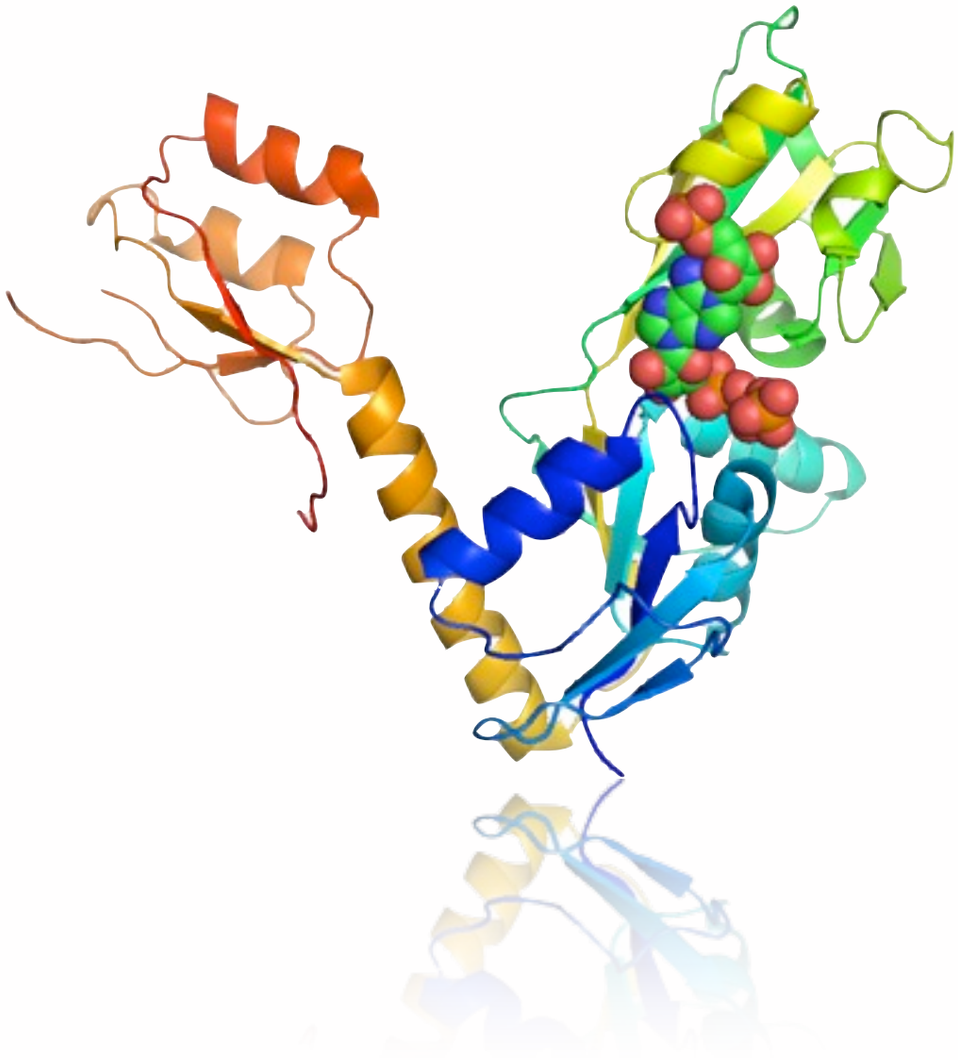
Structural Bioinformatics

Davide Baù
Staff Scientist
Genome Biology Group (CNAG)
Structural Genomics Group (CRG)

dbau@pcb.ub.cat

Proteins



Summary

Protein structural levels

Primary Secondary Tertiary Quaternary

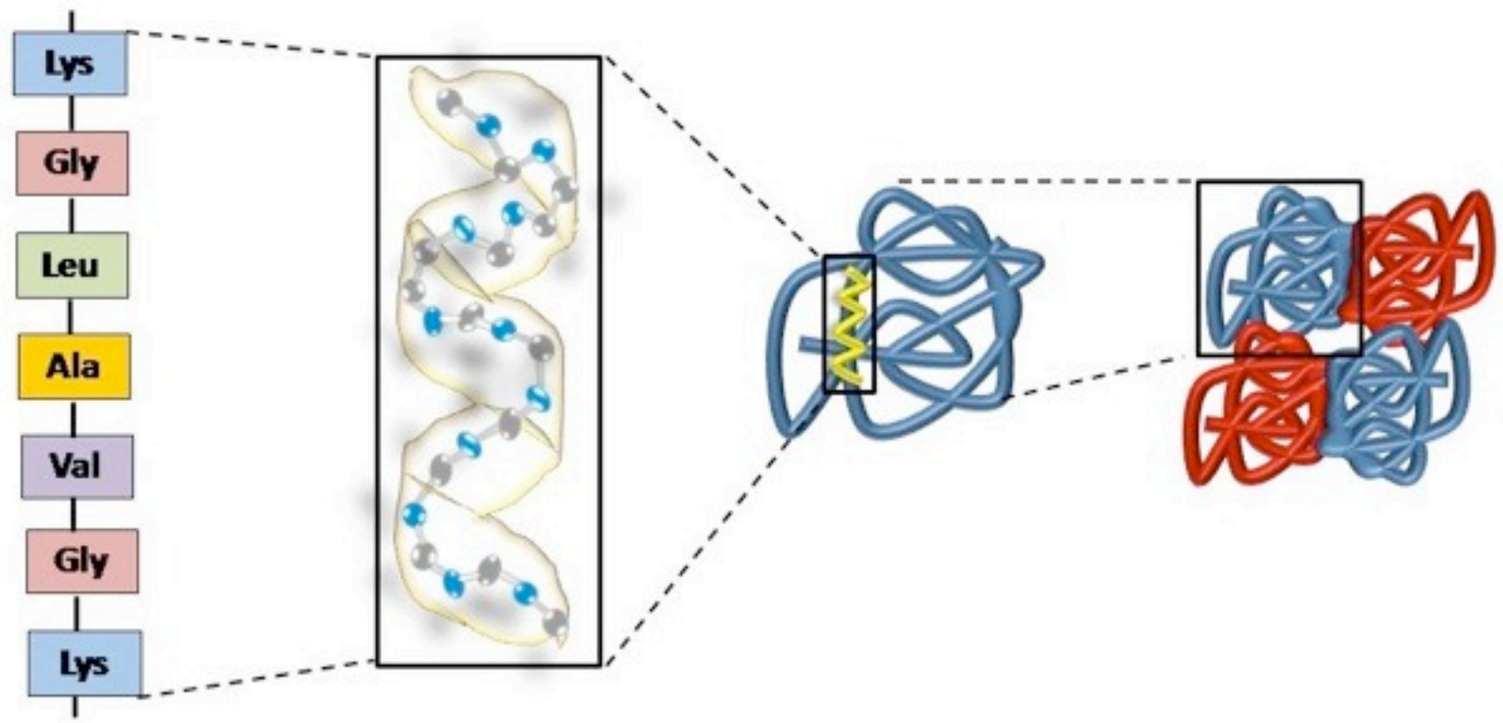


Image credits: http://iitb.vlab.co.in/

Take home message

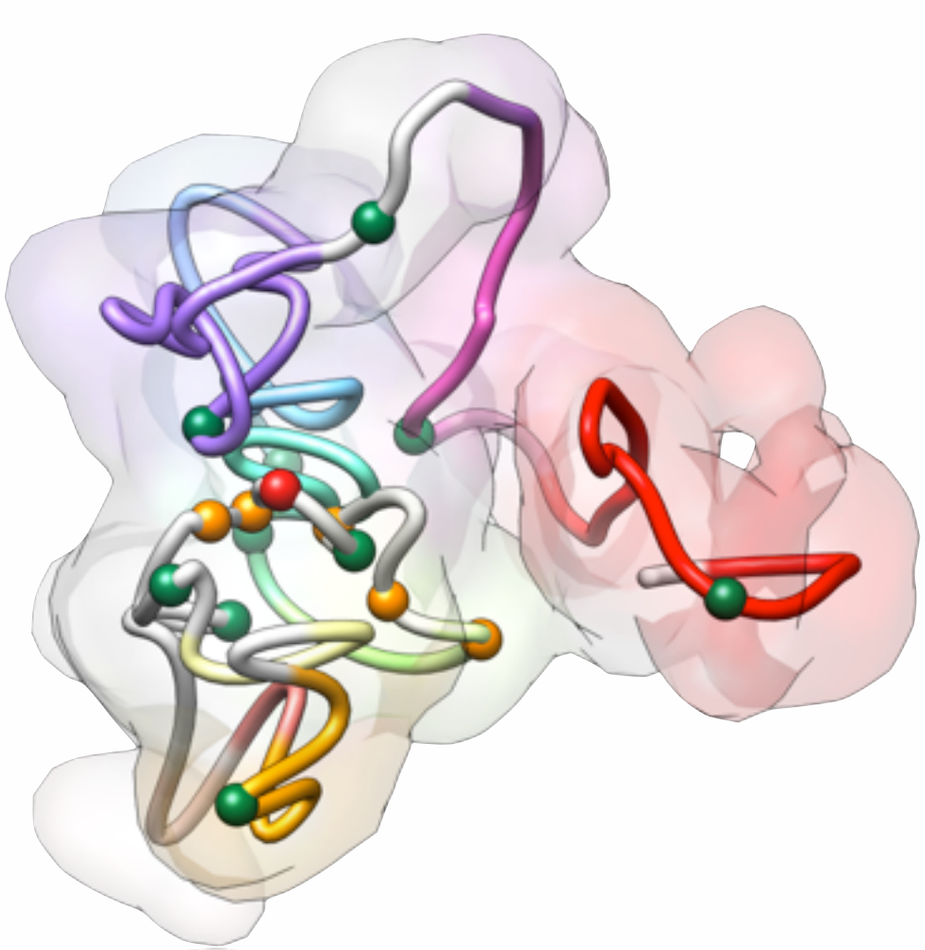
Biochemical function
Activity depends on the 3D structure

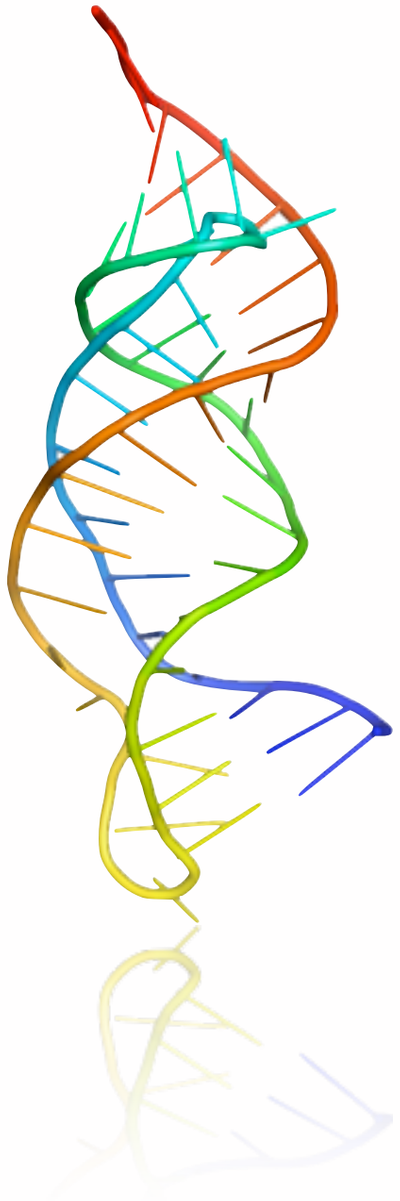
Evolution conserve
Structure is more conserved than sequence

Protein types
Fibrous
Membrane
Globular

Nucleic acids

DNA and RNA





Take home message

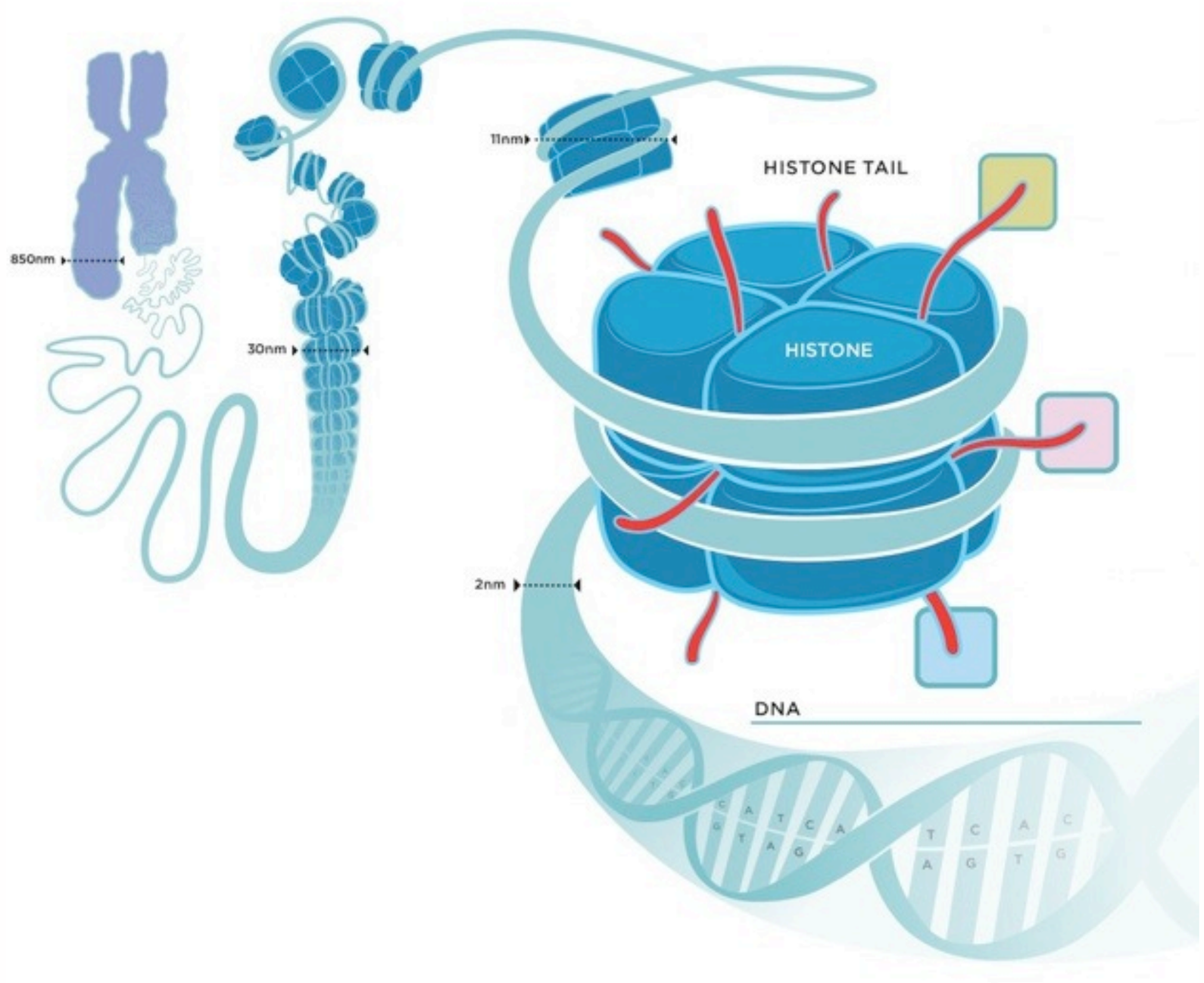
DNA and RNA
Polymers of nucleotide units

Nucleotides
Nucleobase (G,C,A,T - U)
+ sugar +phosphate

DNA
Store the genetic information
RNA
Implicated in various
biological processes

The nuclear organization of DNA

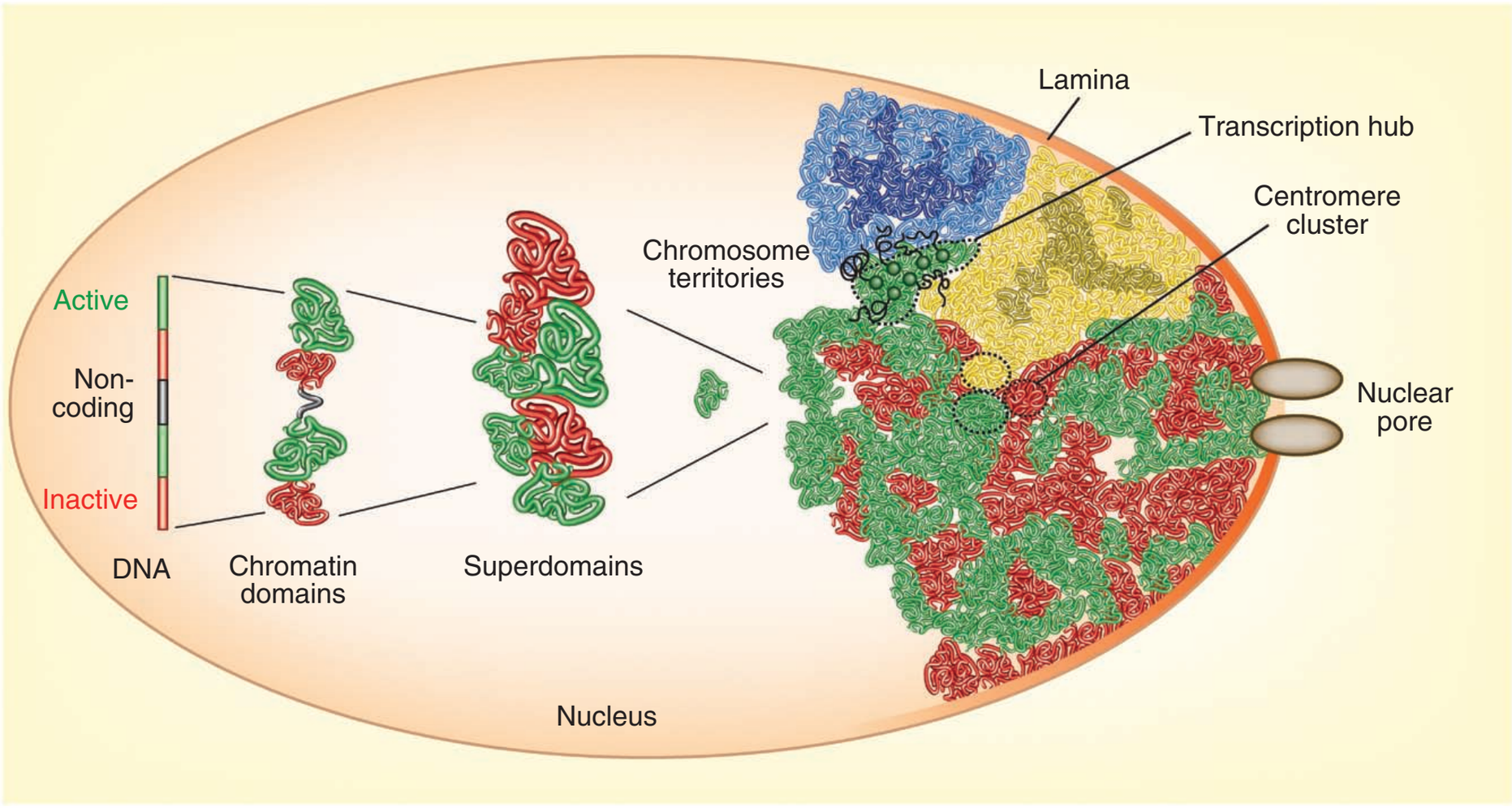
Chromosome **Chromatin fibre** **Nucleosome**



Adapted from Richard E. Ballermann, 2012

Complex genome organization

Cavalli, G. & Misteli, Nat Struct Mol Biol 20, 290–299 (2013)



Marina Corral

Take home message

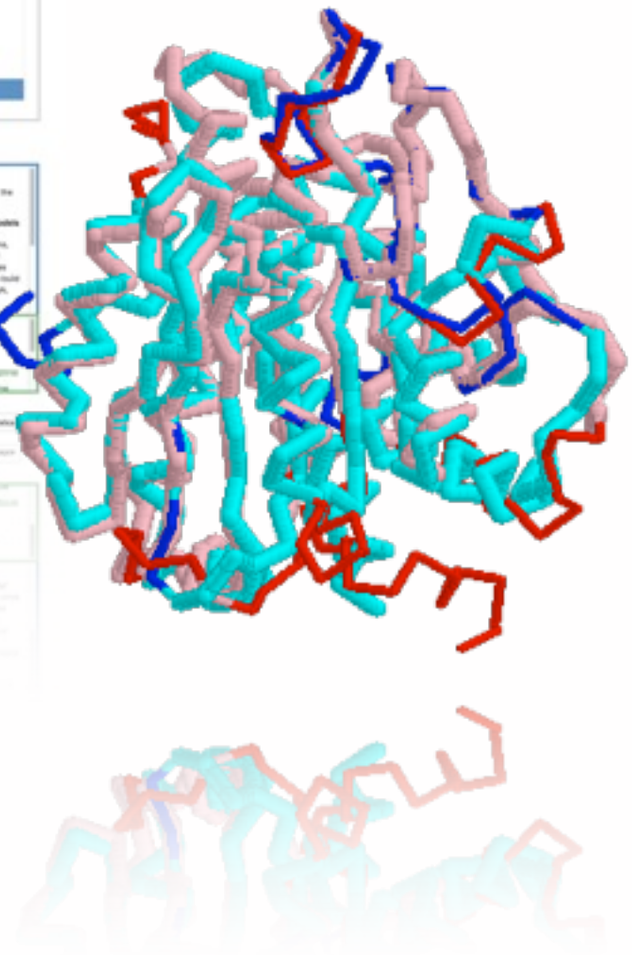
Chromatin = DNA + (histone) proteins + other biomolecules

The genome is well organized and hierarchically packaged

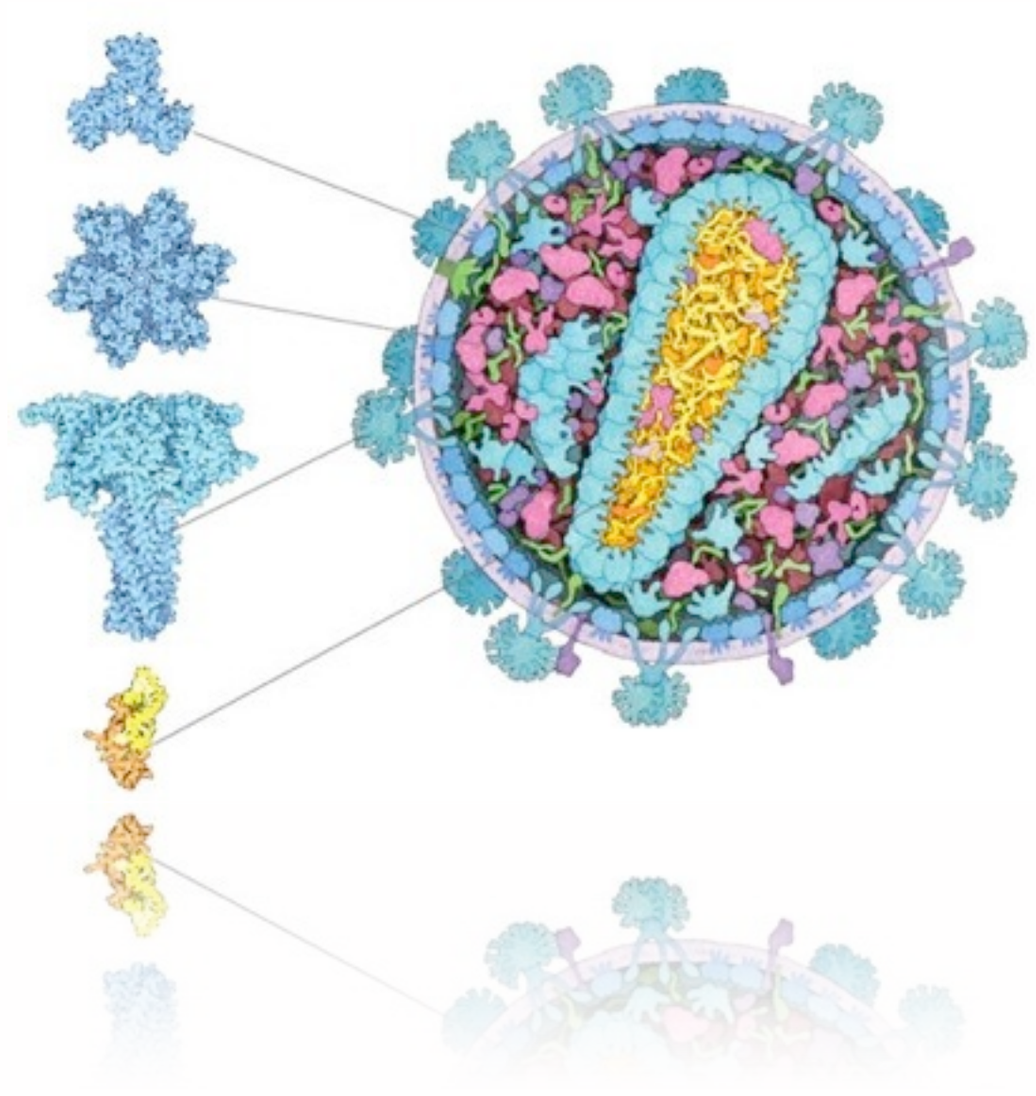
Histone modifications affect chromatin structure and activity

3C-like data measure the frequency of interaction between distant loci

Databases, alignments and structure classification



Known structures



The Protein Data Bank

http://www.pdb.org

PDB History

Structural Biology Highlights

2011 — **PDB40 Symposium**
More than 77000 structures in the archive

2010 — Deposition of chemical shift data mandatory

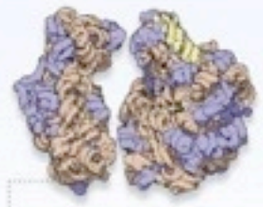
2009 — Nobel prize awarded to V. Ramakrishnan, T.A. Steitz, A.E. Yonath for *Studies of the structure and function of the ribosome*

2008 — Deposition of experimental data mandatory
50000th structure is released

2007 — First release of remediated data

2006 — **BMRB** joins wwPDB

Nobel prize awarded to R.D. Kornberg for *Studies of the molecular basis of eukaryotic transcription*



2005 — Last PDB archive distribution by postal mail (8 DVDs)

2003 — **wwPDB** established by RCSB PDB, PDBe, PDBj

Nobel prize awarded to R. MacKinnon for *Potassium channels*

2002 — **EMDB** established at MSD-EBI

Nobel prize awarded to K. Wüthrich for *Development of NMR spectroscopy for determining the 3D structure of biological macromolecules in solution*

2000 — **Osaka University** opens a PDB data deposition center

First **ribosome structures** determined (Ban *et al.*, 2000; Carter *et al.*, 2000; Schluenzen *et al.*, 2000)

1999 — **10000th structure is released**

1998 — PDB moves to **RCSB PDB**

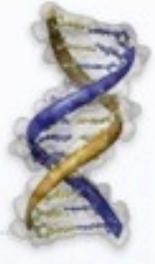
1997 — **MSD** at the EBI becomes a deposition center for PDB data

Nobel prize awarded to P.D. Boyer, J.E. Walker, J.C. Skou for *Elucidation of the enzymatic mechanism underlying the synthesis of adenosine triphosphate (ATP) and discovery of an ion-transporting enzyme*

1991 — First **EM entry** released in PDB: bacteriorhodopsin (Henderson *et al.*, 1990)

1989 — IUCr policy on data deposition published

First **NMR entry** released in PDB: protein BDS-I (Driscoll *et al.*, 1989)



1988 — Nobel prize awarded to J. Deisenhofer, R. Huber, H. Michel for *The determination of the 3D structure of a photosynthetic reaction centre*

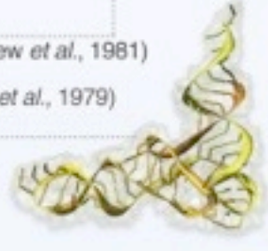
1982 — Nobel prize awarded to A. Klug for *Development of crystallographic electron microscopy and discovery of the structure of biologically important nucleic acid-protein complexes*

Early structures include
carboxypeptidase
chymotrypsin
cytochrome b5
hemoglobin
lactate dehydrogenase
myoglobin
rubredoxin
subtilisin
trypsin inhibitor

1981 — First **B-DNA** structure determined (Drew *et al.*, 1981)

1979 — First **DNA (Z-DNA)** determined (Wang *et al.*, 1979)

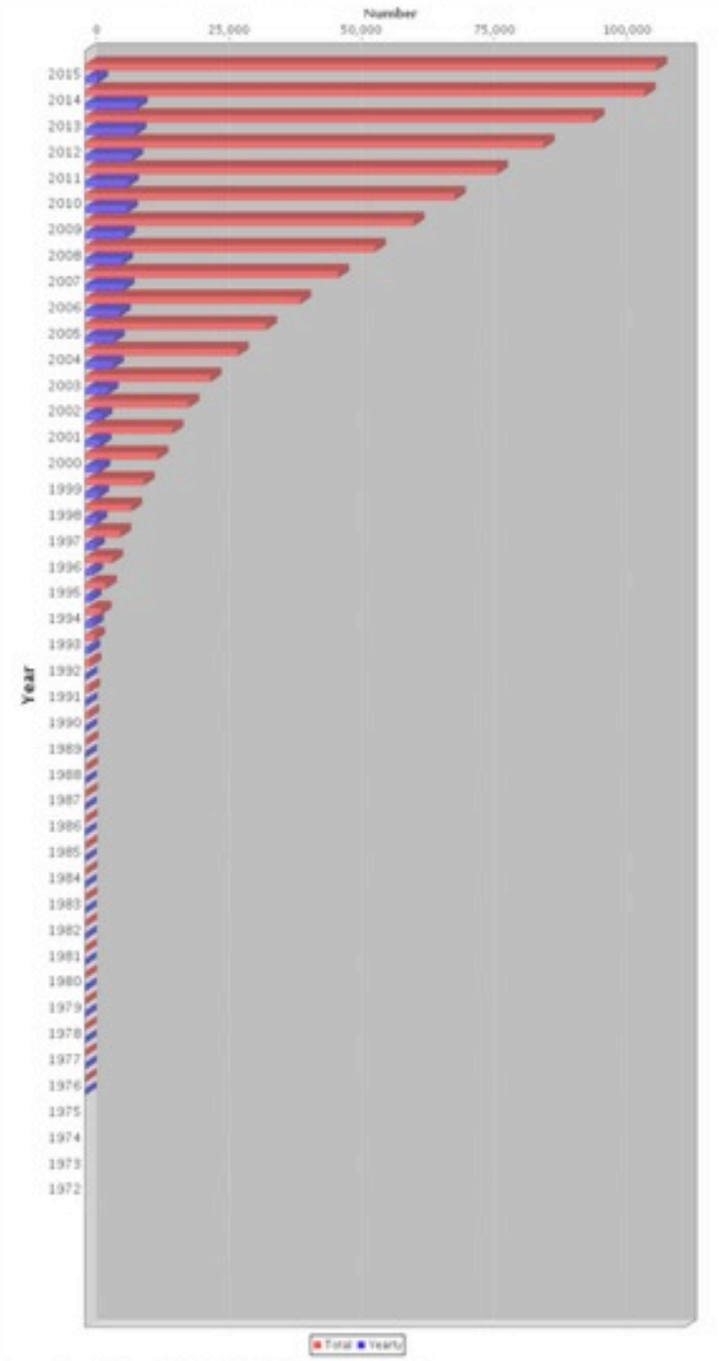
1973 — First **tRNA** structures determined (Kim *et al.*, 1973; Robertus *et al.*, 1974)



1971 — **PDB established**

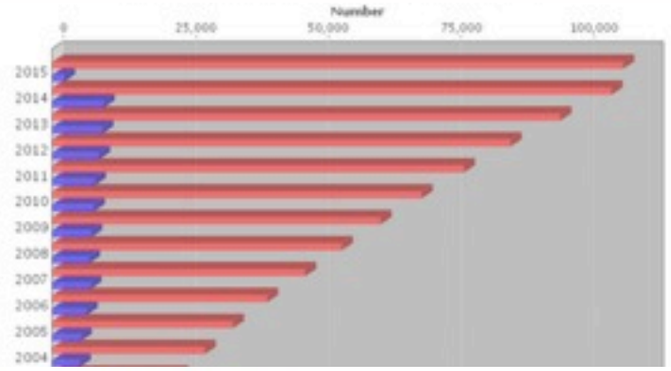
Yearly growth of total structures

http://www.pdb.org



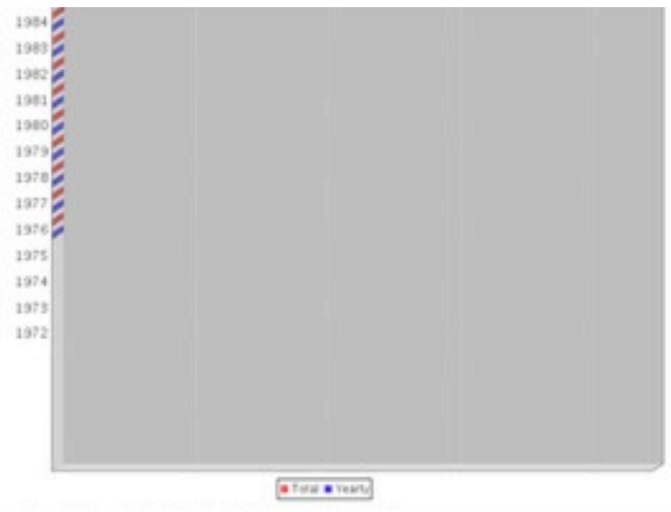
Yearly growth of total structures

http://www.pdb.org



PDB Current Holdings Breakdown

Exp.Method	Proteins	Nucleic Acids	Protein/NA Complexes	Other	Total
X-RAY	89684	1612	4426	4	95726
NMR	9534	1112	224	8	10878
ELECTRON MICROSCOPY	549	29	177	0	755
HYBRID	68	3	2	1	74
other	164	4	6	13	187
Total	99999	2760	4835	26	107620



PDB format

http://www.pdb.org

```
HEADER    EXTRACELLULAR MATRIX                      22-JAN-98   1A3I
TITLE     X-RAY CRYSTALLOGRAPHIC DETERMINATION OF A COLLAGEN-LIKE
TITLE    2 PEPTIDE WITH THE REPEATING SEQUENCE (PRO-PRO-GLY)
...
EXPDTA    X-RAY DIFFRACTION
AUTHOR    R.Z.KRAMER,L.VITAGLIANO,J.BELLA,R.BERISIO,L.MAZZARELLA,
AUTHOR   2 B.BRODSKY,A.ZAGARI,H.M.BERMAN
...
REMARK 350 BIOMOLECULE: 1
REMARK 350 APPLY THE FOLLOWING TO CHAINS: A, B, C
REMARK 350   BIOMT1   1  1.000000  0.000000  0.000000        0.00000
REMARK 350   BIOMT2   1  0.000000  1.000000  0.000000        0.00000
...
SEQRES   1 A    9  PRO PRO GLY PRO PRO GLY PRO PRO GLY
SEQRES   1 B    6  PRO PRO GLY PRO PRO GLY
SEQRES   1 C    6  PRO PRO GLY PRO PRO GLY
...
ATOM      1  N   PRO A   1       8.316  21.206  21.530  1.00 17.44           N
ATOM      2  CA  PRO A   1       7.608  20.729  20.336  1.00 17.44           C
ATOM      3  C   PRO A   1       8.487  20.707  19.092  1.00 17.44           C
ATOM      4  O   PRO A   1       9.466  21.457  19.005  1.00 17.44           O
ATOM      5  CB  PRO A   1       6.460  21.723  20.211  1.00 22.26           C
...
HETATM  130  C   ACY   401       3.682  22.541  11.236  1.00 21.19           C
HETATM  131  O   ACY   401       2.807  23.097  10.553  1.00 21.19           O
HETATM  132  OXT ACY   401       4.306  23.101  12.291  1.00 21.19           O
...
```

PDB format

http://www.pdb.org

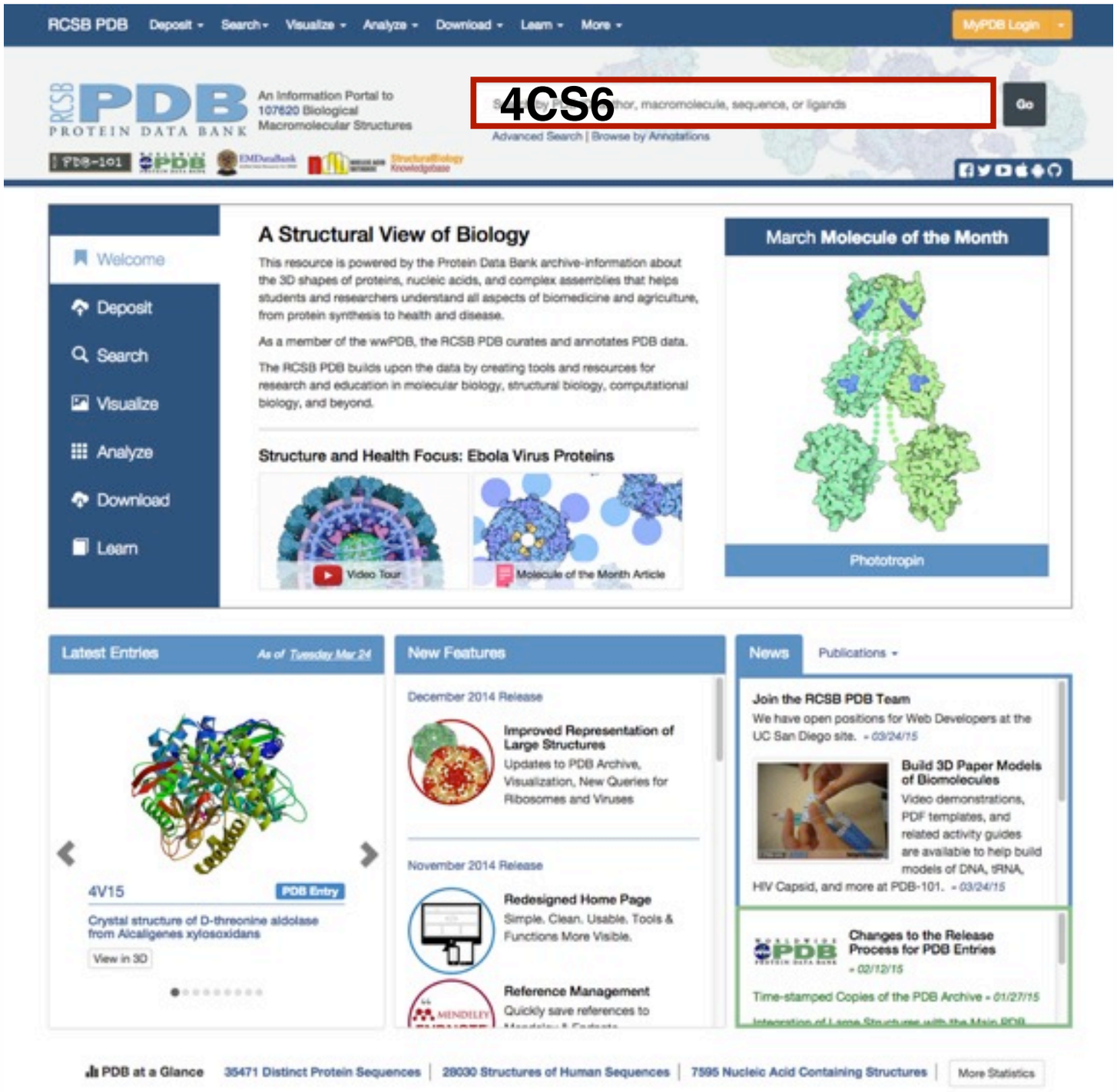
```
ATOM      1  N   GLY A   1      15.740  11.178 -11.733  1.00  0.00           N
ATOM      2  CA  GLY A   1      15.234  10.462 -10.556  1.00  0.00           C
ATOM      3  C   GLY A   1      16.284   9.483  -9.998  1.00  0.00           C
ATOM      4  O   GLY A   1      17.150   8.979 -10.709  1.00  0.00           O
ATOM      5  N   LEU A   2      16.122   9.240  -8.705  1.00  0.00           N
ATOM      6  CA  LEU A   2      16.803   8.164  -7.994  1.00  0.00           C
ATOM      7  C   LEU A   2      17.902   7.481  -8.831  1.00  0.00           C
ATOM      8  O   LEU A   2      19.057   7.424  -8.402  1.00  0.00           O
ATOM      9  CB  LEU A   2      15.755   7.101  -7.594  1.00  0.00           C
ATOM     10  CG  LEU A   2      14.565   7.724  -6.856  1.00  0.00           C
ATOM     11  CD1 LEU A   2      14.958   8.214  -5.453  1.00  0.00           C
ATOM     12  CD2 LEU A   2      13.894   8.850  -7.657  1.00  0.00           C
ATOM     13  N   SER A   3      17.505   6.971  -9.986  1.00  0.00           N
ATOM     14  CA  SER A   3      18.416   6.404 -10.972  1.00  0.00           C
ATOM     15  C   SER A   3      19.535   5.484 -10.470  1.00  0.00           C
ATOM     16  O   SER A   3      19.869   5.398  -9.293  1.00  0.00           O
ATOM     17  CB  SER A   3      19.079   7.578 -11.738  1.00  0.00           C
ATOM     18  OG  SER A   3      19.875   8.293 -10.785  1.00  0.00           O
ATOM     19  N   ASP A   4      20.127   4.842 -11.478  1.00  0.00           N
ATOM     20  CA  ASP A   4      21.268   3.953 -11.276  1.00  0.00           C
ATOM     21  C   ASP A   4      22.229   4.485 -10.197  1.00  0.00           C
ATOM     22  O   ASP A   4      22.347   3.897  -9.121  1.00  0.00           O
ATOM     23  CB  ASP A   4      22.048   3.770 -12.587  1.00  0.00           C
ATOM     24  CG  ASP A   4      21.138   3.738 -13.824  1.00  0.00           C
ATOM     25  OD1 ASP A   4      20.964   4.838 -14.416  1.00  0.00           O
ATOM     26  OD2 ASP A   4      20.667   2.612 -14.138  1.00  0.00           O
ATOM     27  N   GLY A   5      22.907   5.579 -10.530  1.00  0.00           N
ATOM     28  CA  GLY A   5      23.875   6.166  -9.594  1.00  0.00           C
ATOM     29  C   GLY A   5      23.310   6.210  -8.163  1.00  0.00           C
ATOM     30  O   GLY A   5      23.845   5.573  -7.256  1.00  0.00           O
ATOM     31  N   GLU A   6      22.228   6.967  -8.038  1.00  0.00           N
ATOM     32  CA  GLU A   6      21.545   7.090  -6.742  1.00  0.00           C
ATOM     33  C   GLU A   6      21.419   5.695  -6.103  1.00  0.00           C
ATOM     34  O   GLU A   6      21.902   5.451  -4.997  1.00  0.00           O
ATOM     35  CB  GLU A   6      20.162   7.735  -6.956  1.00  0.00           C
ATOM     36  CG  GLU A   6      19.493   8.120  -5.628  1.00  0.00           C
ATOM     37  CD  GLU A   6      18.554   9.328  -5.804  1.00  0.00           C
ATOM     38  OE1 GLU A   6      17.794   9.578  -4.827  1.00  0.00           O
ATOM     39  OE2 GLU A   6      18.634   9.959  -6.891  1.00  0.00           O
ATOM     40  N   TRP A   7      20.694   4.832  -6.811  1.00  0.00           N
ATOM     41  CA  TRP A   7      20.454   3.477  -6.298  1.00  0.00           C
ATOM     42  C   TRP A   7      21.756   2.847  -5.765  1.00  0.00           C
ATOM     43  O   TRP A   7      21.824   2.453  -4.599  1.00  0.00           O
ATOM     44  CB  TRP A   7      19.878   2.574  -7.407  1.00  0.00           C
ATOM     45  CG  TRP A   7      18.589   3.146  -7.920  1.00  0.00           C
ATOM     46  CD1 TRP A   7      18.358   3.602  -9.158  1.00  0.00           C
ATOM     47  CD2 TRP A   7      17.391   3.325  -7.161  1.00  0.00           C
ATOM     48  NE1 TRP A   7      17.049   4.030  -9.246  1.00  0.00           N
ATOM     49  CE2 TRP A   7      16.475   3.892  -8.055  1.00  0.00           C
ATOM     50  CE3 TRP A   7      17.004   3.075  -5.840  1.00  0.00           C
```

PDB search

http://www.pdb.org

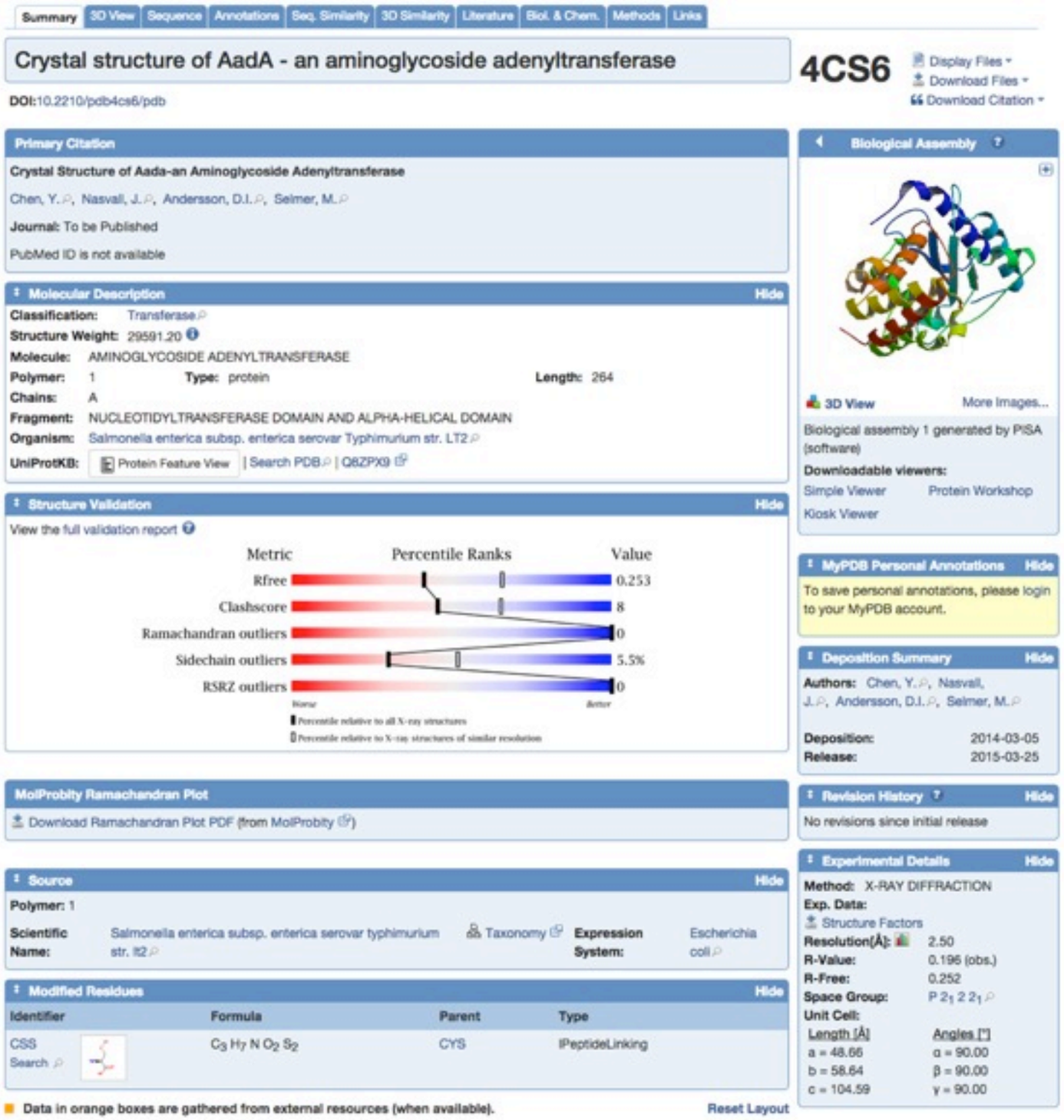
PDB search

http://www.pdb.org



PDB search

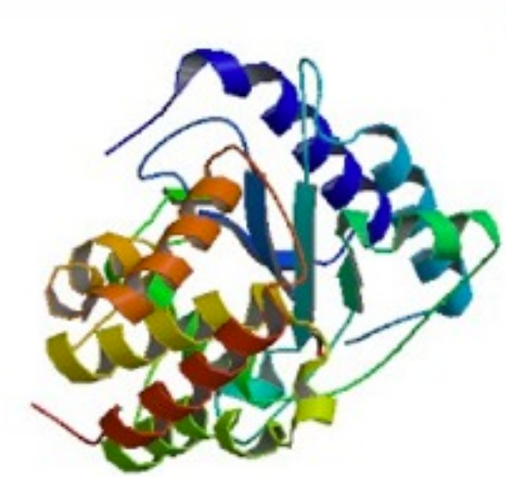
http://www.pdb.org



PDB search

http://www.pdb.org

Biological Assembly ?



3D View More Images...

Biological assembly 1 generated by PISA (software)

Downloadable viewers:

Simple Viewer Protein Workshop

Kiosk Viewer

Experimental Details Hide

Method: X-RAY DIFFRACTION

Exp. Data:

Structure Factors

Resolution[Å]: 2.50

R-Value: 0.196 (obs.)

R-Free: 0.252

Space Group: P 2_1 2 2_1

Unit Cell:

Length [Å]	Angles [°]
a = 48.66	α = 90.00
b = 58.64	β = 90.00
c = 104.59	γ = 90.00

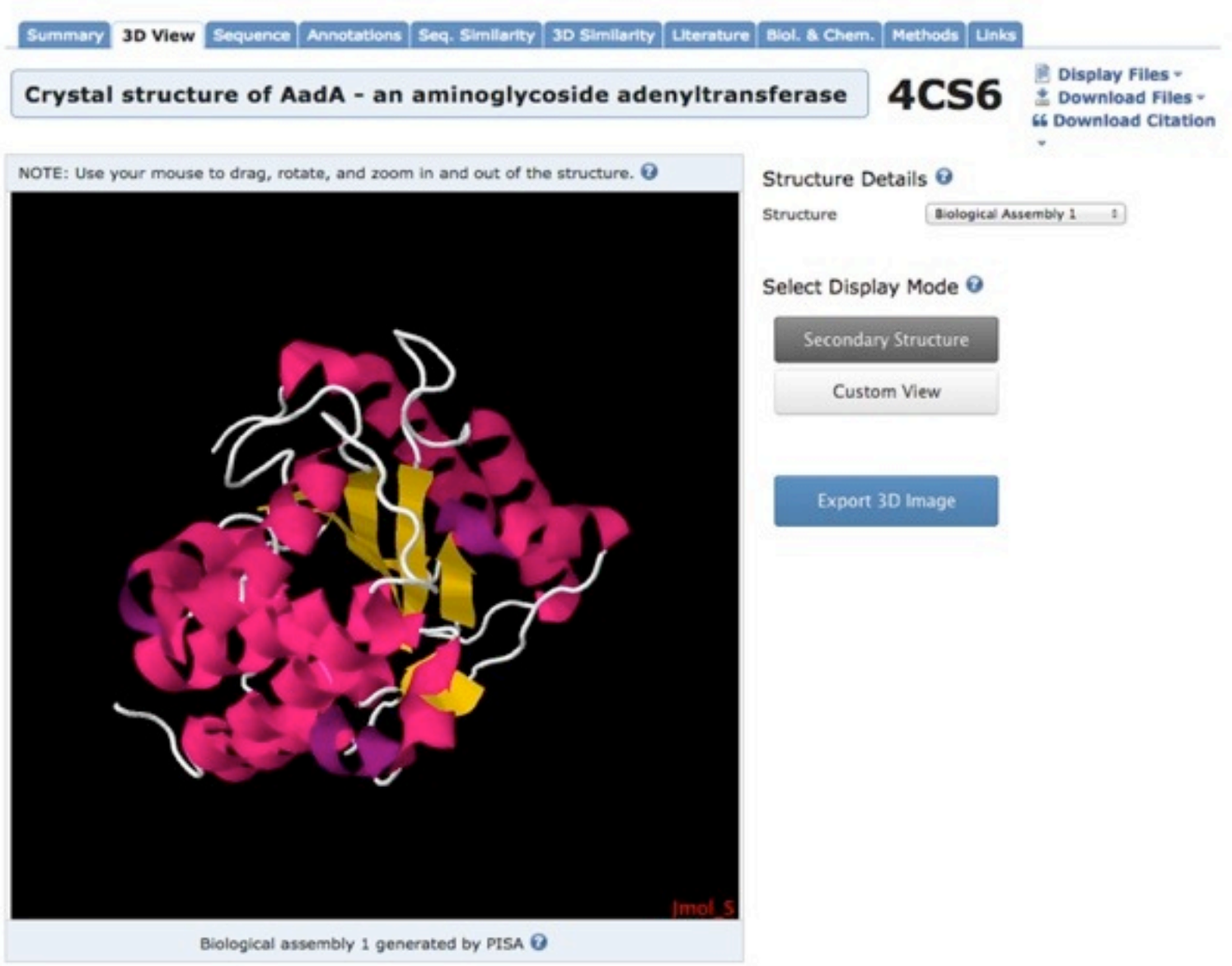
PDB search

http://www.pdb.org

```
HEADER    TRANSFERASE                             05-MAR-14   4CS6
TITLE     CRYSTAL STRUCTURE OF AADA - AN AMINOGLYCOSIDE ADENYLTRANSFERASE
COMPND    MOL_ID: 1;
COMPND   2 MOLECULE: AMINOGLYCOSIDE ADENYLTRANSFERASE;
COMPND   3 CHAIN: A;
COMPND   4 FRAGMENT: NUCLEOTIDYLTRANSFERASE DOMAIN AND ALPHA-HELICAL DOMAIN;
COMPND   5 ENGINEERED: YES
SOURCE    MOL_ID: 1;
SOURCE   2 ORGANISM_SCIENTIFIC: SALMONELLA ENTERICA SUBSP. ENTERICA SEROVAR
SOURCE   3  TYPHIMURIUM STR. LT2;
SOURCE   4 ORGANISM_TAXID: 99287;
SOURCE   5 EXPRESSION_SYSTEM: ESCHERICHIA COLI;
SOURCE   6 EXPRESSION_SYSTEM_TAXID: 469008;
SOURCE   7 EXPRESSION_SYSTEM_STRAIN: BL21(DE3);
SOURCE   8 EXPRESSION_SYSTEM_VARIANT: STAR
KEYWDS    TRANSFERASE, AMINOGLYCOSIDE ADENYLTRANSFERASE, ANT(3'')
EXPDTA    X-RAY DIFFRACTION
AUTHOR    Y.CHEN,J.NASVALL,D.I.ANDERSSON,M.SELMER
REVDAT   1   25-MAR-15 4CS6    0
JRNL        AUTH   Y.CHEN,J.NASVALL,D.I.ANDERSSON,M.SELMER
JRNL        TITL   CRYSTAL STRUCTURE OF AADA-AN AMINOGLYCOSIDE
JRNL        TITL 2 ADENYLTRANSFERASE
JRNL        REF    TO BE PUBLISHED
JRNL        REFN
REMARK   2
REMARK   2 RESOLUTION.    2.50 ANGSTROMS.
REMARK   3
REMARK   3 REFINEMENT.
REMARK   3  PROGRAM     : PHENIX (PHENIX.REFINE)
REMARK   3  AUTHORS     : PAUL ADAMS,PAVEL AFONINE,VICENT CHEN,IAN
REMARK   3              : DAVIS,KRESHNA GOPAL,RALF GROSSE-KUNSTLEVE,
REMARK   3              : LI-WEI HUNG,ROBERT IMMORMINO,TOM IOERGER,
REMARK   3              : AIRLIE MCCOY,ERIK MCKEE,NIGEL MORIARTY,
REMARK   3              : REETAL PAI,RANDY READ,JANE RICHARDSON,
REMARK   3              : DAVID RICHARDSON,TOD ROMO,JIM SACCHETTINI,
REMARK   3              : NICHOLAS SAUTER,JACOB SMITH,LAURENT
REMARK   3              : STORONI,TOM TERWILLIGER,PETER ZWART
REMARK   3
REMARK   3   REFINEMENT TARGET : ML
REMARK   3
REMARK   3  DATA USED IN REFINEMENT.
REMARK   3   RESOLUTION RANGE HIGH (ANGSTROMS) : 2.502
REMARK   3   RESOLUTION RANGE LOW  (ANGSTROMS) : 39.029
REMARK   3   MIN(FOBS/SIGMA_FOBS)              : 1.34
REMARK   3   COMPLETENESS FOR RANGE        (%) : 99.66
REMARK   3   NUMBER OF REFLECTIONS             : 10816
REMARK   3
REMARK   3  FIT TO DATA USED IN REFINEMENT.
REMARK   3   R VALUE     (WORKING + TEST SET) : 0.1962
REMARK   3   R VALUE           (WORKING SET) : 0.1935
REMARK   3   FREE R VALUE                     : 0.2521
REMARK   3   FREE R VALUE TEST SET SIZE   (%) : 4.8
REMARK   3   FREE R VALUE TEST SET COUNT      : 517
REMARK   3
REMARK   3  FIT TO DATA USED IN REFINEMENT (IN BINS).
REMARK   3   BIN  RESOLUTION RANGE  COMPL.    NWORK NFREE   RWORK  RFREE
REMARK   3     1 39.0340 -  3.9704    0.99     2685   128  0.1698 0.2001
REMARK   3     2  3.9704 -  3.1519    1.00     2571   120  0.1898 0.2684
REMARK   3     3  3.1519 -  2.7536    1.00     2543   126  0.2505 0.3378
REMARK   3     4  2.7536 -  2.5019    0.99     2500   143  0.2754 0.3924
REMARK   3
REMARK   3  BULK SOLVENT MODELLING.
REMARK   3   METHOD USED        : FLAT BULK SOLVENT MODEL
REMARK   3   SOLVENT RADIUS     : 1.11
REMARK   3   SHRINKAGE RADIUS   : 0.90
REMARK   3   K_SOL              : NULL
REMARK   3   B_SOL              : NULL
```

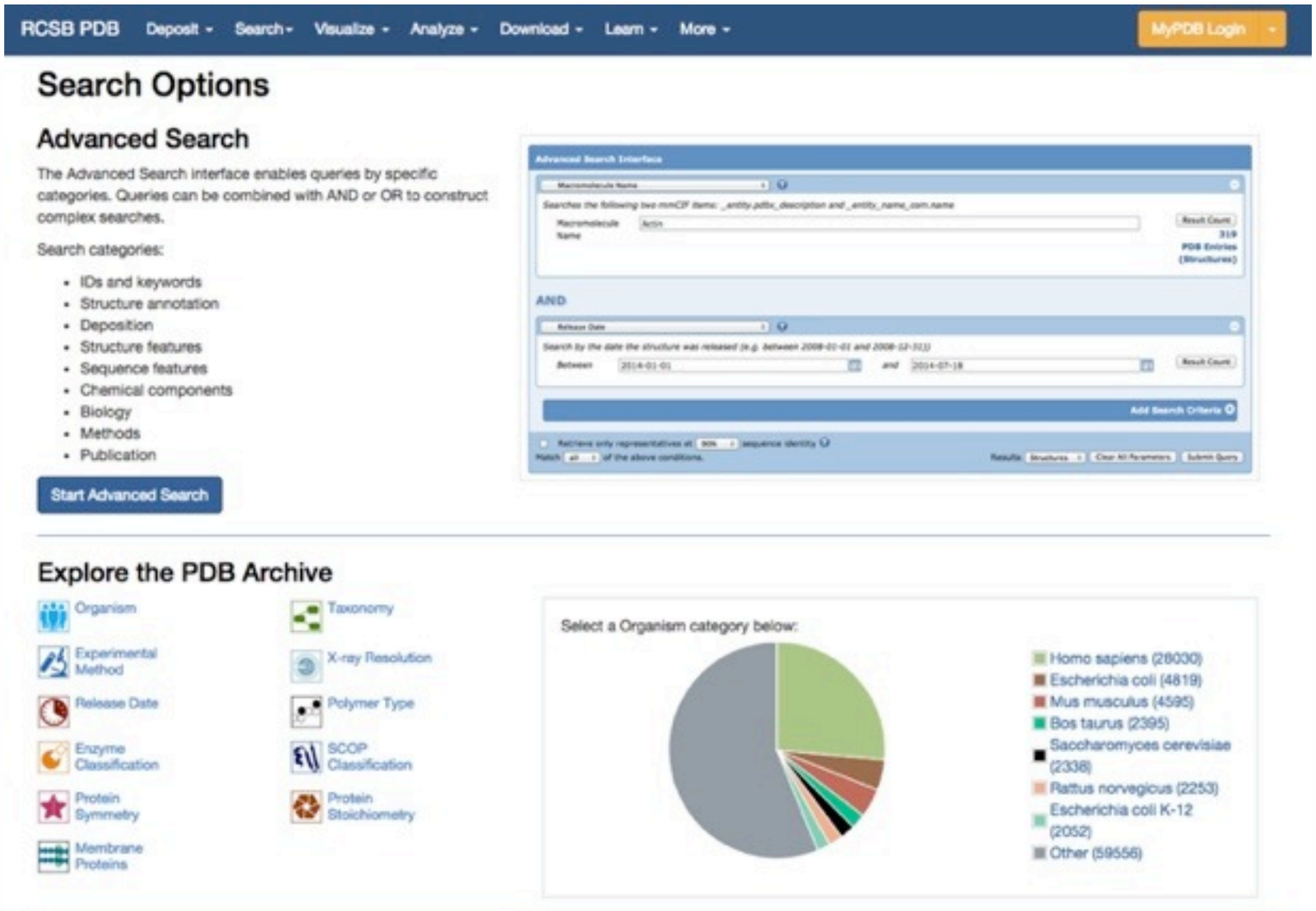
PDB search

http://www.pdb.org

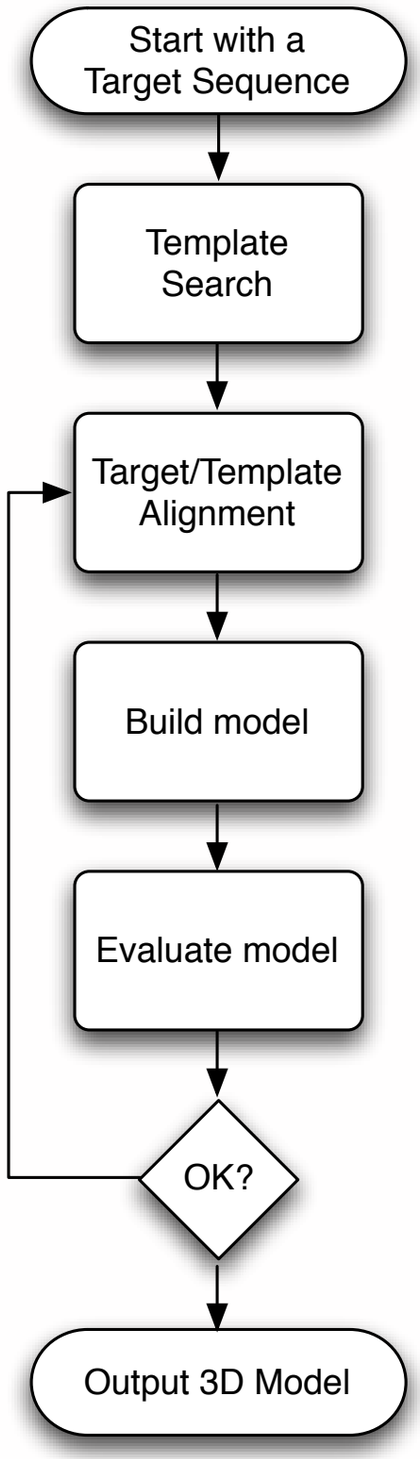


PDB advanced search

http://www.pdb.org



Alignments



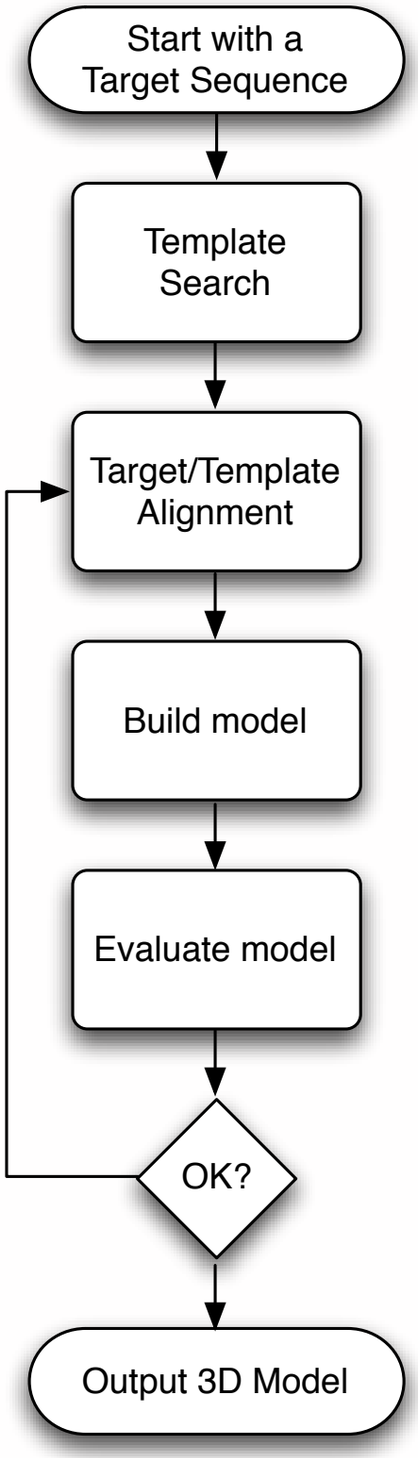
As in any other bioinformatics problem…

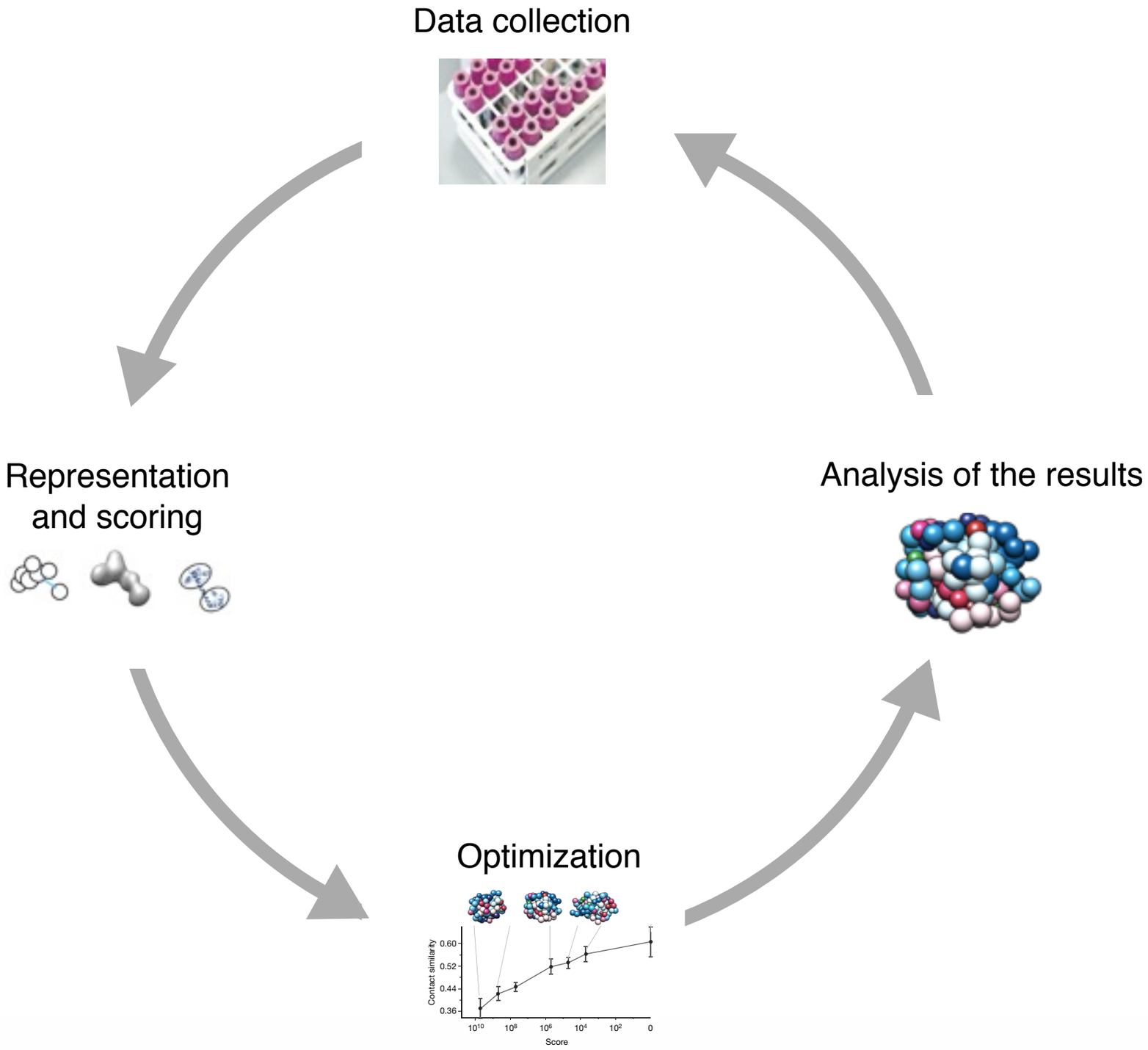
1. Representation

2. Scoring

3. Optimization

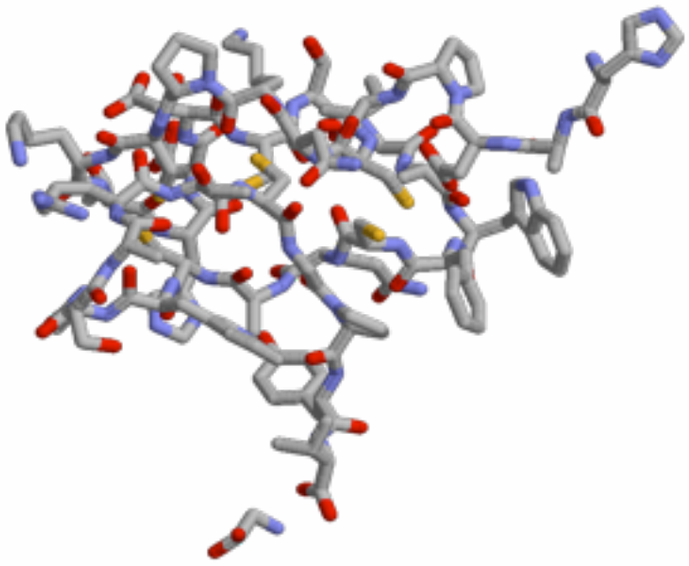
Alignments


Start with a Target Sequence
Template Search
Target/Template Alignment
Build model
Evaluate model
OK?
Output 3D Model

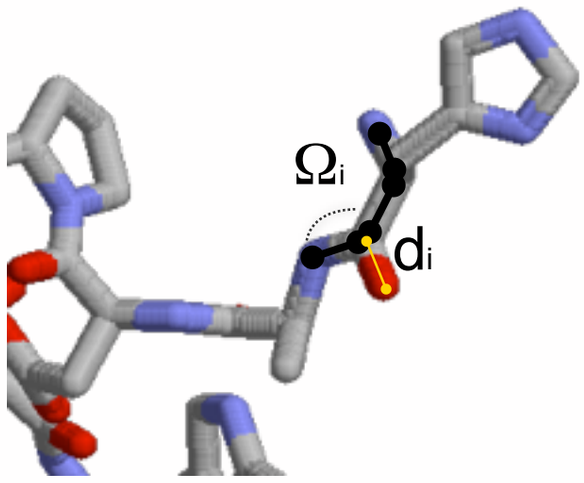

Data collection
Representation and scoring
Optimization
Contact similarity
0.60
0.52
0.44
0.36
10^{10}
10^{8}
10^{6}
10^{4}
10^{2}
0
Score
Analysis of the results

Structures

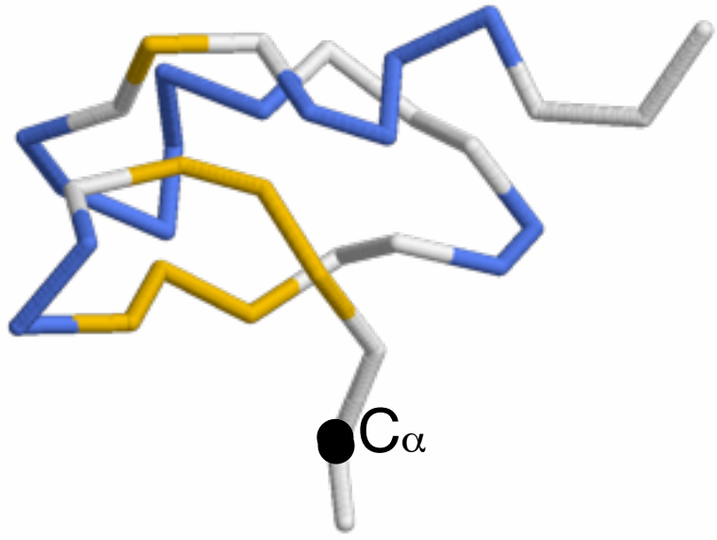
Representation



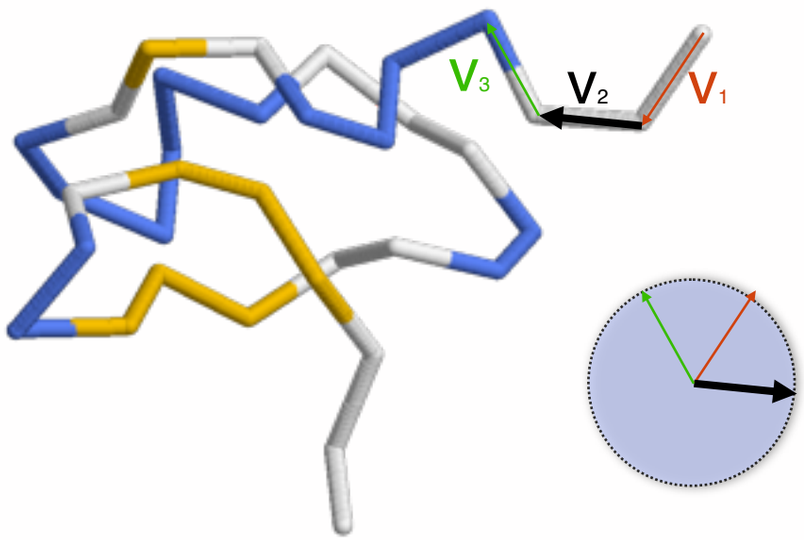
All atoms and coordinates



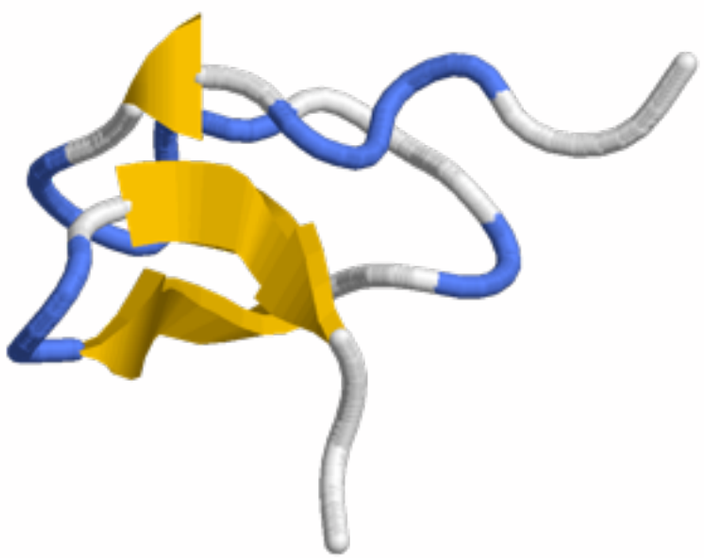
Dihedral space or distance space



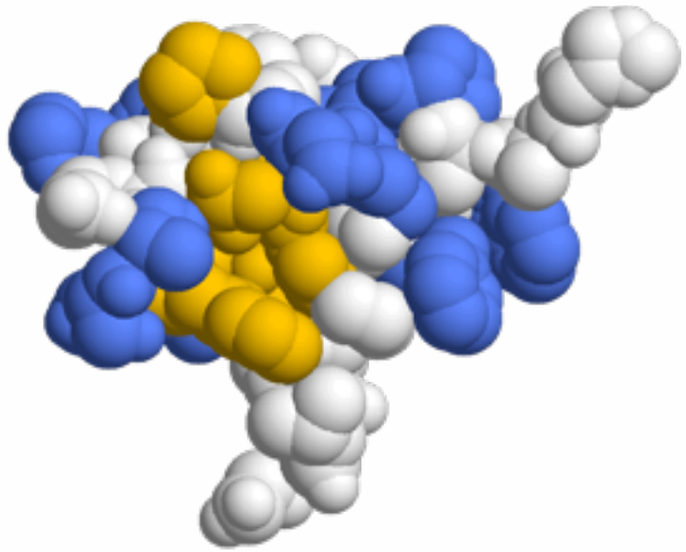
Reduced atom representation



Vector representation



Secondary Structure



Accessible surface (and others)

Raw scores

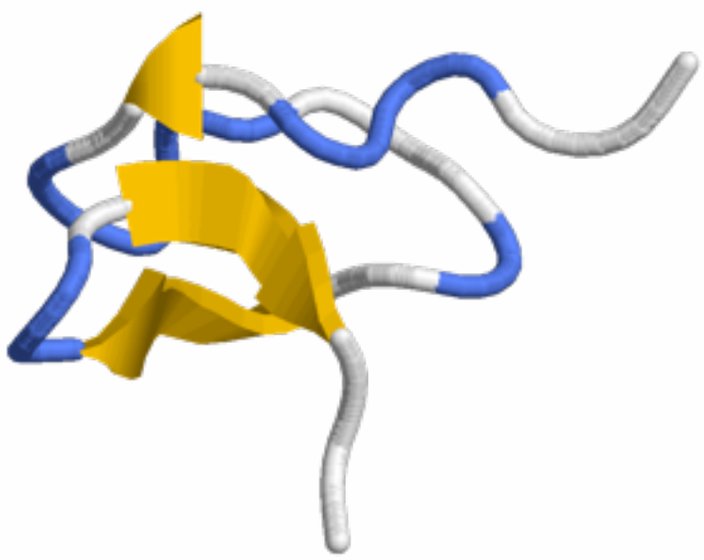
Scoring

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

Aminoacid substitutions

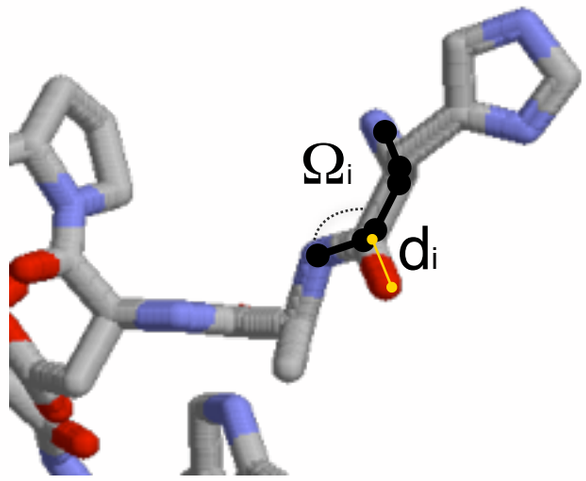
$$\mathrm{RMSD} = \sqrt{\frac{1}{N}\sum_{i}^{N}\left(m_i - m_i^*\right)}$$

Root Mean Square Deviation



Secondary Structure (H,B,C)

Accessible surface (B,A [%])



Angles or distances

The Root Mean Square Deviation

$$\mathrm{RMSD} = \sqrt{\frac{1}{N}\sum_{i}^{N}\left(m_i - m_i^{*}\right)}$$

Number of atoms

3D coordinates of
atom *i* in the first model

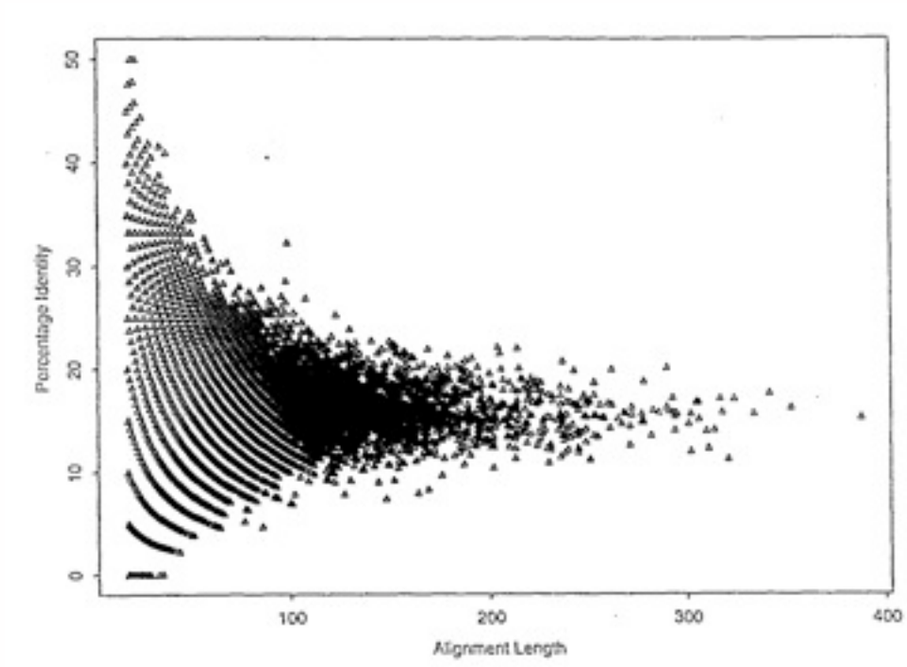
3D coordinates of
atom *i* in the second model

Significance of an alignment (score)

Scoring

Probability that the optimal alignment of two random sequences/structures of the same length and composition as the aligned sequences/structures have at least as good a score as the evaluated alignment.

Empirical:



Analytic:

$$P(S>x) = 1 - \exp(-Kmne^{-\lambda e})$$

Karlin and Altschul, 1990 *PNAS* **87**, pp2264

Global dynamic programming alignment

Optimizer

Needleman-Wunsch

match = 1 mismatch = -1 gap = -1

		G	C	A	T	G	C	U
	0	-1	-2	-3	-4	-5	-6	-7
G	-1	1	0	-1	-2	-3	-4	-5
A	-2	0	0	1	0	-1	-2	-3
T	-3	-1	-1	0	2	1	0	-1
T	-4	-2	-2	-1	1	1	0	-1
A	-5	-3	-3	-1	0	0	0	-1
C	-6	-4	-2	-2	-1	-1	1	0
A	-7	-5	-3	-1	-1	-1	0	0

Backtracking to get the best alignment

```
Sequences     Best Alignments
---------     ----------------------
GCATGCU       GCATG-CU       GCA-TGCU       GCAT-GCU
GATTACA       G-ATTACA       G-ATTACA       G-ATTACA
```

Needleman and Wunsch (1970) *J. Mol Biol*, **3** pp443

Local dynamic programming alignment

Optimizer

Smith-Waterman

match = 2 mismatch = -1 gap = -1

		A	C	A	C	A	C	T	A
	0	0	0	0	0	0	0	0	0
A	0	2	1	2	1	2	1	0	2
G	0	1	1	1	1	1	1	0	1
C	0	0	3	2	3	2	3	2	1
A	0	2	2	5	4	5	4	3	4
C	0	1	4	4	7	6	7	6	5
A	0	2	3	6	6	9	8	7	8
C	0	1	4	5	8	8	11	10	9
A	0	2	3	6	7	10	10	10	12

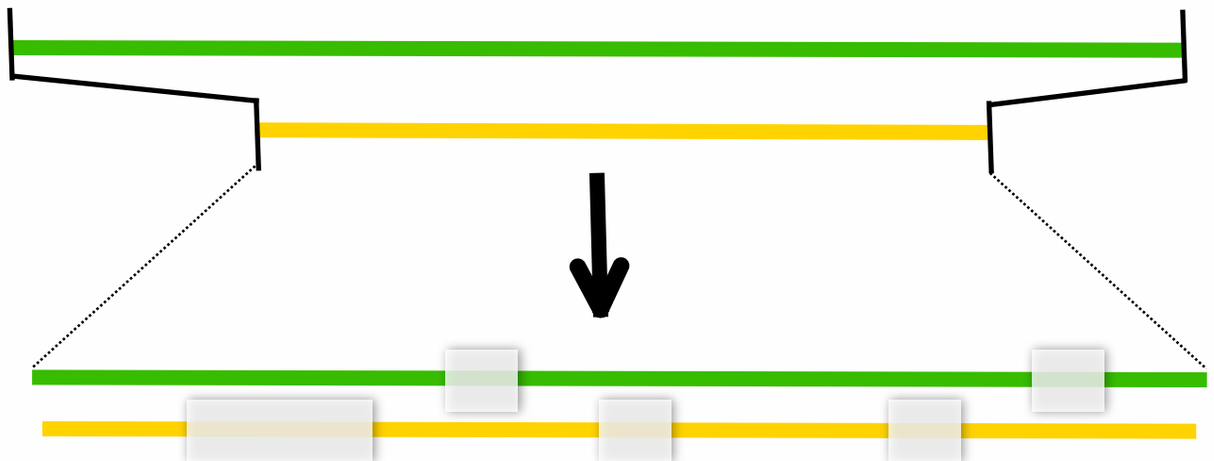
Backtracking to get the best alignment

```
Sequences     Best Alignment
---------     ----------------------
ACACACTA       A-CACACTA
AGCACACA       AGCACAC-A
```

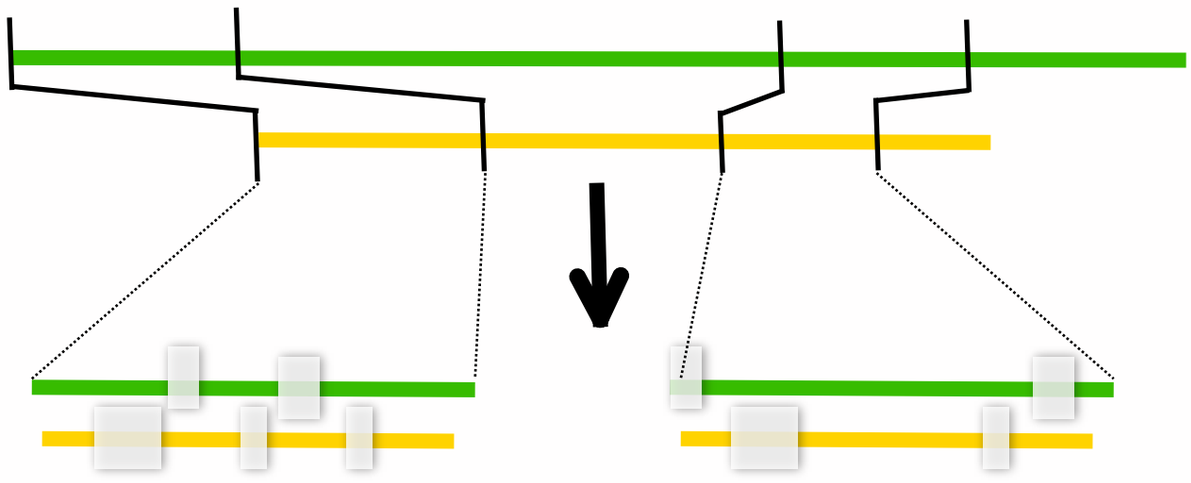
Smith and Waterman (1981) J. Mol Biol, 147 pp195

Global vs local alignment

Optimizer



Global alignment



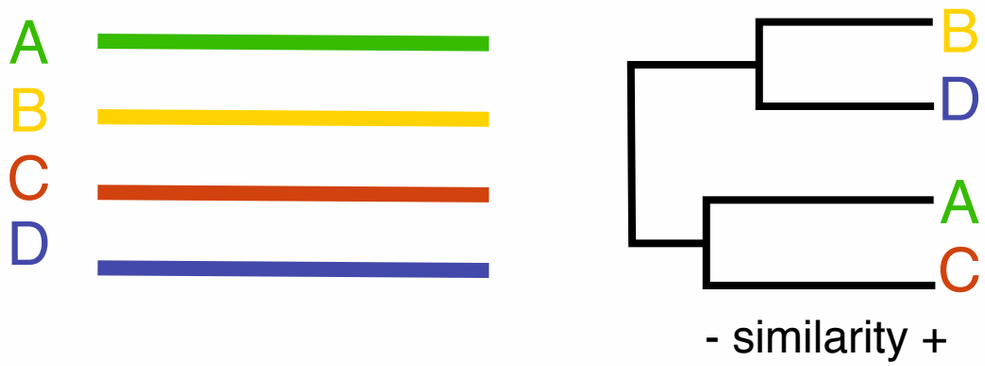
Local alignment

Multiple alignment

Optimizer

Pairwise alignments

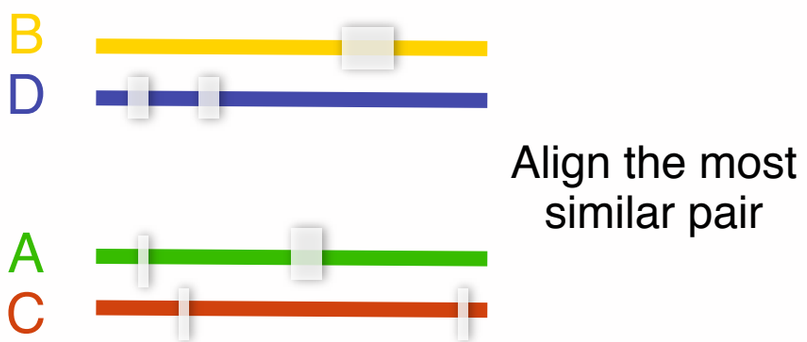
Example: 4 sequences A, B, C, D

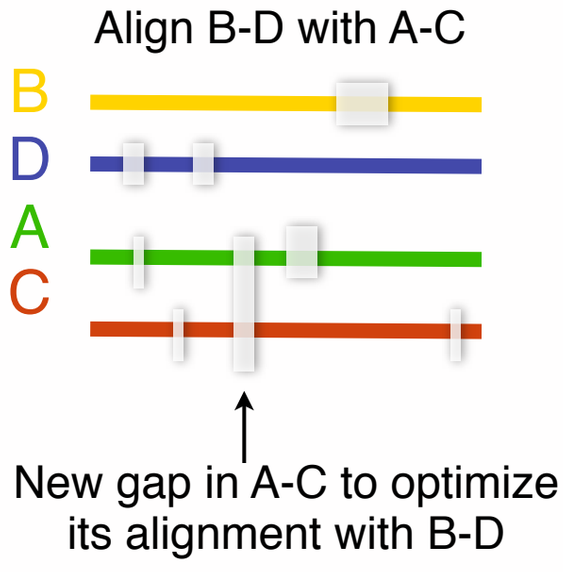


6 pairwise comparisons
then cluster analysis

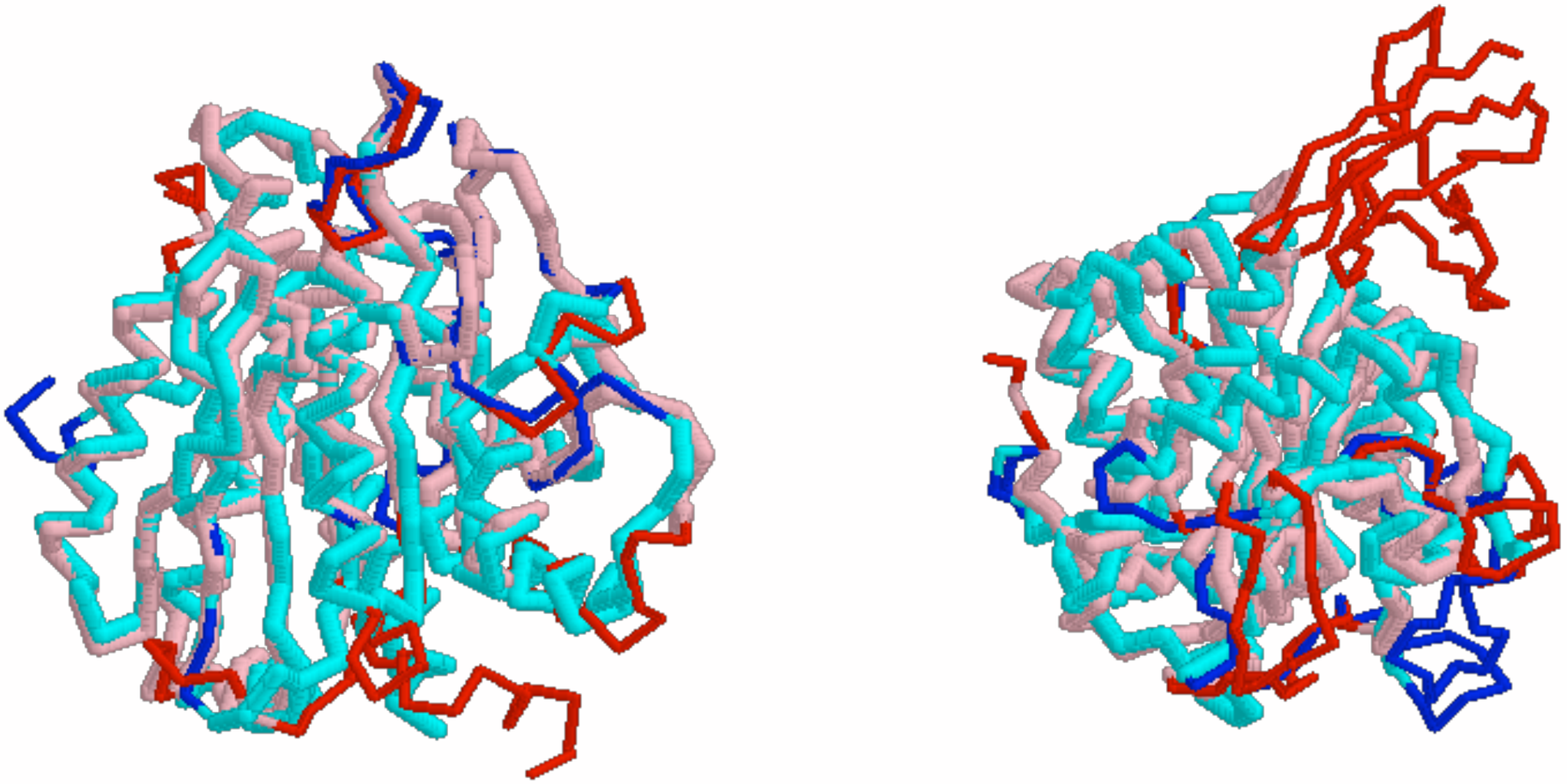
Multiple alignments

Following the tree from step 1





Coverage vs Accuracy

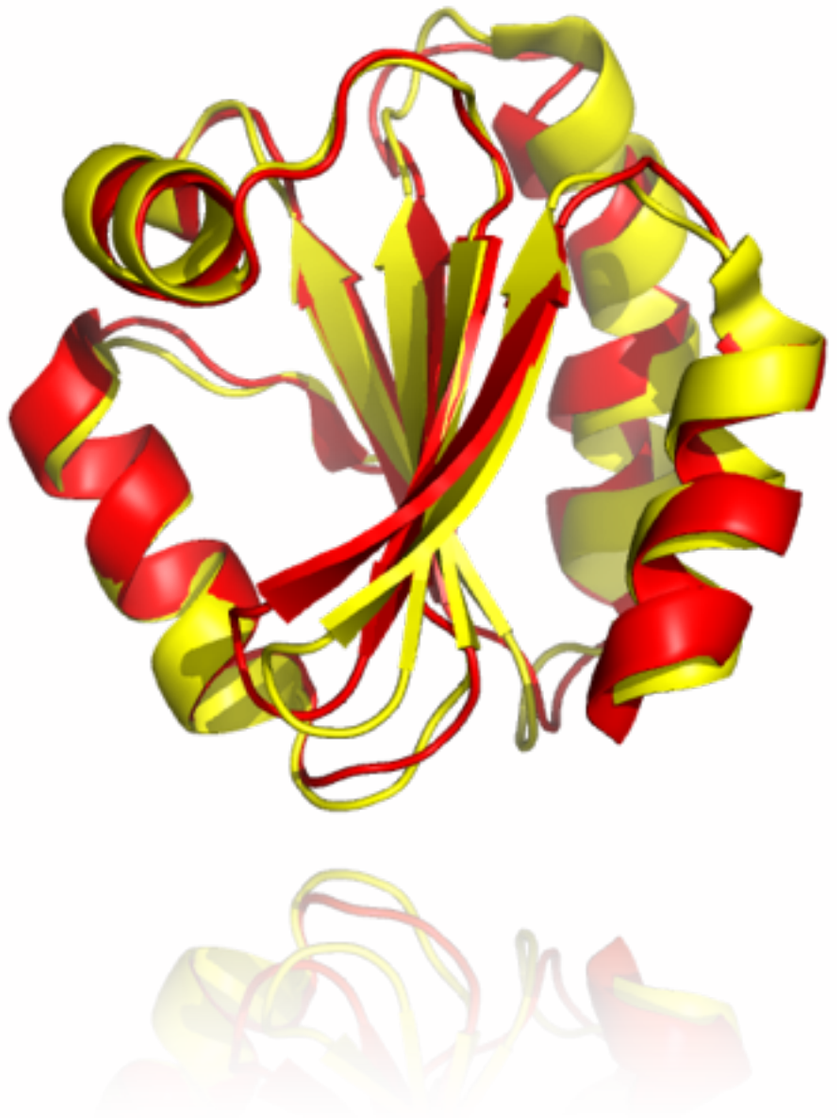


Same RMSD ~ 2.5Å

Coverage ~90% Cα

Coverage ~75% Cα

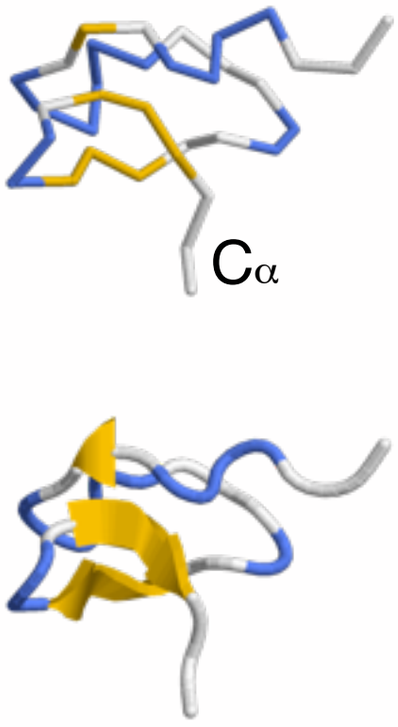
Structural alignment

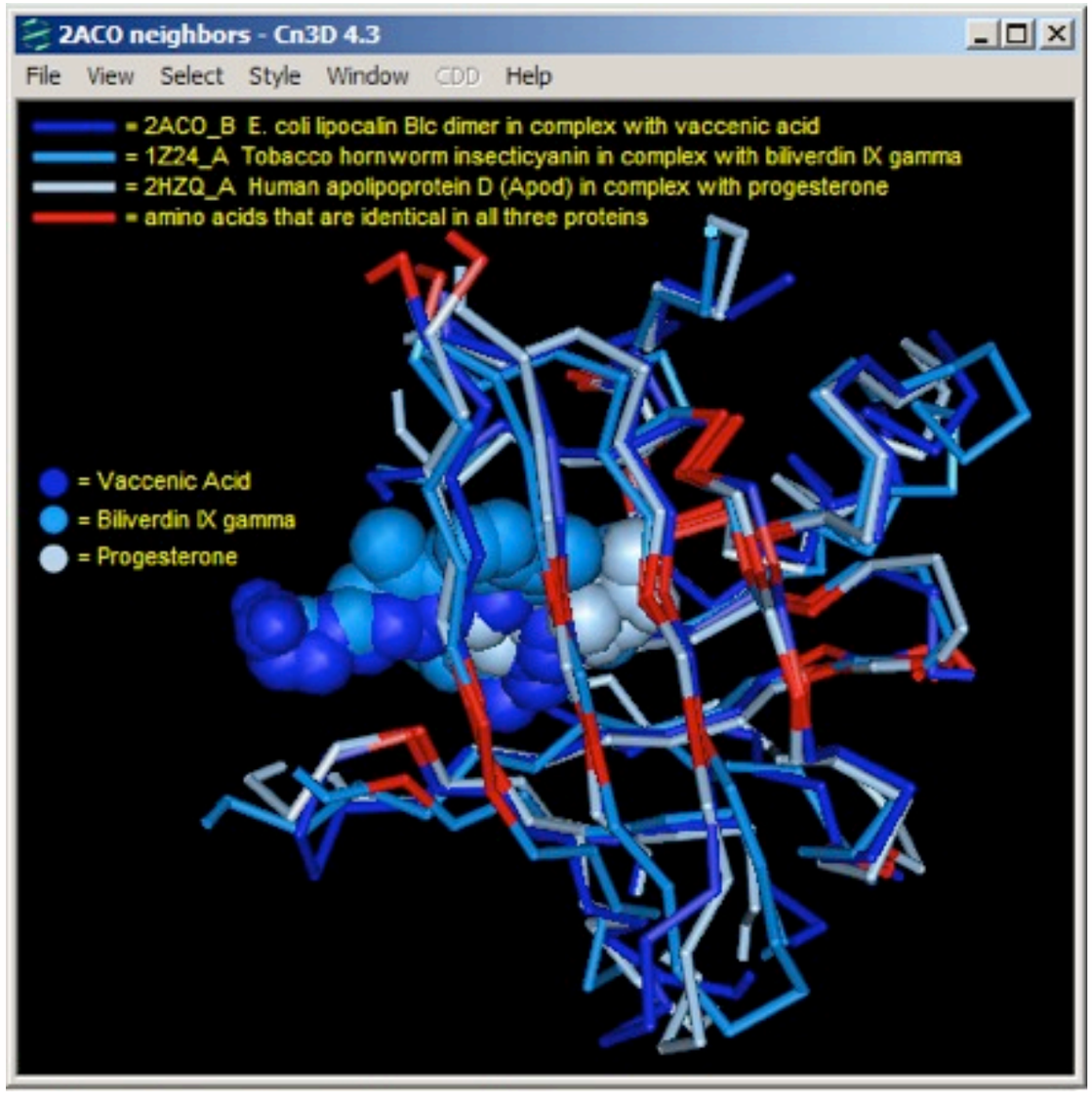


Vector Alignment Search Tool (VAST)

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

Vectorial representation of secondary structure elements





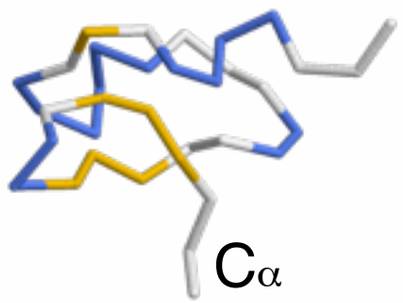
Gibrat JF et al. (1996) *Curr Opin Struct Biol* **3** pp377

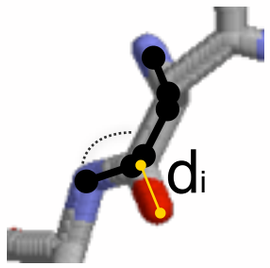
cnag CRG

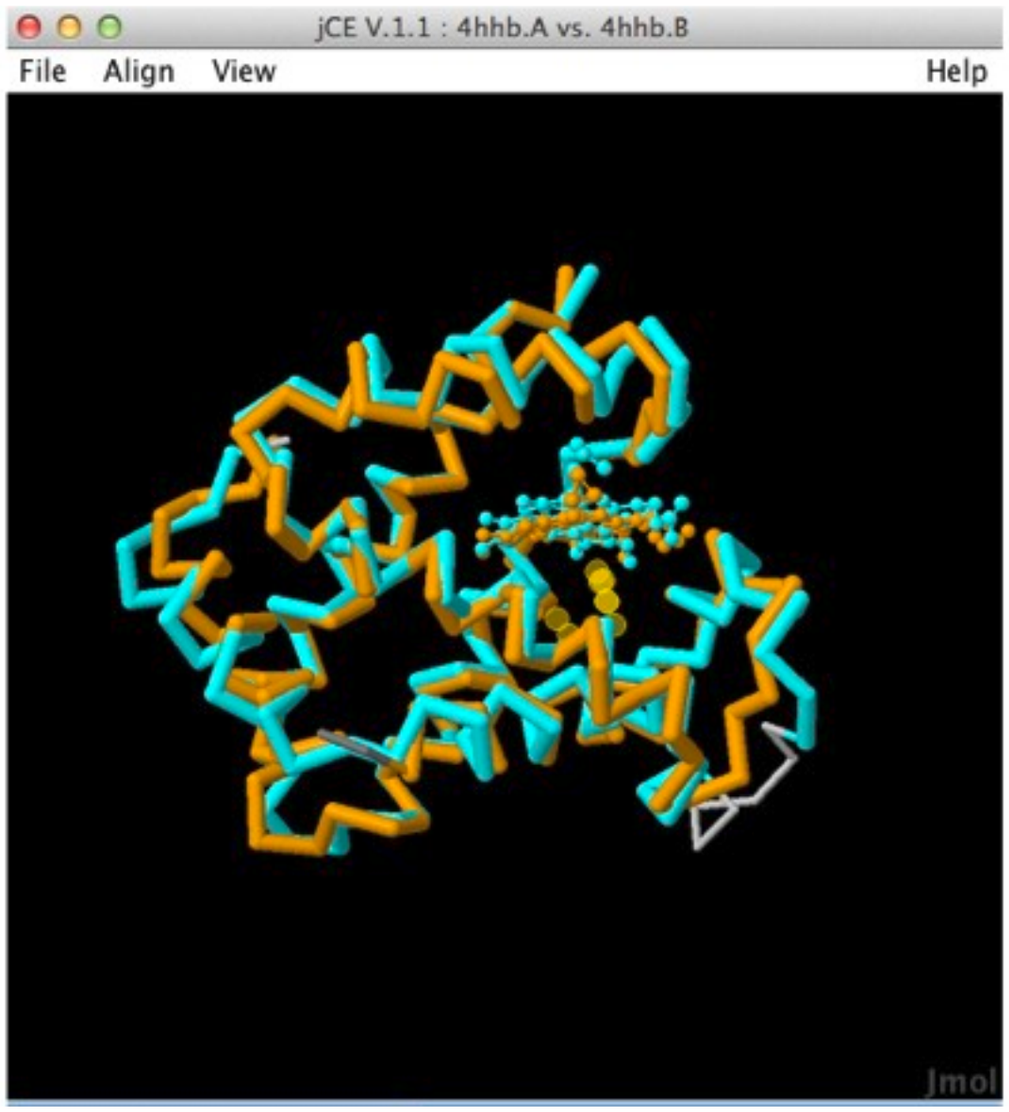
Incremental combinatorial extension (CE)

http://source.rcsb.org/jfatcatserver/ceHome.jsp

8 residues peptides





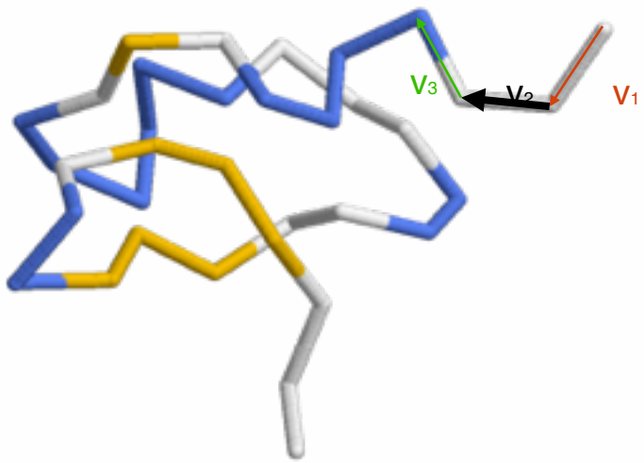


Shindyalov IN, amd Bourne PE. (1998) *Protein Eng.* **9** pp739

Matching molecular models obtained from theory (MAMMOTH)

http://ub.cbm.uam.es/software/mammoth.php

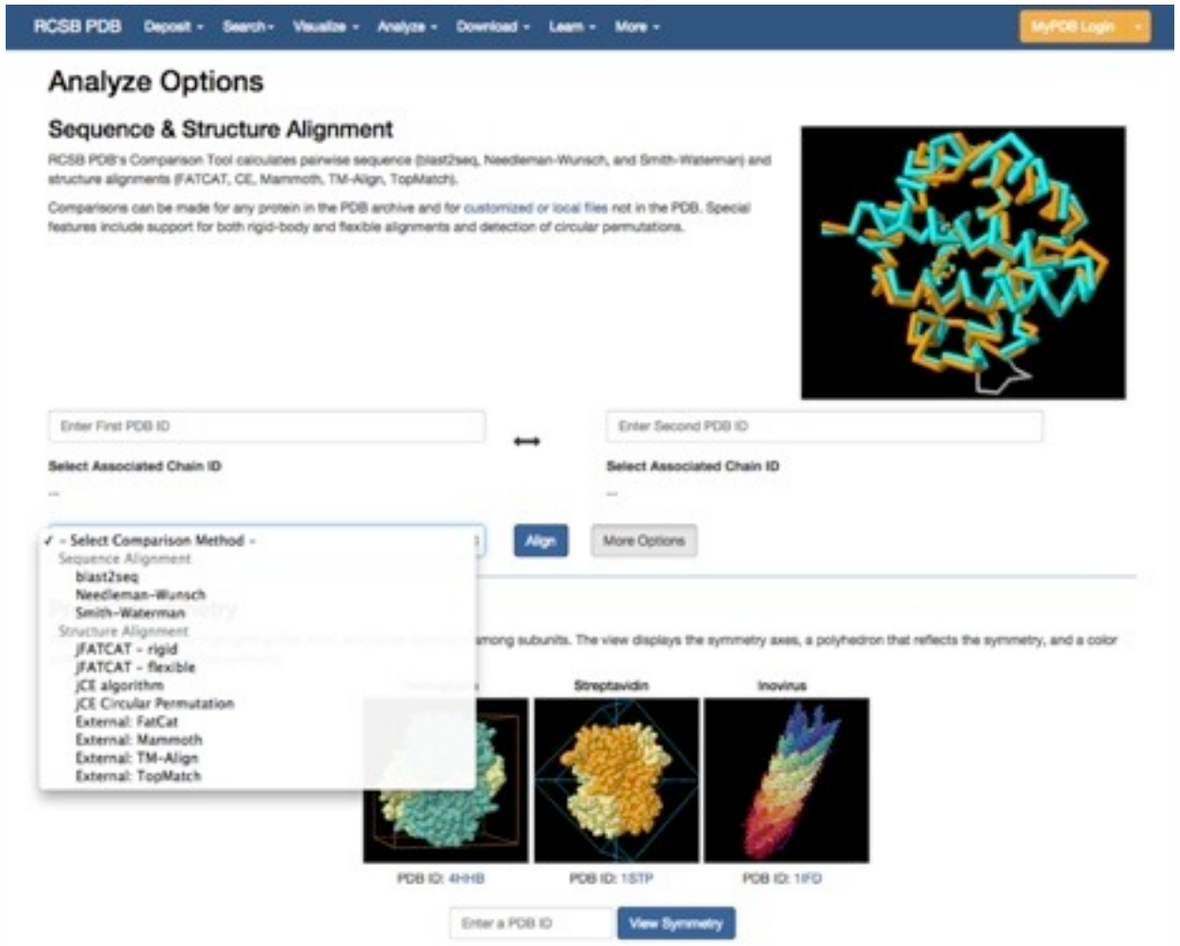
7 residues peptides



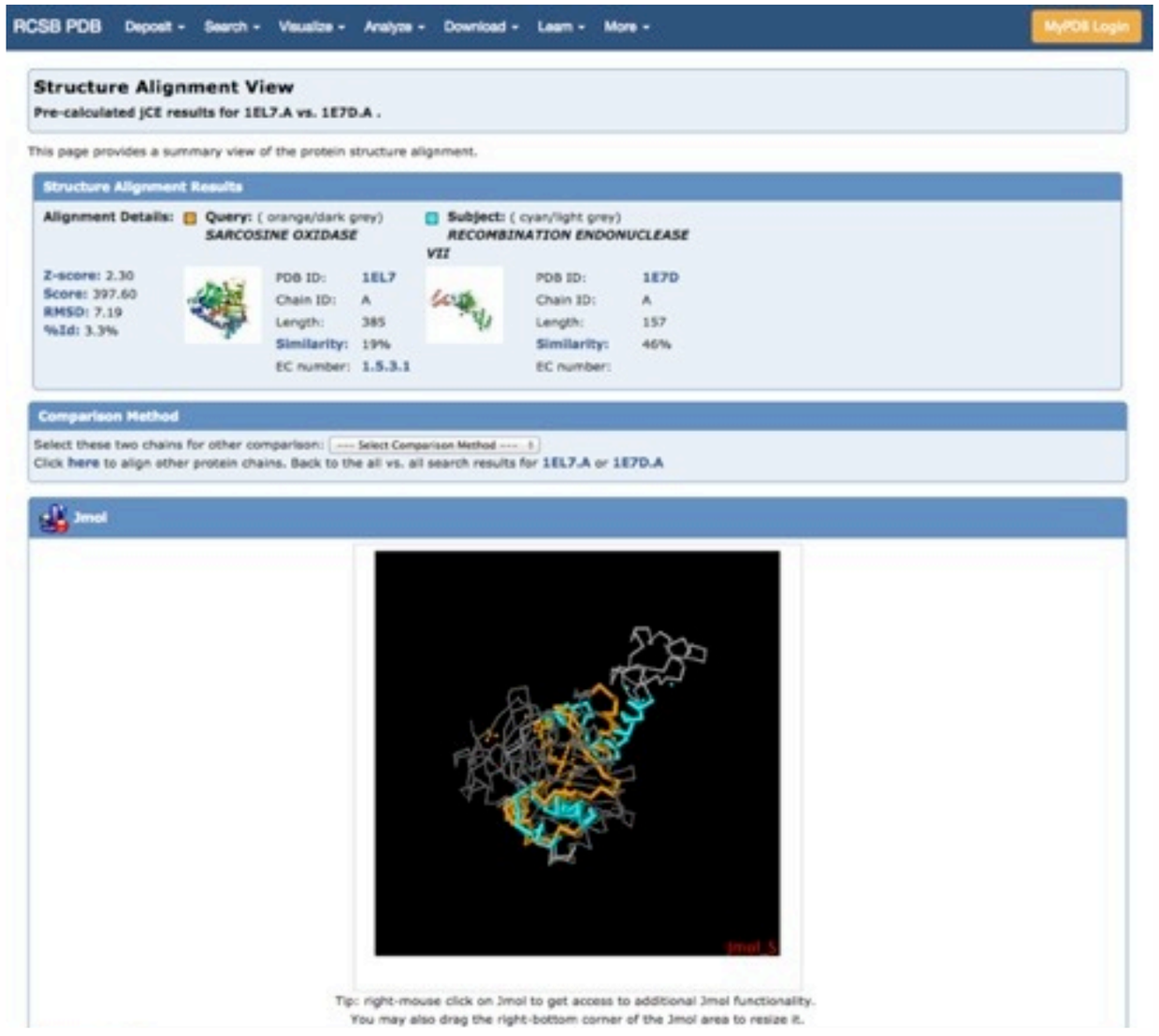
URMS instead of RMSD

Ortiz AR, (2002) *Protein Sci.* **11** pp2606

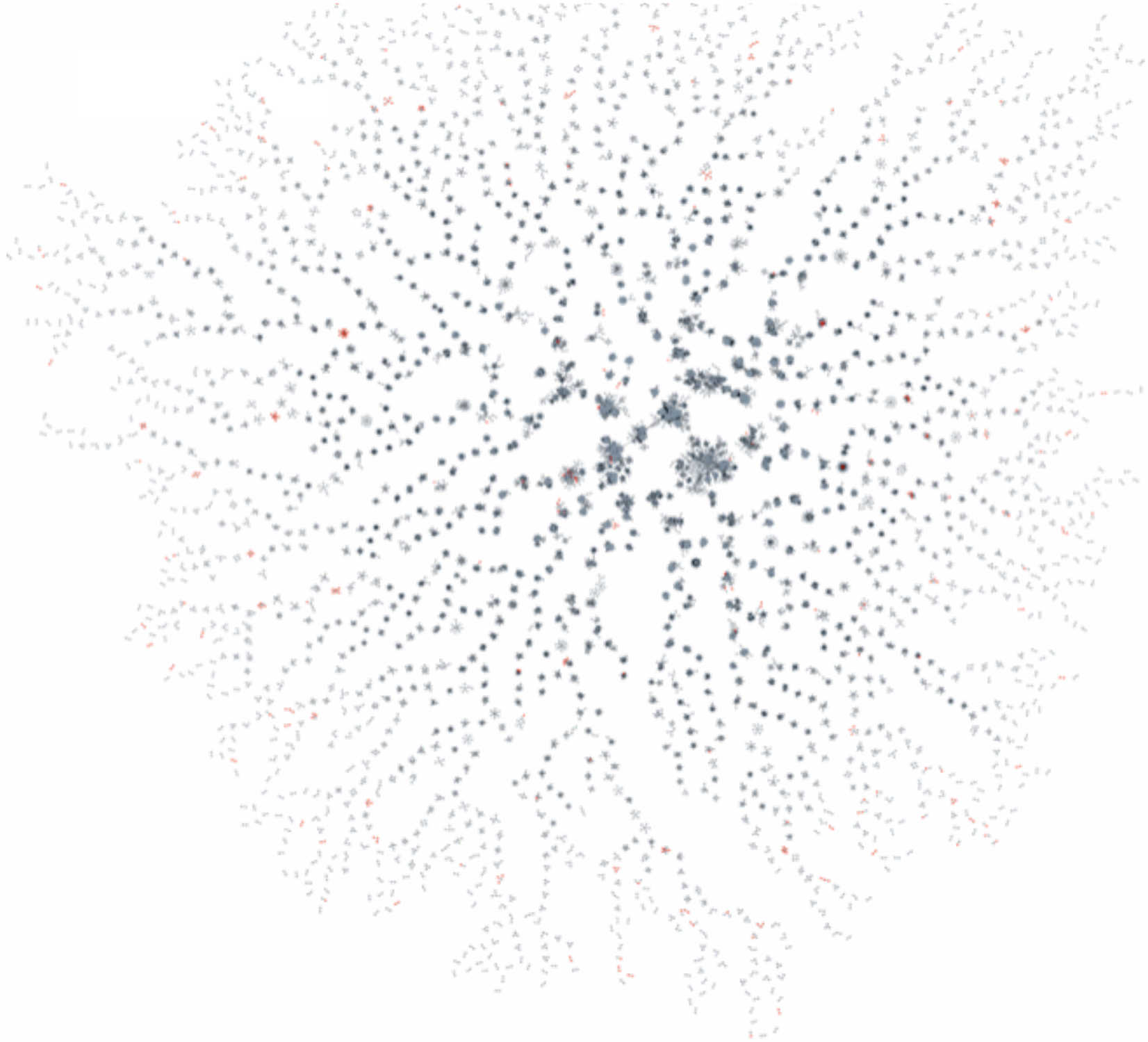
Structural alignment in the PDB



Structural alignment in the PDB

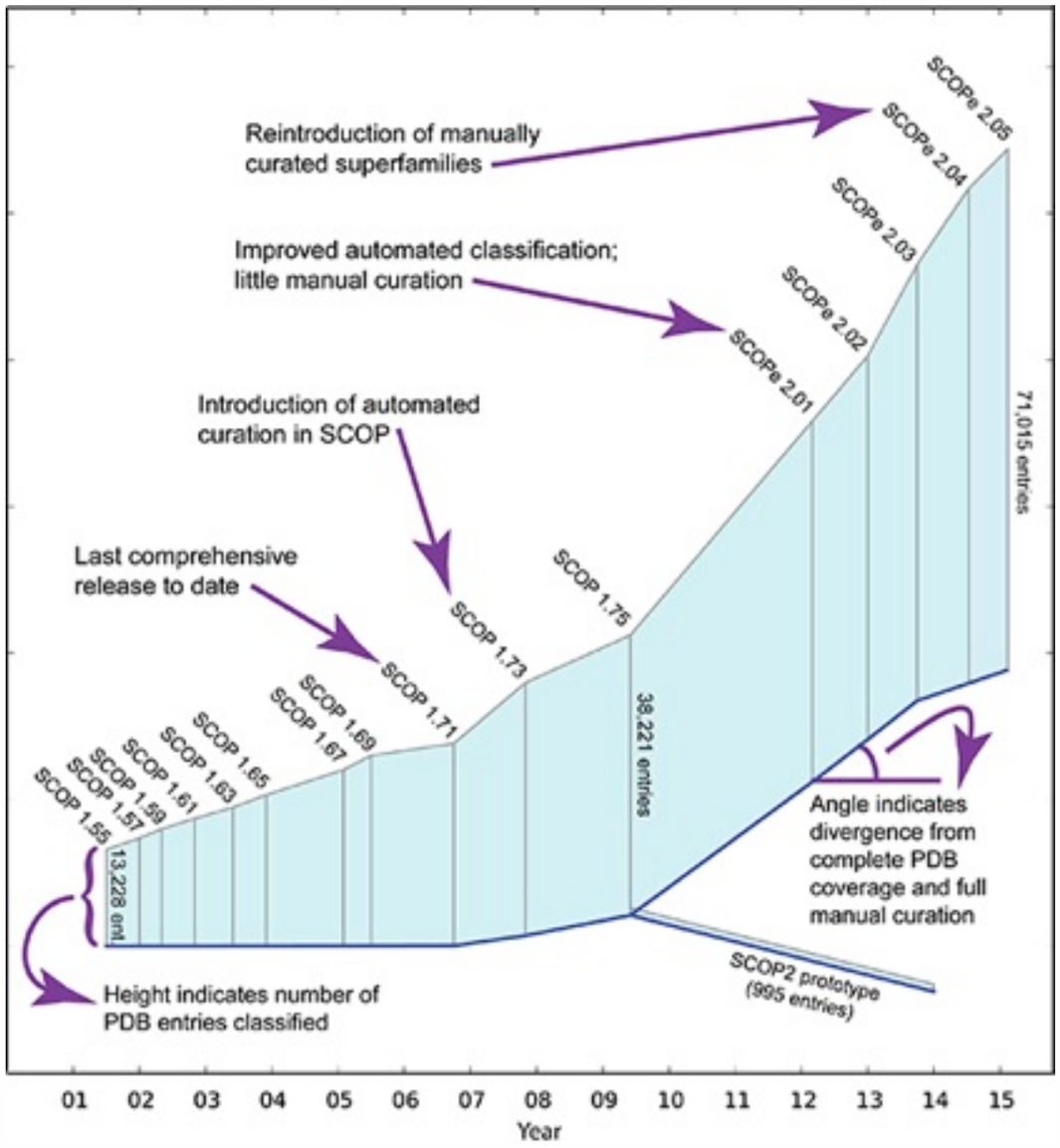


Classification of the structural space



SCOPe2.05 database

http://scop.berkeley.edu/statistics/ver=2.05



Class	Number of folds	Number of superfamilies	Number of families
All alpha proteins	286	509	1037
All beta proteins	176	359	931
Alpha and beta proteins (a/b)	148	245	965
Alpha and beta proteins (a+b)	381	558	1301
Multi-domain proteins	68	68	109
Membrane and cell surface proteins	57	113	153
Small proteins	92	132	260
Total	1208	1984	4756

Clear classification of structures in:

- CLASS
- FOLD
- SUPER-FAMILY
- FAMILY

Murzin A. G.,el at. (1995). J. Mol. Biol. ***247****, 536-540.*

cnag CRG

SCOP2 database

http://scop2.mrc-lmb.cam.ac.uk/

SCOP

Class

Fold

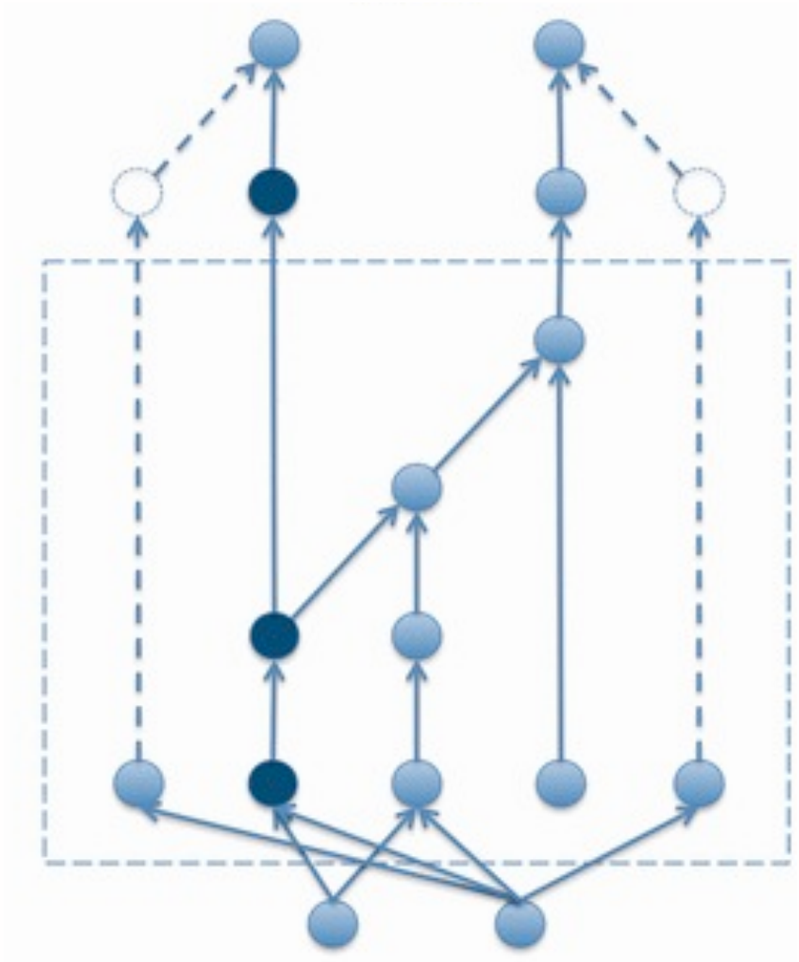
Super-Family

Family

Protein

Species

SCOP2

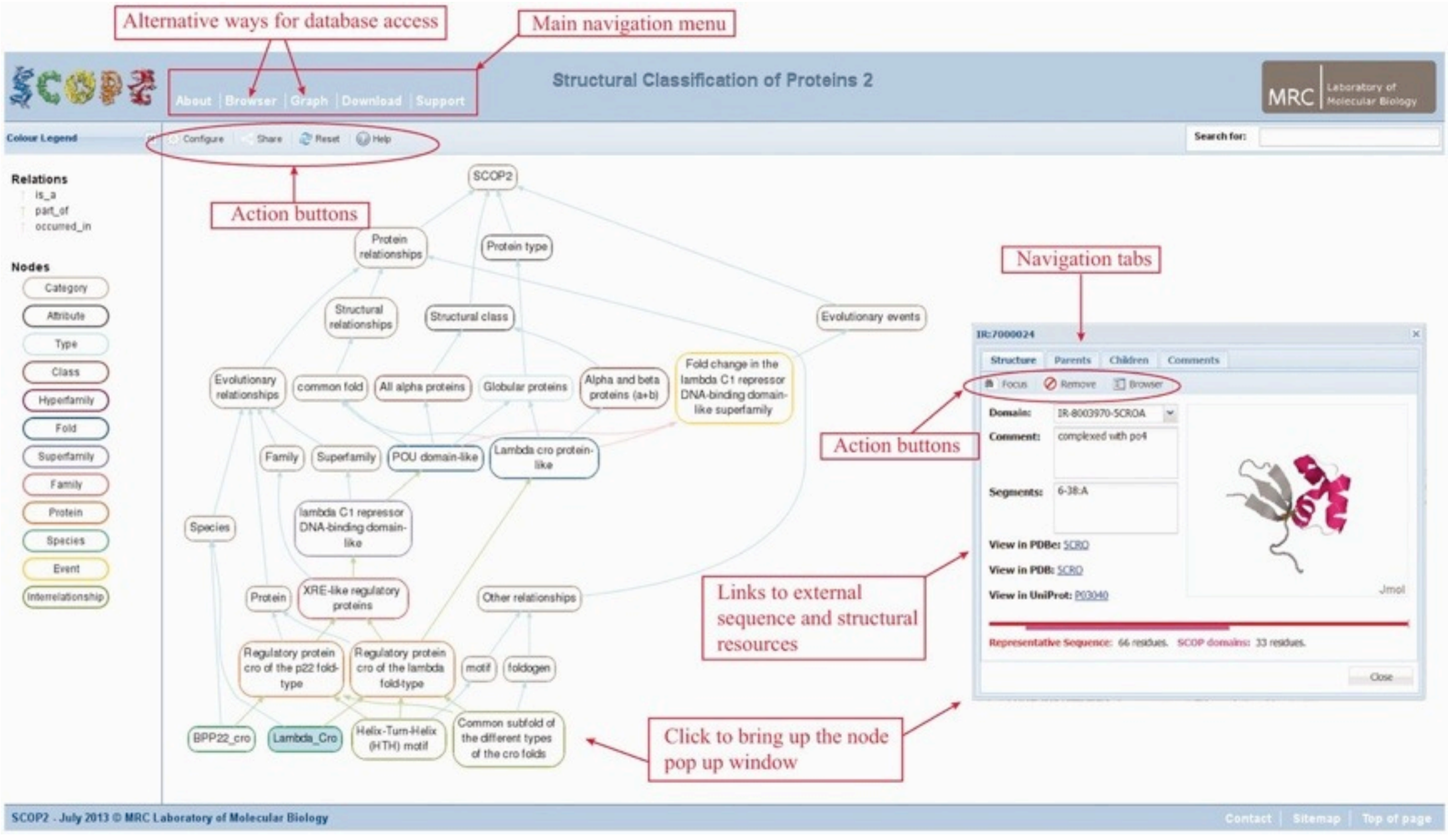


Nucl. Acids Res. (1 January 2014)42 (D1): D310-D314.

cnag CRG

SCOP2 database

http://scop2.mrc-lmb.cam.ac.uk/

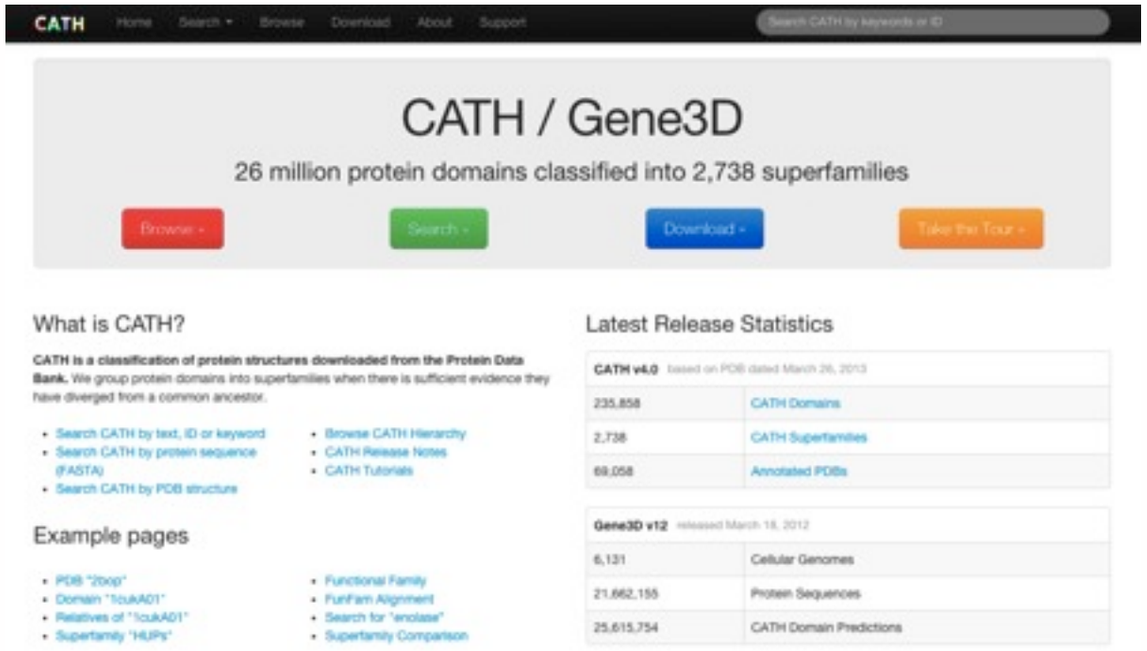


Nucl. Acids Res. (1 January 2014)42 (D1): D310-D314.

cnag CRG

CATH4 database

http://www.cathdb.info/



Class: SS composition (mostly alpha, mostly beta, mixed alpha/beta or few secondary structures)

Architecture: overall shape SSe orientations in 3D space but ignores the connectivity between them

Topology (fold family): fold groups depending on both the overall shape and connectivity of the SSe

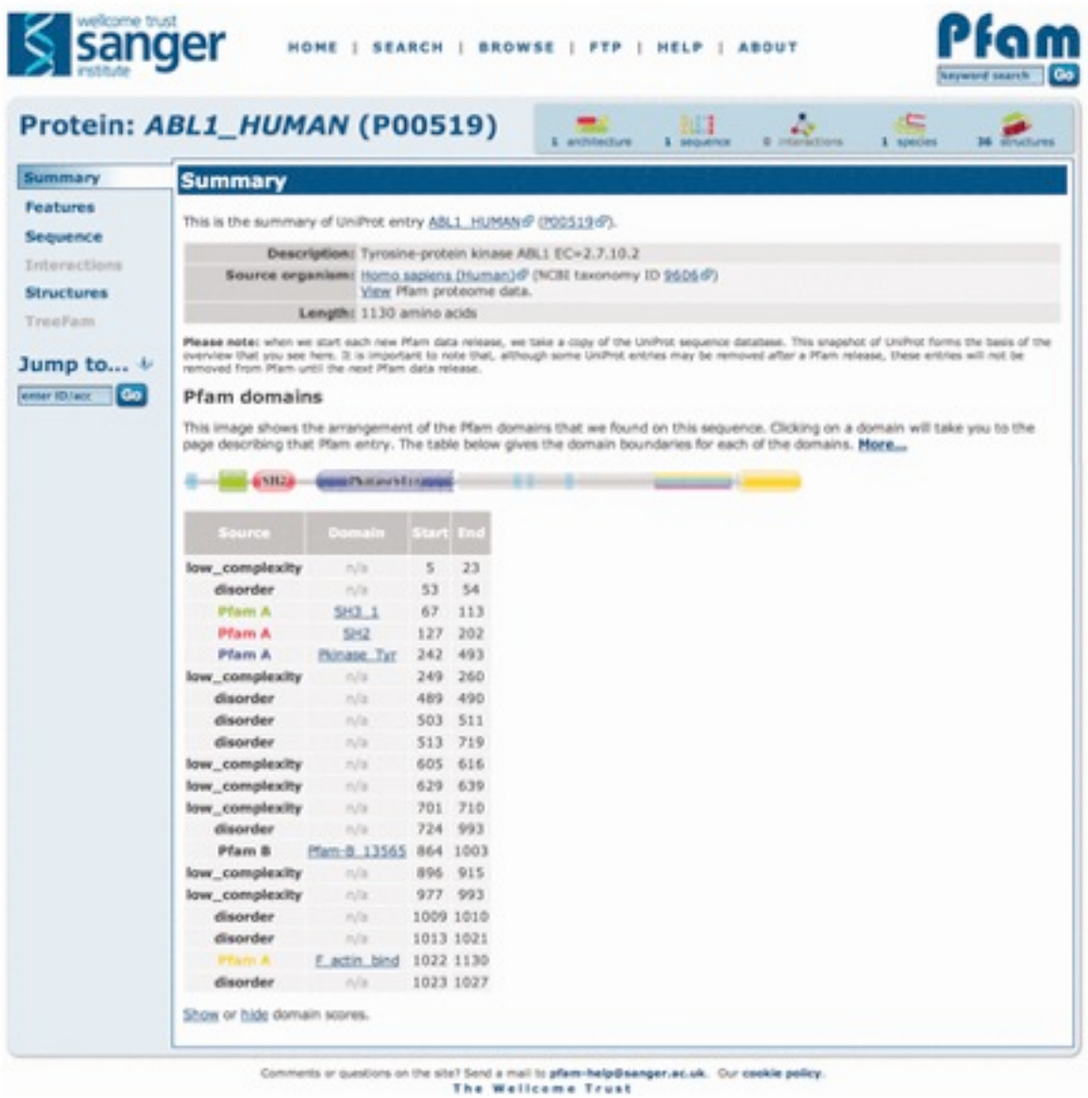
Homologous superfamily: groups together homologous protein domains

Class	Architecture	Topology	Homologous Superfamily	S35 Family	S60 Family	S95 Family	S100 Family	Domains
Class 1	5	397	907	3879	5118	6737	13368	48121
Class 2	20	241	547	3650	5221	8373	14526	58944
Class 3	14	626	1158	9171	13415	17047	35313	125772
Class 4	1	111	126	233	293	410	651	3021
TOTAL	**40**	**1375**	**2738**	**16933**	**24047**	**32567**	**63858**	**235858**

*Orengo, C.A., et al. (1997) Structure. **5**. 1093-1108.*

PFAM

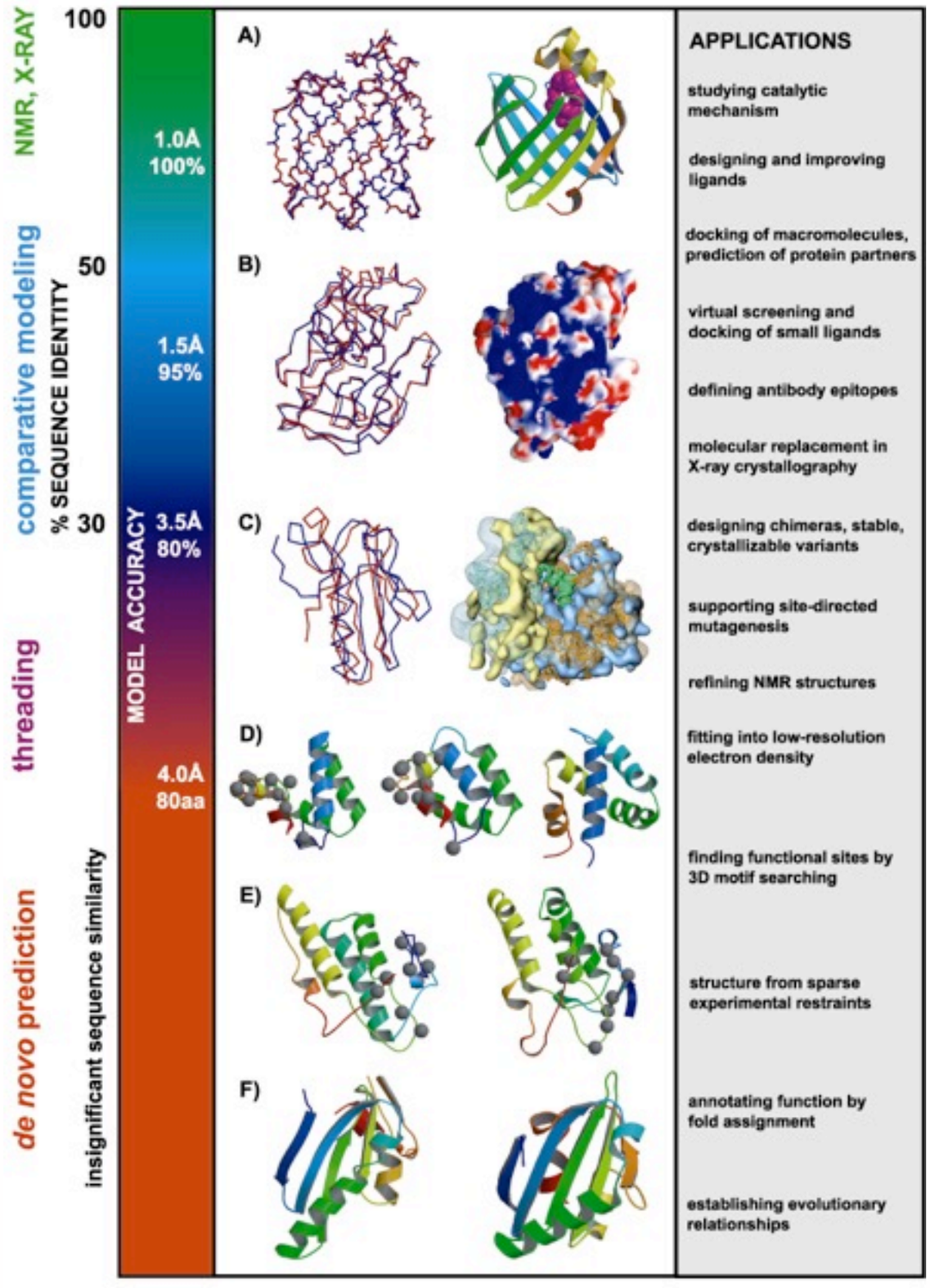
http://pfam.xfam.org/



Pfam domains:

Source	Domain	Start	End
disorder	n/a	1	306
low_complexity	n/a	2	43
low_complexity	n/a	51	89
low_complexity	n/a	117	134
low_complexity	n/a	205	224
low_complexity	n/a	252	307
coiled_coil	n/a	254	276
Pfam A	zf-H2C2_2	386	411
Pfam A	zf-C2H2	430	453
low_complexity	n/a	457	469
disorder	n/a	477	496
low_complexity	n/a	489	500
disorder	n/a	547	549
disorder	n/a	559	562
disorder	n/a	564	570

Utility of protein structure models, despite errors



D. Baker & A. Sali. Science 294, 93, 2001

Take home message

Protein and Nucleic acids structures are stored in publicly available databases

Proteins are aligned by sequence and by structure

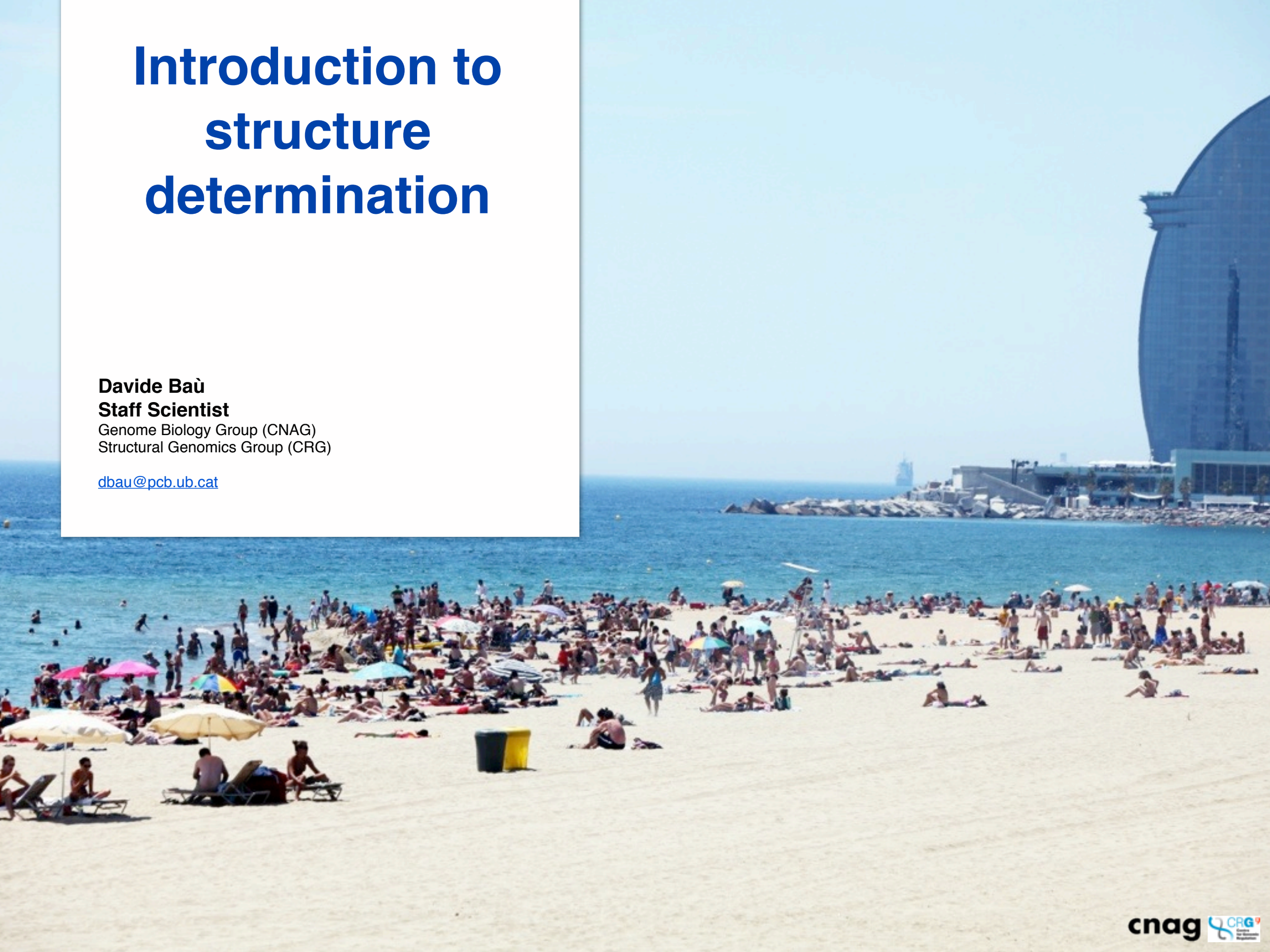
Structural alignments might identify distant homologs that cannot be recognized by sequence comparison

There are several databases that classify protein structures

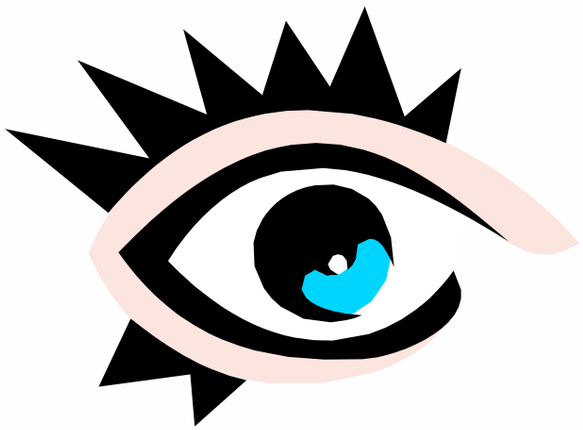
Introduction to structure determination

Davide Baù
Staff Scientist
Genome Biology Group (CNAG)
Structural Genomics Group (CRG)

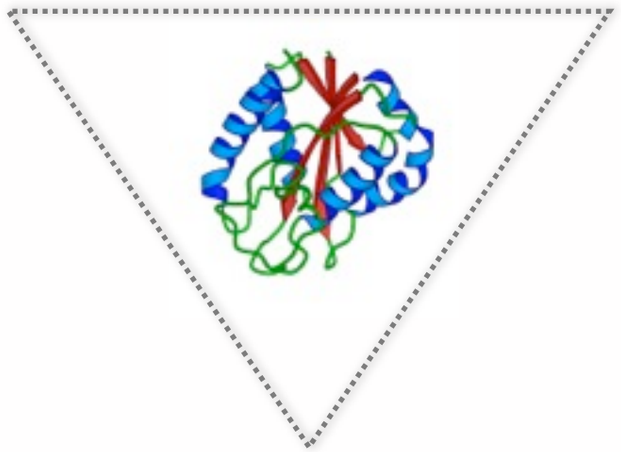
dbau@pcb.ub.cat

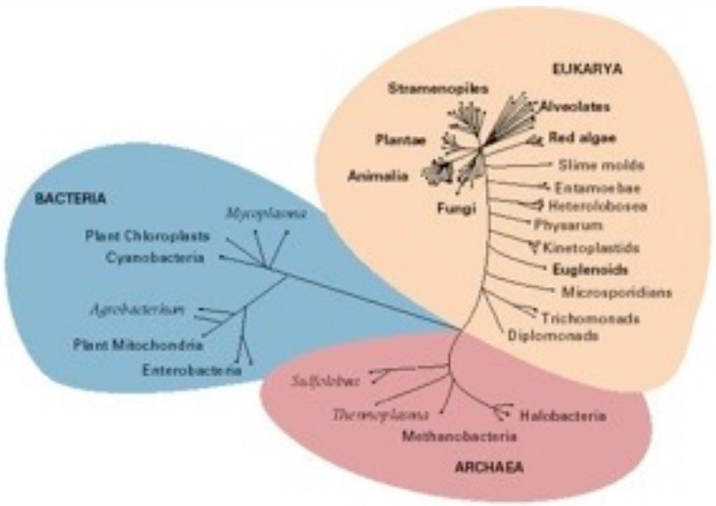


Data groups

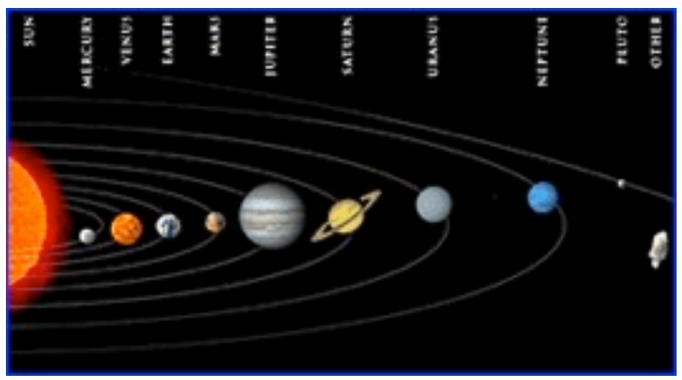


Experimental observations



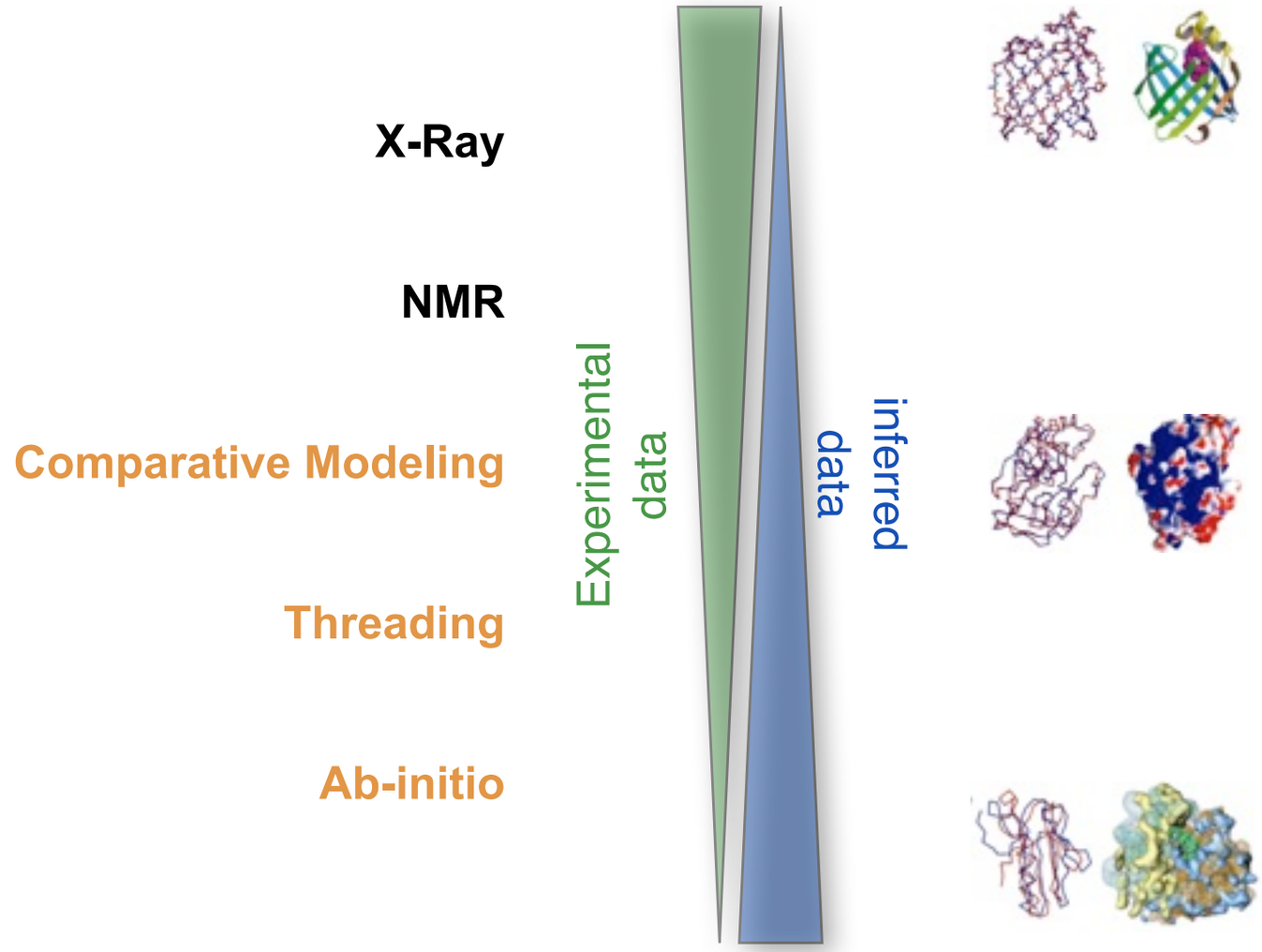


Statistical rules



Laws of physics

Structure prediction vs determination

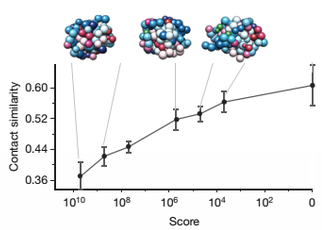


The four stages of integrative modeling

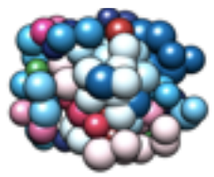
Data collection

Representation and scoring

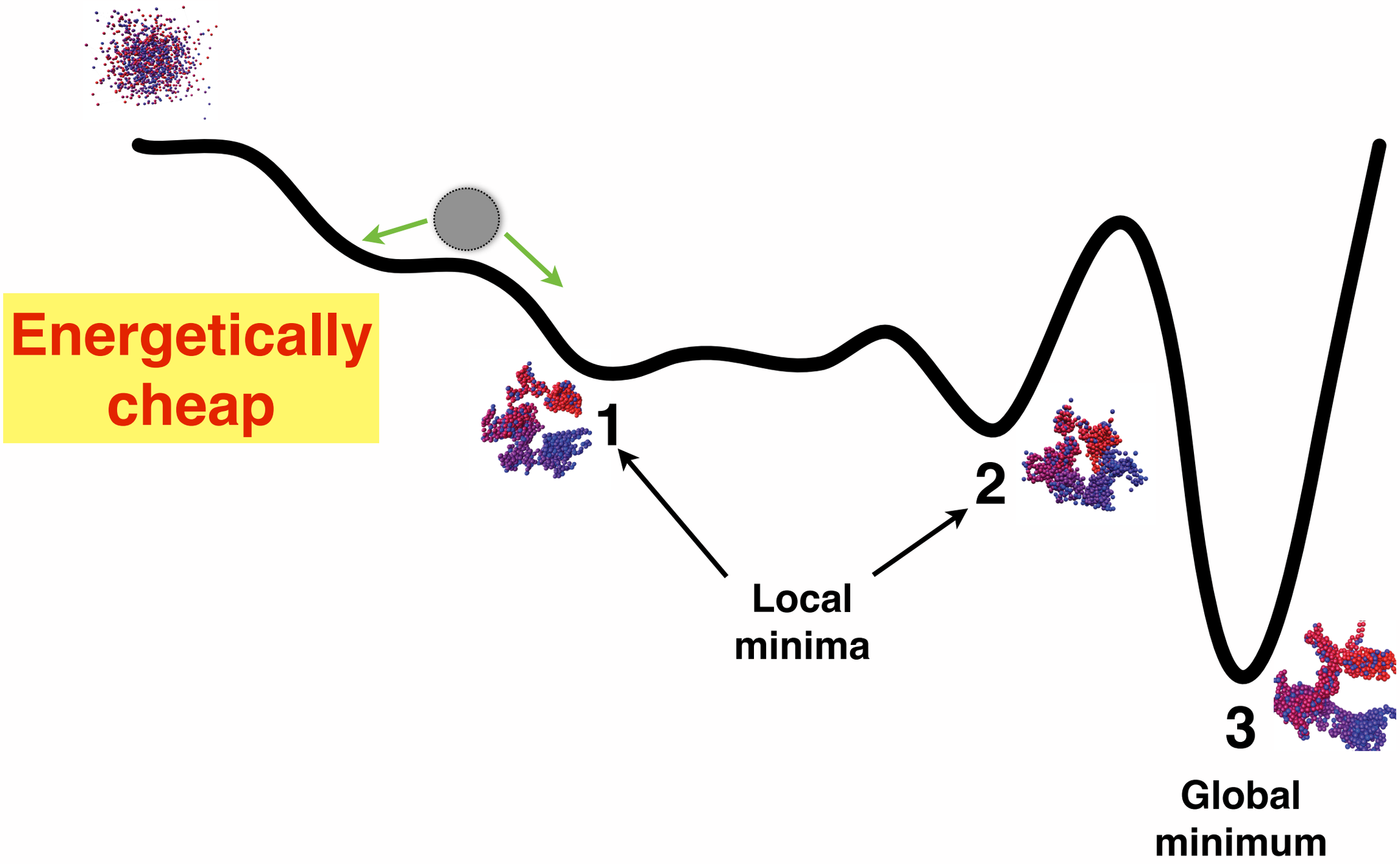
Optimization



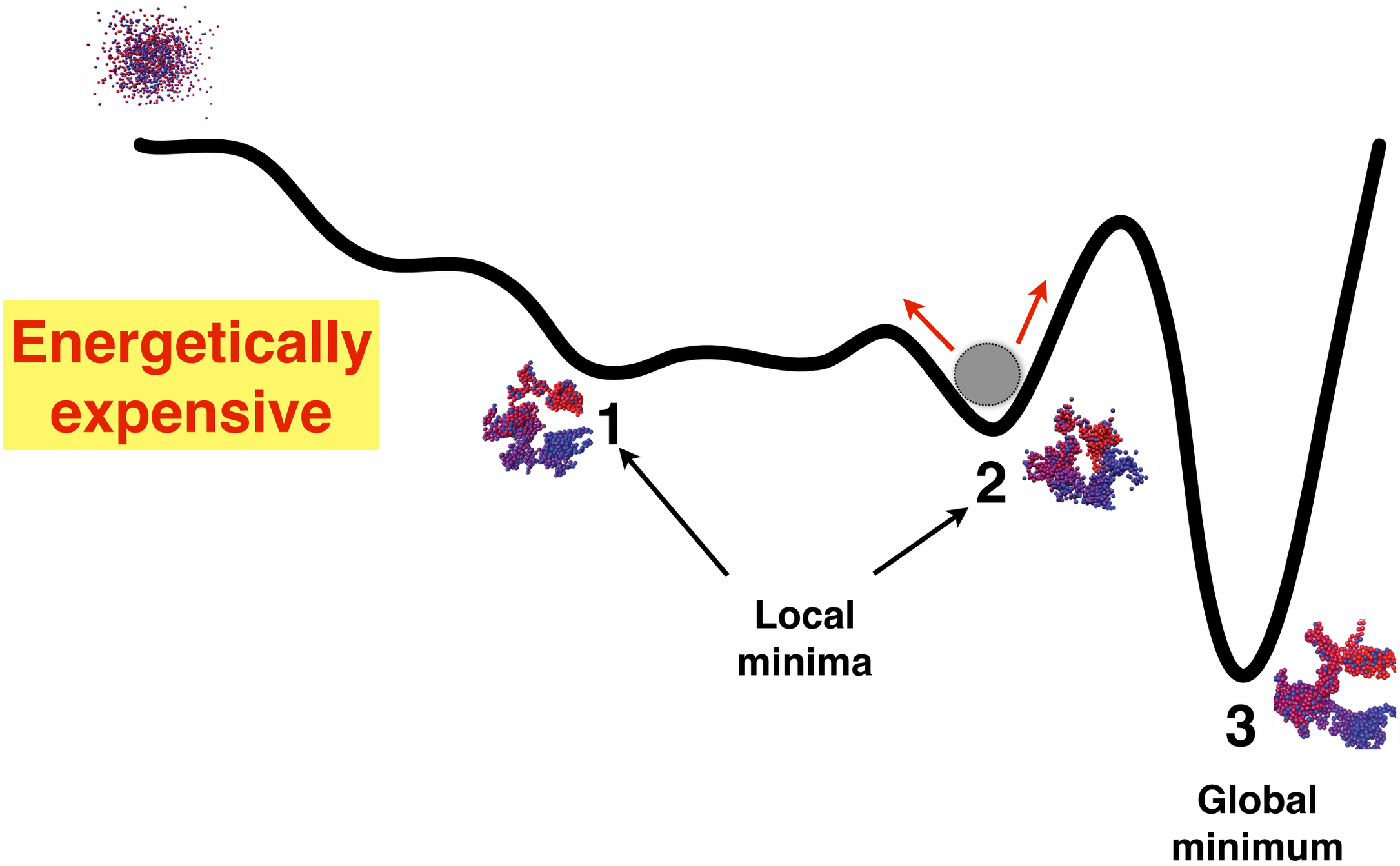
Analysis of the results



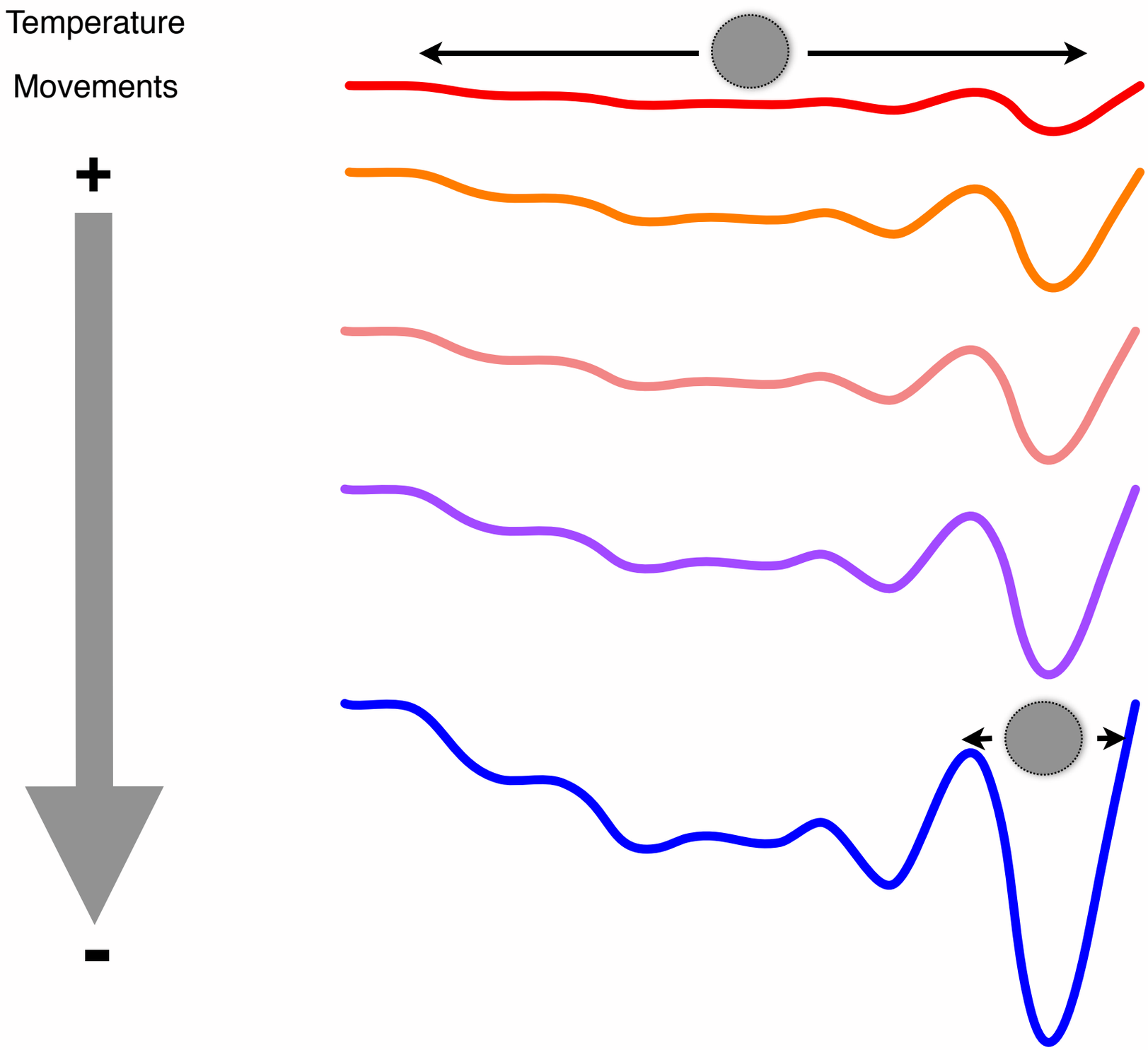
Energy landscape



Energy landscape



The simulating annealing procedure

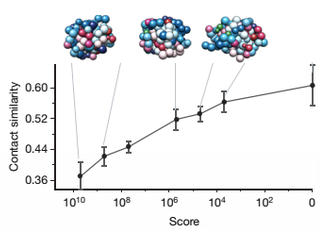


The four stages of integrative modeling

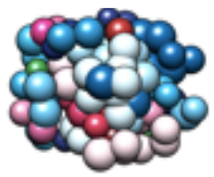
Data collection

Representation and scoring

Optimization



Analysis of the results

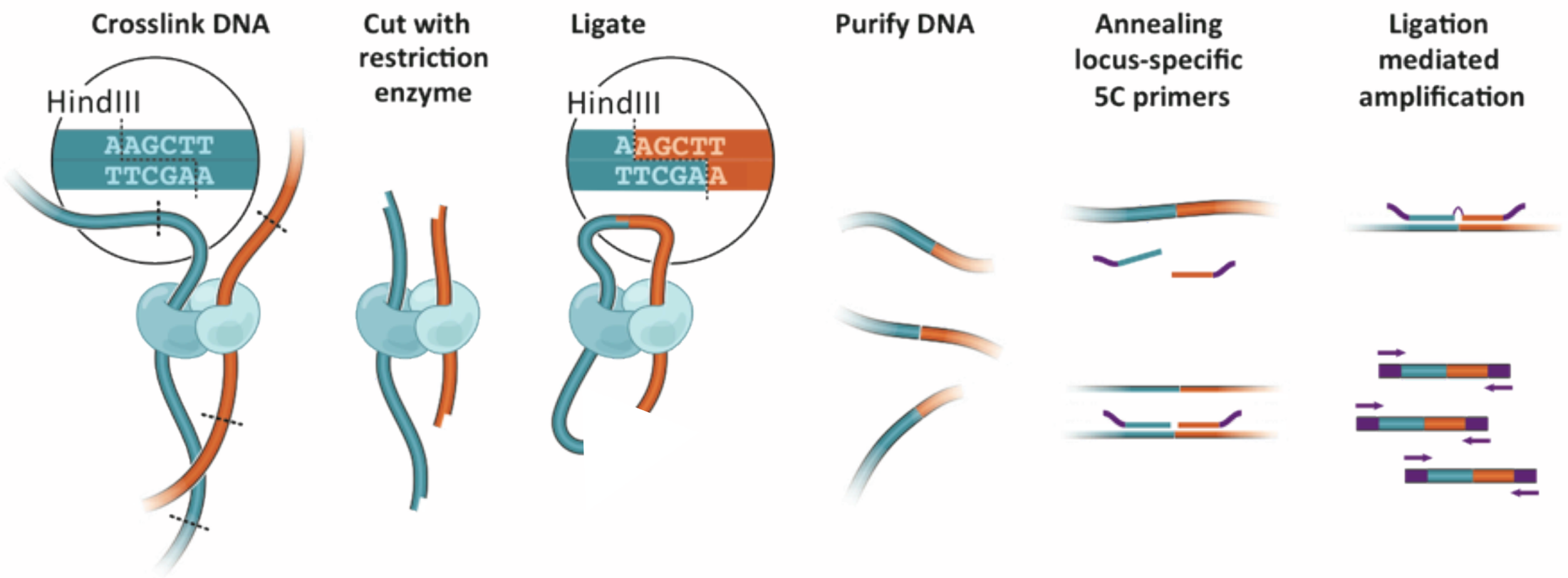


Job Dekker

5C technology

http://my5C.umassmed.edu

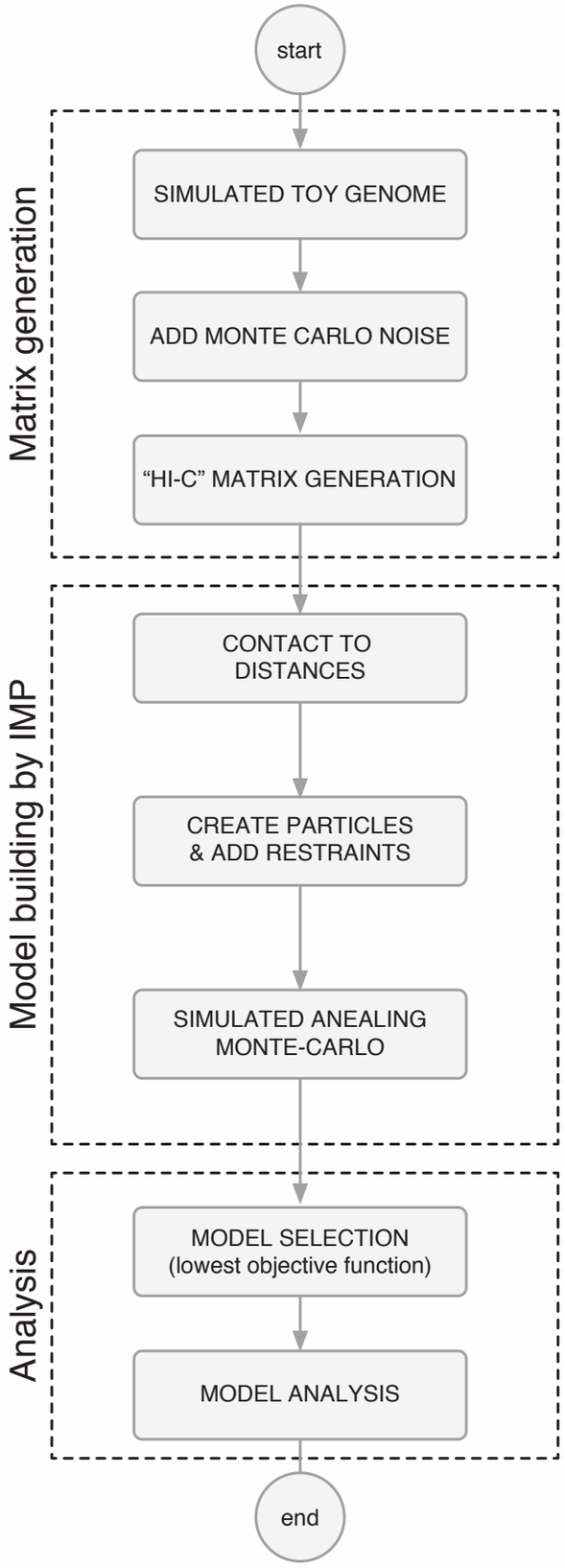
UMASS

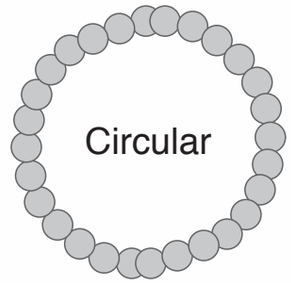


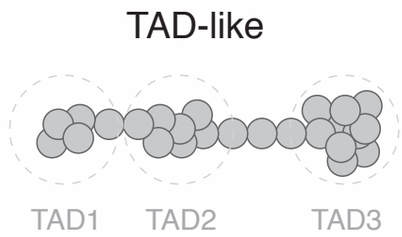
Dostie et al. Genome Res (2006) vol. 16 (10) pp. 1299-309

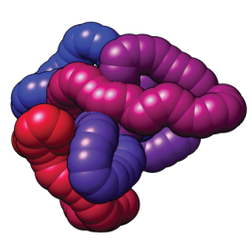
cnag CRG

Toy models

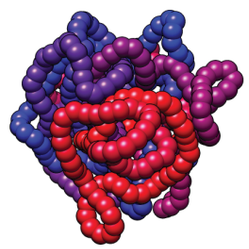




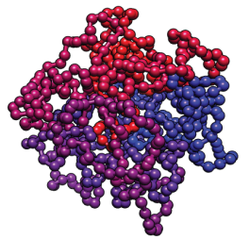




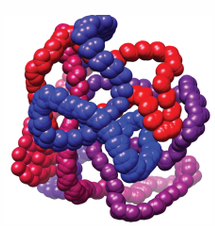
150 bp/nm



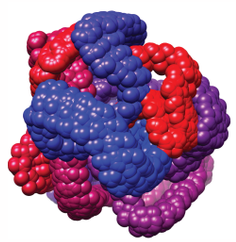
75 bp/nm



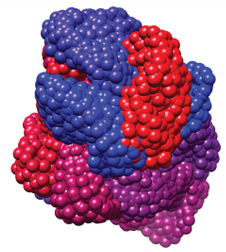
40 bp/nm



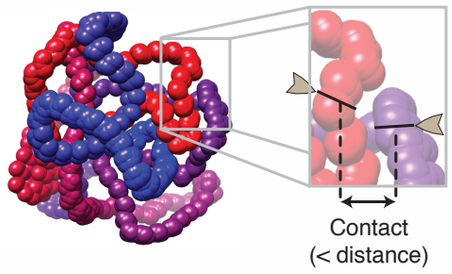
ts = 10^0

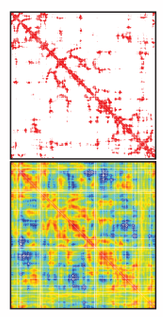


ts = 10^1



ts = 10^2



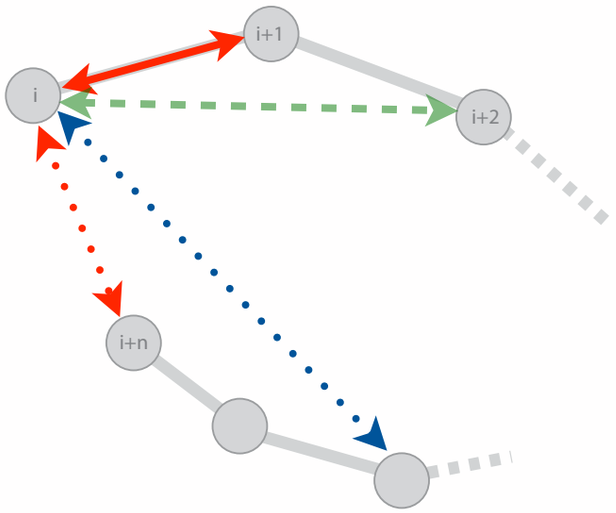


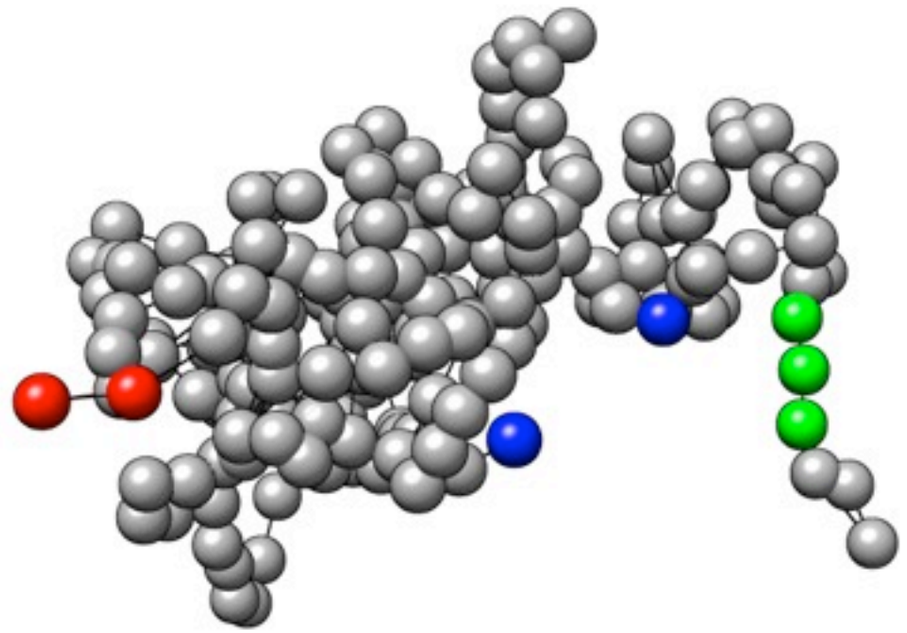
Contact Map

"Hi-C" matrix with noise

Model representation and scoring

Constituent parts of the molecule





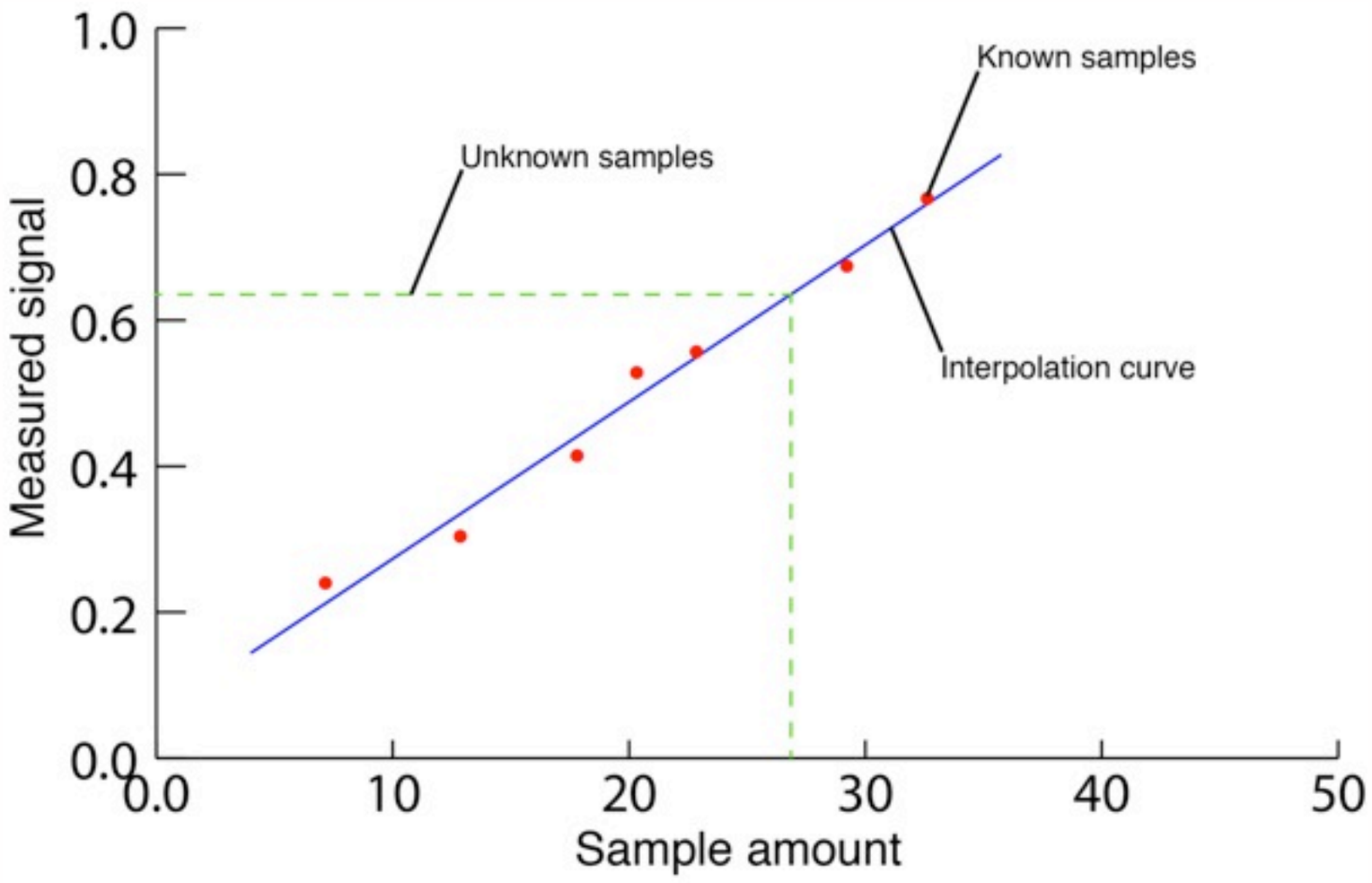
$d < d_0$

$d = d_0$

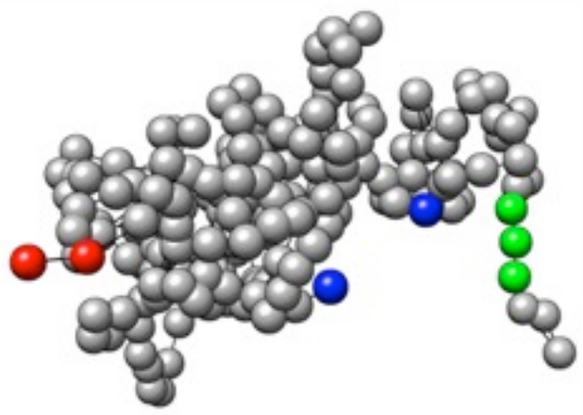
$d > d_0$

From 5C data to spatial distances

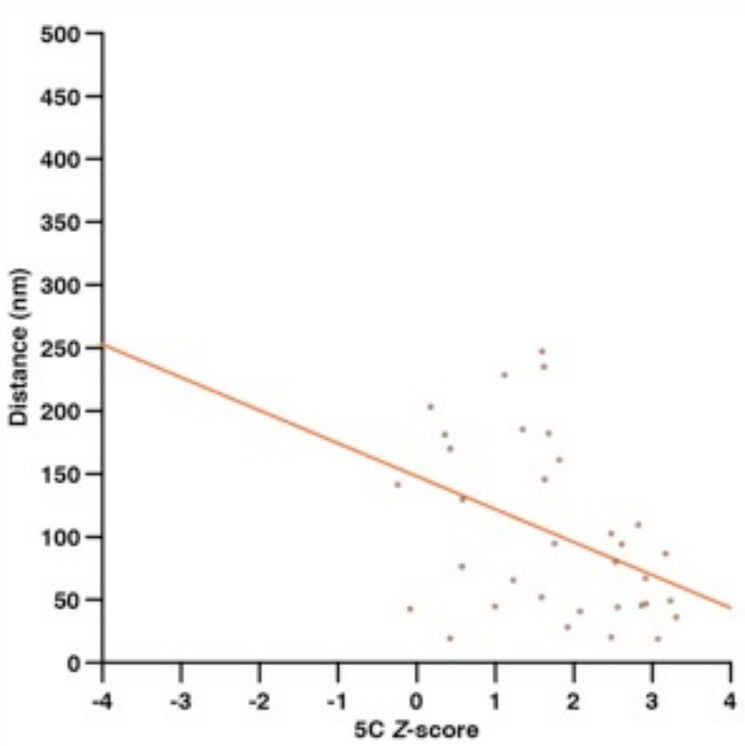
The sample curve



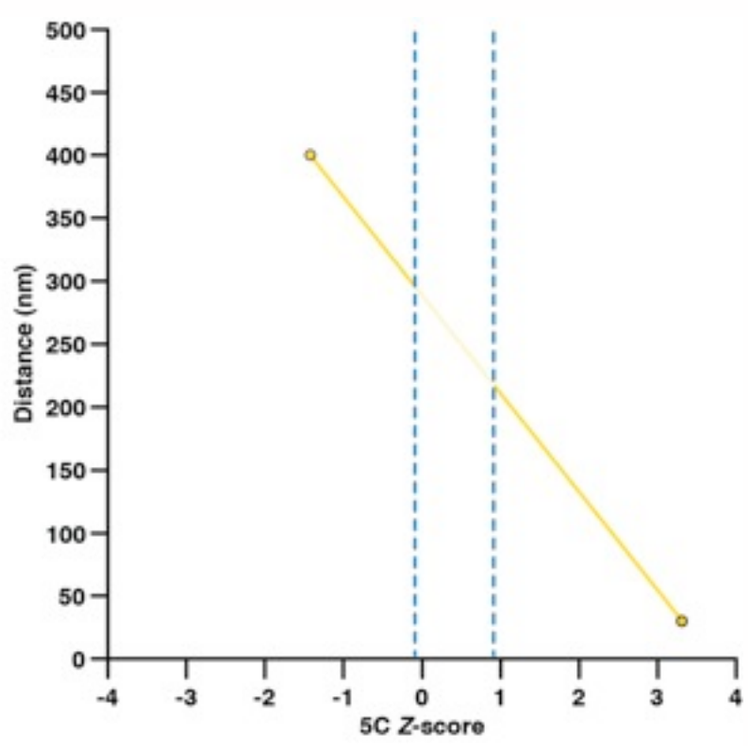
From 5C data to spatial distances



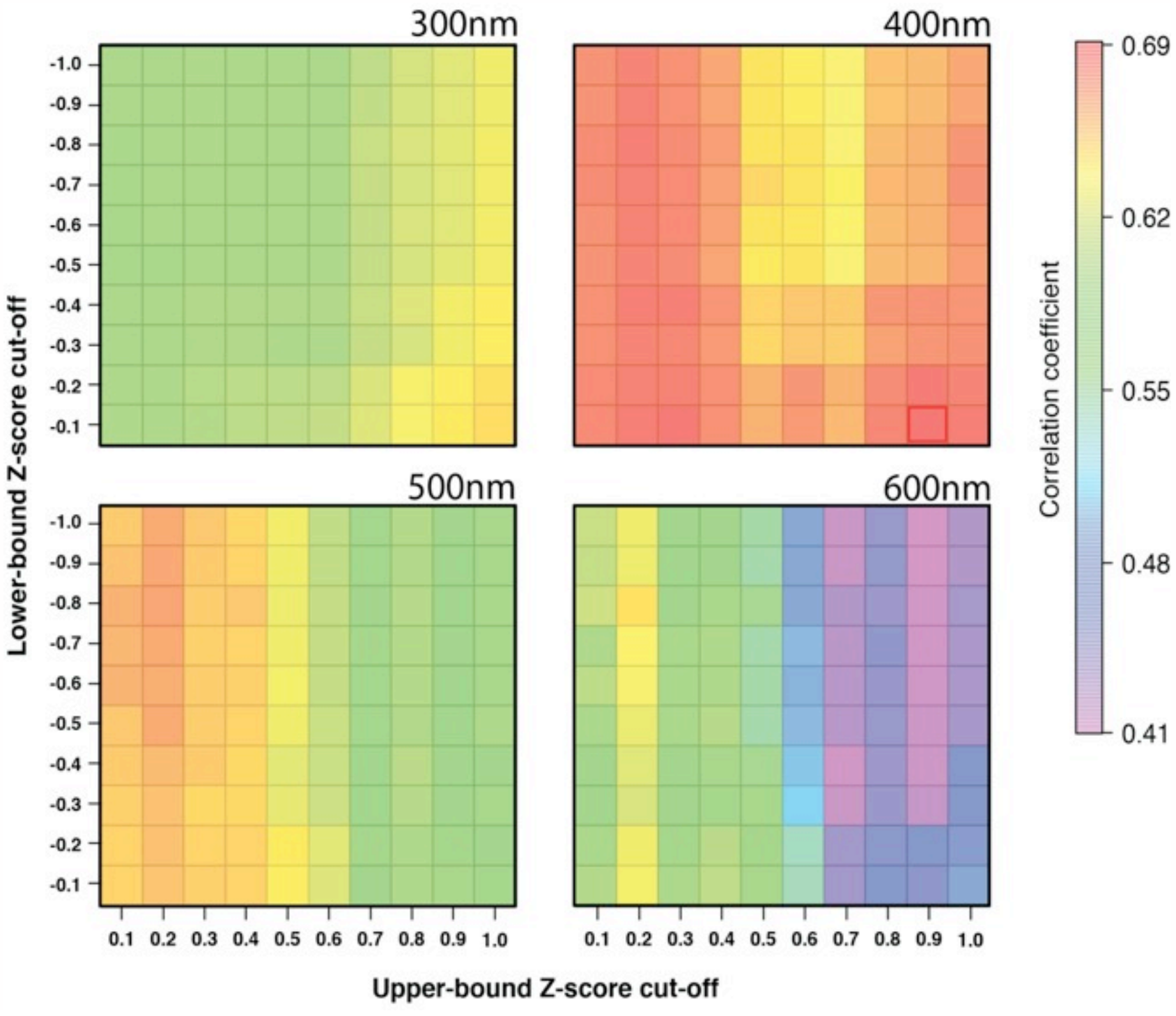
Neighbor fragments



Non-Neighbor fragments

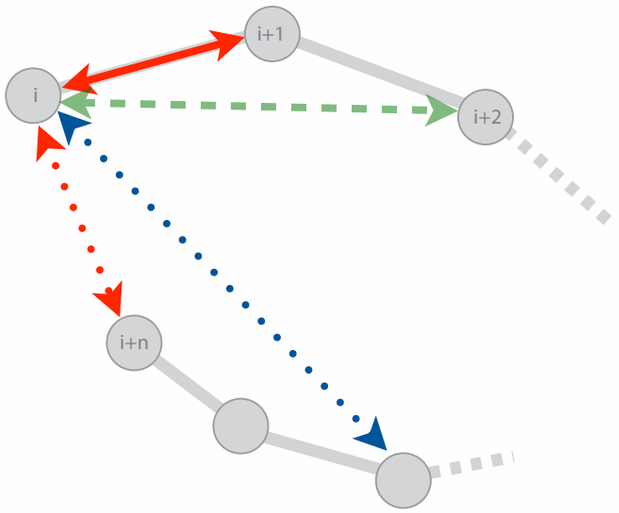


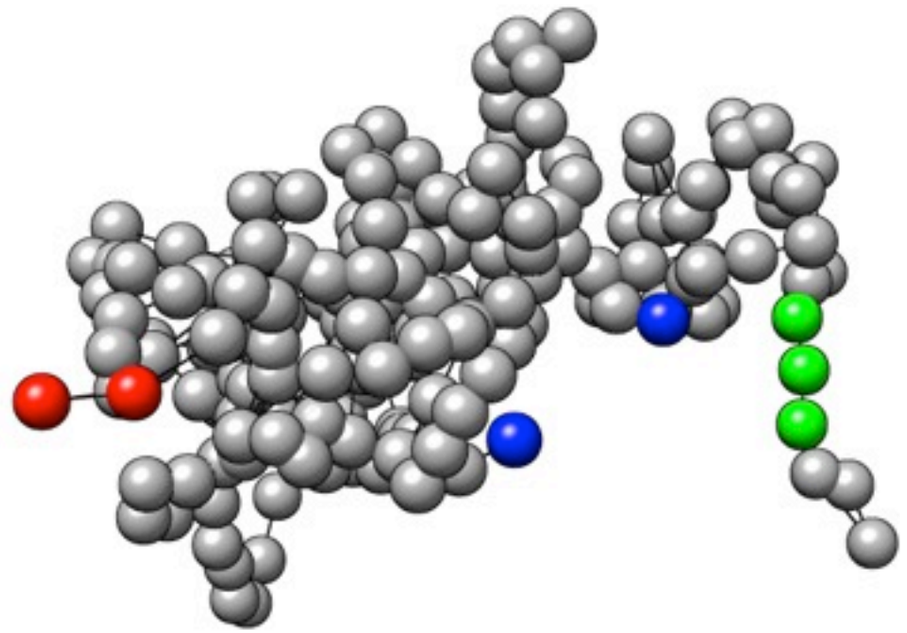
Parameter optimization



Model representation and scoring

Constituent parts of the molecule





$d < d_0$

$d = d_0$

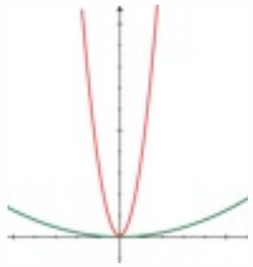
$d > d_0$

Model representation and scoring

Constituent parts of the molecule

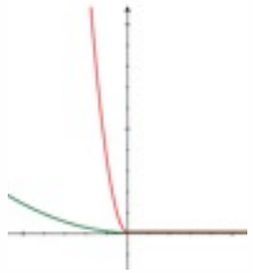
Harmonic

$$H_{i,j} = k\left(d_{i,j} - d^0_{i,j}\right)^2$$



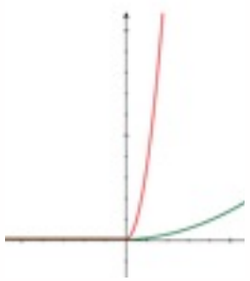
Harmonic Lower Bound

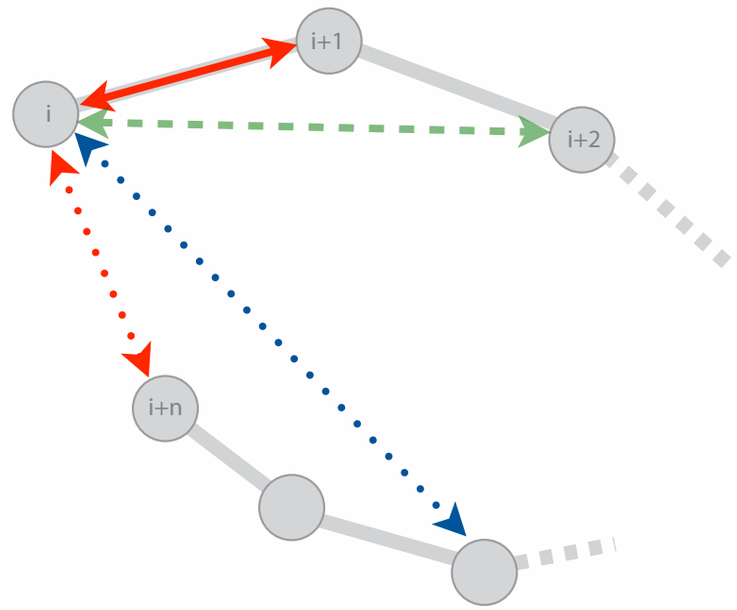
$$\begin{cases} if\ \ d_{i,j} \le d^0_{i,j}; \quad lbH_{i,j} = k\left(d_{i,j} - d^0_{i,j}\right)^2 \\ if\ \ d_{i,j} > d^0_{i,j}; \quad lbH_{i,j} = 0 \end{cases}$$



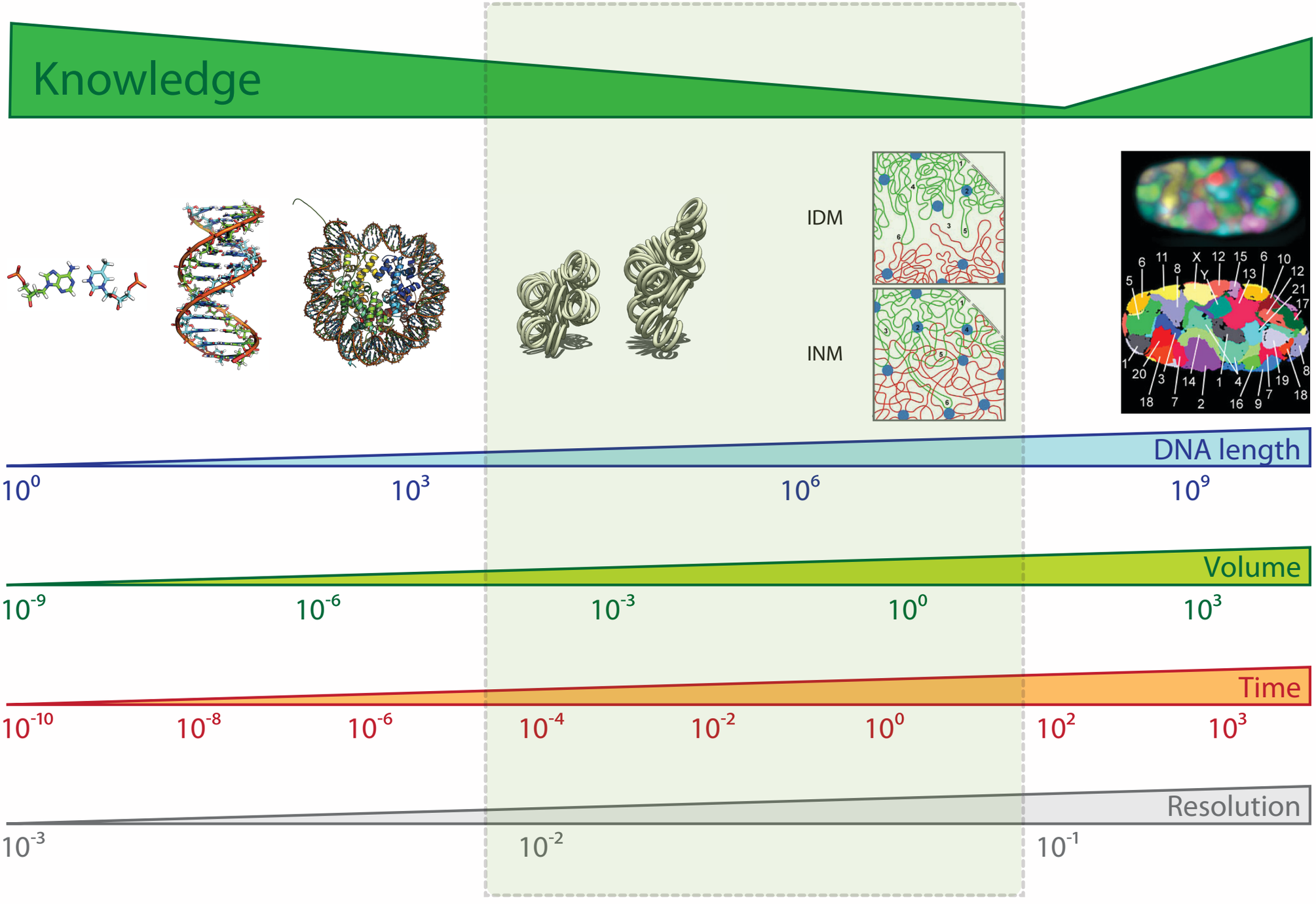
Harmonic Upper Bound

$$\begin{cases} if\ \ d_{i,j} \ge d^0_{i,j}; \quad ubH_{i,j} = k\left(d_{i,j} - d^0_{i,j}\right)^2 \\ if\ \ d_{i,j} < d^0_{i,j}; \quad ubH_{i,j} = 0 \end{cases}$$

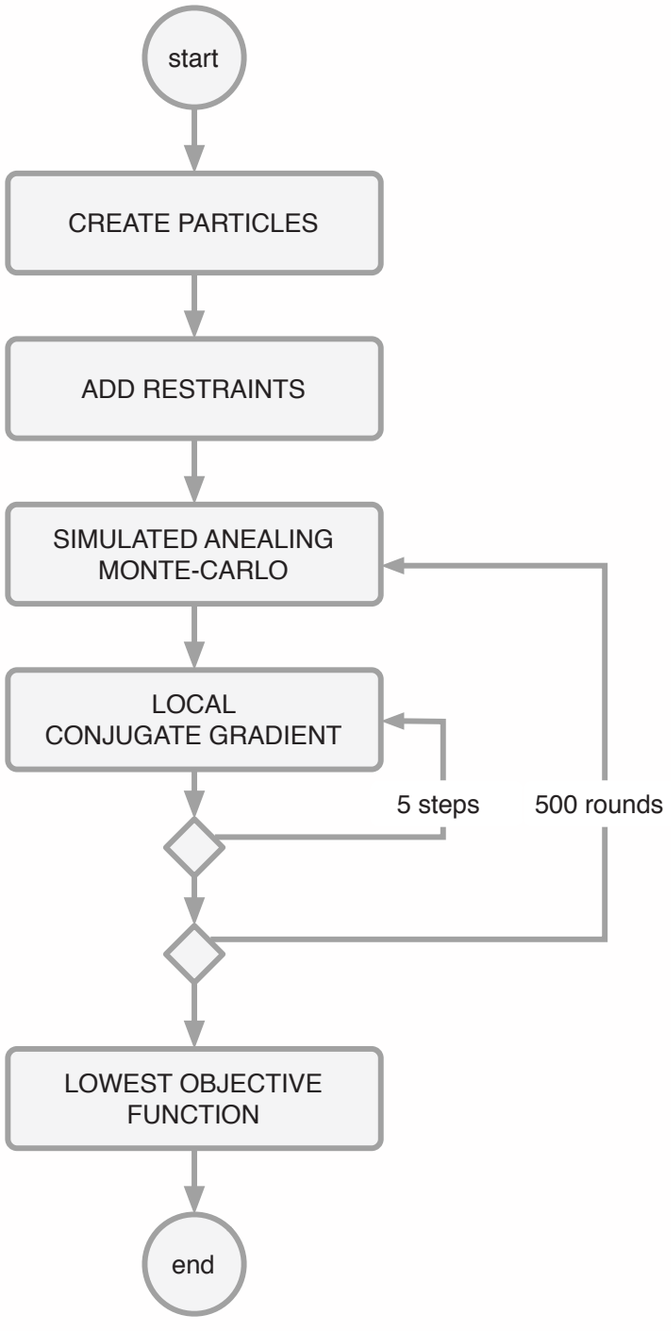


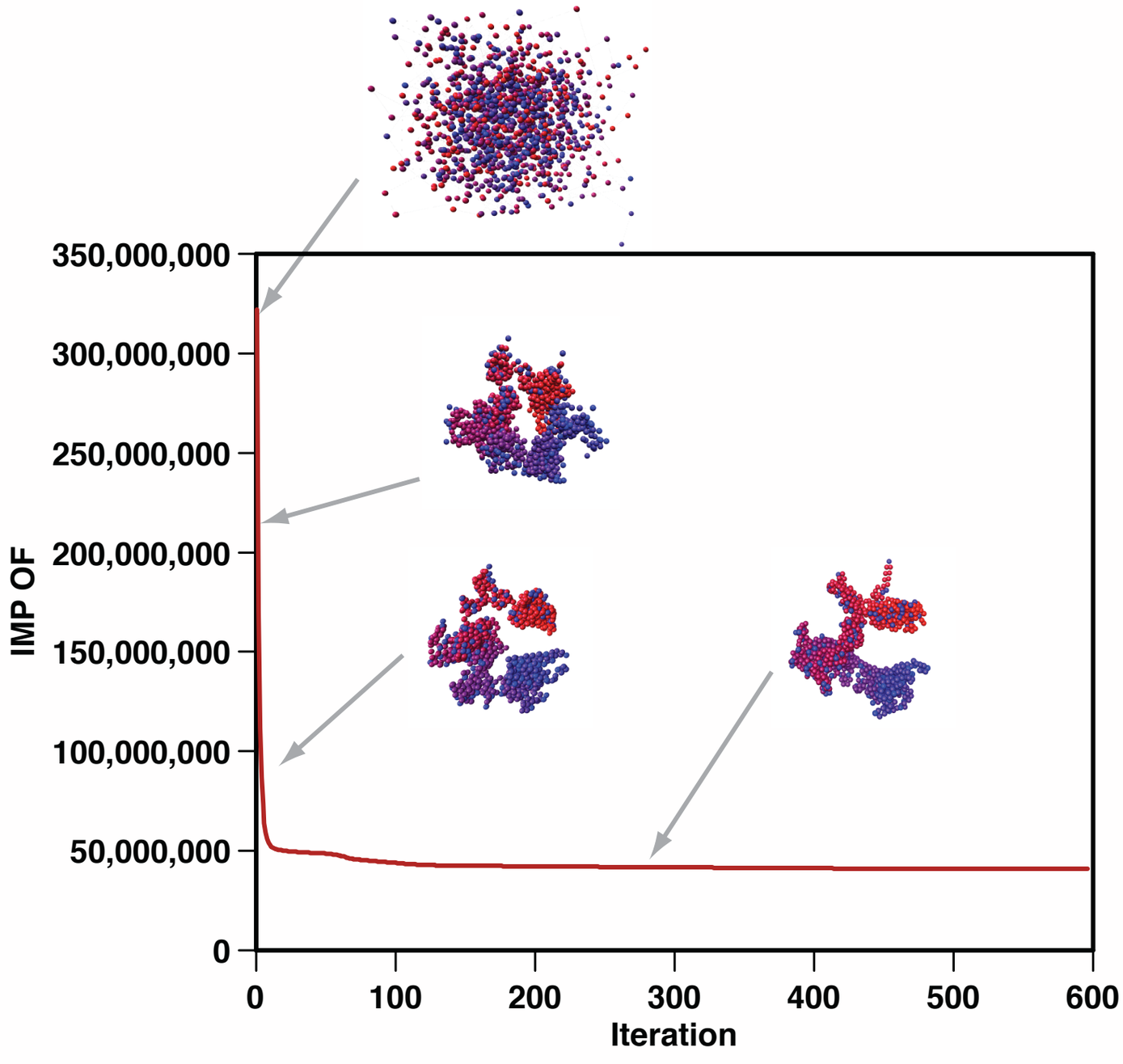


The resolution gap



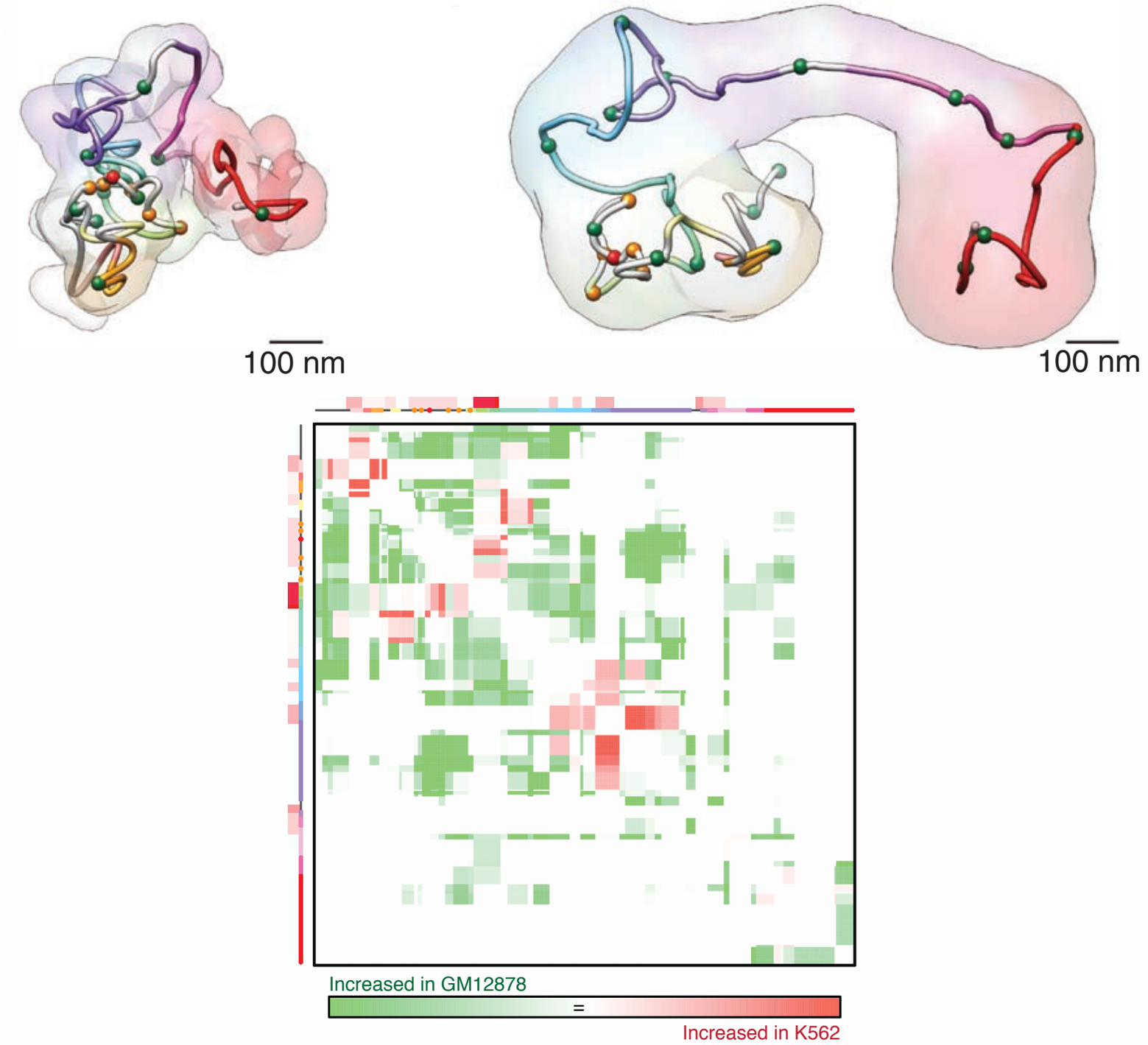
Optimization of the scoring function



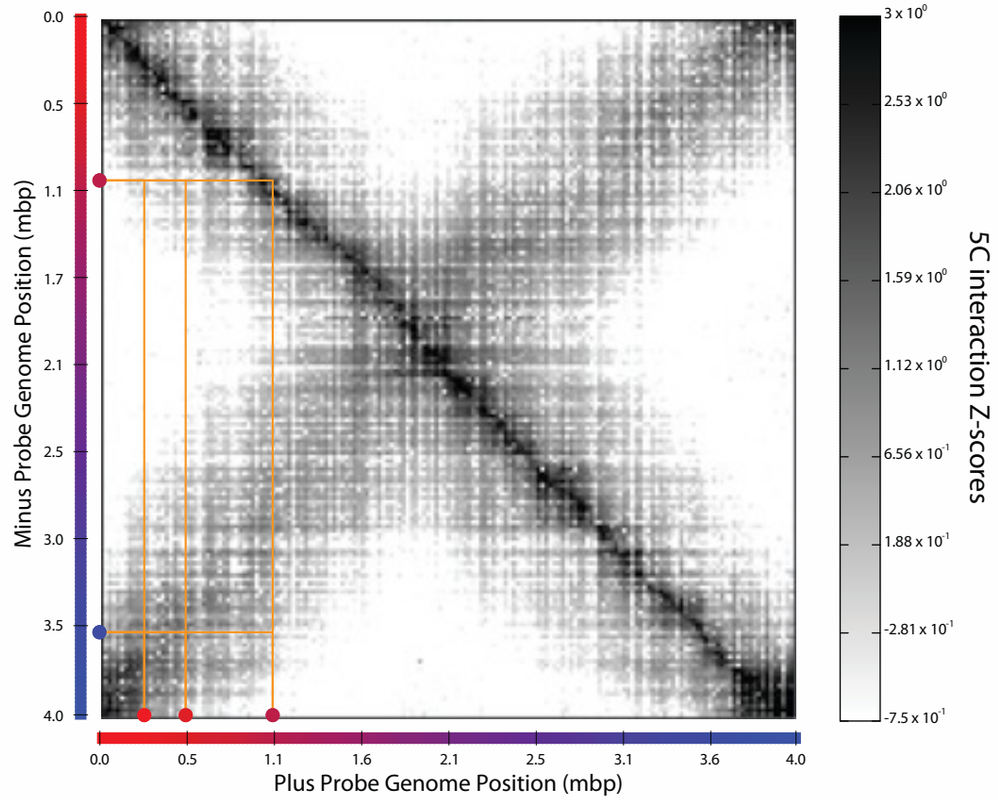


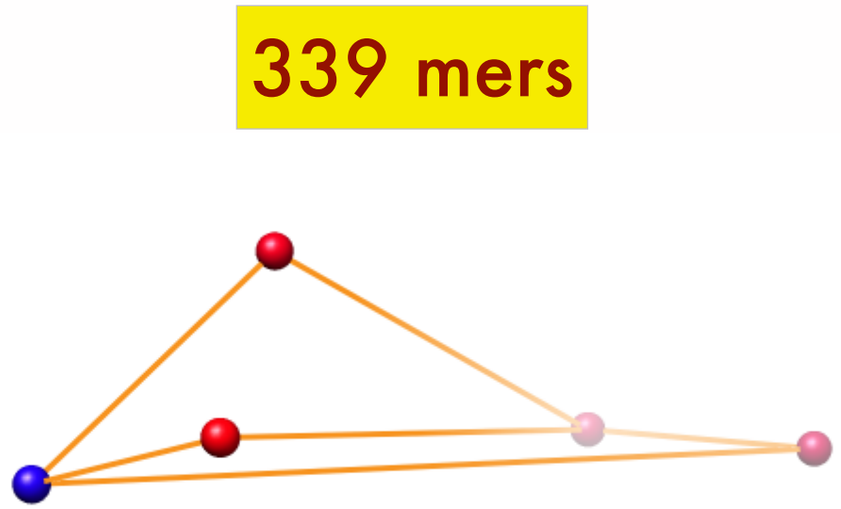
Chromatin globules

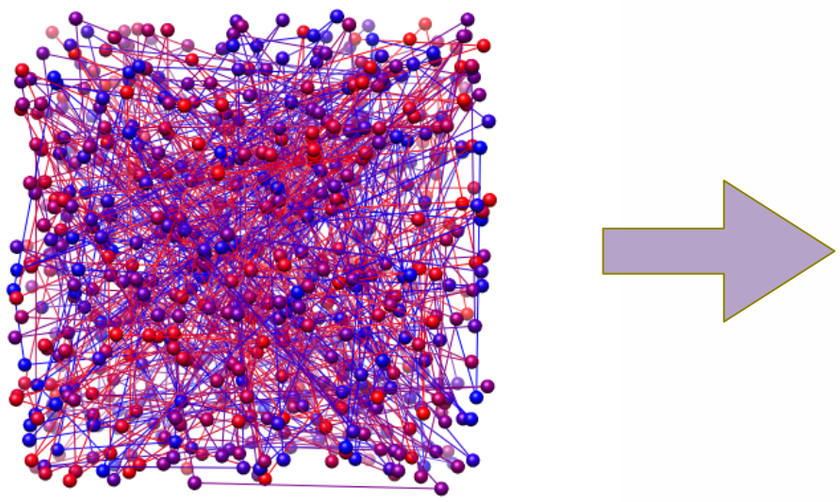
Frequency contact map differences

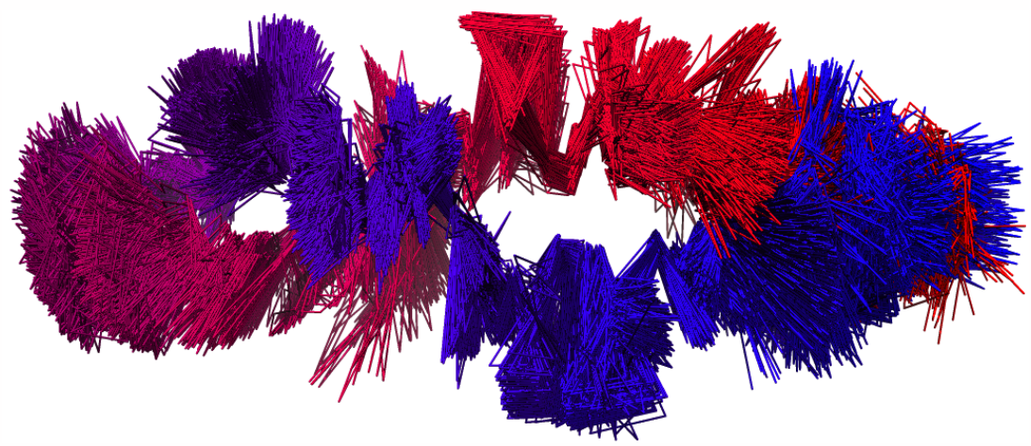


3D model building with the 5C + IMP approach



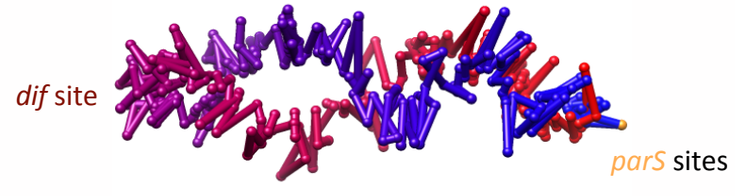




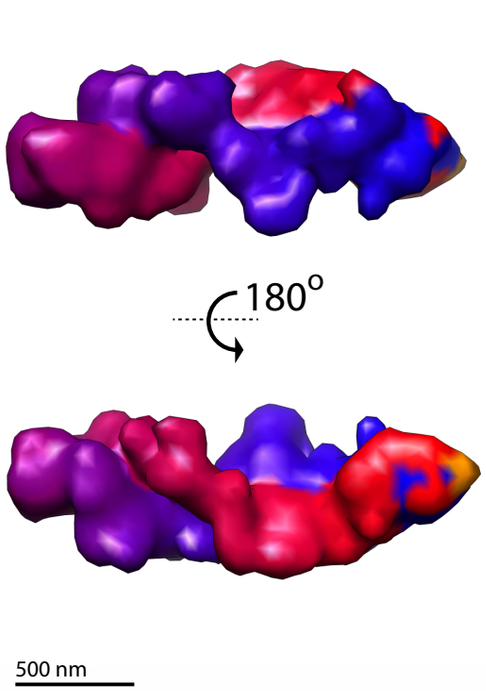


Genome organization in *Caulobacter Crescentus*

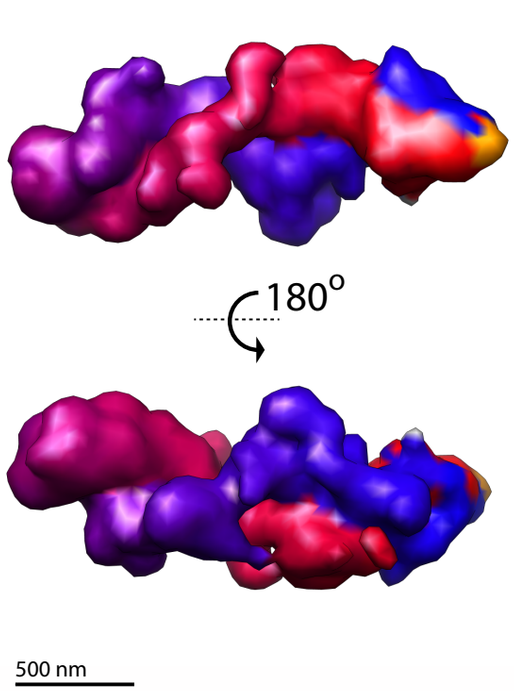
Arms are helical



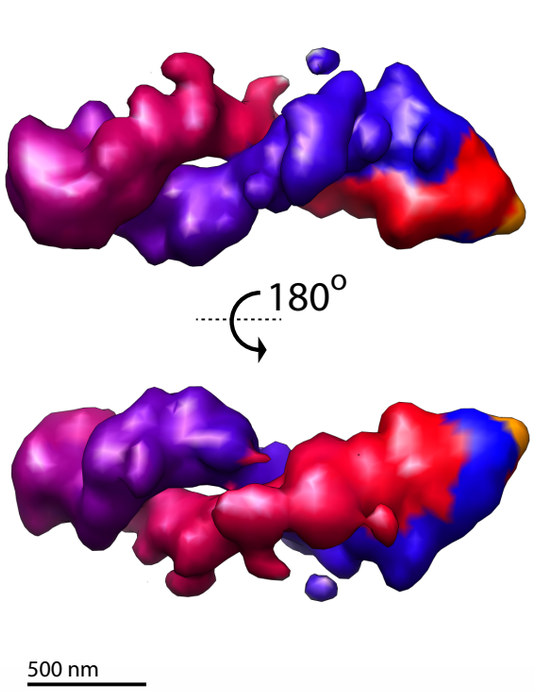
Cluster 1



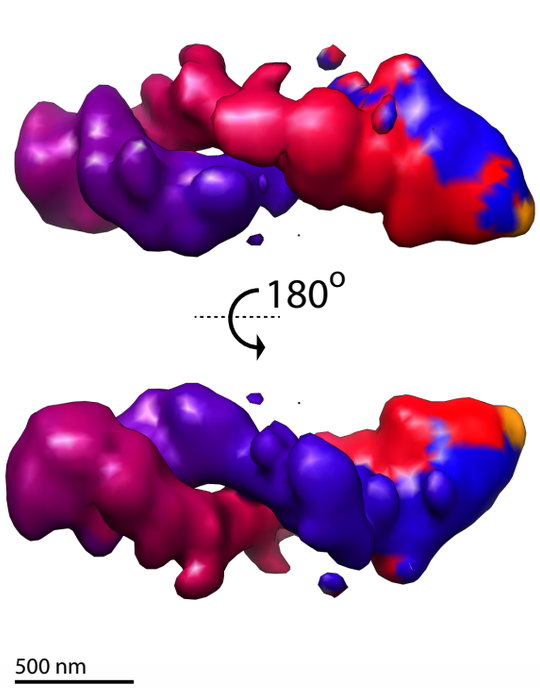
Cluster 2



Cluster 3

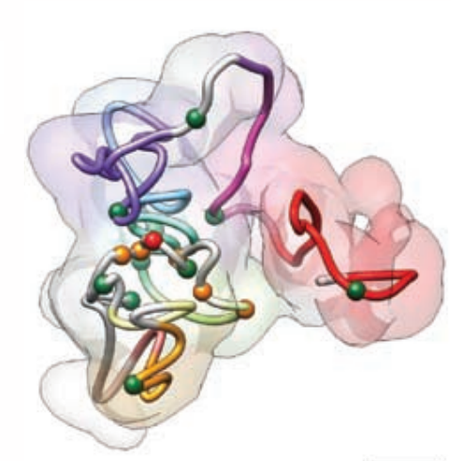


Cluster 4

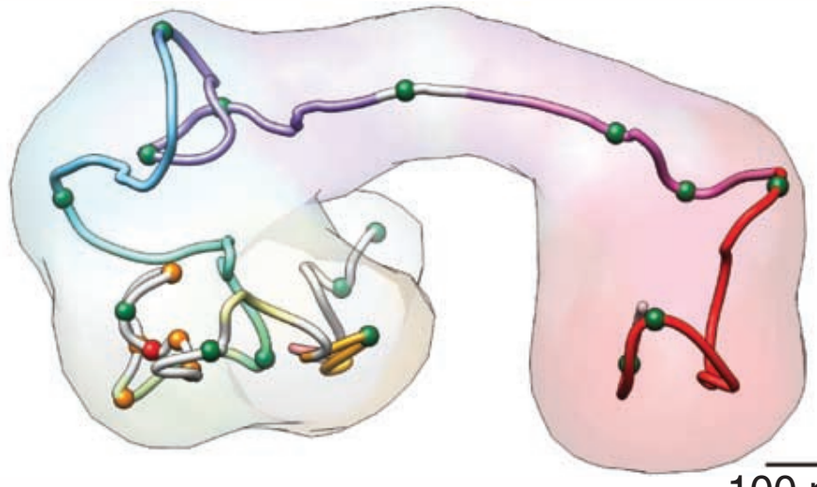


MIRRORS!

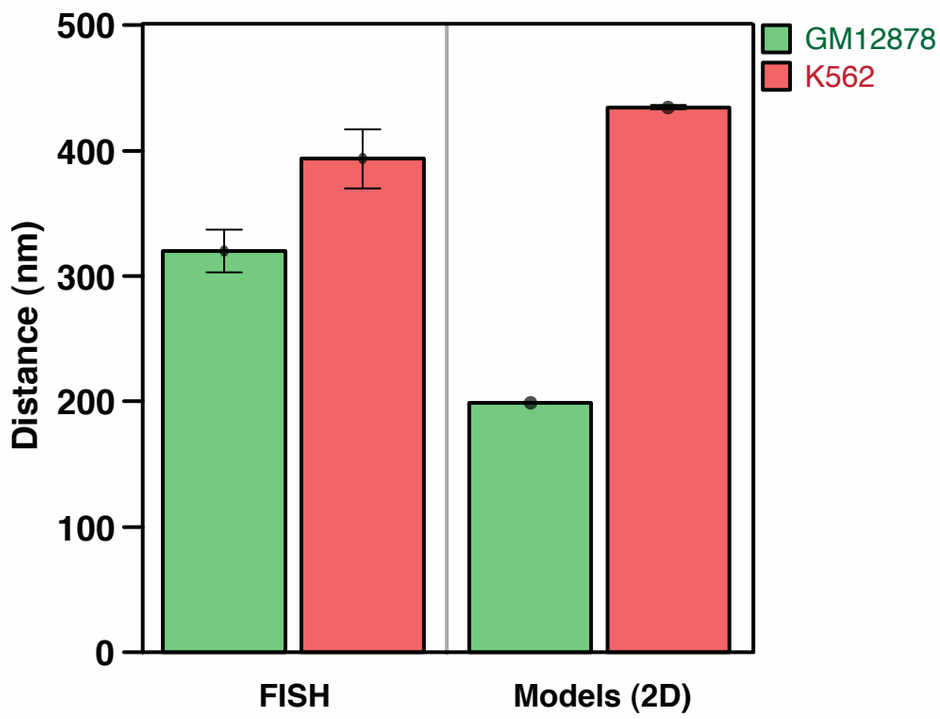
Model validation

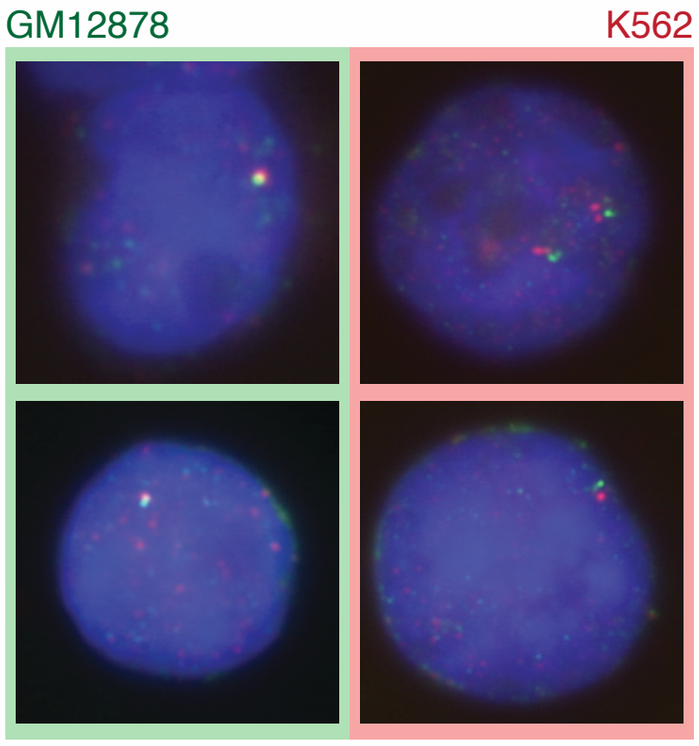


100 nm



100 nm





Take home message