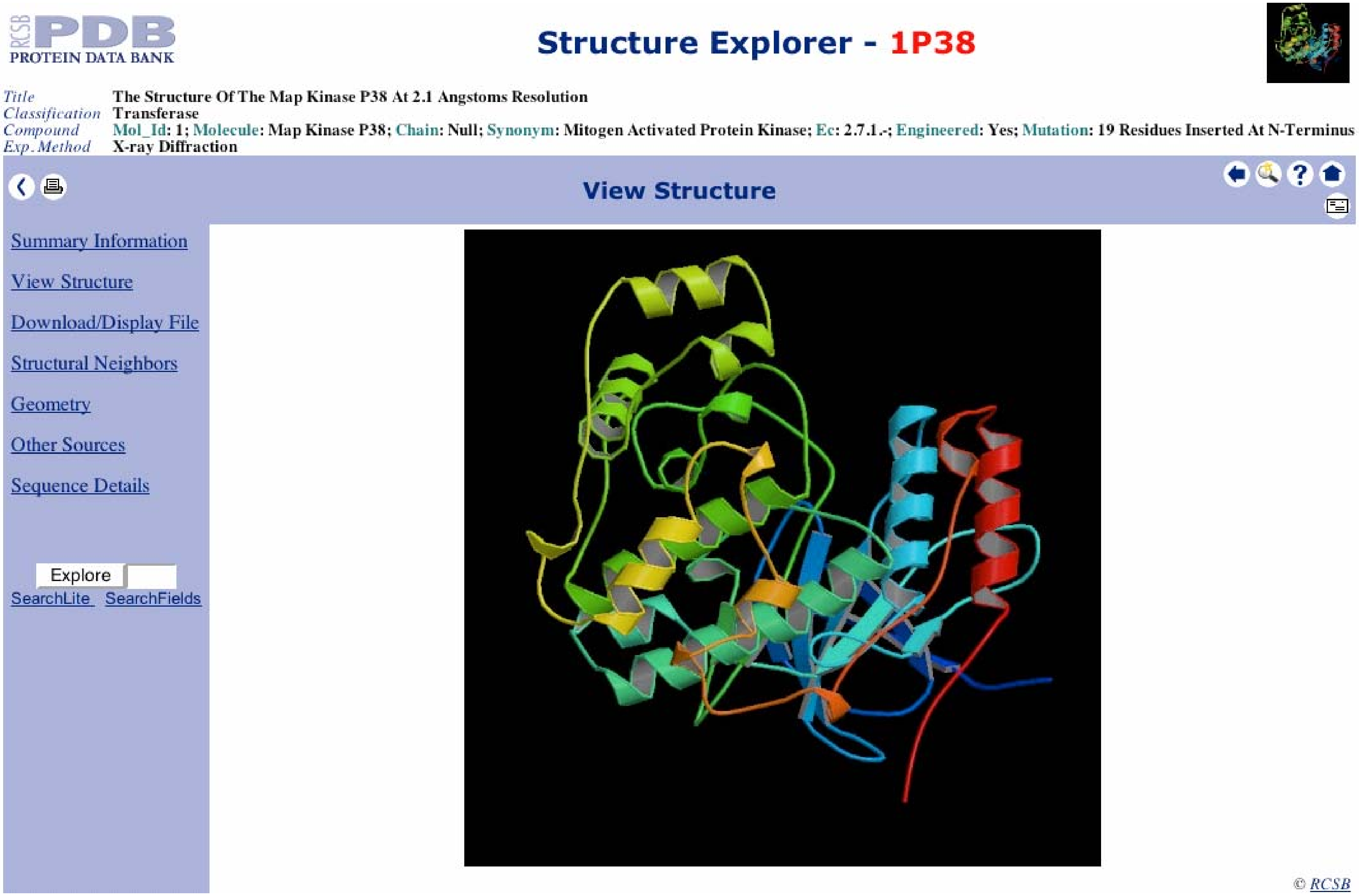
SCOP: Structural Classification

© RCSB

R-value

THE RESOLUTION

Exploring structures in the PDB



RCSB PDB PROTEIN DATA BANK

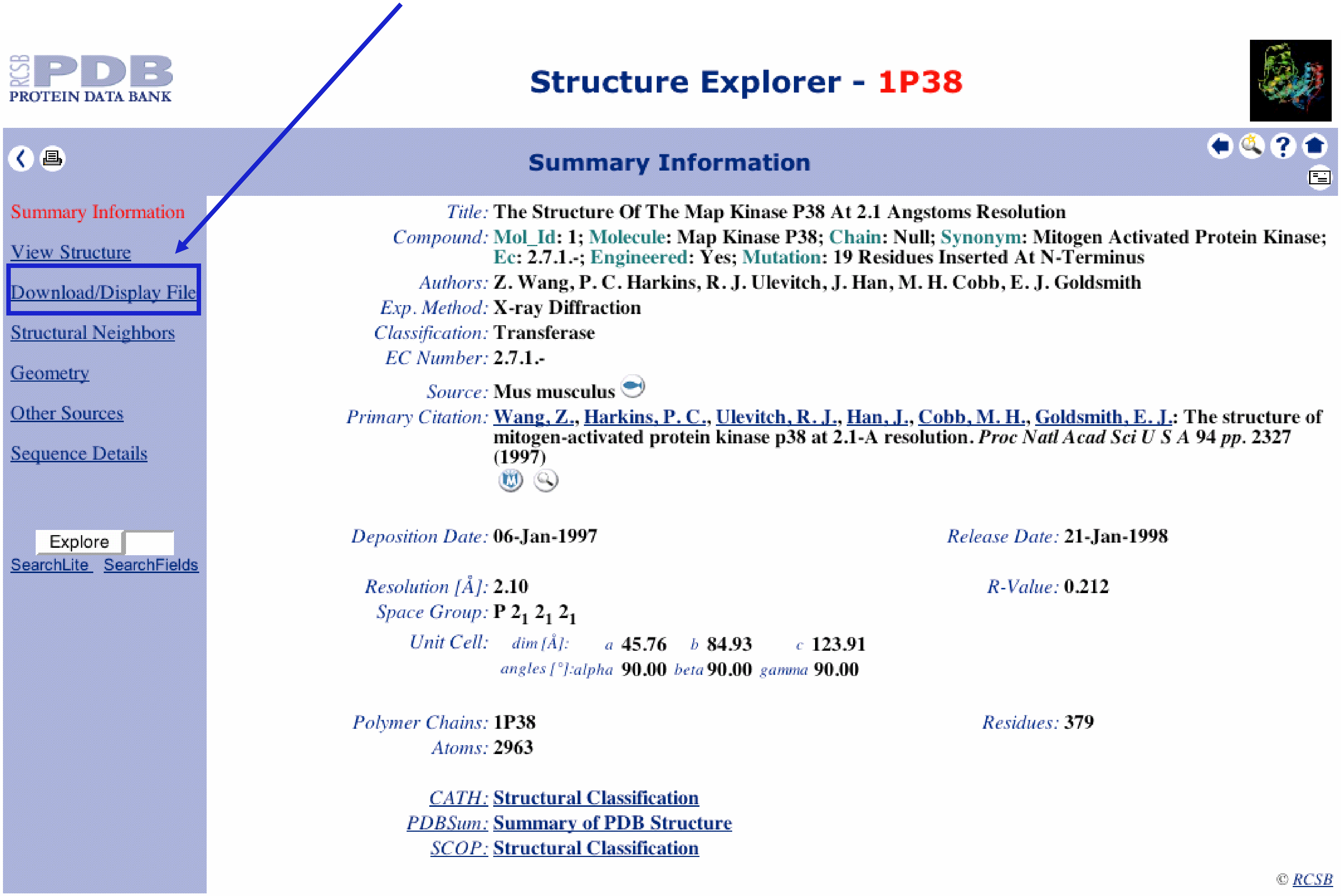
Structure Explorer - 1P38

Title	The Structure Of The Map Kinase P38 At 2.1 Angstoms Resolution
Classification	Transferase
Compound	Mol_Id: 1; Molecule: Map Kinase P38; Chain: Null; Synonym: Mitogen Activated Protein Kinase; Ec: 2.7.1.-; Engineered: Yes; Mutation: 19 Residues Inserted At N-Terminus
Exp. Method	X-ray Diffraction

View Structure

© RCSB

GET THE PDB FILE



Structure Explorer - 1P38

Summary Information

Title: The Structure Of The Map Kinase P38 At 2.1 Angstoms Resolution

Compound: Mol_Id: 1; Molecule: Map Kinase P38; Chain: Null; Synonym: Mitogen Activated Protein Kinase; Ec: 2.7.1.-; Engineered: Yes; Mutation: 19 Residues Inserted At N-Terminus

Authors: Z. Wang, P. C. Harkins, R. J. Ulevitch, J. Han, M. H. Cobb, E. J. Goldsmith

Exp. Method: X-ray Diffraction

Classification: Transferase

EC Number: 2.7.1.-

Source: Mus musculus

Primary Citation: Wang, Z., Harkins, P. C., Ulevitch, R. J., Han, J., Cobb, M. H., Goldsmith, E. J.: The structure of mitogen-activated protein kinase p38 at 2.1-A resolution. *Proc Natl Acad Sci U S A* 94 *pp.* 2327 (1997)

Deposition Date: 06-Jan-1997 *Release Date:* 21-Jan-1998

Resolution [Å]: 2.10 *R-Value:* 0.212

Space Group: P 2_1 2_1 2_1

Unit Cell:

dim [Å]:	*a* 45.76	*b* 84.93	*c* 123.91
angles [°]:	*alpha* 90.00	*beta* 90.00	*gamma* 90.00

Polymer Chains: 1P38 *Residues:* 379

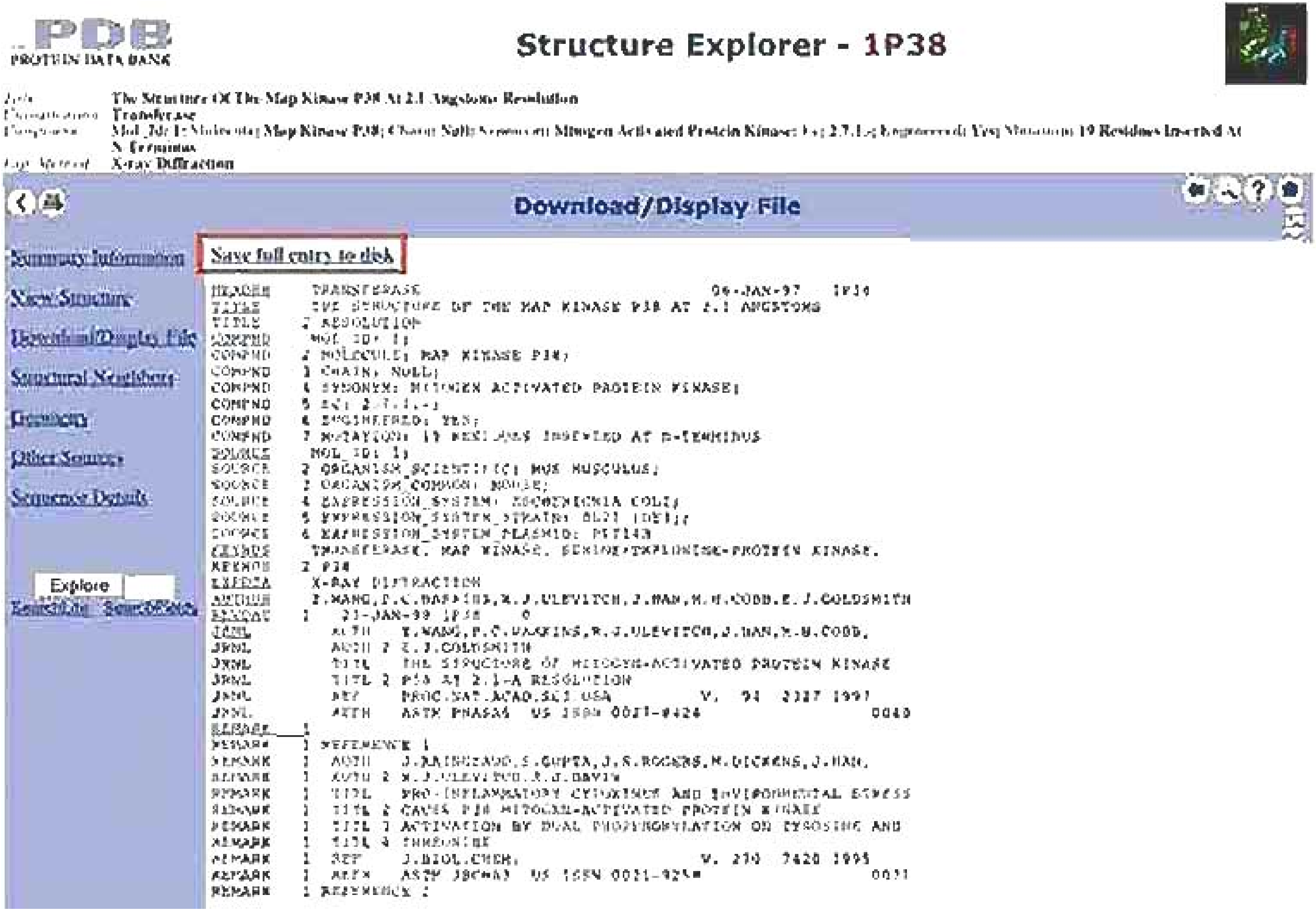
Atoms: 2963

CATH: Structural Classification

PDBSum: Summary of PDB Structure

SCOP: Structural Classification

© RCSB



PDB
PROTEIN DATA BANK

Structure Explorer - 1P38

Title The Structure Of The Map Kinase P38 At 2.1 Angstoms Resolution
Classification Transferase
Compound Mol_Id: 1; Molecule: Map Kinase P38; Chain: Null; Synonym: Mitogen Activated Protein Kinase; Ec: 2.7.1.-; Engineered: Yes; Mutation: 19 Residues Inserted At N-Terminus
Exp. Method X-ray Diffraction

Download/Display File

Save full entry to disk

```
HEADER    TRANSFERASE                             06-JAN-97   1P38
TITLE     THE STRUCTURE OF THE MAP KINASE P38 AT 2.1 ANGSTOMS
TITLE    2 RESOLUTION
COMPND    MOL_ID: 1;
COMPND   2 MOLECULE: MAP KINASE P38;
COMPND   3 CHAIN: NULL;
COMPND   4 SYNONYM: MITOGEN ACTIVATED PROTEIN KINASE;
COMPND   5 EC: 2.7.1.-;
COMPND   6 ENGINEERED: YES;
COMPND   7 MUTATION: 19 RESIDUES INSERTED AT N-TERMINUS
SOURCE    MOL_ID: 1;
SOURCE   2 ORGANISM_SCIENTIFIC: MUS MUSCULUS;
SOURCE   3 ORGANISM_COMMON: MOUSE;
SOURCE   4 EXPRESSION_SYSTEM: ESCHERICHIA COLI;
SOURCE   5 EXPRESSION_SYSTEM_STRAIN: BL21 (DE3);
SOURCE   6 EXPRESSION_SYSTEM_PLASMID: PET14B
KEYWDS    TRANSFERASE, MAP KINASE, SERINE/THREONINE-PROTEIN KINASE,
KEYWDS   2 P38
EXPDTA    X-RAY DIFFRACTION
AUTHOR    Z.WANG,P.C.HARKINS,R.J.ULEVITCH,J.HAN,M.H.COBB,E.J.GOLDSMITH
REVDAT   1   21-JAN-98 1P38    0
JRNL        AUTH   Z.WANG,P.C.HARKINS,R.J.ULEVITCH,J.HAN,M.H.COBB,
JRNL        AUTH 2 E.J.GOLDSMITH
JRNL        TITL   THE STRUCTURE OF MITOGEN-ACTIVATED PROTEIN KINASE
JRNL        TITL 2 P38 AT 2.1-A RESOLUTION
JRNL        REF    PROC.NAT.ACAD.SCI.USA         V.  94  2327 1997
JRNL        REFN   ASTM PNASA6  US ISSN 0027-8424                 0040
REMARK   1
REMARK   1 REFERENCE 1
REMARK   1  AUTH   J.RAINGEAUD,S.GUPTA,J.S.ROGERS,M.DICKENS,J.HAN,
REMARK   1  AUTH 2 R.J.ULEVITCH,R.J.DAVIS
REMARK   1  TITL   PRO-INFLAMMATORY CYTOKINES AND ENVIRONMENTAL STRESS
REMARK   1  TITL 2 CAUSE P38 MITOGEN-ACTIVATED PROTEIN KINASE
REMARK   1  TITL 3 ACTIVATION BY DUAL PHOSPHORYLATION ON TYROSINE AND
REMARK   1  TITL 4 THREONINE
REMARK   1  REF    J.BIOL.CHEM.                  V. 270  7420 1995
REMARK   1  REFN   ASTM JBCHA3  US ISSN 0021-9258                 0071
REMARK   1 REFERENCE 2
```

Useful information in the PDB header

```
REMARK 280 CRYSTAL
REMARK 280 SOLVENT CONTENT, VS   (%): 58.0
REMARK 280 MATTHEWS COEFFICIENT, VM (ANGSTROMS**3/DA): 2.92
REMARK 280
REMARK 280 CRYSTALLIZATION CONDITIONS: THE PROTEIN CRYSTALLIZED IN 18%
REMARK 280 PEG 8000, 0.2M MG(OAC)2, 0.1M HEPES, PH7.0.  THE PROTEIN
REMARK 280 CONCENTRATION WAS ~ 10MG/ML IN A BUFFER OF 50MM NACL,
REMARK 280 1MM EDTA, 10MM DTT, 1MM BENZAMIDINE, 1UM PEPSTATIN, 10UG/ML
REMARK 280 LEUPEPTIN, 25MM HEPES,PH7.4.
```

```
REMARK 999 SEQUENCE
REMARK 999 1P38       SWS     P47811       1 -     3 NOT IN ATOMS LIST
REMARK 999 1P38       SWS     P47811     355 -   360 NOT IN ATOMS LIST
DBREF  1P38      4   354  SWS    P47811   MP38_MOUSE       4    354
SEQRES   1    379  GLY SER SER HIS HIS HIS HIS HIS HIS SER SER GLY LEU
SEQRES   2    379  VAL PRO ARG GLY SER HIS MET SER GLN GLU ARG PRO THR
SEQRES   3    379  PHE TYR ARG GLN GLU LEU ASN LYS THR ILE TRP GLU VAL
SEQRES   4    379  PRO GLU ARG TYR GLN ASN LEU SER PRO VAL GLY SER GLY
```

Useful information in the PDB header

```
REMARK   3  FIT TO DATA USED IN REFINEMENT.
REMARK   3   CROSS-VALIDATION METHOD          : NULL
REMARK   3   FREE R VALUE TEST SET SELECTION  : RANDOM
REMARK   3   R VALUE            (WORKING SET) : 0.212
REMARK   3   FREE R VALUE                     : 0.244
REMARK   3   FREE R VALUE TEST SET SIZE   (%) : 10.
REMARK   3   FREE R VALUE TEST SET COUNT      : NULL
REMARK   3   ESTIMATED ERROR OF FREE R VALUE  : NULL

REMARK   3  RMS DEVIATIONS FROM IDEAL VALUES.
REMARK   3   BOND LENGTHS                 (A) : 0.010
REMARK   3   BOND ANGLES            (DEGREES) : 1.58
REMARK   3   DIHEDRAL ANGLES        (DEGREES) : NULL
REMARK   3   IMPROPER ANGLES        (DEGREES) : NULL


REMARK   3  B VALUES.
REMARK   3   FROM WILSON PLOT           (A**2) : NULL
REMARK   3   MEAN B VALUE      (OVERALL, A**2) : 29.7
```

Atomic coordinates in the PDB file

					X	**Y**	**Z**	**occ**	**B**
ATOM	1	N	GLU	4	28.492	3.212	23.465	1.00	70.88
ATOM	2	CA	GLU	4	27.552	4.354	23.629	1.00	69.99
ATOM	3	C	GLU	4	26.545	4.432	22.489	0.00	67.56
ATOM	4	O	GLU	4	26.915	4.250	21.328	0.00	68.09
ATOM	5	CB	GLU	4	28.326	5.683	23.680	0.00	72.34
ATOM	6	CG	GLU	4	27.447	6.910	23.973	0.00	75.98
ATOM	7	CD	GLU	4	28.123	8.247	23.659	0.00	78.43
ATOM	8	OE1	GLU	4	29.375	8.299	23.604	0.00	79.32
ATOM	9	OE2	GLU	4	27.393	9.251	23.468	0.00	79.58
ATOM	10	N	ARG	5	25.274	4.610	22.852	1.00	63.77
ATOM	11	CA	ARG	5	24.179	4.807	21.907	1.00	59.83
ATOM	12	C	ARG	5	23.411	3.698	21.219	1.00	56.20
ATOM	13	O	ARG	5	23.987	2.808	20.596	1.00	57.33
ATOM	14	CB	ARG	5	24.604	5.784	20.812	1.00	60.86
ATOM	15	CG	ARG	5	23.926	7.127	20.866	1.00	61.89
ATOM	16	CD	ARG	5	24.295	7.944	19.647	1.00	62.21

Looking at Protein Structures

Quick and dirty

Rasmol

Chime

Cn3D (NCBI)

More powerful

Swiss PDB Viewer, PyMol (free! Many platforms)

Insight, Quanta ($$$, nice interface, powerful)

Publication quality graphics, but not easy to manipulate

Molscript/Raster3D

Comparing Protein Structures

Why?

Reading: Mount, Chapter 9

Comparing Protein Structures

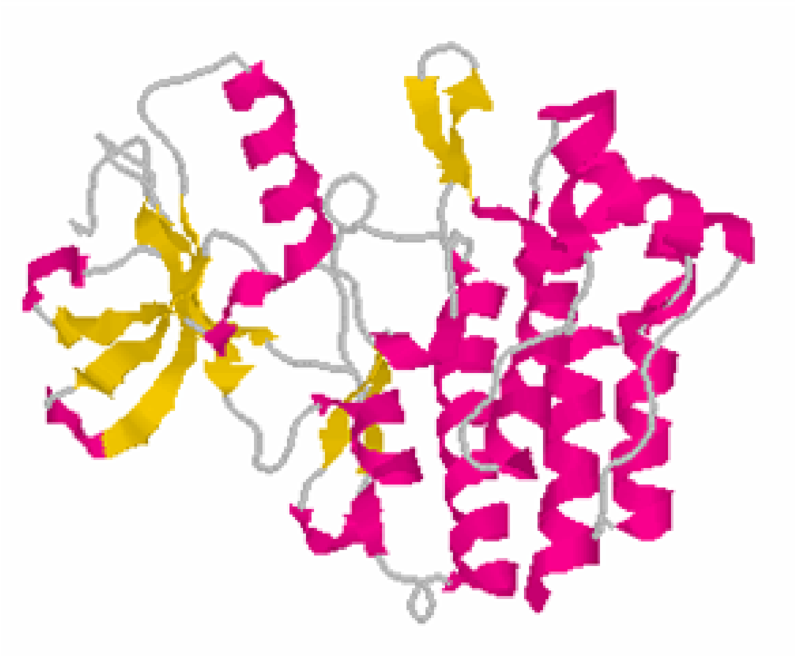
Why?

detect evolutionary relationships
identify recurring motifs
detect structure/function relationships
predict function
assess predicted structures
classify structures - used for many purposes

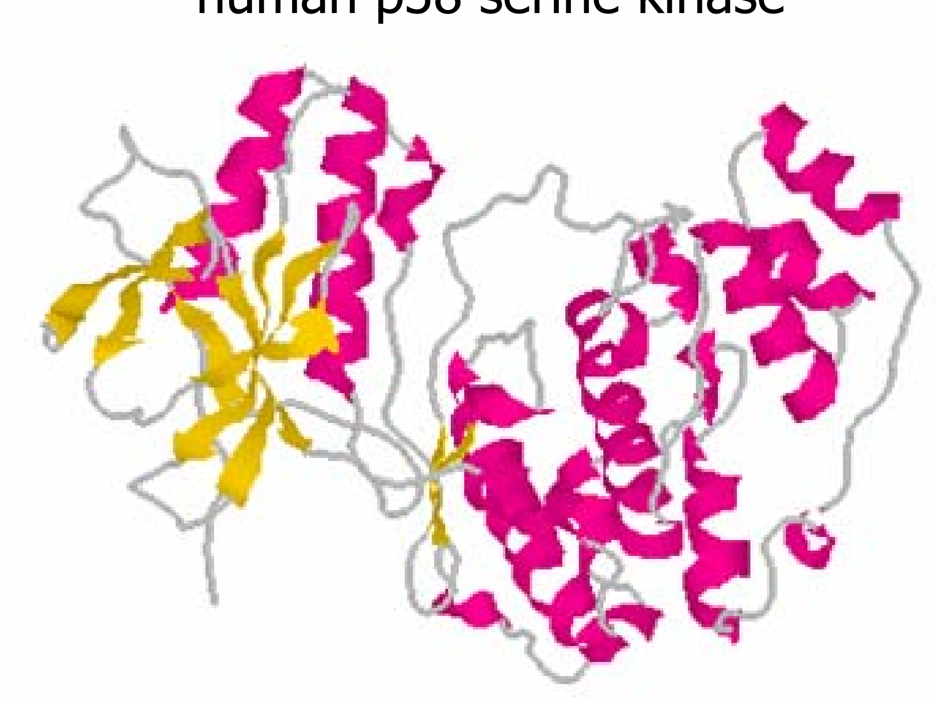
Structure is more conserved than sequence

28% sequence identity

mouse Abl tyrosine kinase



human p38 serine kinase



Detecting substructures is challenging

Please see figure 1 of

Ortiz, Angel R., Charlie E. M. Strauss, and Osvaldo Olmea. "MAMMOTH (Matching Molecular Models Obtained from Theory): An Automated Method for Model Comparison." *Protein Sci* 11 (2002): 2606-2621.

Recognizing Structural Similarity

GOAL: Of all solved structures, find the structure or substructure most similar to a protein of interest

By eye - tried and true! requires an expert viewer with a GREAT memory!

Automated detection - good for database searching

How would you do this?

Features of automated structure comparison

1. What representation will you use for the protein?
2. How will you assess structural similarity?
3. How will you search the possible comparisons?
4. How significant is a “hit”?

Example: Superposition to minimize RMSD

1. Define measure of similarity

   RMSD = $\{\Sigma|x_i\text{-}x_j|^2)/N\}^{1/2}$
2. Determine correspondence between residues of each protein (e.g. by sequence alignment, or a guess)
3. Align centers of mass
4. Use matrix methods to solve for the rotation that gives minimal RMSD (variety of methods available)
5. Evaluate the resulting number
6. Refine the alignment
7. iterate

Very useful. Commonly used for comparing similar structures.

But…

Example: Superposition to minimize RMSD

1. Define measure of similarity
   RMSD = $\{\Sigma|x_i - x_j|^2)/N\}^{1/2}$
2. Determine correspondence between residues of each protein (e.g. by sequence alignment, or a guess)
3. Align centers of mass
4. Use matrix methods to solve for the rotation that gives minimal RMSD (variety of methods available)
5. Evaluate the resulting number
6. Refine the alignment
7. iterate

Very useful. Commonly used for comparing similar structures.

But…

Not a good choice when proteins are only partially similar. Why?

Also, points far from center of mass are weighted more heavily.

Algorithms for detecting structure similarity

Dynamic Programming

- works on 1D strings - reduce problem to this
- can’t accommodate topological changes
- example: Secondary Structure Alignment Program (SSAP)

3D Comparison/Clustering

- identify secondary structure elements or fragments
- look for a similar arrangement of these between different structures
- allows for different topology, large insertions
- example: Vector Alignment Search Tool (VAST)

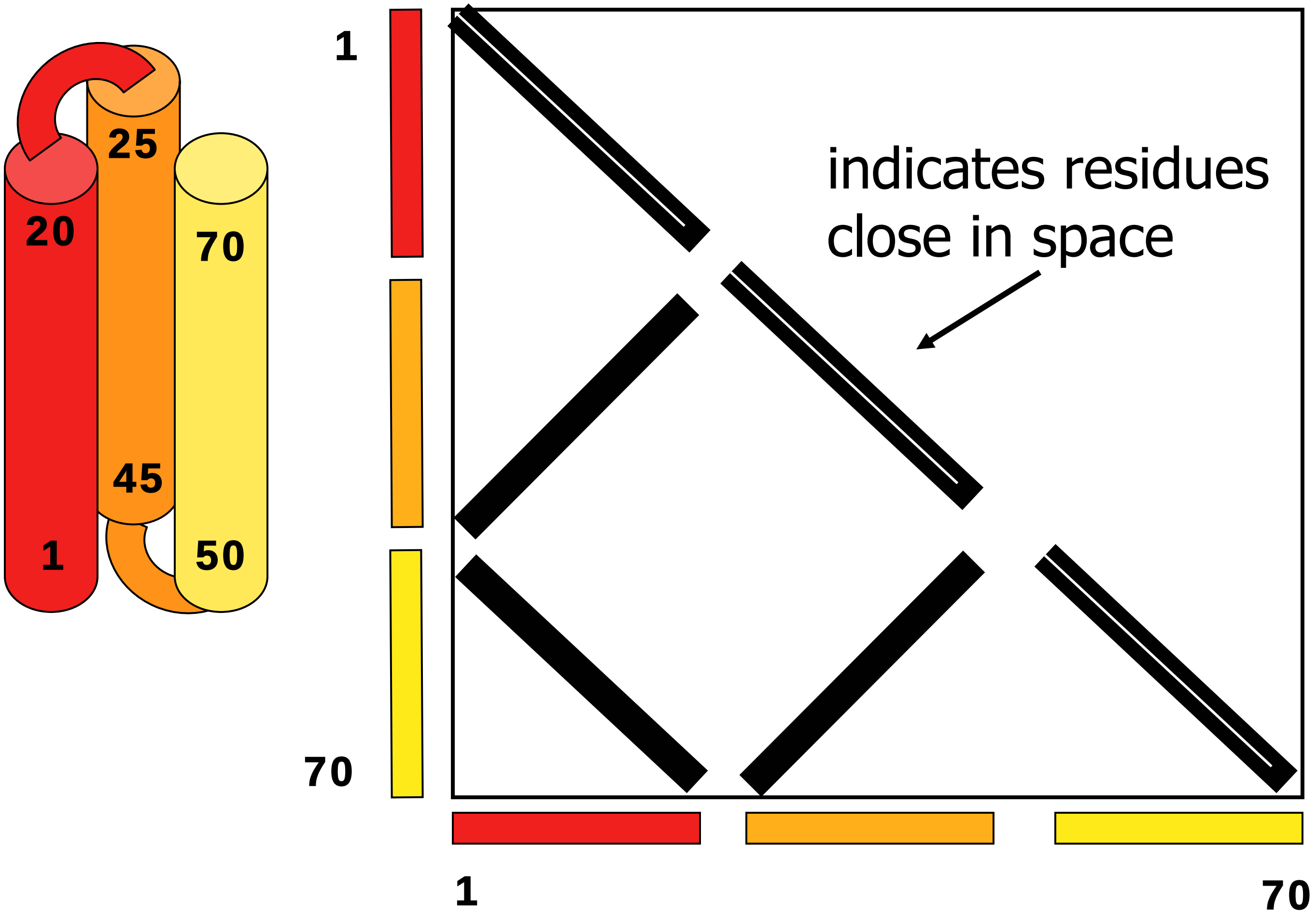
Distance Matrix

- identify contact patterns of groups that are close together
- compare these for different structures
- fast, insensitive to insertions
- example: Distance ALIgnment Tool (DALI)

Unit vector RMS

- map structure to sphere of vectors
- minimize the difference between spheres
- fast, insensitive to outliers
- example: Matching Molecular Models Obtained from Theory (MAMMOTH)

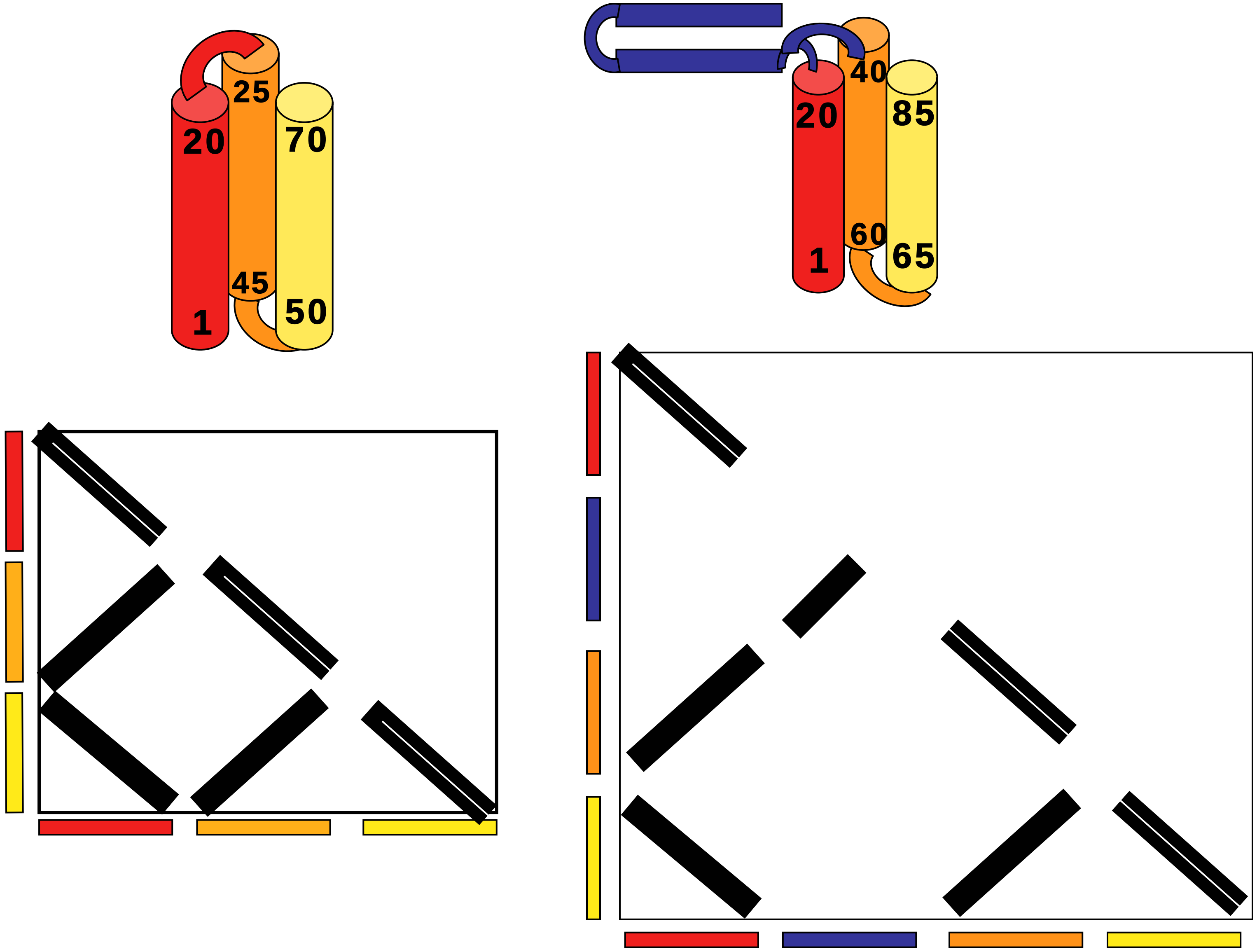
DALI represents proteins at the residue level; look for similarities using a distance matrix



indicates residues close in space

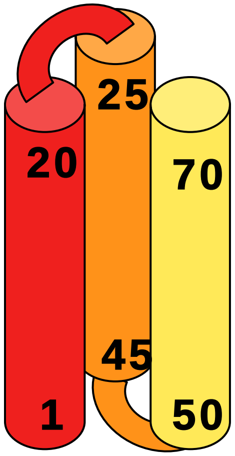
Images based on Holm, L, and C Sander. "Protein Structure Comparison by Alignment of Distance Matrices." *J Mol Biol.* 233, no. 1 (5 September 1993): 123-38.

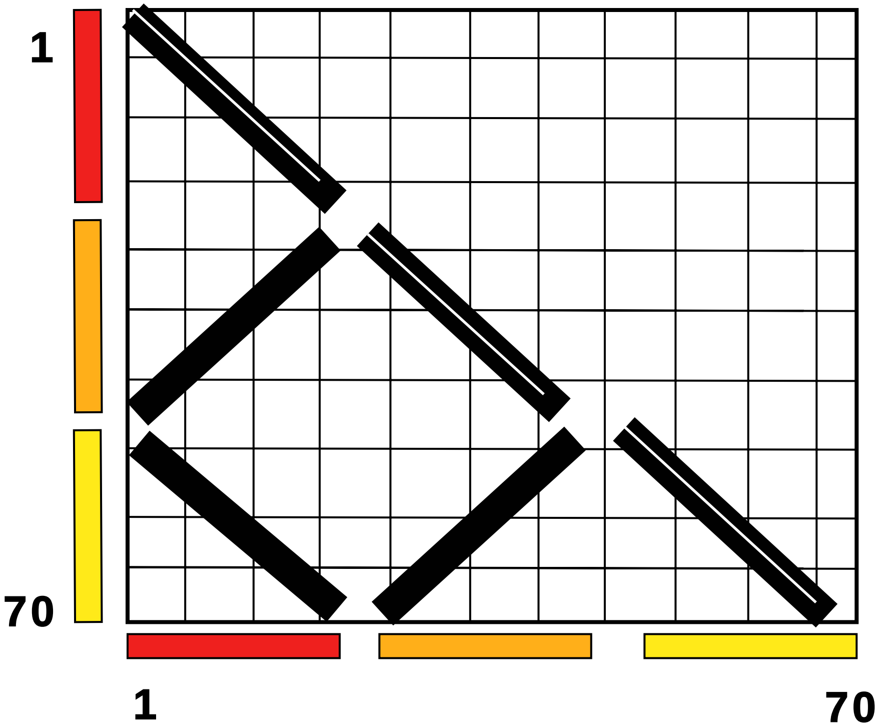
Compare contact patterns of different proteins



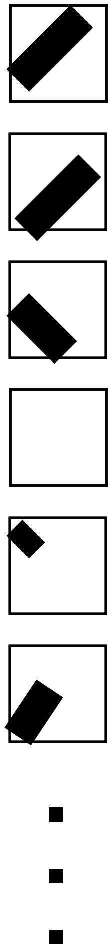
Images based on Holm, L, and C Sander. "Protein Structure Comparison by Alignment of Distance Matrices." *J Mol Biol.* 233, no. 1 (5 September 1993): 123-38.

Break distance matrix into hexapeptide regions



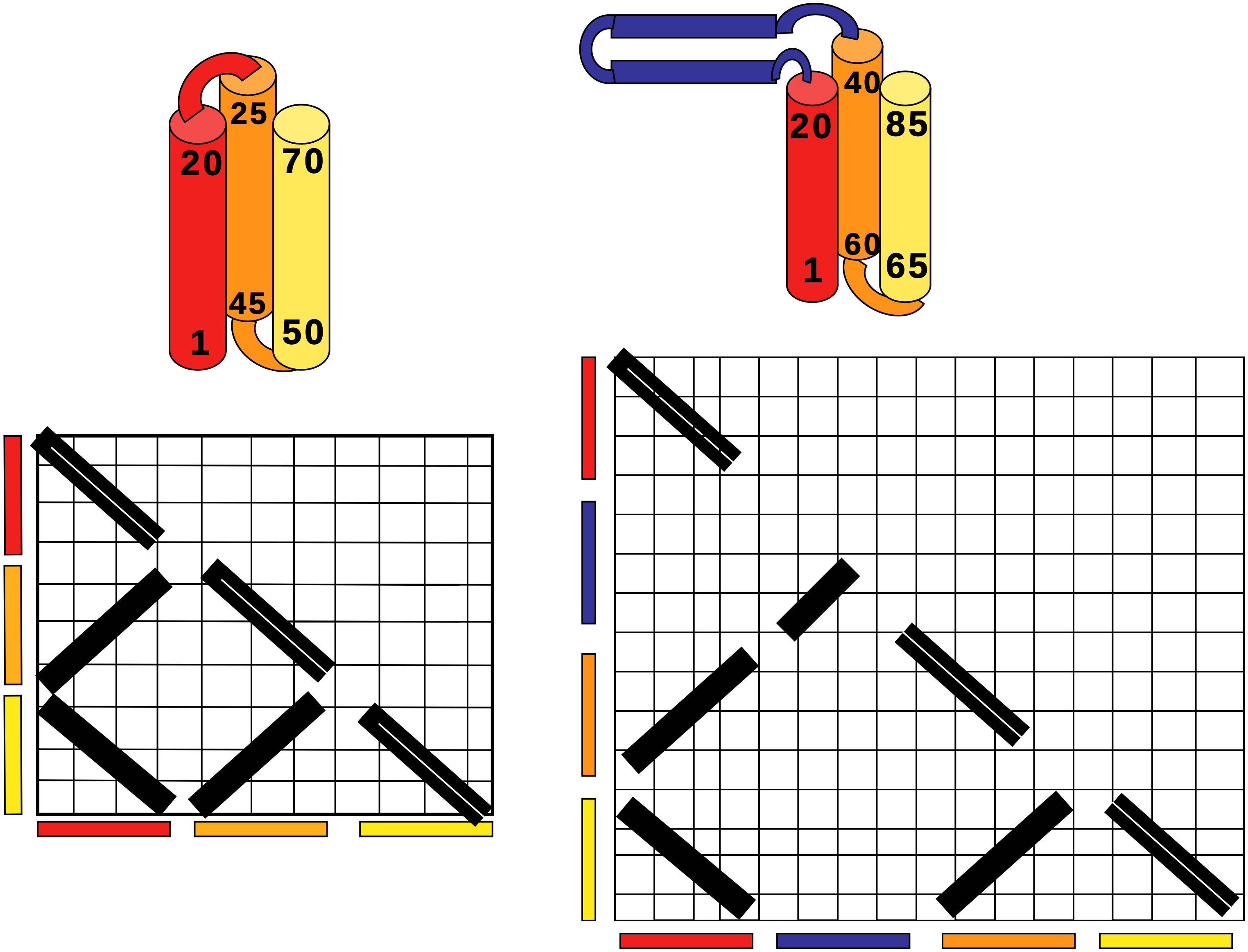


list of contact patterns



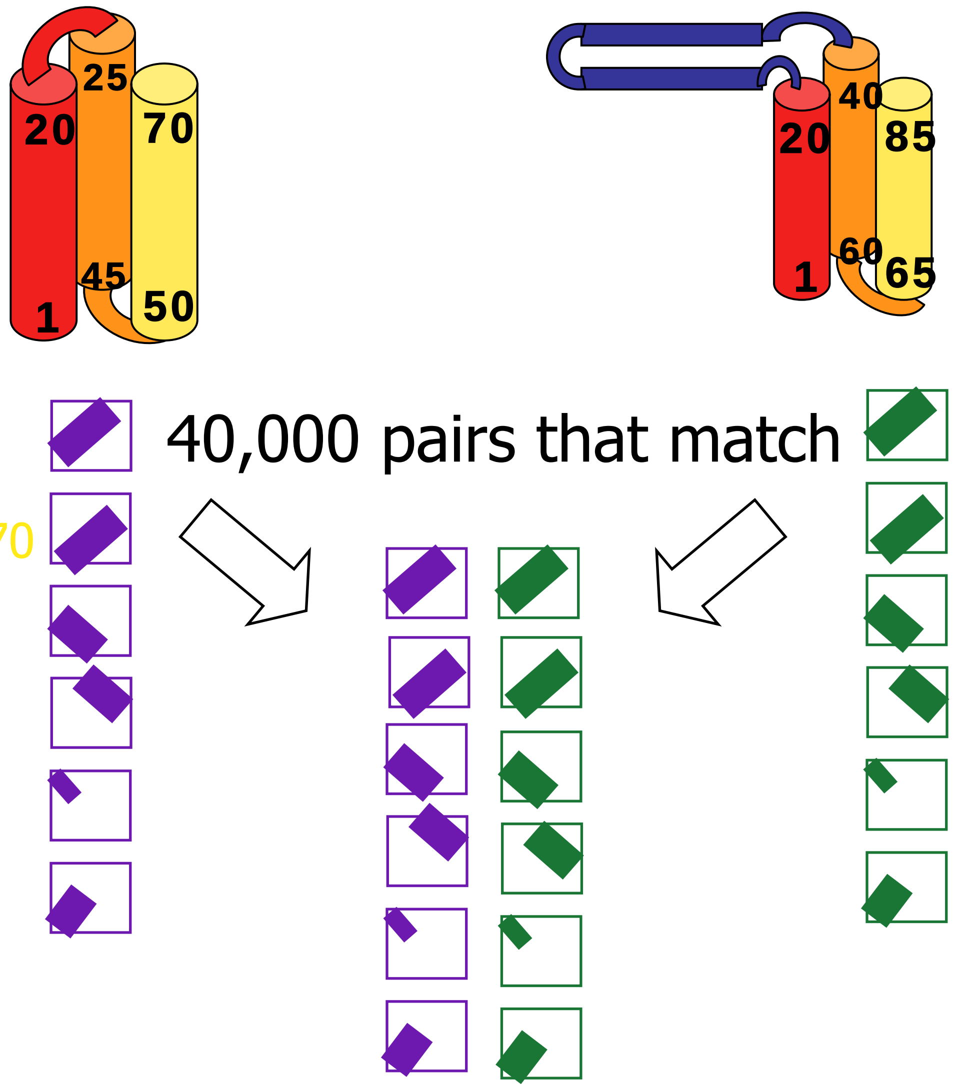
Images based on Holm, L, and C Sander. "Protein Structure Comparison by Alignment of Distance Matrices." *J Mol Biol.* 233, no. 1 (5 September 1993): 123-38.

Compare contact patterns of different proteins



Images based on Holm, L, and C Sander. "Protein Structure Comparison by Alignment of Distance Matrices." *J Mol Biol.* 233, no. 1 (5 September 1993): 123-38.

Compare contact patterns of different proteins



1-6 with 50-55

15-20 with 65-70

1-6 with 40-45

40,000 pairs that match

1-6 with 65-70

15-20 with 80-85

1-6 with 55-60

Images based on Holm, L, and C Sander. "Protein Structure Comparison by Alignment of Distance Matrices." *J Mol Biol.* 233, no. 1 (5 September 1993): 123-38.

How do you compare assemblies?

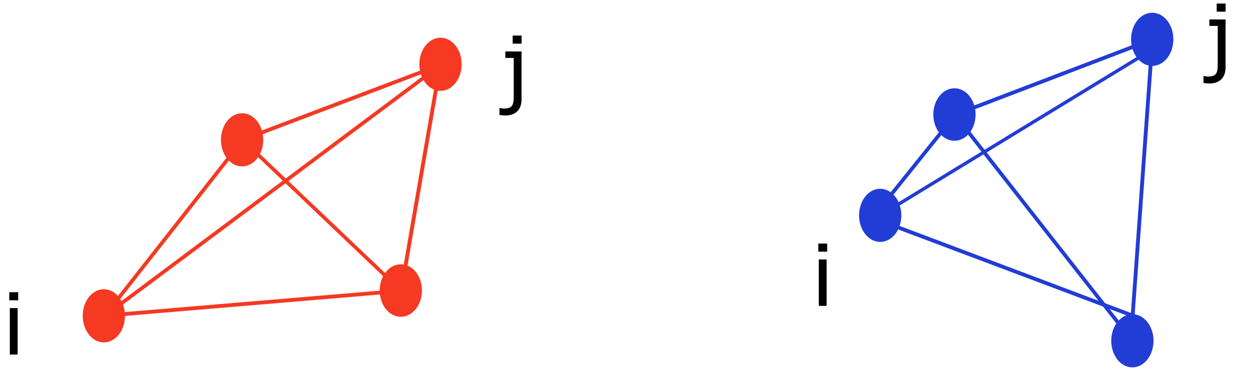
$S = \Sigma_i \Sigma_j \phi(i,j)$, where (i, j) is a pair of matches residues

distance between i and j in A (get from matrix)

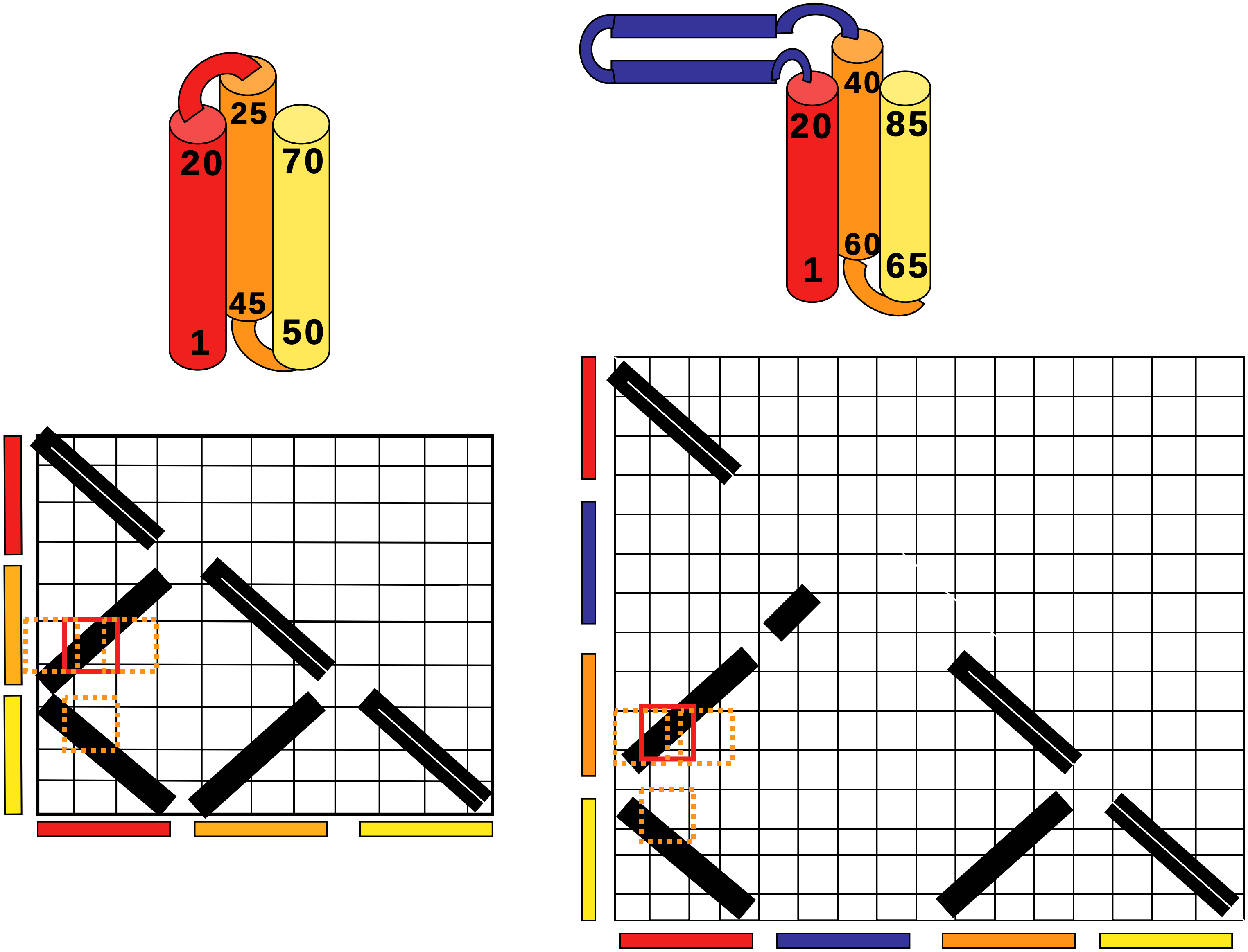
distance between i and j in B

$$\Phi(i,j) = \left(0.2 - \frac{|d^A_{ij} - d^B_{ij}|}{\mathrm{avg}(d^A_{ij}, d^B_{ij})}\right) e^{-r^2/a^2}$$

down-weight pairs that are far

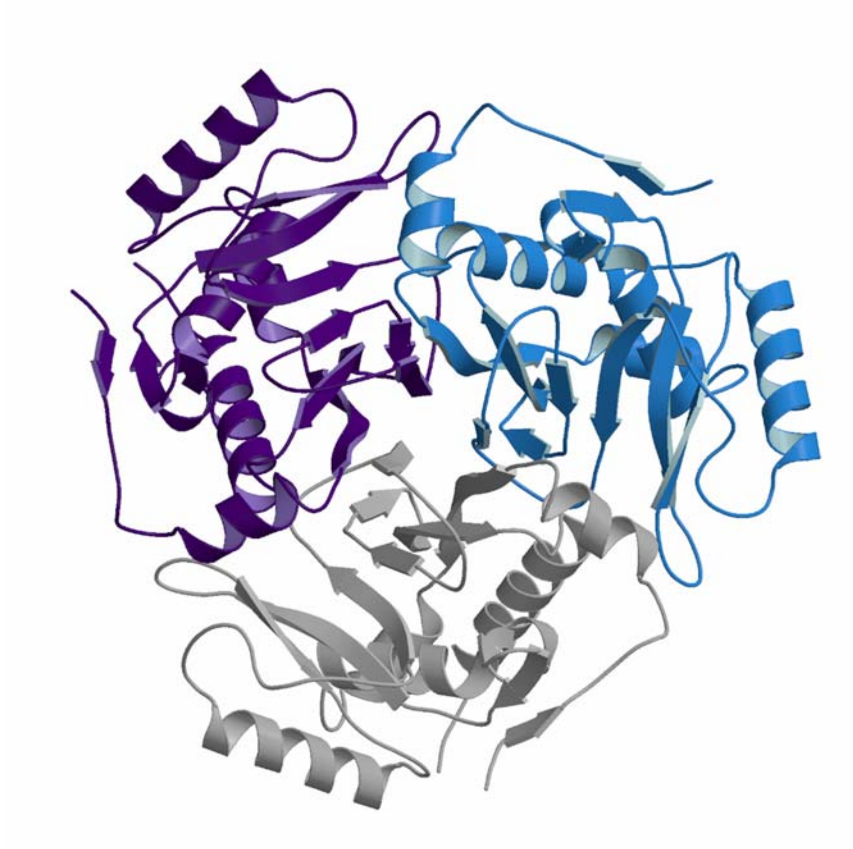


Monte Carlo assembly of fragments



Example of structural similarity detected by DALI

10-18% sequence identity



chloramphenicol acetyl transferase

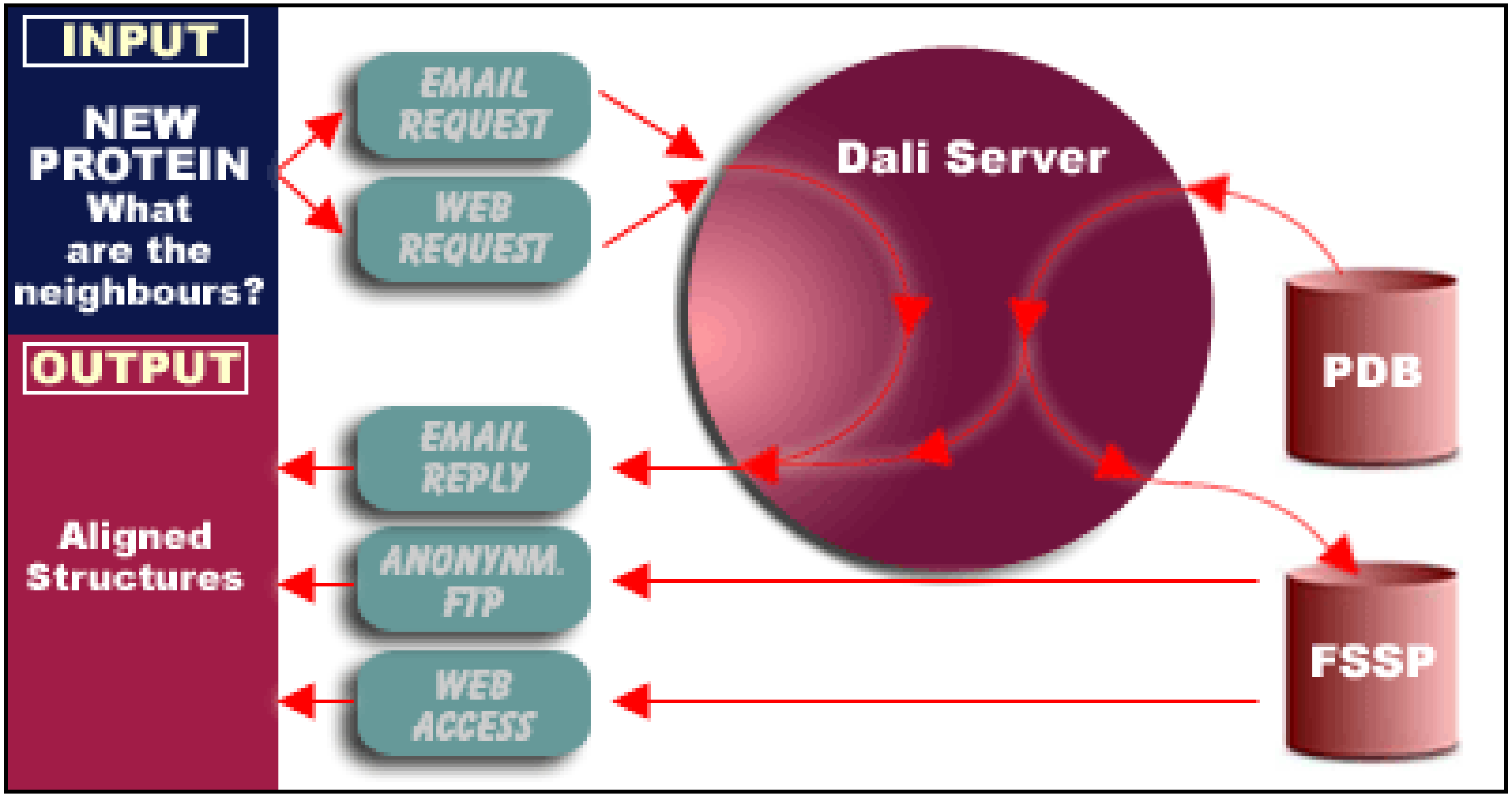
Keating et al. Nat. Struct. Biol. (2002) 9, 522-526

Advantages of DALI 3D matrix similarity search

- Can accommodate:
  - gaps/insertions
  - altered connectivity
  - chain reversal
- Fast enough for database comparisons
- Coordinate-frame invariant
- Pre-processing of distance matrices gives fast alignment performance
- Sensitive and accurate, even in presence of distortions

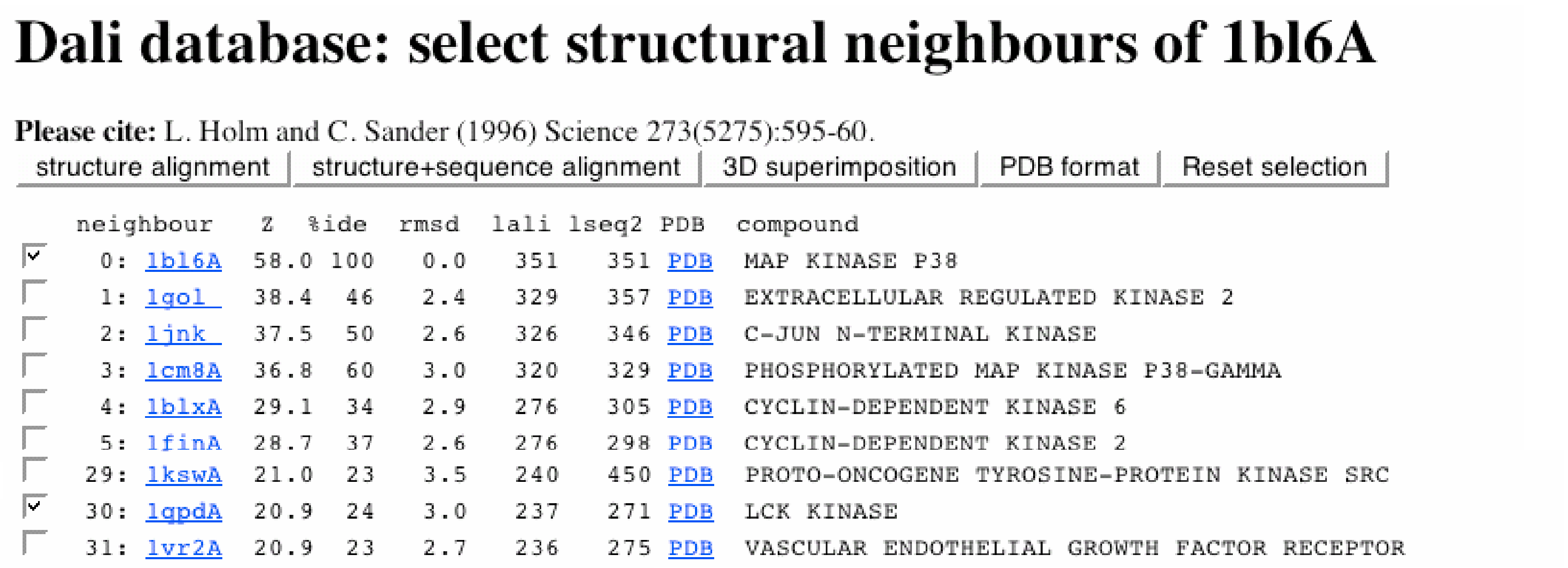
- CONVENIENT WEB INTERFACE!!

www.ebi.ac.uk/dali/



Fold classificatiion based on **S**tructure-**S**tructure Alignment of **P**roteins

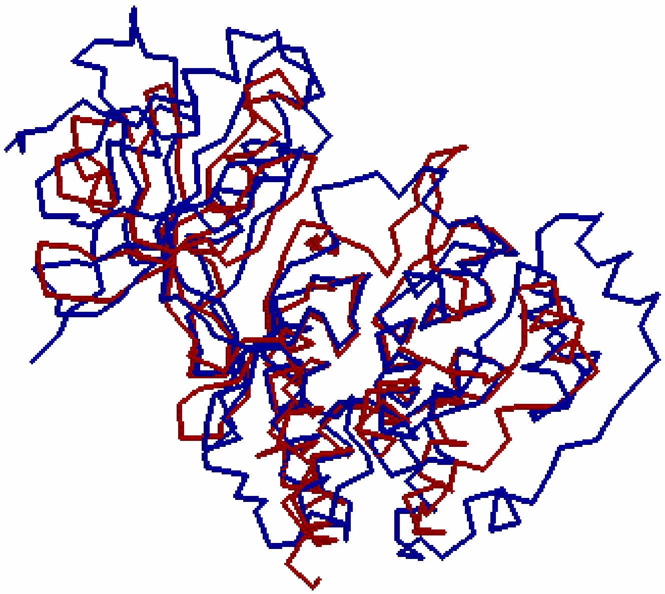
Pre-computed similarities of proteins in the pdb



Dali database: select structural neighbours of 1bl6A

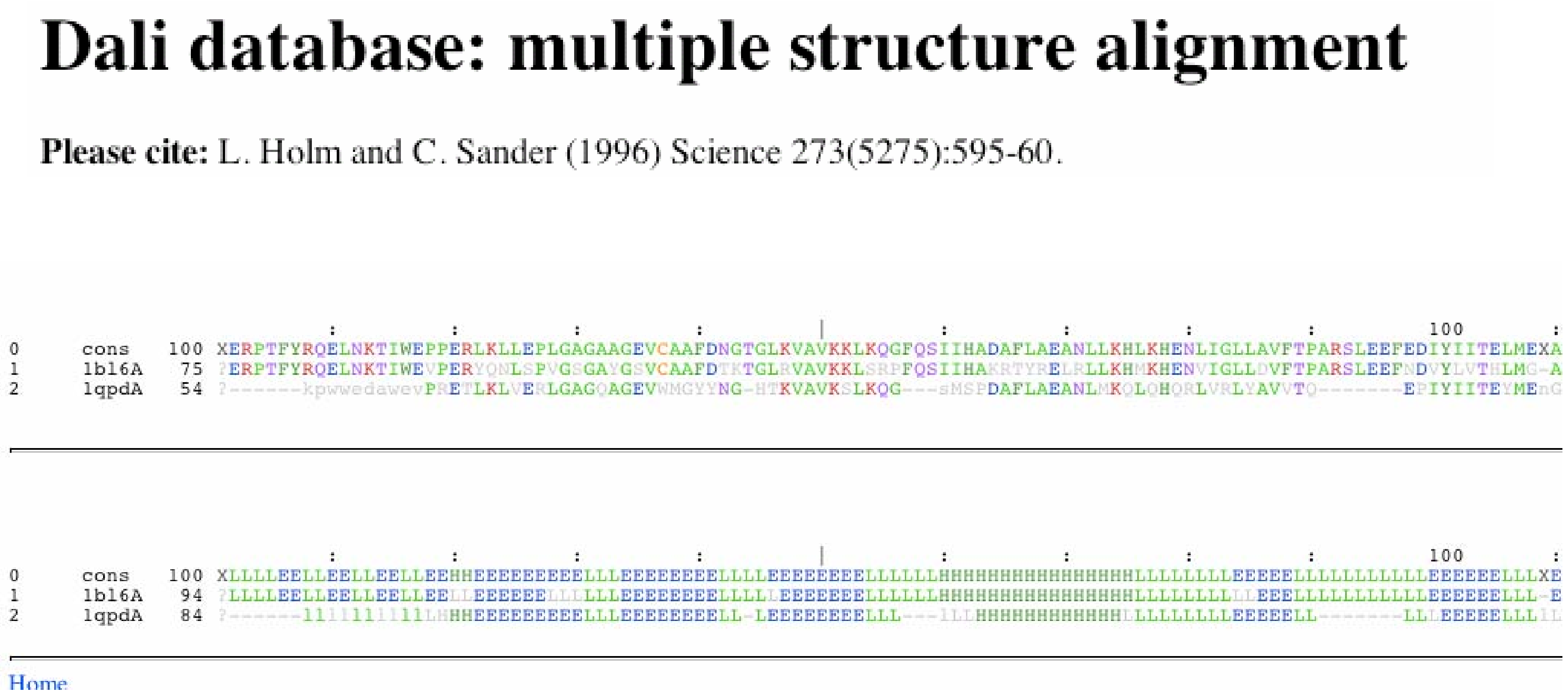
Please cite: L. Holm and C. Sander (1996) Science 273(5275):595-60.

	neighbour	Z	%ide	rmsd	lali	lseq2	PDB	compound
☑	0: 1bl6A	58.0	100	0.0	351	351	PDB	MAP KINASE P38
□	1: 1gol	38.4	46	2.4	329	357	PDB	EXTRACELLULAR REGULATED KINASE 2
□	2: 1jnk	37.5	50	2.6	326	346	PDB	C-JUN N-TERMINAL KINASE
□	3: 1cm8A	36.8	60	3.0	320	329	PDB	PHOSPHORYLATED MAP KINASE P38-GAMMA
□	4: 1blxA	29.1	34	2.9	276	305	PDB	CYCLIN-DEPENDENT KINASE 6
□	5: 1finA	28.7	37	2.6	276	298	PDB	CYCLIN-DEPENDENT KINASE 2
□	29: 1kswA	21.0	23	3.5	240	450	PDB	PROTO-ONCOGENE TYROSINE-PROTEIN KINASE SRC
☑	30: 1qpdA	20.9	24	3.0	237	271	PDB	LCK KINASE
□	31: 1vr2A	20.9	23	2.7	236	275	PDB	VASCULAR ENDOTHELIAL GROWTH FACTOR RECEPTOR



24% sequence ID, rmsd = 3.0 Å

http://www.ebi.ac.uk/dali/



Dali database: multiple structure alignment

Please cite: L. Holm and C. Sander (1996) Science 273(5275):595-60.

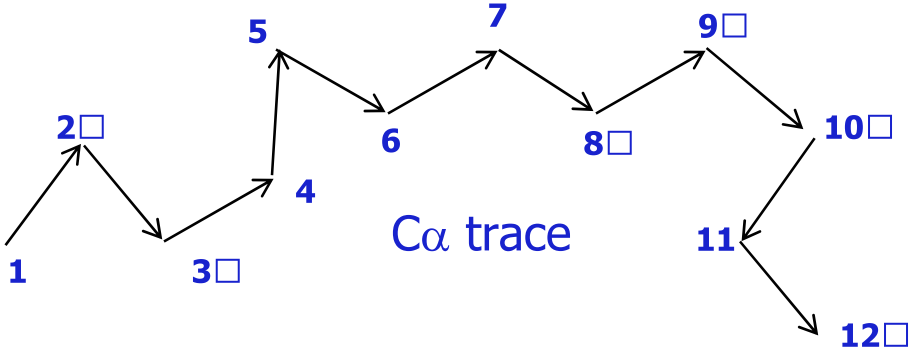
```
                           :         :         :         :         |         :         :         :         :        100        :
0     cons    100 XERPTFYRQELNKTIWEPPERLKLLEPLGAGAAGEVCAAFDNGTGLKVAVKKLKQGFQSIIHADAFLAEANLLKHLKHENLIGLLAVFTPARSLEEFEDIYIITELMEXA
1     1bl6A    75 ?ERPTFYRQELNKTIWEVPERYQNLSPVGSGAYGSVCAAFDTKTGLRVAVKKLSRPFQSIIHAKRTYRELRLLKHMKHENVIGLLDVFTPARSLEEFNDVYLVTHLMG-A
2     1qpdA    54 ?------kpwwedawevPRETLKLVERLGAGQAGEVWMGYYNG-HTKVAVKSLKQG---sMSPDAFLAEANLMKQLQHQRLVRLYAVVTQ-------EPIYIITEYMEnG
```

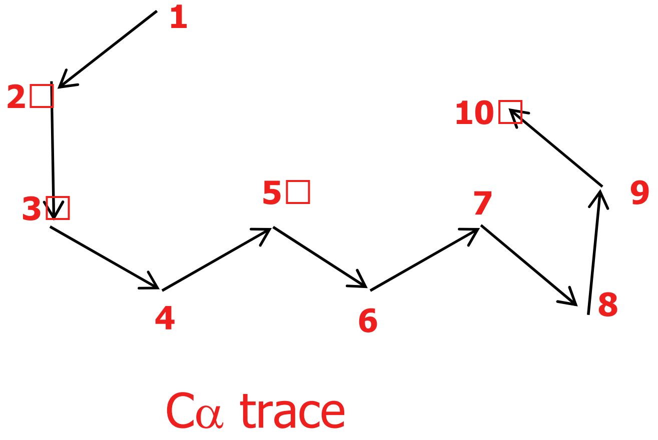
```
                           :         :         :         :         |         :         :         :         :        100        :
0     cons    100 XLLLLEELLEELLEELLEEHHEEEEEEEEELLLEEEEEEEELLLLEEEEEEEELLLLLLHHHHHHHHHHHHHHHHLLLLLLLLEEEEELLLLLLLLLLLEEEEEELLLXE
1     1bl6A    94 ?LLLLEELLEELLEELLEELLEEEEEELLLLLLEEEEEEEELLLLLEEEEEEELLLLLLHHHHHHHHHHHHHHHHLLLLLLLLLLEEELLLLLLLLLLLEEEEEELLL-E
2     1qpdA    84 ?------1111111111LHHHEEEEEEEEELLLEEEEEEEELL-LEEEEEEEELLL---1LLHHHHHHHHHHHHLLLLLLLLLEEEEELL-------LLLEEEEELLL1L
```

Home

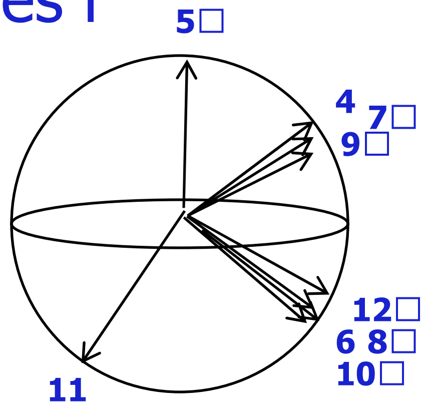
structure-based sequence alignment

unitRMS

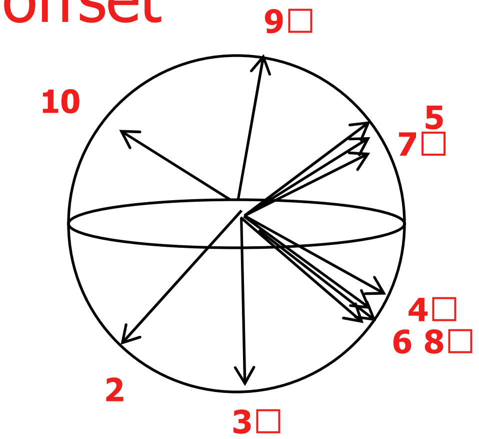




indices i



j = i + offset



$$\mathrm{URMS} = \text{min\_over\_rotations}\left(\Sigma(\mathbf{V}_i - \mathbf{V}_j)^2\right)^{1/2}$$

Chew et al, RECOMB (1999)
Kedem et al. PROTEINS 37, 554 (1999)

URMS advantages

1. Insensitive to outliers
   $\text{URMS}_{\text{max}} = 2$

2. Weighs all parts of protein equally

3. URMS_{min} is bounded - not very sensitive to length of protein

4. More compact representation - O(n), compared to $O(n^2)$ for distance matrices

5. Fast to compute: O(nlogn) for searching for substructures