Protein Structure Prediction and Analysis Tools

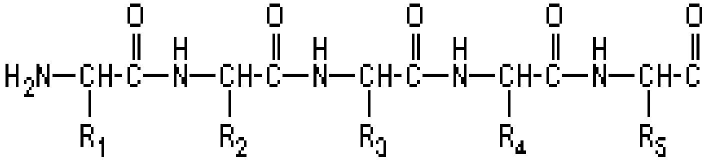
Jianlin Cheng, PhD

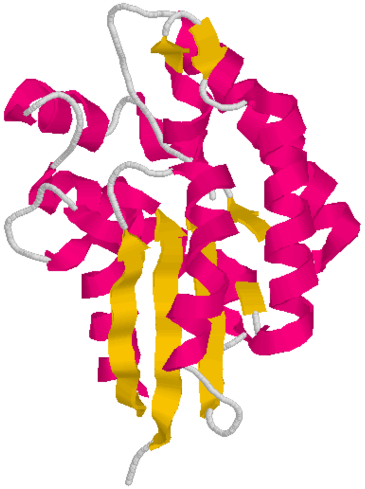
Assistant Professor
Department of Computer Science & Informatics Institute
University of Missouri, Columbia

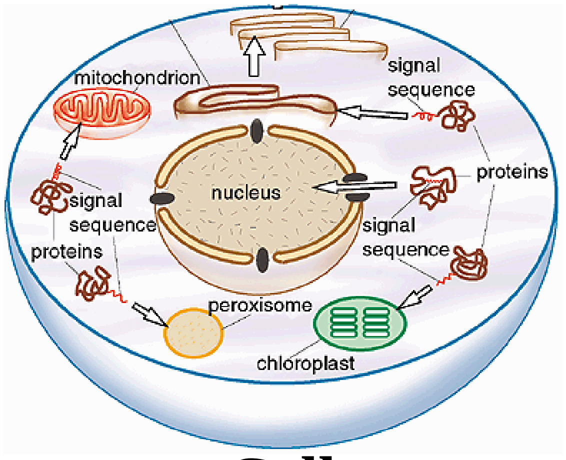
2011

Sequence, Structure and Function

AGCWY……





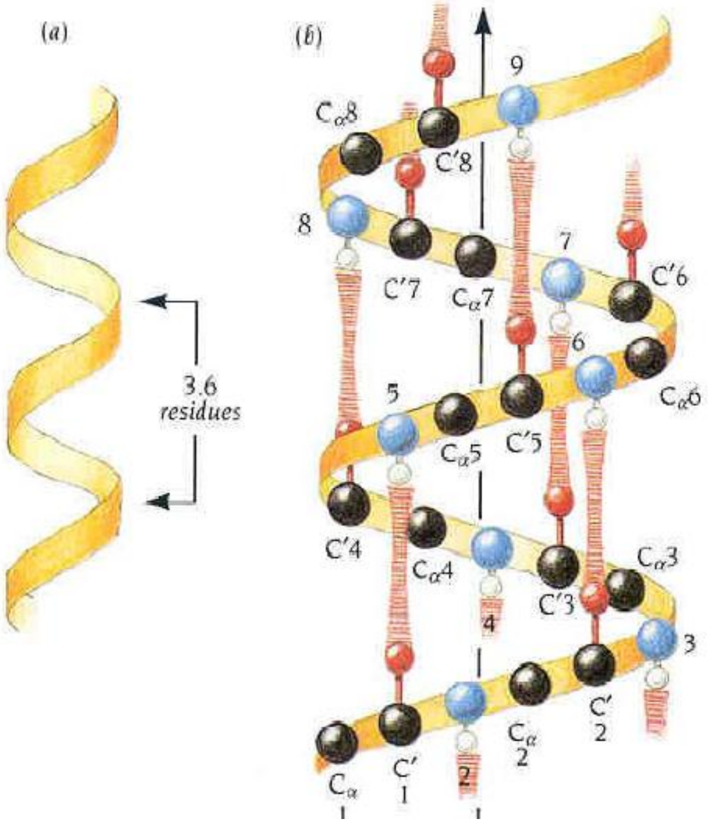


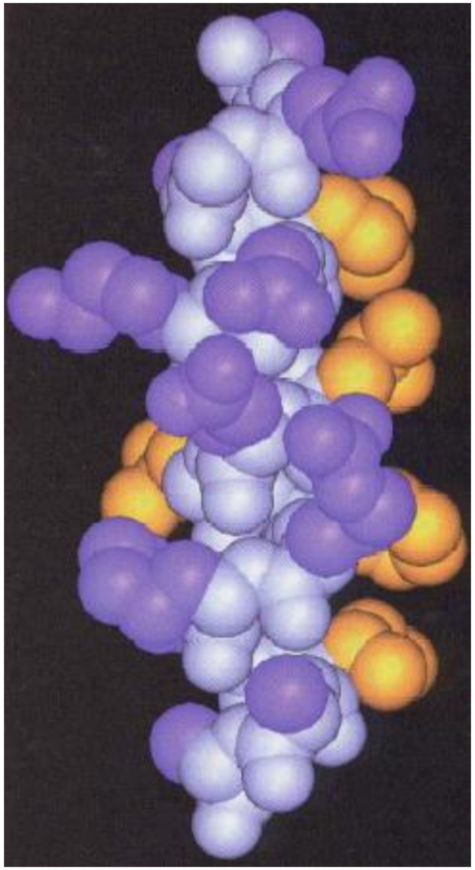
Cell

Protein Folding Movie

http://www.youtube.com/watch?v=fvBO3TqJ6FE&feature=fvw

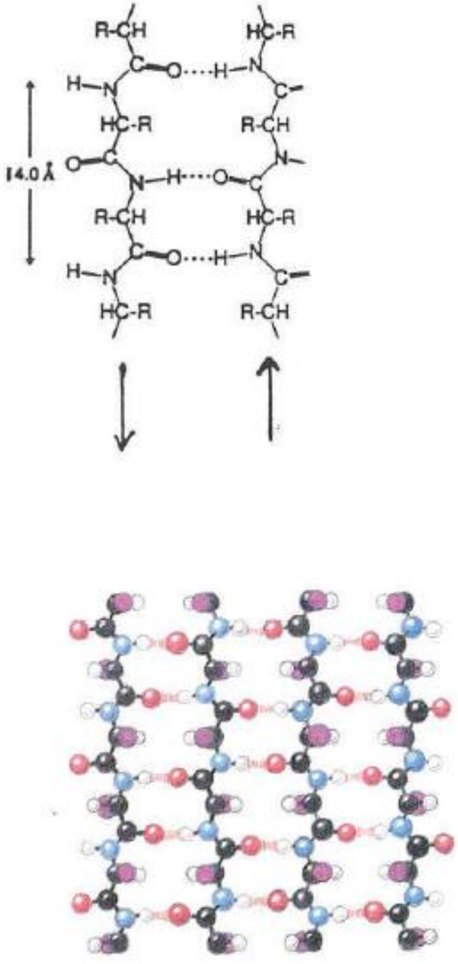
Alpha-Helix



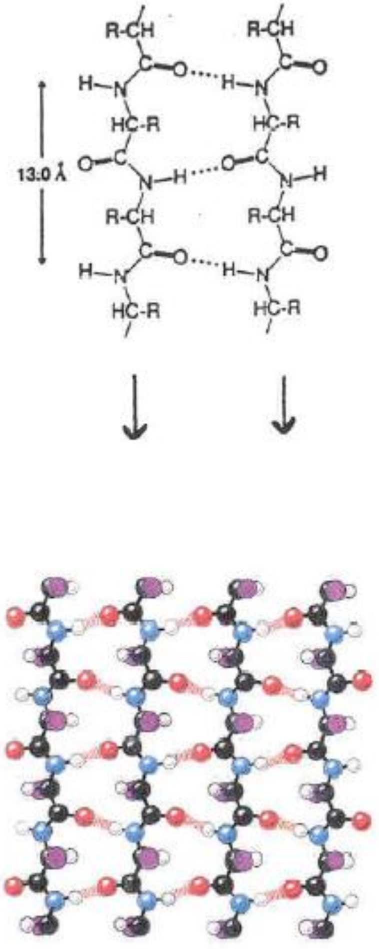


Jurnak, 2003

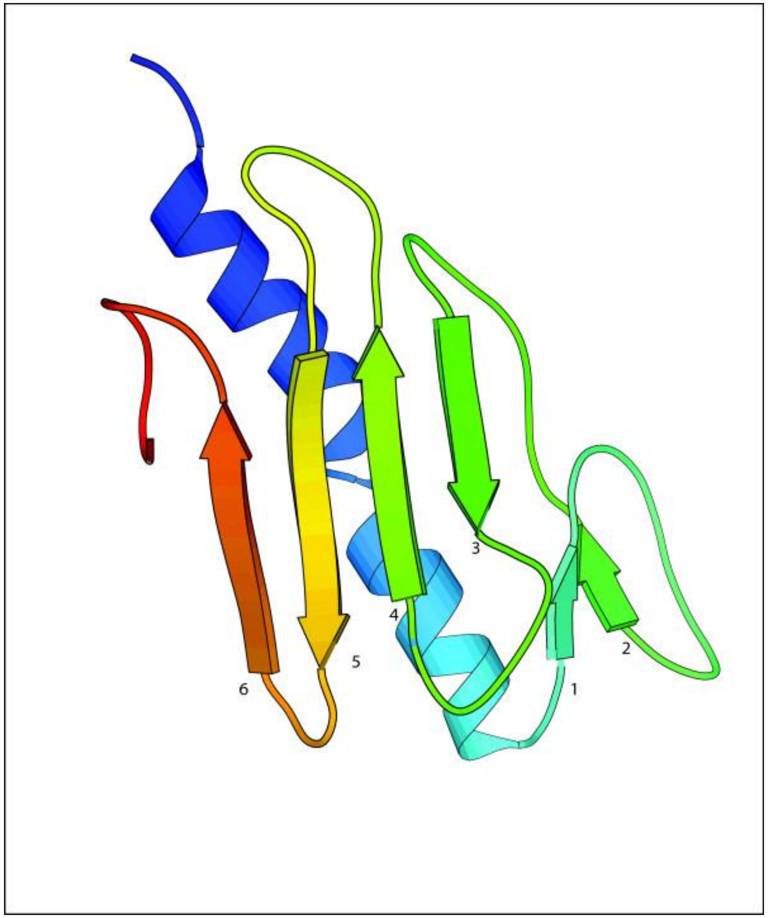
Beta-Sheet



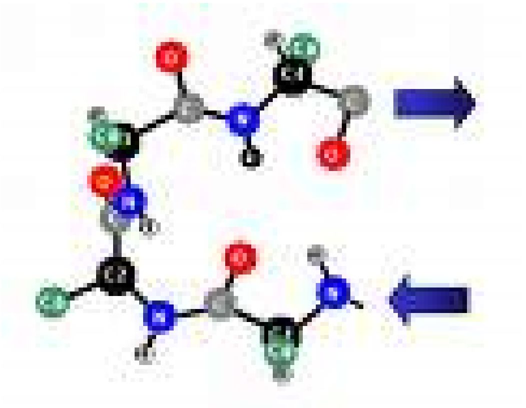
Anti-Parallel



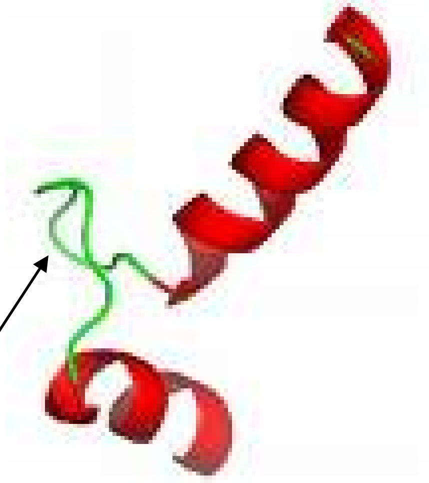
Parallel



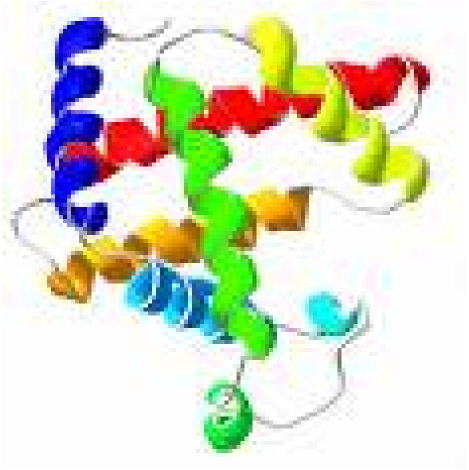
Non-Repetitive Secondary Structure



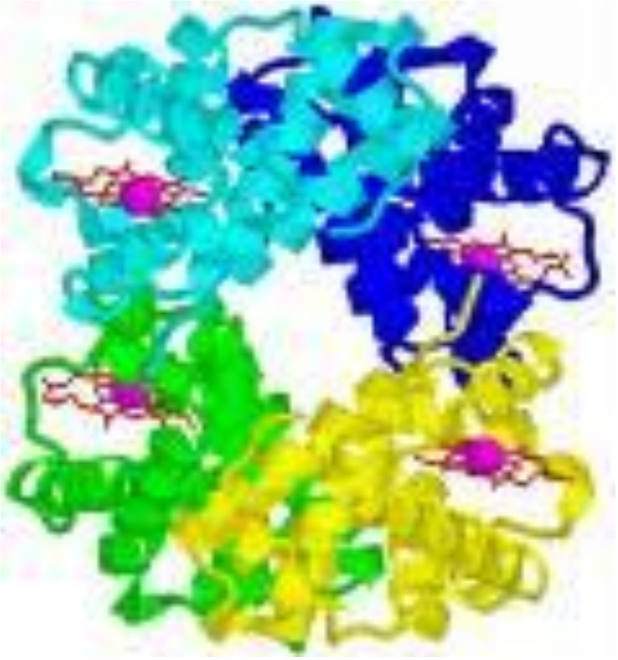
Beta-Turn



Loop

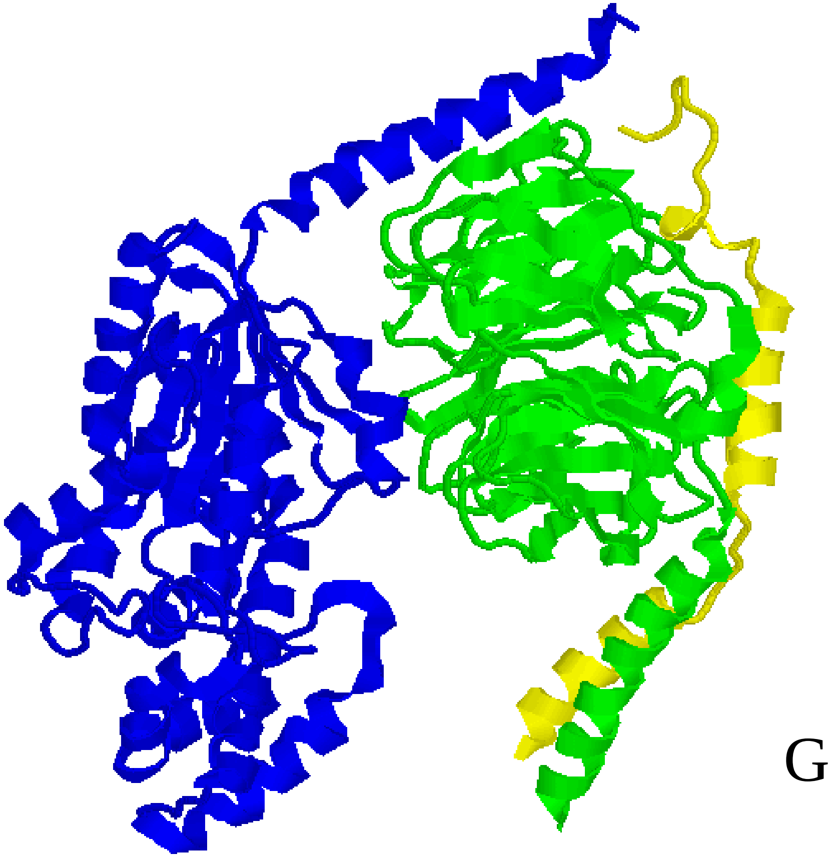


myoglobin



haemoglobin

Quaternary Structure: Complex



G-Protein Complex

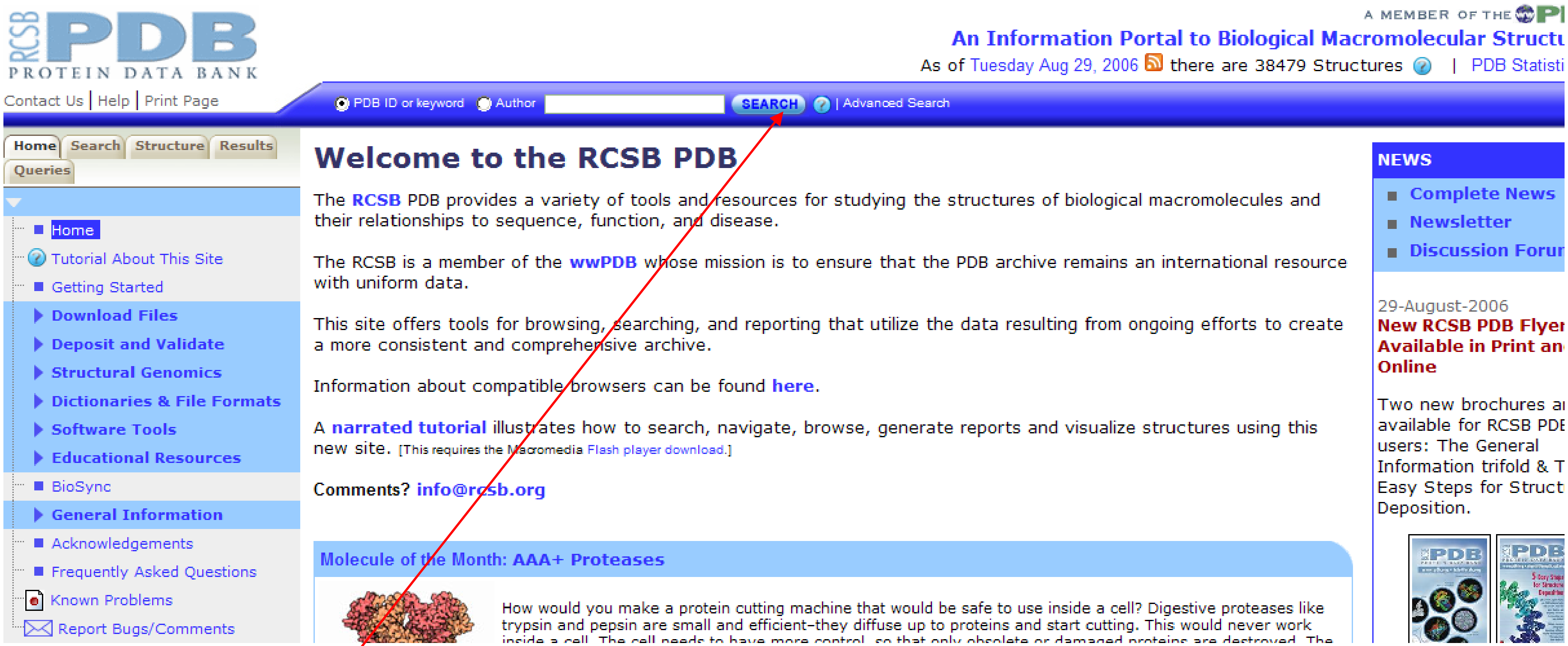
Protein Structure Determination

- X-ray crystallography
- Nuclear Magnetic Resonance (NMR) Spectroscopy
- X-ray: any size, accurate (1-3 Angstrom (10^{-10} m)), sometime hard to grow crystal
- NMR: small to medium size, moderate accuracy, structure in solution

Pacific Northwest National Laboratory's high magnetic field (800 MHz, 18.8 T) NMR spectrometer being loaded with a sample.

Wikipedia, the free encyclopedia

Storage in Protein Data Bank



RCSB PDB PROTEIN DATA BANK

A MEMBER OF THE PI

An Information Portal to Biological Macromolecular Structu

As of Tuesday Aug 29, 2006 there are 38479 Structures | PDB Statisti

Contact Us | Help | Print Page

● PDB ID or keyword ○ Author SEARCH | Advanced Search

- Home
- Tutorial About This Site
- Getting Started
- Download Files
- Deposit and Validate
- Structural Genomics
- Dictionaries & File Formats
- Software Tools
- Educational Resources
- BioSync
- General Information
- Acknowledgements
- Frequently Asked Questions
- Known Problems
- Report Bugs/Comments

Welcome to the RCSB PDB

The RCSB PDB provides a variety of tools and resources for studying the structures of biological macromolecules and their relationships to sequence, function, and disease.

The RCSB is a member of the wwPDB whose mission is to ensure that the PDB archive remains an international resource with uniform data.

This site offers tools for browsing, searching, and reporting that utilize the data resulting from ongoing efforts to create a more consistent and comprehensive archive.

Information about compatible browsers can be found here.

A narrated tutorial illustrates how to search, navigate, browse, generate reports and visualize structures using this new site. [This requires the Macromedia Flash player download.]

Comments? info@rcsb.org

Molecule of the Month: AAA+ Proteases

How would you make a protein cutting machine that would be safe to use inside a cell? Digestive proteases like trypsin and pepsin are small and efficient-they diffuse up to proteins and start cutting. This would never work inside a cell. The cell needs to have more control, so that only obsolete or damaged proteins are destroyed. The

NEWS

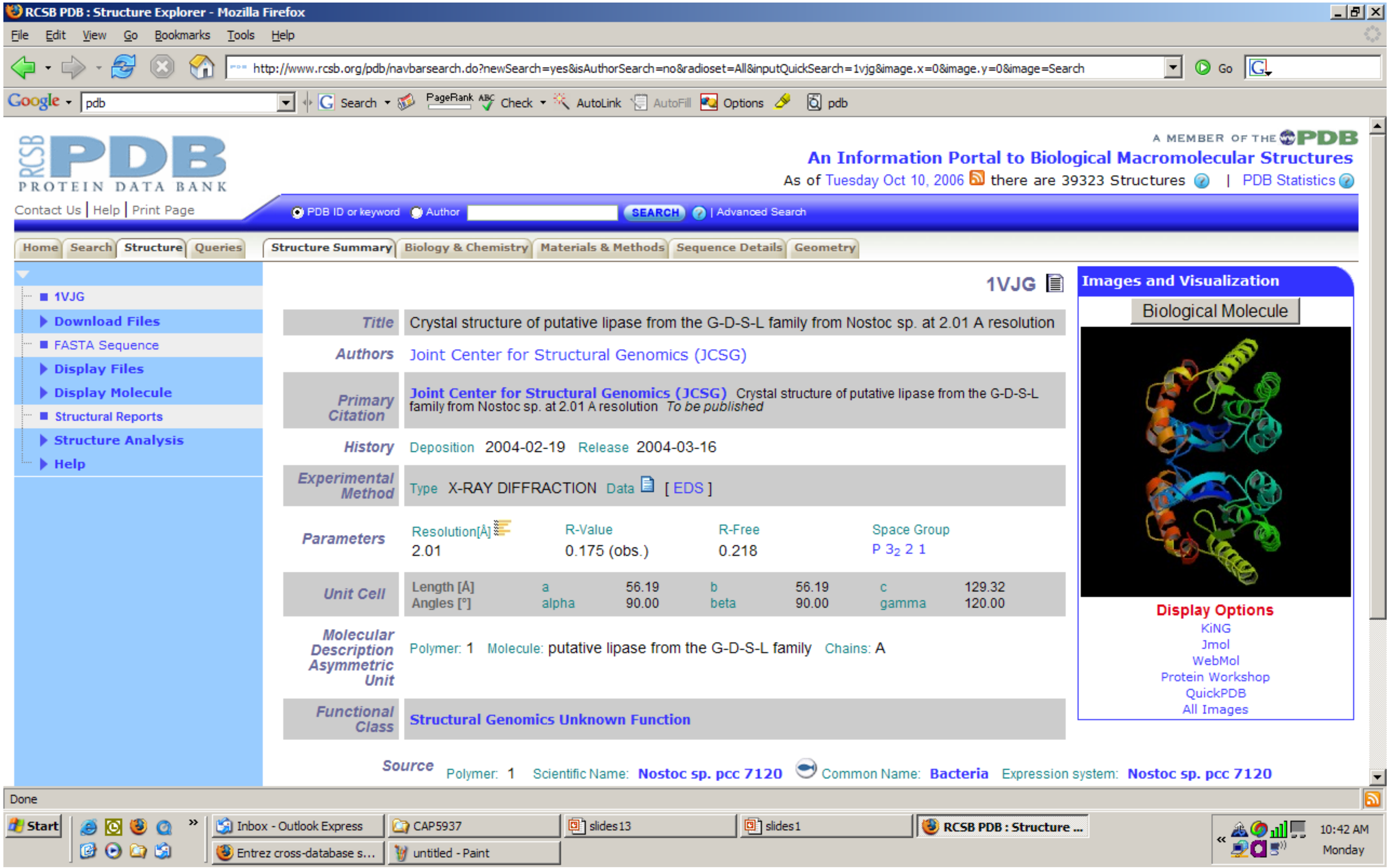
- Complete News
- Newsletter
- Discussion Forur

29-August-2006

New RCSB PDB Flyer Available in Print an Online

Two new brochures ar available for RCSB PDE users: The General Information trifold & T Easy Steps for Struct Deposition.

Search database



http://www.rcsb.org/pdb/navbarsearch.do?newSearch=yes&isAuthorSearch=no&radioset=All&inputQuickSearch=1vjg&image.x=0&image.y=0&image=Search

A MEMBER OF THE PDB

An Information Portal to Biological Macromolecular Structures

As of Tuesday Oct 10, 2006 there are 39323 Structures | PDB Statistics

1VJG

Title	Crystal structure of putative lipase from the G-D-S-L family from Nostoc sp. at 2.01 A resolution
Authors	Joint Center for Structural Genomics (JCSG)
Primary Citation	**Joint Center for Structural Genomics (JCSG)** Crystal structure of putative lipase from the G-D-S-L family from Nostoc sp. at 2.01 A resolution *To be published*
History	Deposition 2004-02-19 Release 2004-03-16
Experimental Method	Type X-RAY DIFFRACTION Data [EDS]

Parameters	Resolution[Å]	R-Value	R-Free	Space Group
	2.01	0.175 (obs.)	0.218	P 3_2 2 1

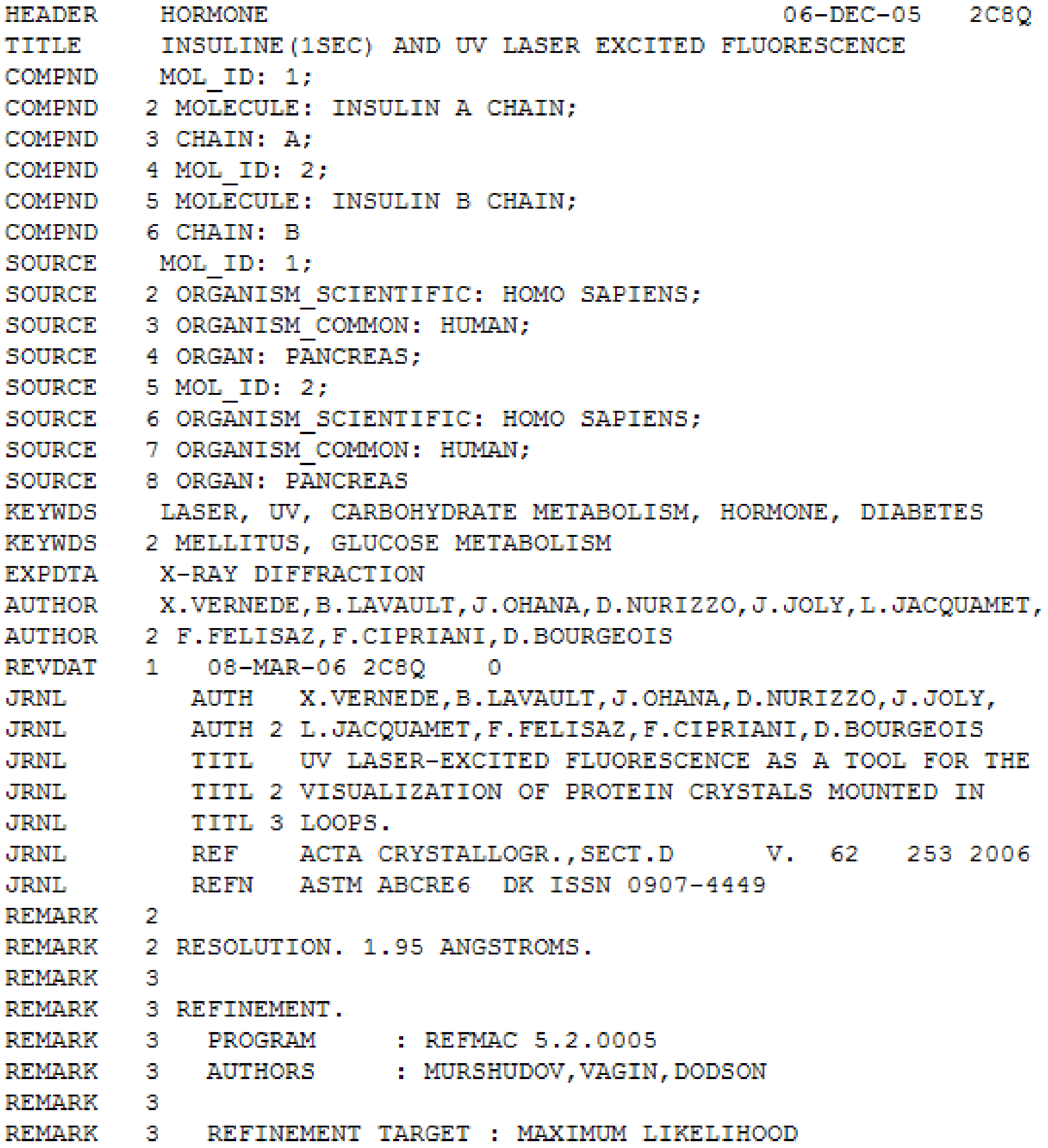
Unit Cell							
	Length [Å]	a	56.19	b	56.19	c	129.32
	Angles [°]	alpha	90.00	beta	90.00	gamma	120.00

Molecular Description Asymmetric Unit	Polymer: 1 Molecule: putative lipase from the G-D-S-L family Chains: A
Functional Class	**Structural Genomics Unknown Function**

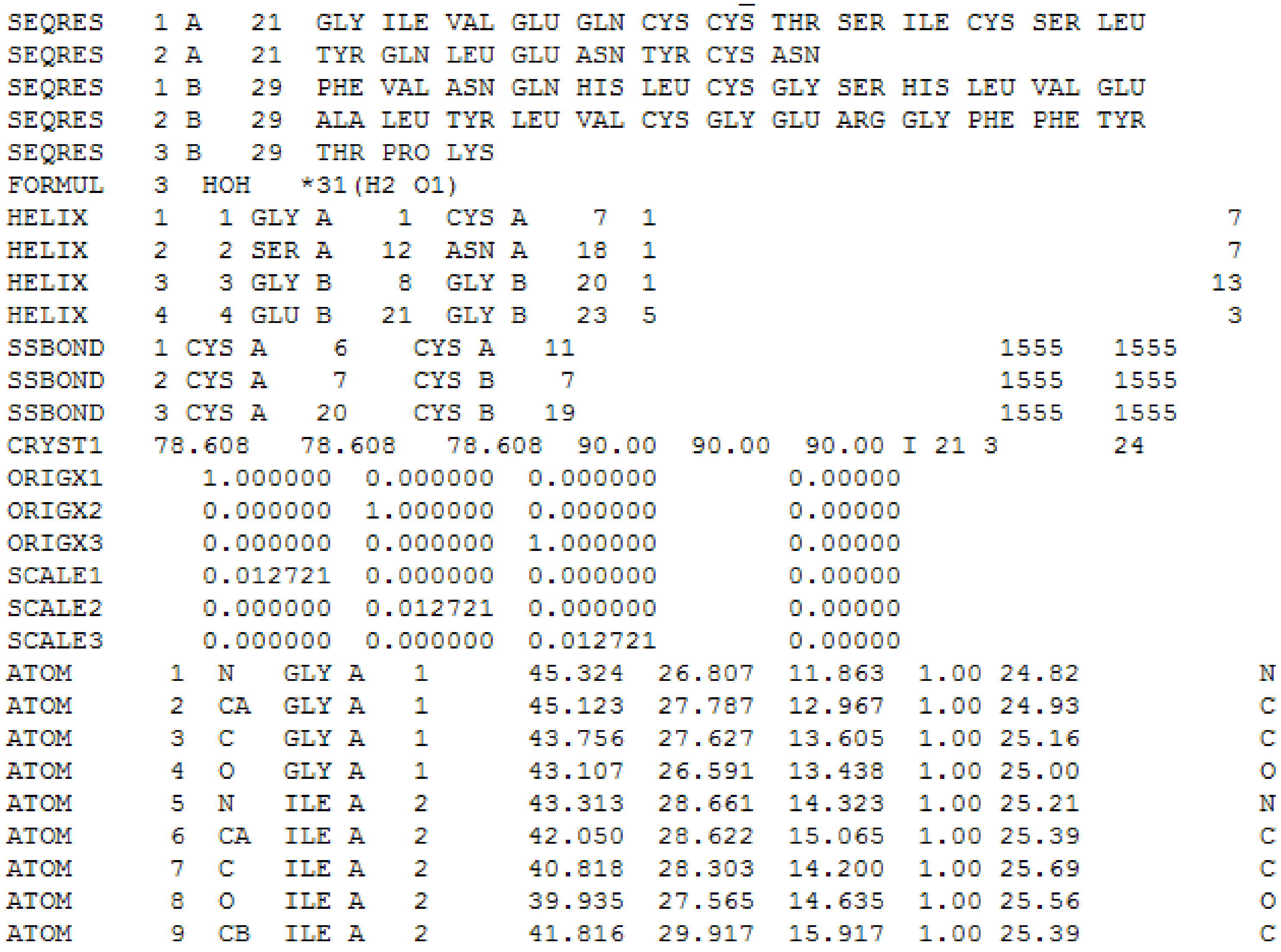
Source Polymer: 1 Scientific Name: **Nostoc sp. pcc 7120** Common Name: **Bacteria** Expression system: **Nostoc sp. pcc 7120**

Search protein 1VJG

PDB Format (2C8Q, insulin)



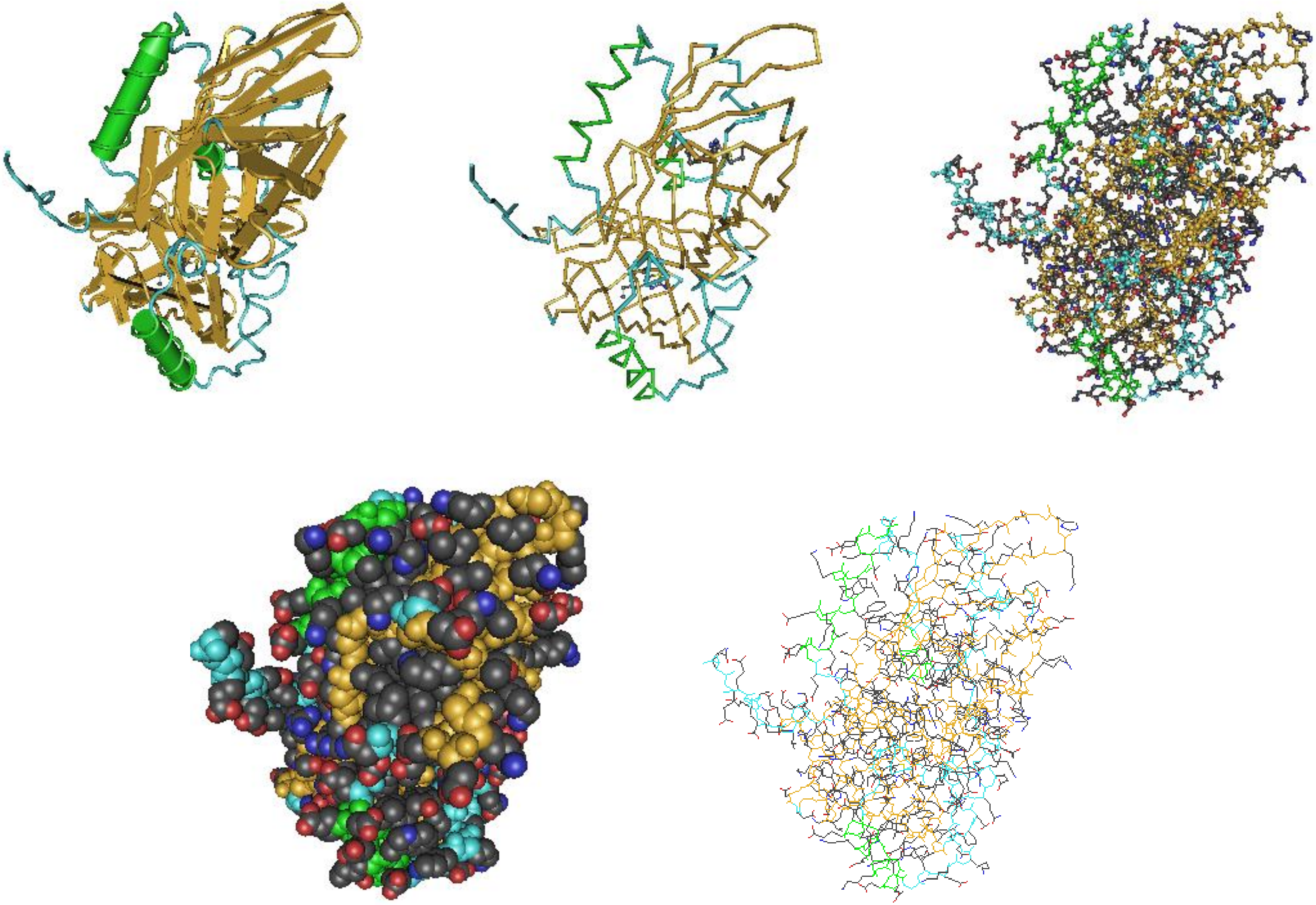
```
HEADER    HORMONE                                 06-DEC-05   2C8Q
TITLE     INSULINE(1SEC) AND UV LASER EXCITED FLUORESCENCE
COMPND    MOL_ID: 1;
COMPND   2 MOLECULE: INSULIN A CHAIN;
COMPND   3 CHAIN: A;
COMPND   4 MOL_ID: 2;
COMPND   5 MOLECULE: INSULIN B CHAIN;
COMPND   6 CHAIN: B
SOURCE    MOL_ID: 1;
SOURCE   2 ORGANISM_SCIENTIFIC: HOMO SAPIENS;
SOURCE   3 ORGANISM_COMMON: HUMAN;
SOURCE   4 ORGAN: PANCREAS;
SOURCE   5 MOL_ID: 2;
SOURCE   6 ORGANISM_SCIENTIFIC: HOMO SAPIENS;
SOURCE   7 ORGANISM_COMMON: HUMAN;
SOURCE   8 ORGAN: PANCREAS
KEYWDS    LASER, UV, CARBOHYDRATE METABOLISM, HORMONE, DIABETES
KEYWDS   2 MELLITUS, GLUCOSE METABOLISM
EXPDTA    X-RAY DIFFRACTION
AUTHOR    X.VERNEDE,B.LAVAULT,J.OHANA,D.NURIZZO,J.JOLY,L.JACQUAMET,
AUTHOR   2 F.FELISAZ,F.CIPRIANI,D.BOURGEOIS
REVDAT   1   08-MAR-06 2C8Q    0
JRNL        AUTH   X.VERNEDE,B.LAVAULT,J.OHANA,D.NURIZZO,J.JOLY,
JRNL        AUTH 2 L.JACQUAMET,F.FELISAZ,F.CIPRIANI,D.BOURGEOIS
JRNL        TITL   UV LASER-EXCITED FLUORESCENCE AS A TOOL FOR THE
JRNL        TITL 2 VISUALIZATION OF PROTEIN CRYSTALS MOUNTED IN
JRNL        TITL 3 LOOPS.
JRNL        REF    ACTA CRYSTALLOGR.,SECT.D      V.  62   253 2006
JRNL        REFN   ASTM ABCRE6  DK ISSN 0907-4449
REMARK   2
REMARK   2 RESOLUTION. 1.95 ANGSTROMS.
REMARK   3
REMARK   3 REFINEMENT.
REMARK   3   PROGRAM     : REFMAC 5.2.0005
REMARK   3   AUTHORS     : MURSHUDOV,VAGIN,DODSON
REMARK   3
REMARK   3   REFINEMENT TARGET : MAXIMUM LIKELIHOOD
```



```
SEQRES   1 A   21  GLY ILE VAL GLU GLN CYS CYS THR SER ILE CYS SER LEU
SEQRES   2 A   21  TYR GLN LEU GLU ASN TYR CYS ASN
SEQRES   1 B   29  PHE VAL ASN GLN HIS LEU CYS GLY SER HIS LEU VAL GLU
SEQRES   2 B   29  ALA LEU TYR LEU VAL CYS GLY GLU ARG GLY PHE PHE TYR
SEQRES   3 B   29  THR PRO LYS
FORMUL   3  HOH   *31(H2 O1)
HELIX    1   1 GLY A    1  CYS A    7  1                                   7
HELIX    2   2 SER A   12  ASN A   18  1                                   7
HELIX    3   3 GLY B    8  GLY B   20  1                                  13
HELIX    4   4 GLU B   21  GLY B   23  5                                   3
SSBOND   1 CYS A    6    CYS A   11                          1555   1555
SSBOND   2 CYS A    7    CYS B    7                          1555   1555
SSBOND   3 CYS A   20    CYS B   19                          1555   1555
CRYST1   78.608   78.608   78.608  90.00  90.00  90.00 I 21 3      24
ORIGX1      1.000000  0.000000  0.000000        0.00000
ORIGX2      0.000000  1.000000  0.000000        0.00000
ORIGX3      0.000000  0.000000  1.000000        0.00000
SCALE1      0.012721  0.000000  0.000000        0.00000
SCALE2      0.000000  0.012721  0.000000        0.00000
SCALE3      0.000000  0.000000  0.012721        0.00000
ATOM      1  N   GLY A   1      45.324  26.807  11.863  1.00 24.82           N
ATOM      2  CA  GLY A   1      45.123  27.787  12.967  1.00 24.93           C
ATOM      3  C   GLY A   1      43.756  27.627  13.605  1.00 25.16           C
ATOM      4  O   GLY A   1      43.107  26.591  13.438  1.00 25.00           O
ATOM      5  N   ILE A   2      43.313  28.661  14.323  1.00 25.21           N
ATOM      6  CA  ILE A   2      42.050  28.622  15.065  1.00 25.39           C
ATOM      7  C   ILE A   2      40.818  28.303  14.200  1.00 25.69           C
ATOM      8  O   ILE A   2      39.935  27.565  14.635  1.00 25.56           O
ATOM      9  CB  ILE A   2      41.816  29.917  15.917  1.00 25.39           C
```

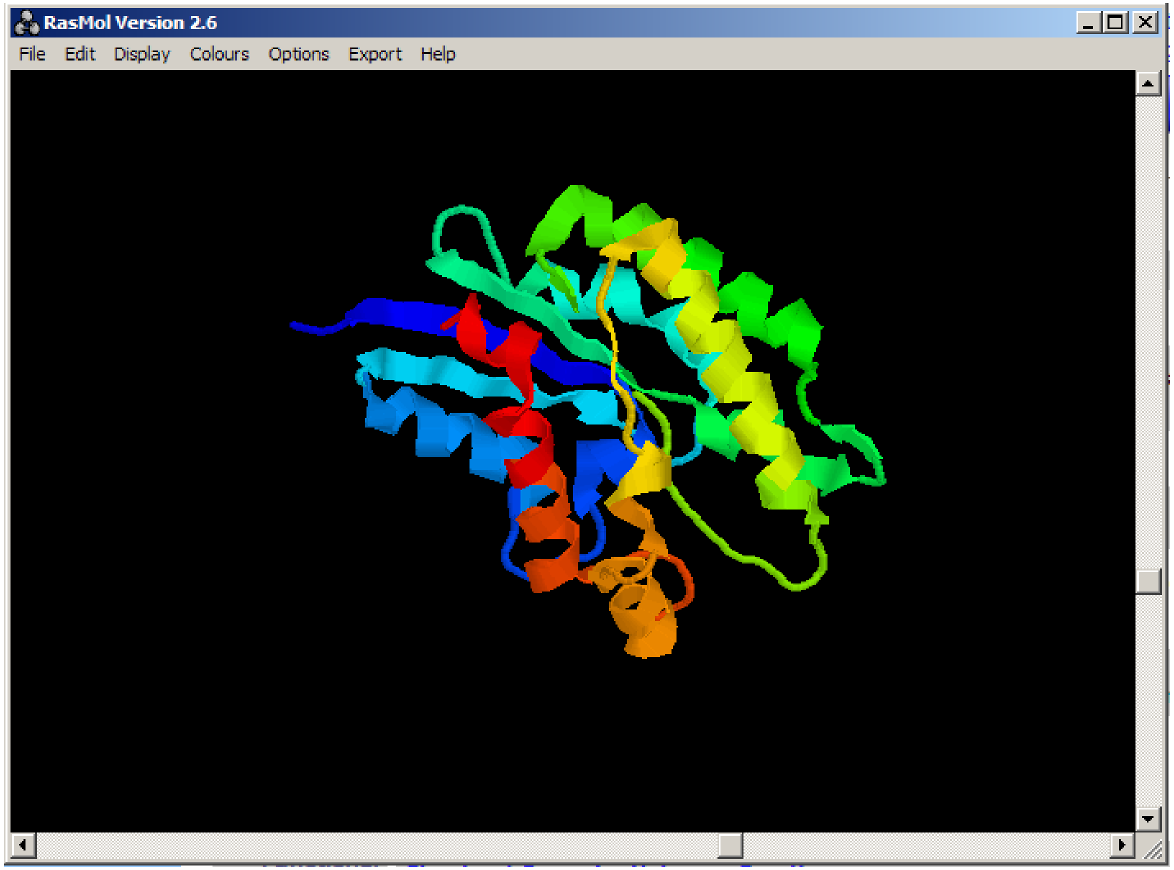
Structure Visualization

- Rasmol (http://www.umass.edu/microbio/rasmol/getras.htm)
- MDL Chime (plug-in) (http://www.mdl.com/products/framework/chime/)
- Protein Explorer (http://molvis.sdsc.edu/protexpl/frntdoor.htm)
- Jmol: http://jmol.sourceforge.net/
- Pymol: http://pymol.sourceforge.net/



J. Pevsner, 2005

Rasmol (1VJG)


RasMol Version 2.6
File
Edit
Display
Colours
Options
Export
Help

Structure Analysis

- Assign secondary structure for amino acids from 3D structure
- Generate solvent accessible area for amino acids from 3D structure
- Most widely used tool: DSSP (Dictionary of Protein Secondary Structure: Pattern Recognition of Hydrogen-Bonded and Geometrical Features. **Kabsch and Sander, 1983**)

DSSP server: http://bioweb.pasteur.fr/seqanal/interfaces/dssp-simple.html

DSSP download: http://swift.cmbi.ru.nl/gv/dssp/

DSSP Code:

H = alpha helix

G = 3-helix (3/10 helix)

I = 5 helix (pi helix)

B = residue in isolated beta-bridge

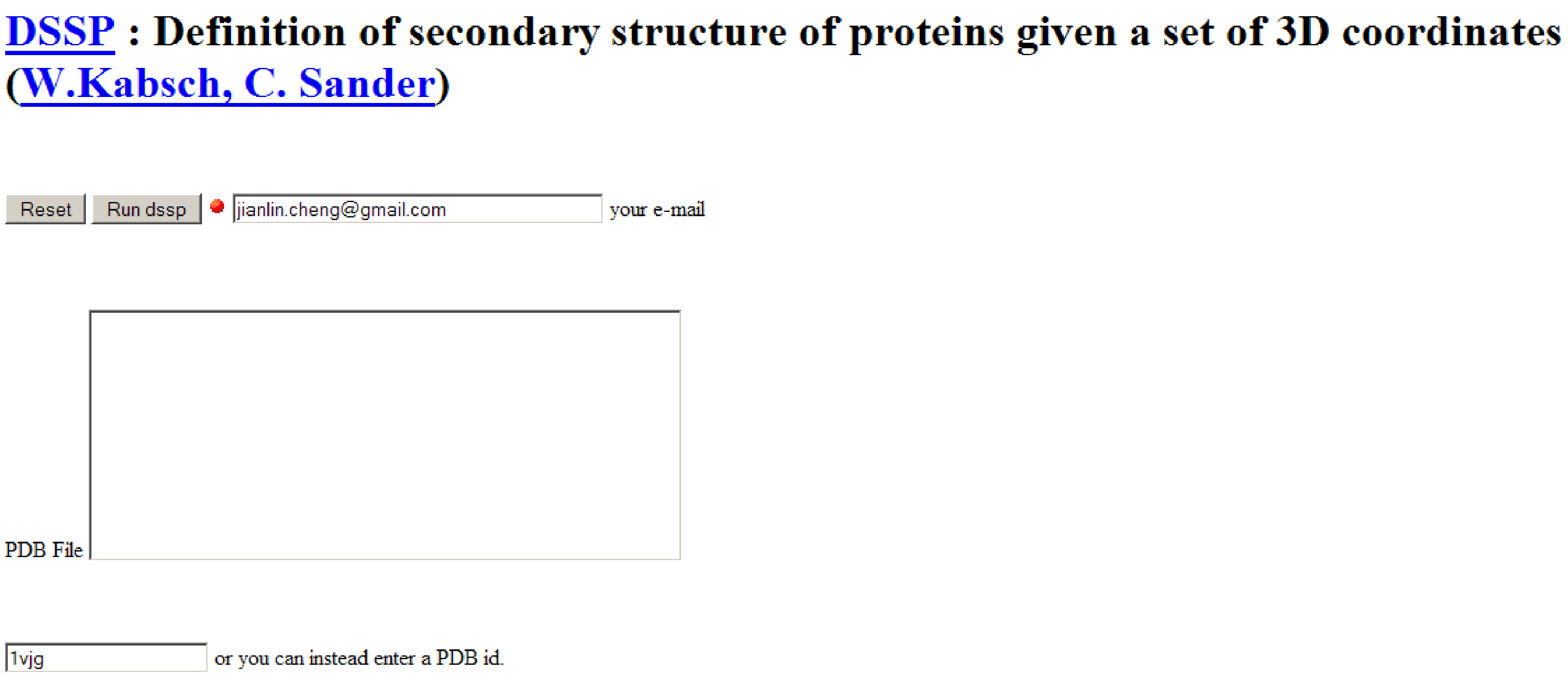
E = extended strand, participates in beta ladder

T = hydrogen bonded turn

S = bend

Blank = loop

DSSP Web Service



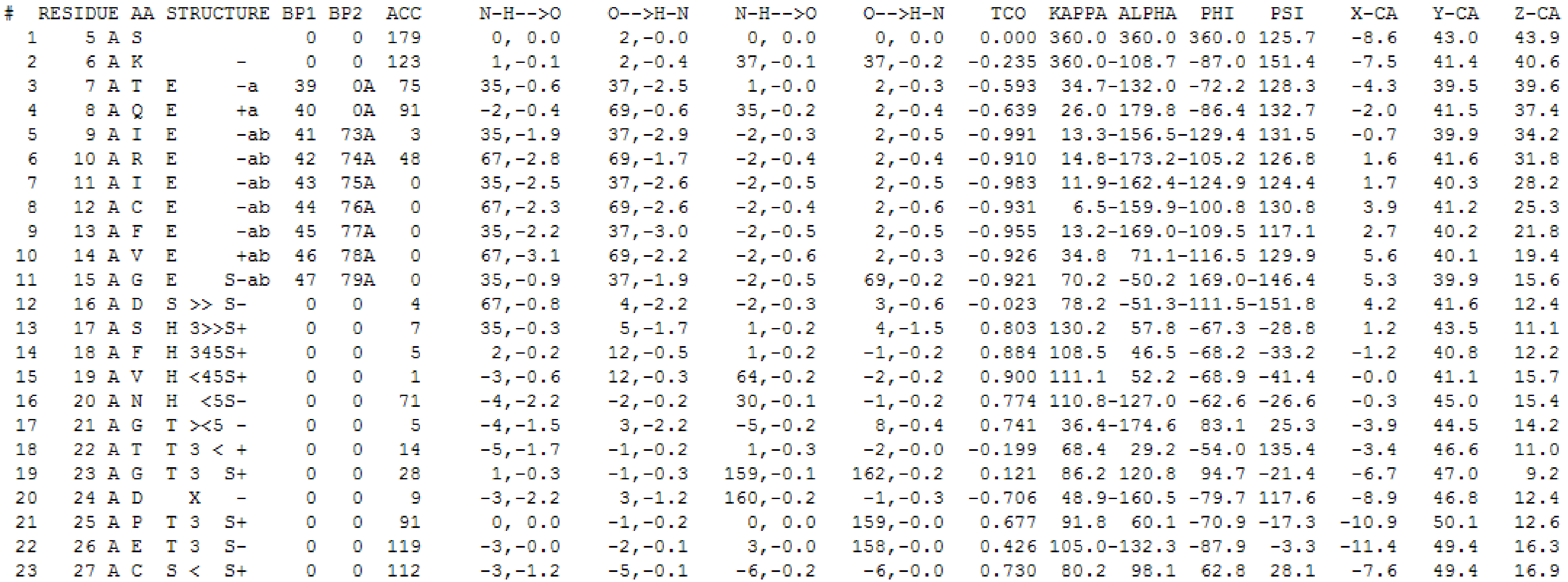
DSSP : Definition of secondary structure of proteins given a set of 3D coordinates (W.Kabsch, C. Sander)

● jianlin.cheng@gmail.com your e-mail

PDB File

1vjg or you can instead enter a PDB id.

http://bioweb.pasteur.fr/seqanal/interfaces/dssp-simple.html



```
  #  RESIDUE AA STRUCTURE BP1 BP2  ACC     N-H-->O    O-->H-N    N-H-->O    O-->H-N    TCO  KAPPA ALPHA  PHI   PSI    X-CA   Y-CA   Z-CA
    1    5 A S              0   0  179      0, 0.0     2,-0.0     0, 0.0     0, 0.0   0.000 360.0 360.0 360.0 125.7   -8.6   43.0   43.9
    2    6 A K        -     0   0  123      1,-0.1     2,-0.4    37,-0.1    37,-0.2  -0.235 360.0-108.7 -87.0 151.4   -7.5   41.4   40.6
    3    7 A T  E     -a   39   0A  75     35,-0.6    37,-2.5     1,-0.0     2,-0.3  -0.593  34.7-132.0 -72.2 128.3   -4.3   39.5   39.6
    4    8 A Q  E     +a   40   0A  91     -2,-0.4    69,-0.6    35,-0.2     2,-0.4  -0.639  26.0 179.8 -86.4 132.7   -2.0   41.5   37.4
    5    9 A I  E     -ab  41  73A   3     35,-1.9    37,-2.9    -2,-0.3     2,-0.5  -0.991  13.3-156.5-129.4 131.5   -0.7   39.9   34.2
    6   10 A R  E     -ab  42  74A  48     67,-2.8    69,-1.7    -2,-0.4     2,-0.4  -0.910  14.8-173.2-105.2 126.8    1.6   41.6   31.8
    7   11 A I  E     -ab  43  75A   0     35,-2.5    37,-2.6    -2,-0.5     2,-0.5  -0.983  11.9-162.4-124.9 124.4    1.7   40.3   28.2
    8   12 A C  E     -ab  44  76A   0     67,-2.3    69,-2.6    -2,-0.4     2,-0.6  -0.931   6.5-159.9-100.8 130.8    3.9   41.2   25.3
    9   13 A F  E     -ab  45  77A   0     35,-2.2    37,-3.0    -2,-0.5     2,-0.5  -0.955  13.2-169.0-109.5 117.1    2.7   40.2   21.8
   10   14 A V  E     +ab  46  78A   0     67,-3.1    69,-2.2    -2,-0.6     2,-0.3  -0.926  34.8  71.1-116.5 129.9    5.6   40.1   19.4
   11   15 A G  E    S-ab  47  79A   0     35,-0.9    37,-1.9    -2,-0.5    69,-0.2  -0.921  70.2 -50.2 169.0-146.4    5.3   39.9   15.6
   12   16 A D  S >> S-     0   0    4     67,-0.8     4,-2.2    -2,-0.3     3,-0.6  -0.023  78.2 -51.3-111.5-151.8    4.2   41.6   12.4
   13   17 A S  H 3>>S+     0   0    7     35,-0.3     5,-1.7     1,-0.2     4,-1.5   0.803 130.2  57.8 -67.3 -28.8    1.2   43.5   11.1
   14   18 A F  H 345S+     0   0    5      2,-0.2    12,-0.5     1,-0.2    -1,-0.2   0.884 108.5  46.5 -68.2 -33.2   -1.2   40.8   12.2
   15   19 A V  H <45S+     0   0    1     -3,-0.6    12,-0.3    64,-0.2    -2,-0.2   0.900 111.1  52.2 -68.9 -41.4   -0.0   41.1   15.7
   16   20 A N  H  <5S-     0   0   71     -4,-2.2    -2,-0.2    30,-0.1    -1,-0.2   0.774 110.8-127.0 -62.6 -26.6   -0.3   45.0   15.4
   17   21 A G  T ><5 -     0   0    5     -4,-1.5     3,-2.2    -5,-0.2     8,-0.4   0.741  36.4-174.6  83.1  25.3   -3.9   44.5   14.2
   18   22 A T  T 3 < +     0   0   14     -5,-1.7    -1,-0.2     1,-0.3    -2,-0.0  -0.199  68.4  29.2 -54.0 135.4   -3.4   46.6   11.0
   19   23 A G  T 3  S+     0   0   28      1,-0.3    -1,-0.3   159,-0.1   162,-0.2   0.121  86.2 120.8  94.7 -21.4   -6.7   47.0    9.2
   20   24 A D    X   -     0   0    9     -3,-2.2     3,-1.2   160,-0.2    -1,-0.3  -0.706  48.9-160.5 -79.7 117.6   -8.9   46.8   12.4
   21   25 A P  T 3  S+     0   0   91      0, 0.0    -1,-0.2     0, 0.0   159,-0.0   0.677  91.8  60.1 -70.9 -17.3  -10.9   50.1   12.6
   22   26 A E  T 3  S-     0   0  119     -3,-0.0    -2,-0.1     3,-0.0   158,-0.0   0.426 105.0-132.3 -87.9  -3.3  -11.4   49.4   16.3
   23   27 A C  S <  S+     0   0  112     -3,-1.2    -5,-0.1    -6,-0.2    -6,-0.0   0.730  80.2  98.1  62.8  28.1   -7.6   49.4   16.9
```

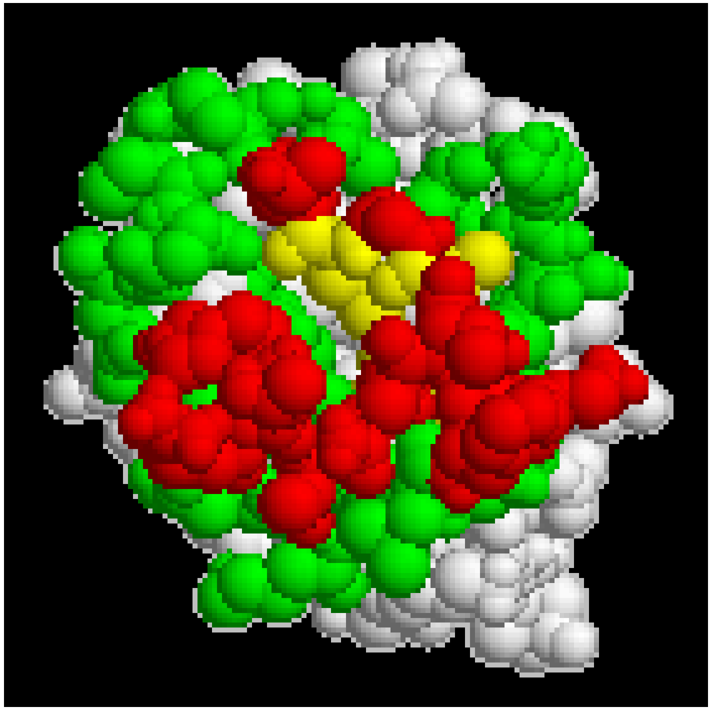
Amino Acids

Secondary Structure

Solvent Accessibility

Solvent Accessibility

Size of the area of an amino acid that is exposed to solvent (water).

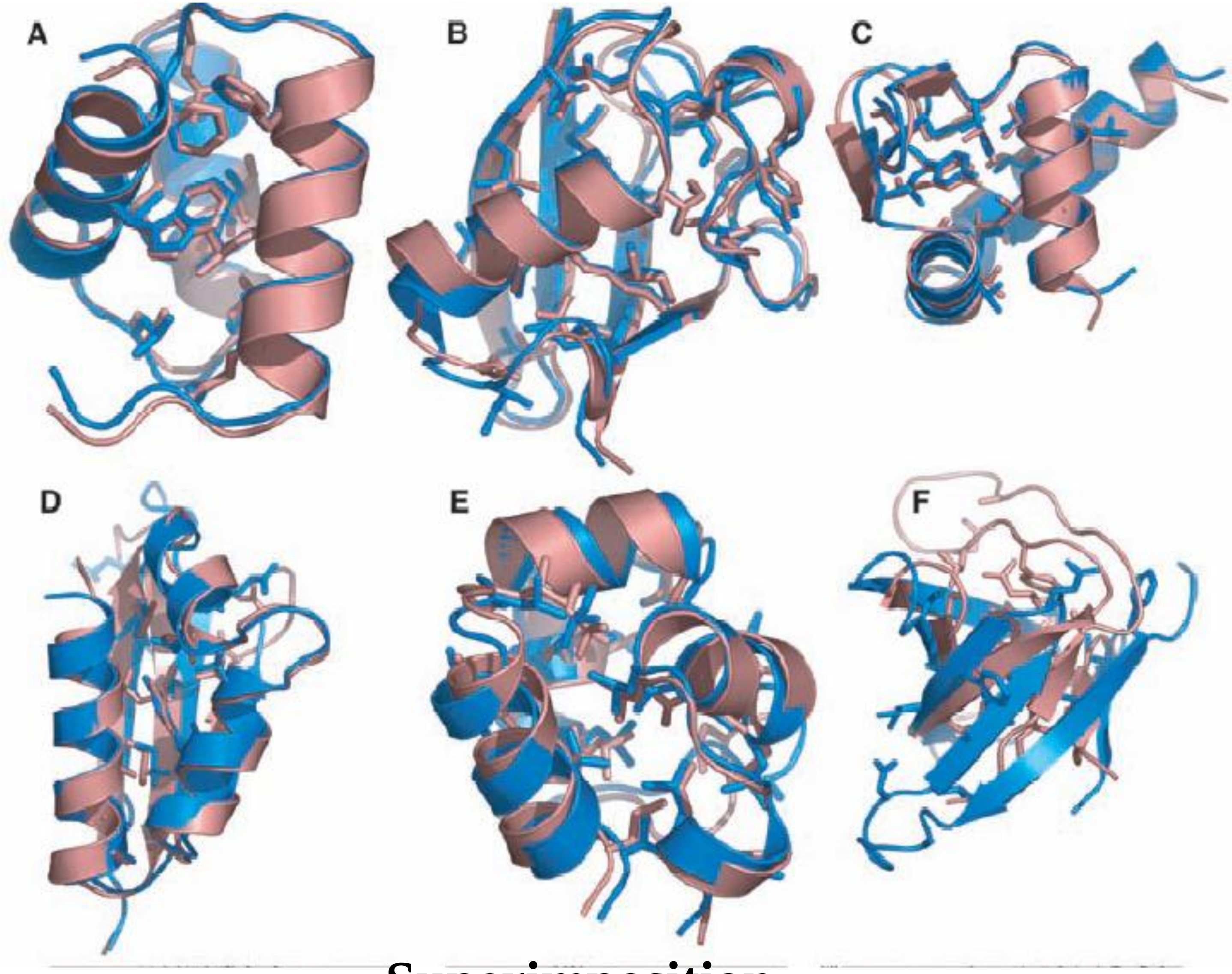


Maximum solvent accessible area for each amino acid is its whole surface area.

Hydrophobic residues like to be Buried inside (interior).
Hydrophilic residues like to be exposed on the surface.

Structure Comparison (Alignment)

- Are the structures of two protein similar?
- Are the two structure models of the same protein similar?
- Different measures (RMSD, GDT-TS (Zemla et al., 1999), MaxSub (Siew et al., 2000), TM score (Zhang and Skolnick,2005))

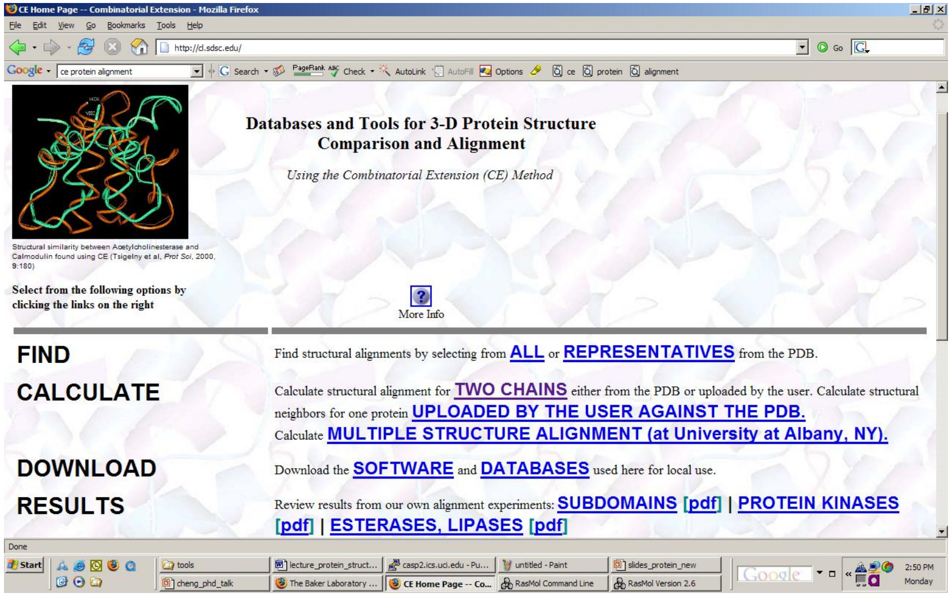


Superimposition

David Baker, 2005

Useful Structure Alignment Tools

- CE (http://cl.sdsc.edu/)
- DALI (http://www.ebi.ac.uk/dali/)
- TM-Align: http://zhang.bioinformatics.ku.edu/TM-align/



Databases and Tools for 3-D Protein Structure Comparison and Alignment

Using the Combinatorial Extension (CE) Method

Structural similarity between Acetylcholinesterase and Calmodulin found using CE (Tsigelny et al, *Prot Sci*, 2000, 9:180)

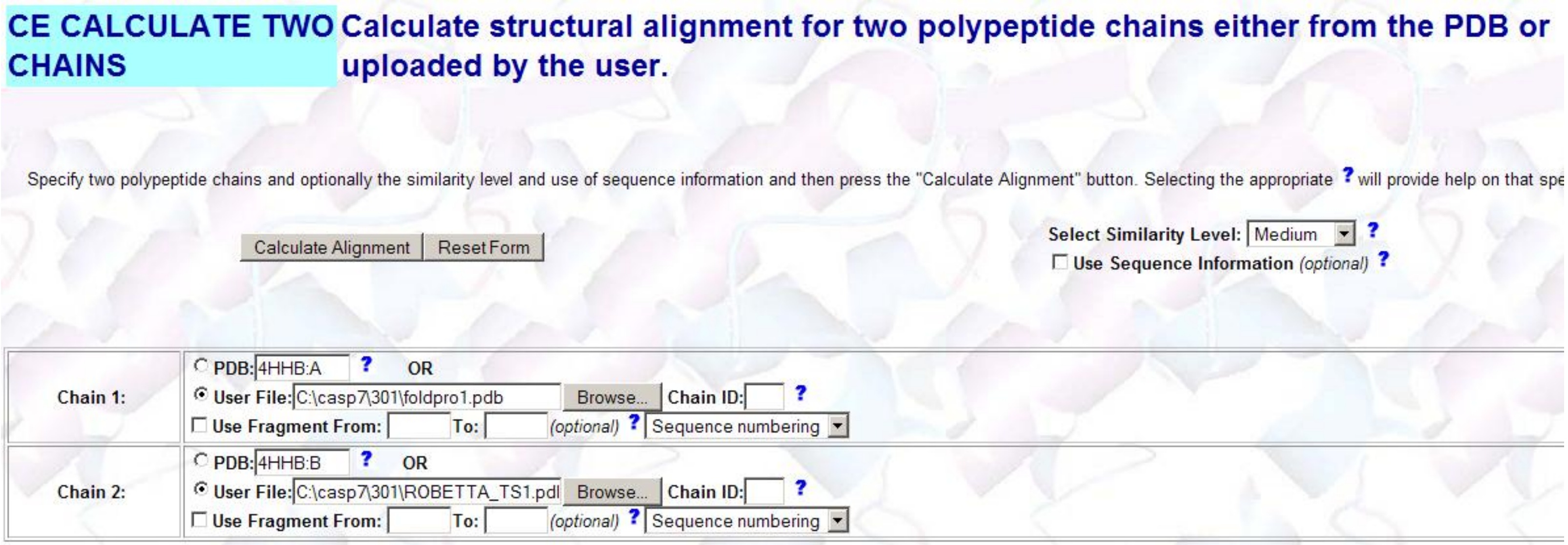
Select from the following options by clicking the links on the right

FIND Find structural alignments by selecting from **ALL** or **REPRESENTATIVES** from the PDB.

CALCULATE Calculate structural alignment for **TWO CHAINS** either from the PDB or uploaded by the user. Calculate structural neighbors for one protein **UPLOADED BY THE USER AGAINST THE PDB.**
Calculate **MULTIPLE STRUCTURE ALIGNMENT (at University at Albany, NY).**

DOWNLOAD Download the **SOFTWARE** and **DATABASES** used here for local use.

RESULTS Review results from our own alignment experiments: **SUBDOMAINS** **[pdf]** | **PROTEIN KINASES** **[pdf]** | **ESTERASES, LIPASES** **[pdf]**

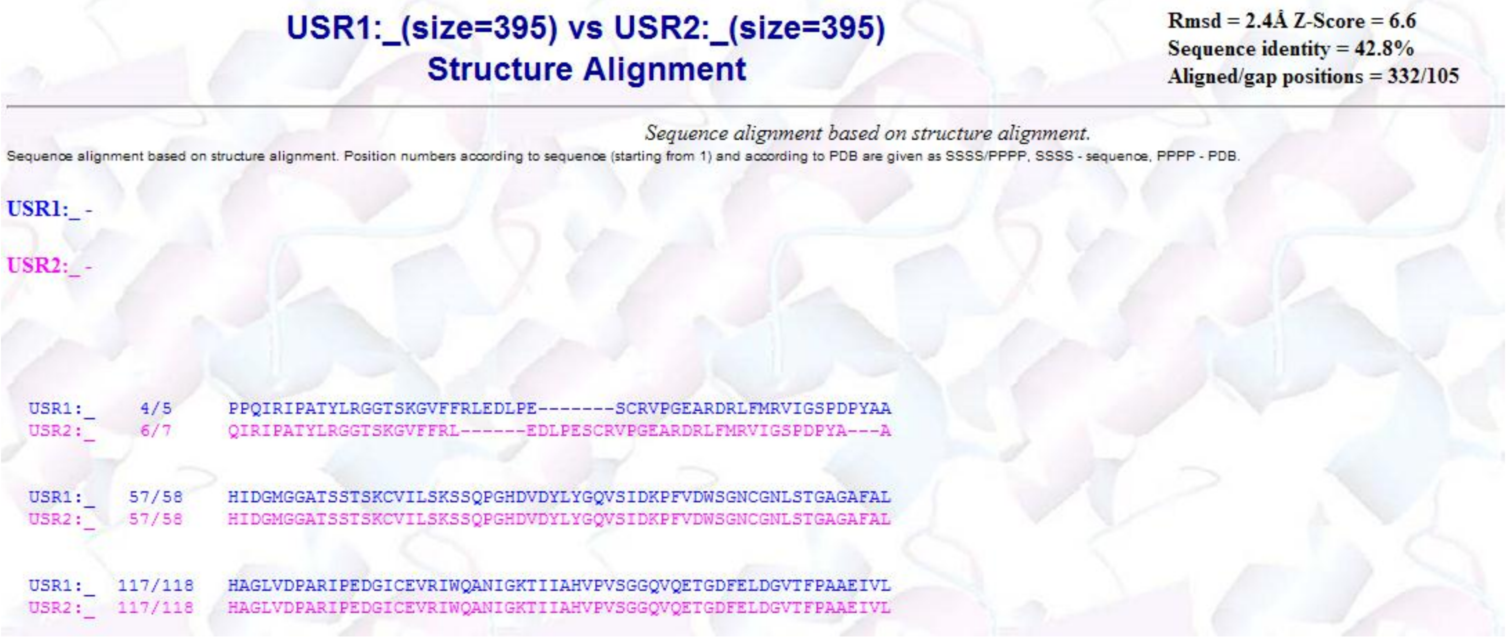


CE CALCULATE TWO CHAINS Calculate structural alignment for two polypeptide chains either from the PDB or uploaded by the user.

Specify two polypeptide chains and optionally the similarity level and use of sequence information and then press the "Calculate Alignment" button. Selecting the appropriate ? will provide help on that spe

Select Similarity Level: Medium ?
□ Use Sequence Information *(optional)* ?

Chain 1:	○ PDB: 4HHB:A ? OR ● User File: C:\casp7\301\foldpro1.pdb Browse... Chain ID: ? □ Use Fragment From: To: *(optional)* ? Sequence numbering
Chain 2:	○ PDB: 4HHB:B ? OR ● User File: C:\casp7\301\ROBETTA_TS1.pdl Browse... Chain ID: ? □ Use Fragment From: To: *(optional)* ? Sequence numbering



USR1:_(size=395) vs USR2:_(size=395)
Structure Alignment

Rmsd = 2.4Å Z-Score = 6.6
Sequence identity = 42.8%
Aligned/gap positions = 332/105

Sequence alignment based on structure alignment.

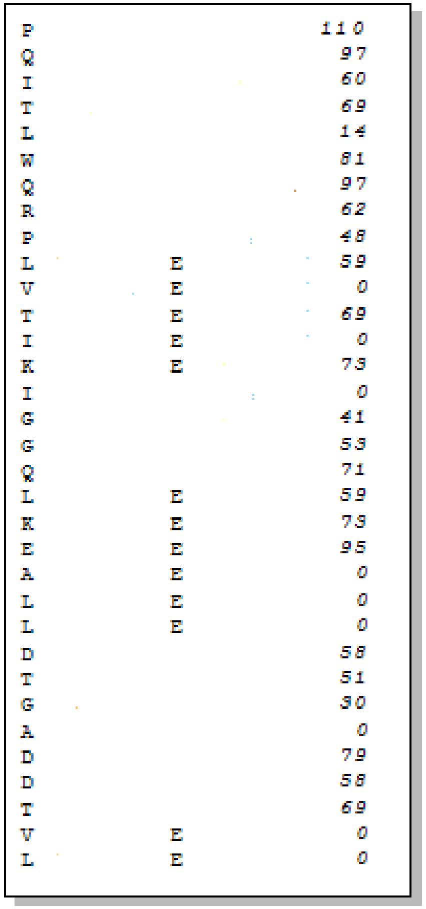
Sequence alignment based on structure alignment. Position numbers according to sequence (starting from 1) and according to PDB are given as SSSS/PPPP, SSSS - sequence, PPPP - PDB.

USR1:_ -

USR2:_ -

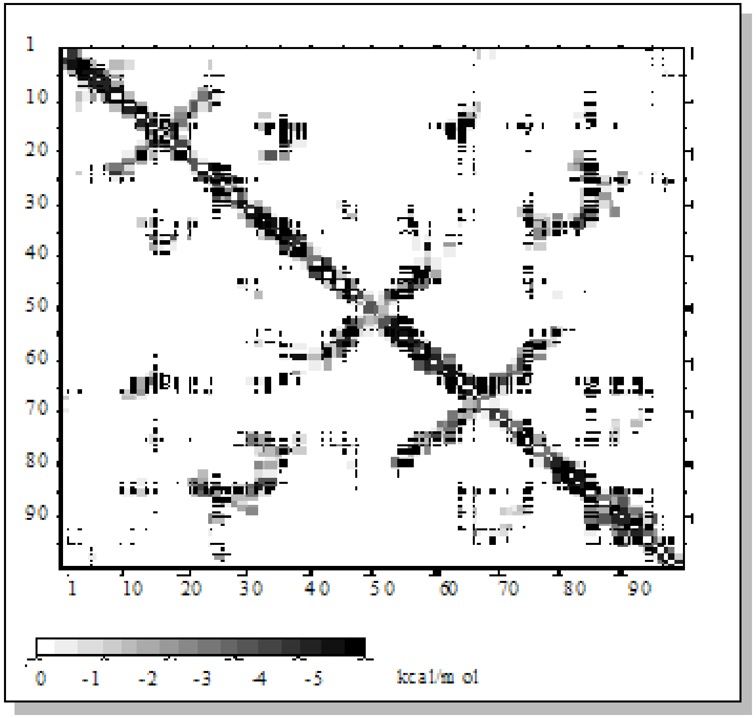
```
USR1:_    4/5     PPQIRIPATYLRGGTSKGVFFRLEDLPE-------SCRVPGEARDRLFMRVIGSPDPYAA
USR2:_    6/7     QIRIPATYLRGGTSKGVFFRL------EDLPESCRVPGEARDRLFMRVIGSPDPYA---A

USR1:_   57/58    HIDGMGGATSSTSKCVILSKSSQPGHDVDYLYGQVSIDKPFVDWSGNCGNLSTGAGAFAL
USR2:_   57/58    HIDGMGGATSSTSKCVILSKSSQPGHDVDYLYGQVSIDKPFVDWSGNCGNLSTGAGAFAL

USR1:_  117/118   HAGLVDPARIPEDGICEVRIWQANIGKTIIAHVPVSGGQVQETGDFELDGVTFPAAEIVL
USR2:_  117/118   HAGLVDPARIPEDGICEVRIWQANIGKTIIAHVPVSGGQVQETGDFELDGVTFPAAEIVL
```

Notation: protein structure 1D, 2D, 3D

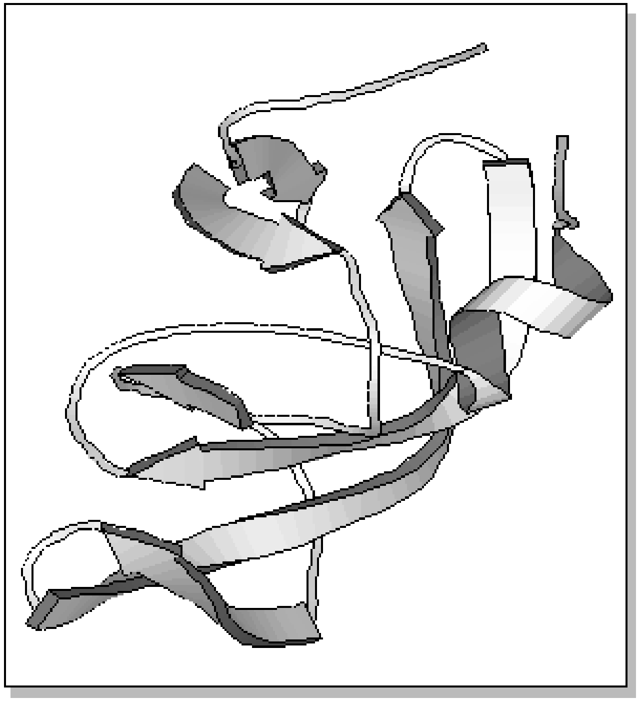


P		110
Q		97
I		60
T		69
L		14
W		81
Q		97
R		62
P		48
L	E	59
V	E	0
T	E	69
I	E	0
K	E	73
I		0
G		41
G		53
Q		71
L	E	59
K	E	73
E	E	95
A	E	0
L	E	0
L	E	0
D		58
T		51
G		30
A		0
D		79
D		58
T		69
V	E	0
L	E	0

1D



2D



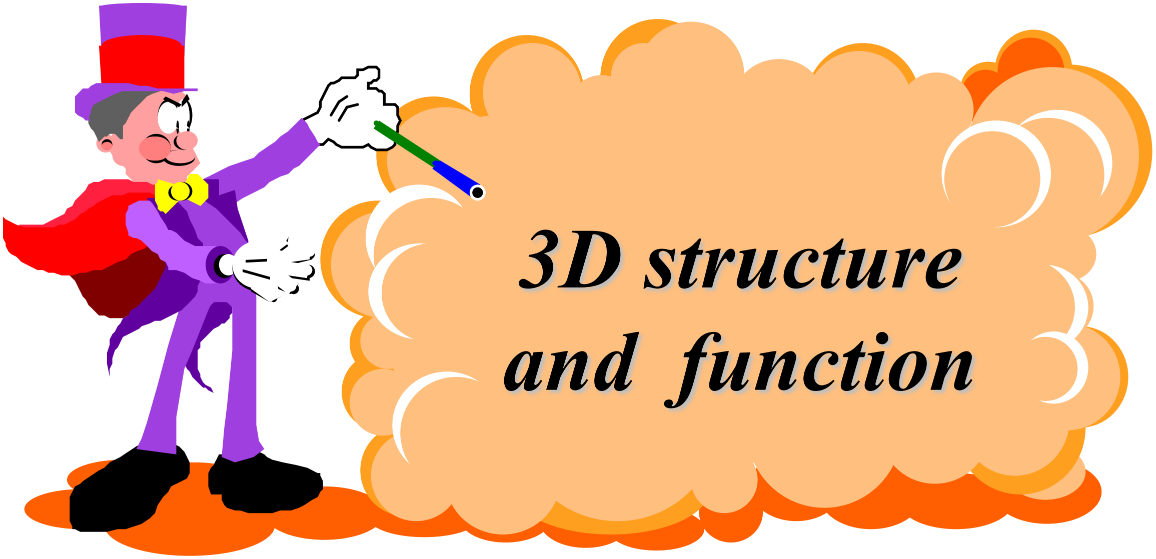
3D

B. Rost, 2005

Goal of structure prediction

- Epstein & Anfinsen, 1961:
  sequence uniquely determines structure

- INPUT: sequence
- OUTPUT:



B. Rost, 2005

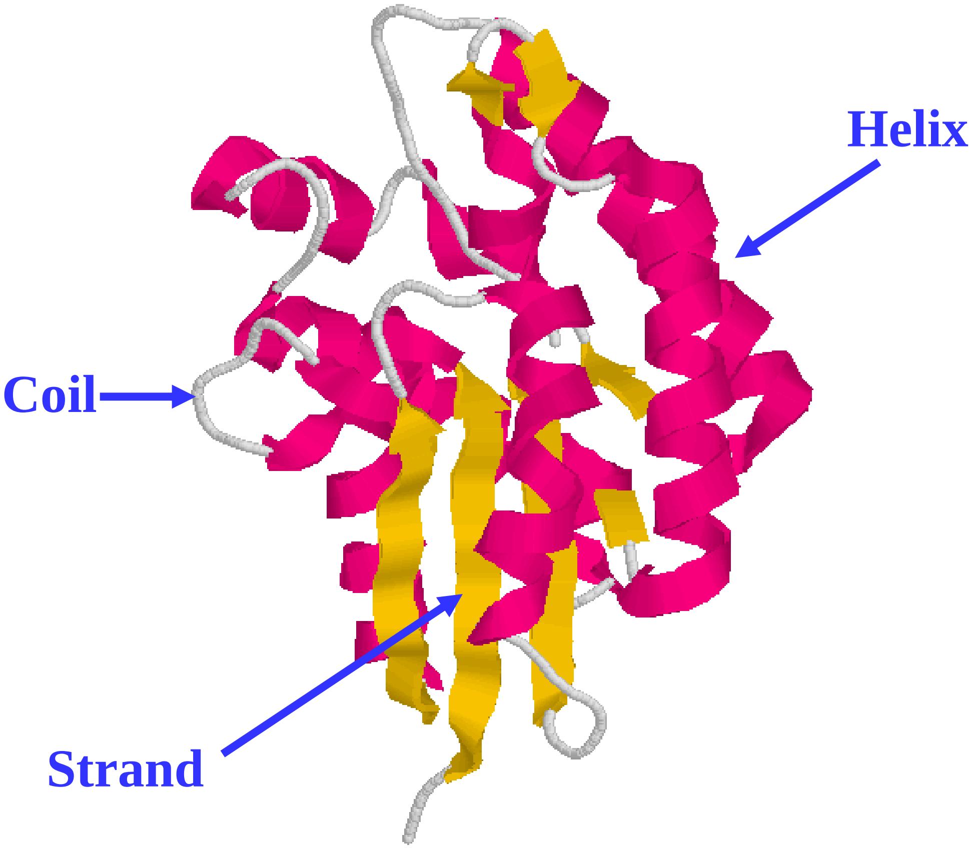
CASP – Olympics of Protein Structure Prediction

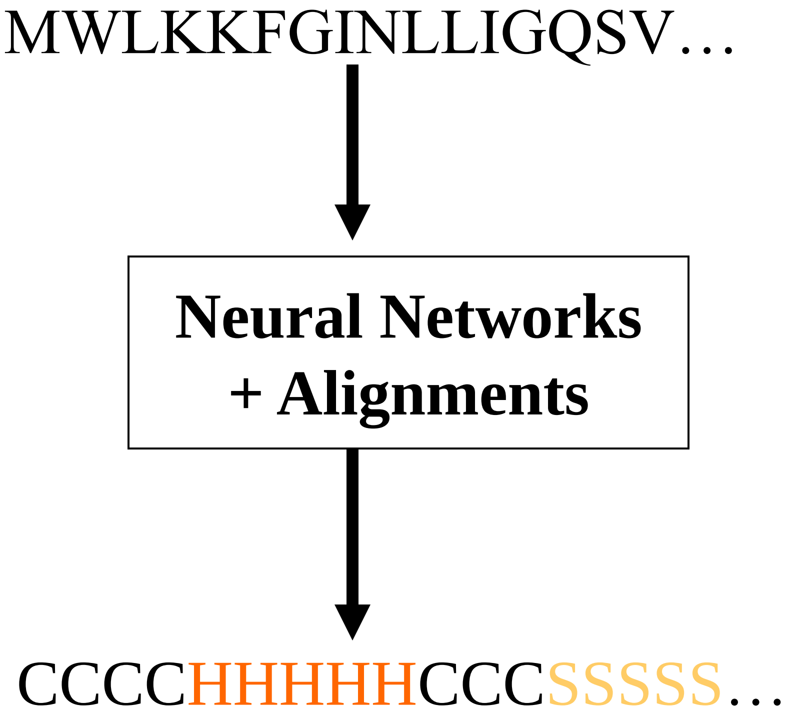
- Critical Assessment of Techniques of Protein Structure Prediction
- 1994,1996,1998,2000,2002,2004,2006, 2008, 2010
- Blind Test, Independent Evaluation
- CASP9: 116 targets

1D Structure Prediction

- Secondary structure
- Solvent accessibility
- Disordered regions
- Domain boundary

1D: Secondary Structure Prediction




MWLKKFGINLLIGQSV…

Neural Networks + Alignments

CCCCHHHHHCCCSSSSS…

Cheng, Randall, Sweredoski, Baldi. ***Nucleic Acid Research*****, 2005**

Widely Used Tools (~78-80%)

SSpro 4.1: http://sysbio.rnet.missouri.edu/multicom_toolbox/

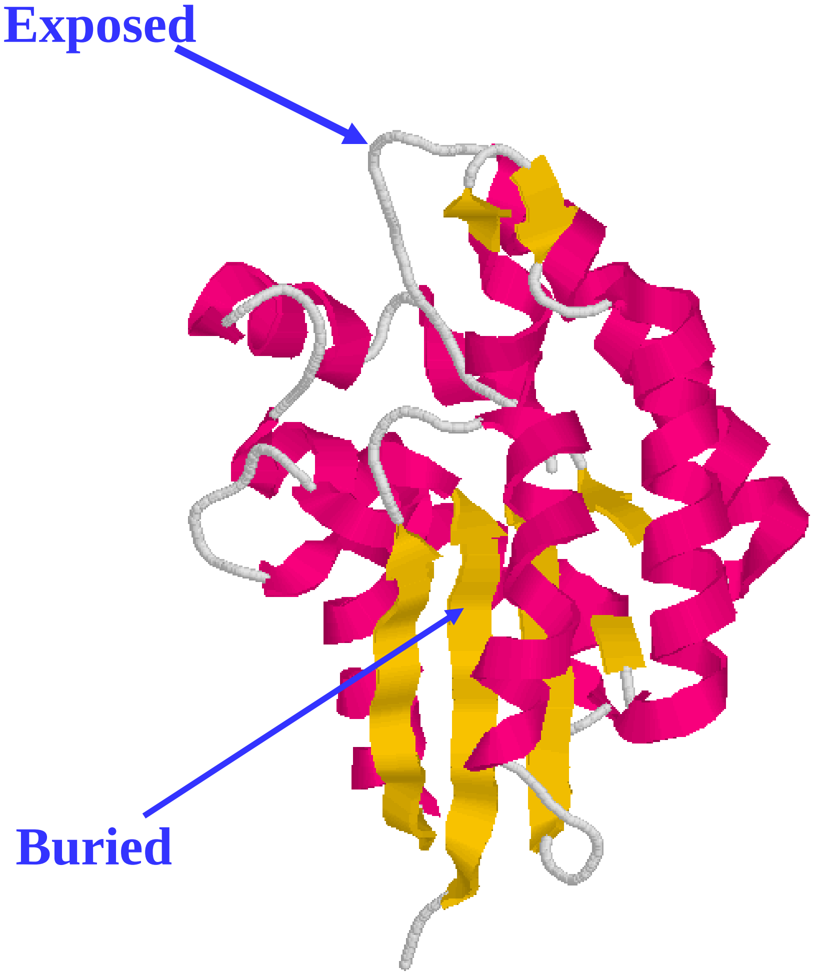
Distill: http://distill.ucd.ie/porter/

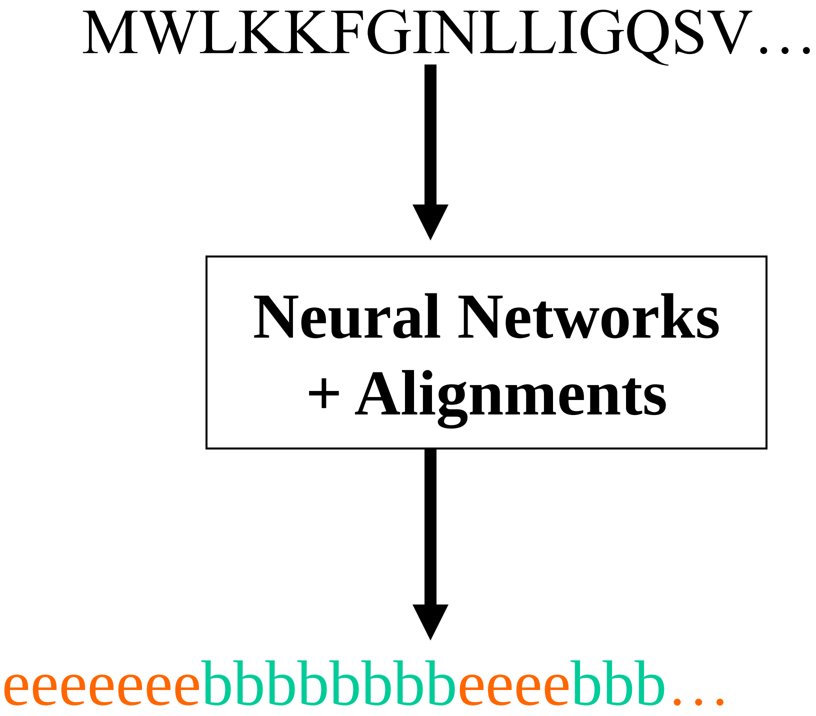
PSI-PRED: http://bioinf.cs.ucl.ac.uk/psipred/psiform.html
software is also available

SAM: http://compbio.soe.ucsc.edu/SAM_T08/T08-query.html

PHD: http://www.predictprotein.org/

1D: Solvent Accessibility Prediction





MWLKKFGINLLIGQSV…

Neural Networks + Alignments

eeeeeeebbbbbbbeeeebbb…

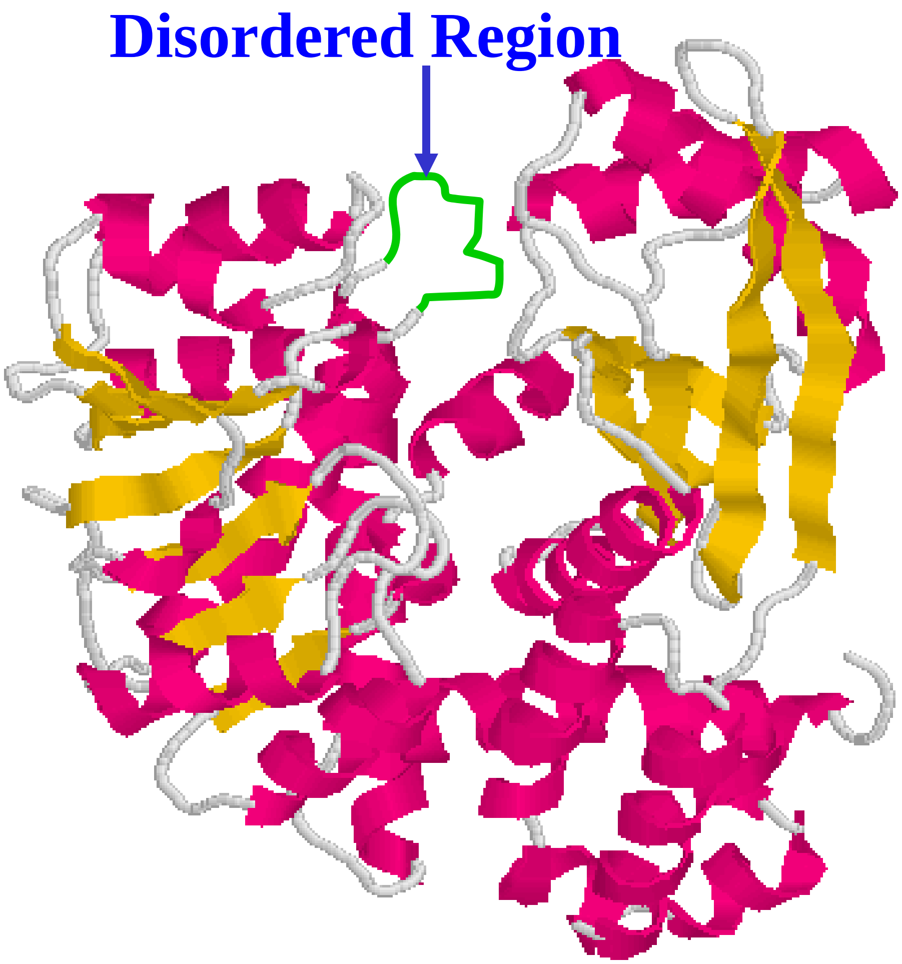
Accuracy: 79% at 25% threshold

Cheng, Randall, Sweredoski, Baldi. ***Nucleic Acid Research*****, 2005**

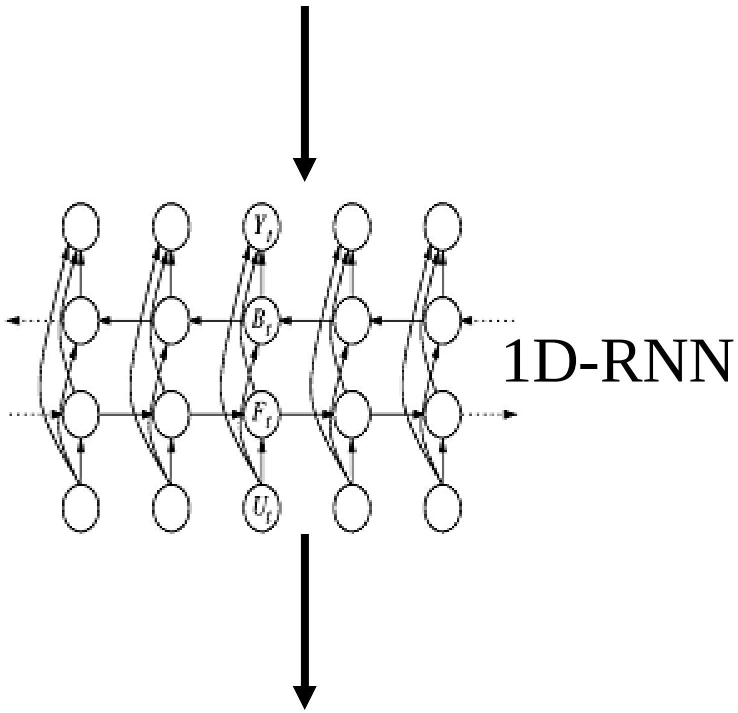
Widely Used Tools (78%)

- ACCpro 4.1: software: http://sysbio.rnet.missouri.edu/multicom_toolbox/
- SCRATCH: http://scratch.proteomics.ics.uci.edu/
- PHD: http://www.predictprotein.org/
- Distill: http://distill.ucd.ie/porter/

1D: Disordered Region Prediction Using Neural Networks



MWLKKFGINLLIGQSV…



OOOOODDDDOOOOO…

93% TP at 5% FP

Deng, Eickholt, Cheng. BMC Bioinformatics, 2009

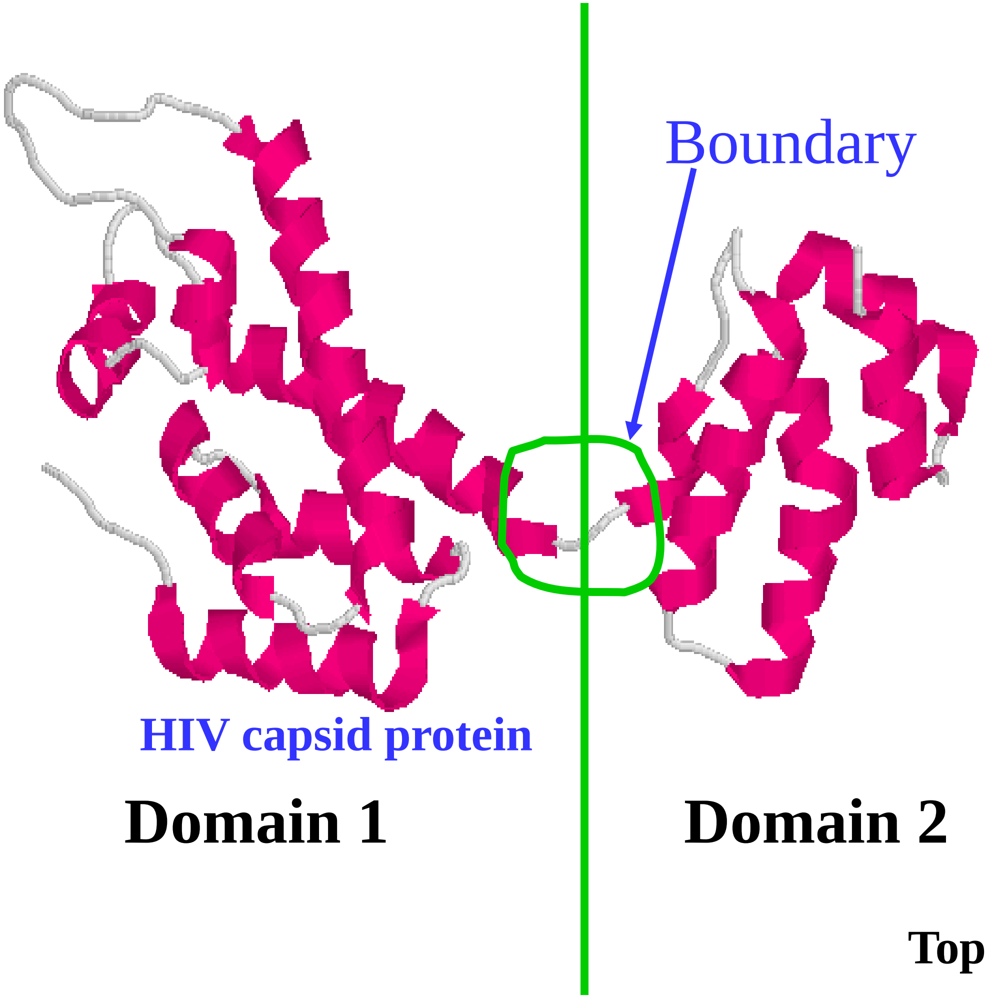
Tools

PreDisorder: http://sysbio.rnet.missouri.edu/multicom_toolbox/

A collection of disorder predictors:
http://www.disprot.org/predictors.php

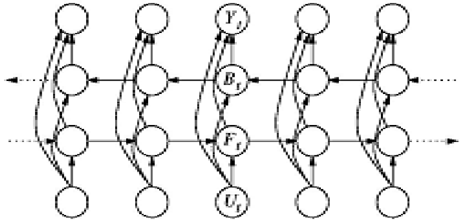
Deng, Eickholt, Cheng. BMC Bioinformatics, 2009

1D: Protein Domain Prediction Using Neural Networks



MWLKKFGINLLIGQSV…

+ SS and SA



1D-RNN

NNNNNNNBBBBBNNNN…

Inference/Cut

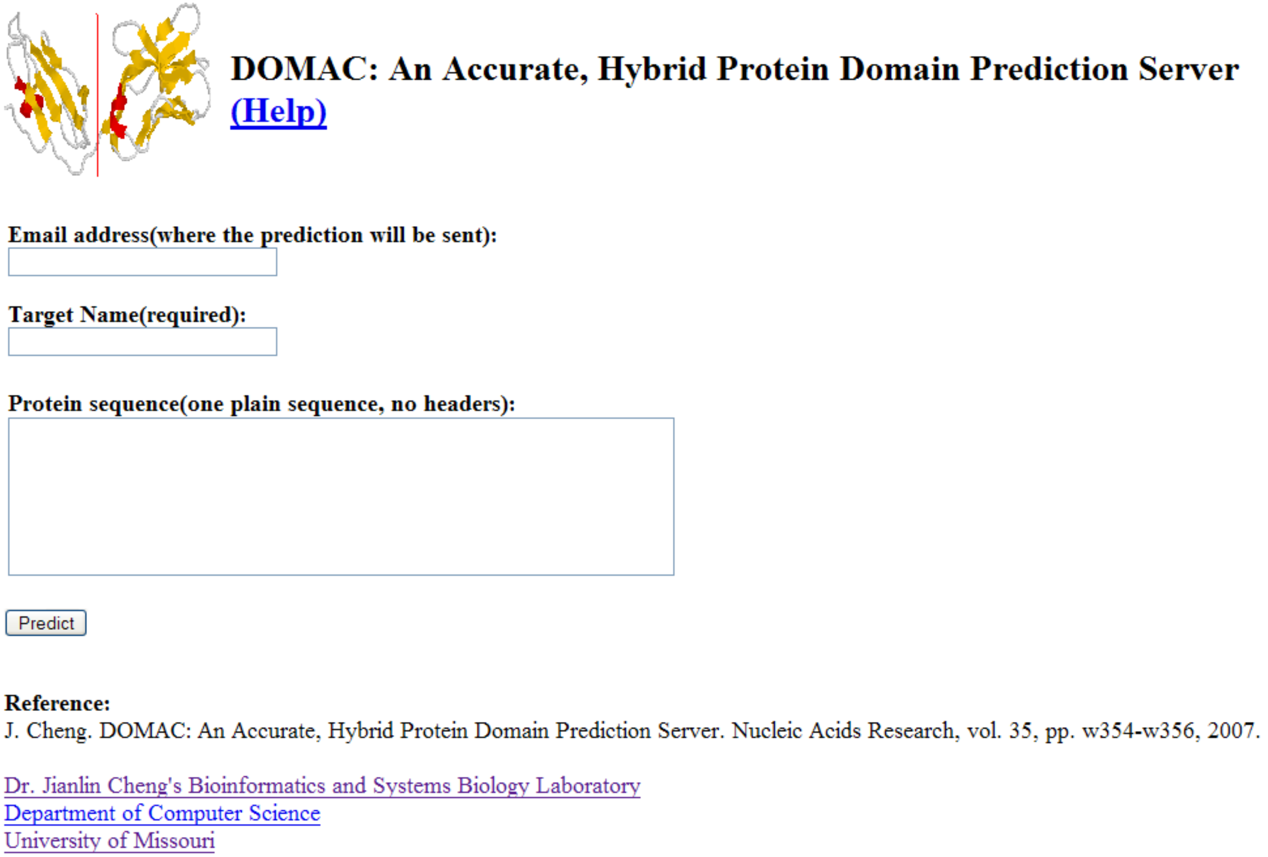
Domains

Top *ab-initio* domain predictor in CAFASP4

Cheng, Sweredoski, Baldi. *Data Mining and Knowledge Discovery*, 2006.

Tools

DOMAC: http://casp.rnet.missouri.net/domac.html



DOMAC: An Accurate, Hybrid Protein Domain Prediction Server
(Help)

Email address(where the prediction will be sent):

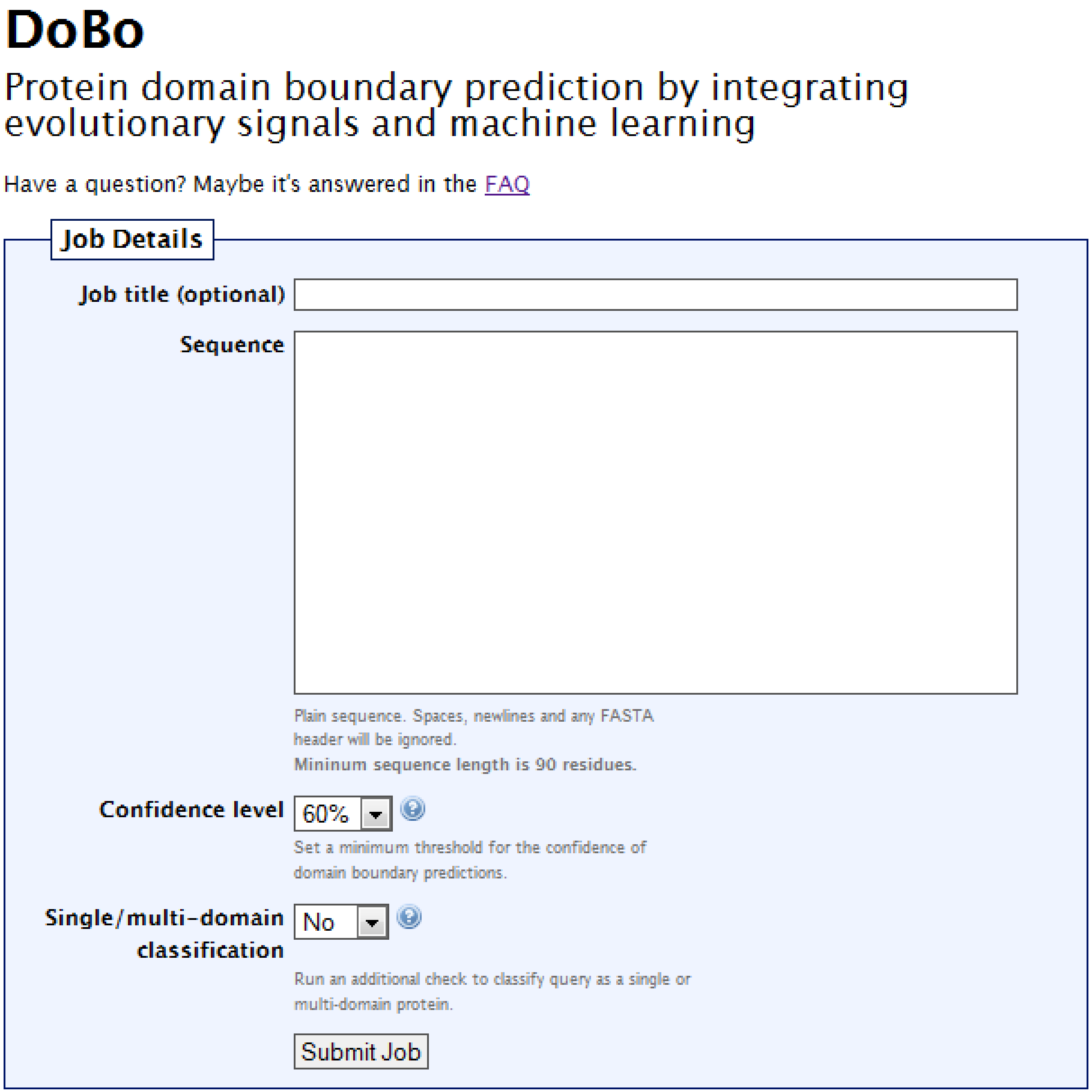
Target Name(required):

Protein sequence(one plain sequence, no headers):

Reference:
J. Cheng. DOMAC: An Accurate, Hybrid Protein Domain Prediction Server. Nucleic Acids Research, vol. 35, pp. w354-w356, 2007.

Dr. Jianlin Cheng's Bioinformatics and Systems Biology Laboratory
Department of Computer Science
University of Missouri

Cheng, Nucleic Acids Research, 2007



DoBo

Protein domain boundary prediction by integrating evolutionary signals and machine learning

Have a question? Maybe it's answered in the FAQ

Job Details

Job title (optional)

Sequence

Plain sequence. Spaces, newlines and any FASTA header will be ignored.
Mininum sequence length is 90 residues.

Confidence level 60%

Set a minimum threshold for the confidence of domain boundary predictions.

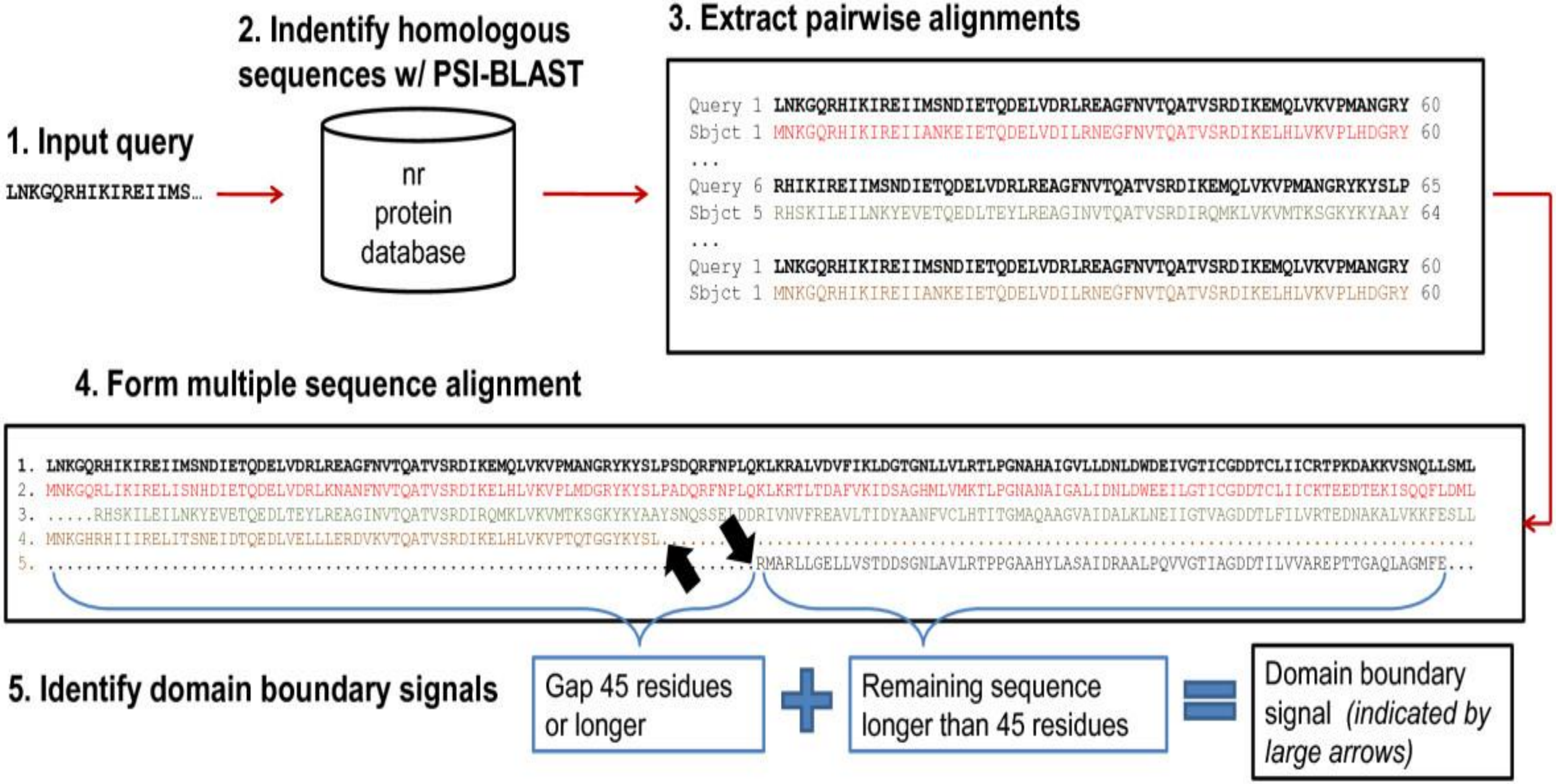
Single/multi-domain classification No

Run an additional check to classify query as a single or multi-domain protein.

Web: http://sysbio.rnet.missouri.edu/multicom_toolbox/index.html

Reference:

J. Eickholt, X. Deng, and J. Cheng. **DoBo: Protein Domain Boundary Prediction by Integrating Evolutionary Signals and Machine Learning.** ***BMC Bioinformatics***. 12:43, 2011.



1. Input query

```
LNKGQRHIKIREIIMS…
```

2. Indentify homologous sequences w/ PSI-BLAST

3. Extract pairwise alignments

```
Query 1 LNKGQRHIKIREIIMSNDIETQDELVDRLREAGFNVTQATVSRDIKEMQLVKVPMANGRY 60
Sbjct 1 MNKGQRHIKIREIIANKEIETQDELVDILRNEGFNVTQATVSRDIKELHLVKVPLHDGRY 60
...
Query 6 RHIKIREIIMSNDIETQDELVDRLREAGFNVTQATVSRDIKEMQLVKVPMANGRYKYSLP 65
Sbjct 5 RHSKILEILNKYEVETQEDLTEYLREAGINVTQATVSRDIRQMKLVKVMTKSGKYKYAAY 64
...
Query 1 LNKGQRHIKIREIIMSNDIETQDELVDRLREAGFNVTQATVSRDIKEMQLVKVPMANGRY 60
Sbjct 1 MNKGQRHIKIREIIANKEIETQDELVDILRNEGFNVTQATVSRDIKELHLVKVPLHDGRY 60
```

4. Form multiple sequence alignment

```
1. LNKGQRHIKIREIIMSNDIETQDELVDRLREAGFNVTQATVSRDIKEMQLVKVPMANGRYKYSLPSDQRFNPLQKLKRALVDVFIKLDGTGNLLVLRTLPGNAHAIGVLLDNLDWDEIVGTICGDDTCLIICRTPKDAKKVSNQLLSML
2. MNKGQRLIKIRELISNHDIETQDELVDRLKNANFNVTQATVSRDIKELHLVKVPLMDGRYKYSLPADQRFNPLQKLKRTLTDAFVKIDSAGHMLVMKTLPGNANAIGALIDNLDWEEILGTICGDDTCLIICKTEEDTEKISQQFLDML
3. .....RHSKILEILNKYEVETQEDLTEYLREAGINVTQATVSRDIRQMKLVKVMTKSGKYKYAAYSNQSSELDDRIVNVFREAVLTIDYAANFVCLHTITGMAQAAGVAIDALKLNEIIGTVAGDDTLFILVRTEDNAKALVKKFESLL
4. MNKGHRHIIIRELITSNEIDTQEDLVELLLERDVKVTQATVSRDIKELHLVKVPTQTGGYKYSL.....................................................................................
5. ..........................................................................RMARLLGELLVSTDDSGNLAVLRTPPGAAHYLASAIDRAALPQVVGTIAGDDTILVVAREPTTGAQLAGMFE...
```

5. Identify domain boundary signals

Gap 45 residues or longer + Remaining sequence longer than 45 residues = Domain boundary signal *(indicated by large arrows)*

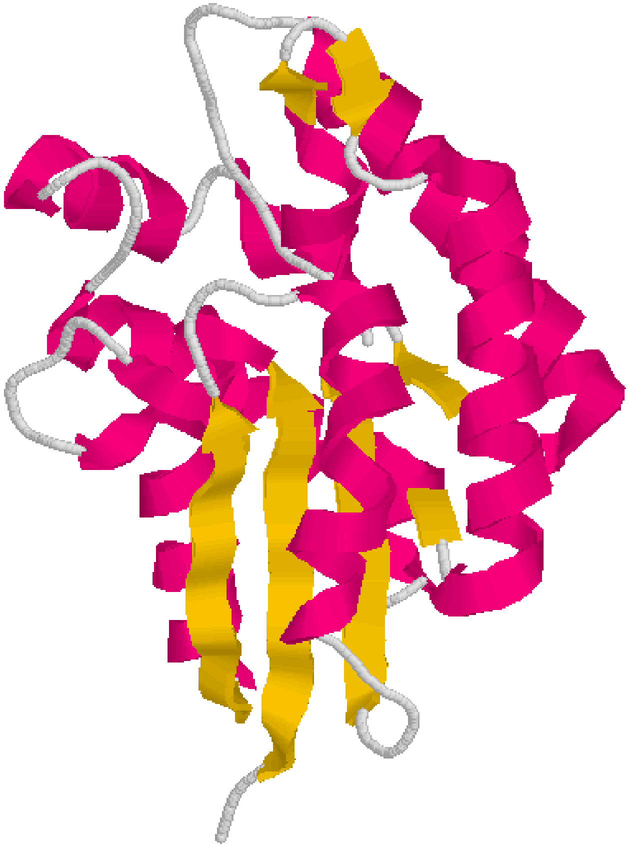
Project 1

- Predict Secondary Structure, Solvent Accessibility, Disorder Regions of soybean transcription factors

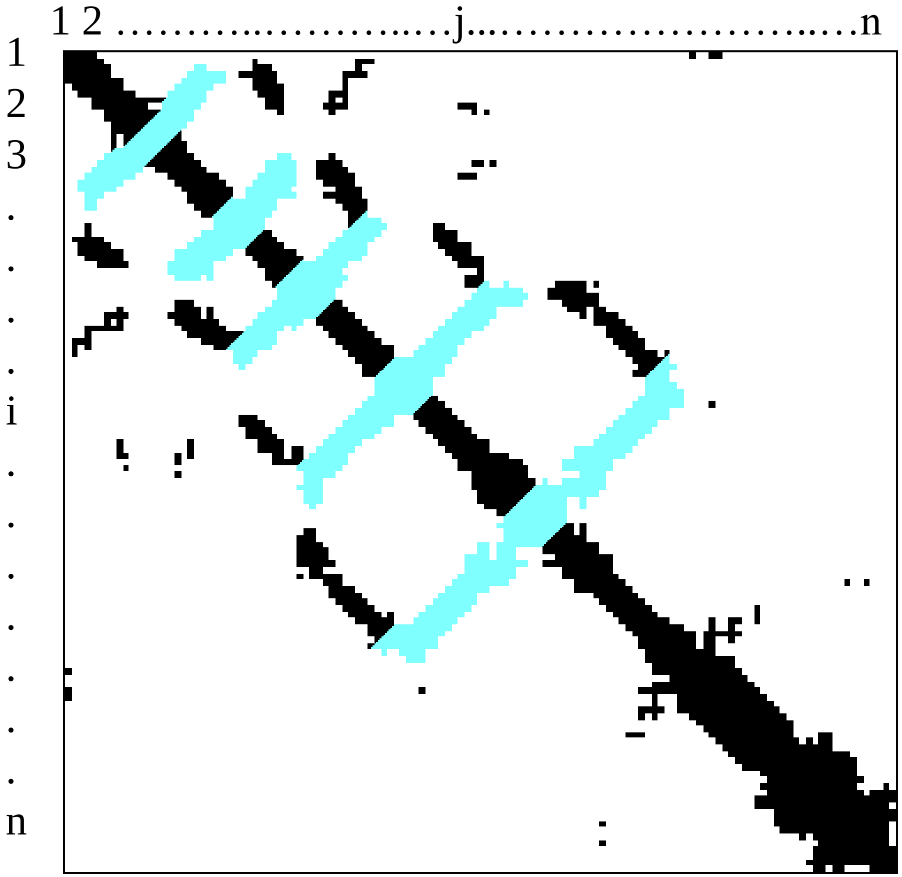
- Data: http://casp.rnet.missouri.edu/marc/muii_7005/SEQ_TFP_90_2500.txt
- Select **10** proteins to make predictions

2D: Contact Map Prediction

3D Structure



2D Contact Map



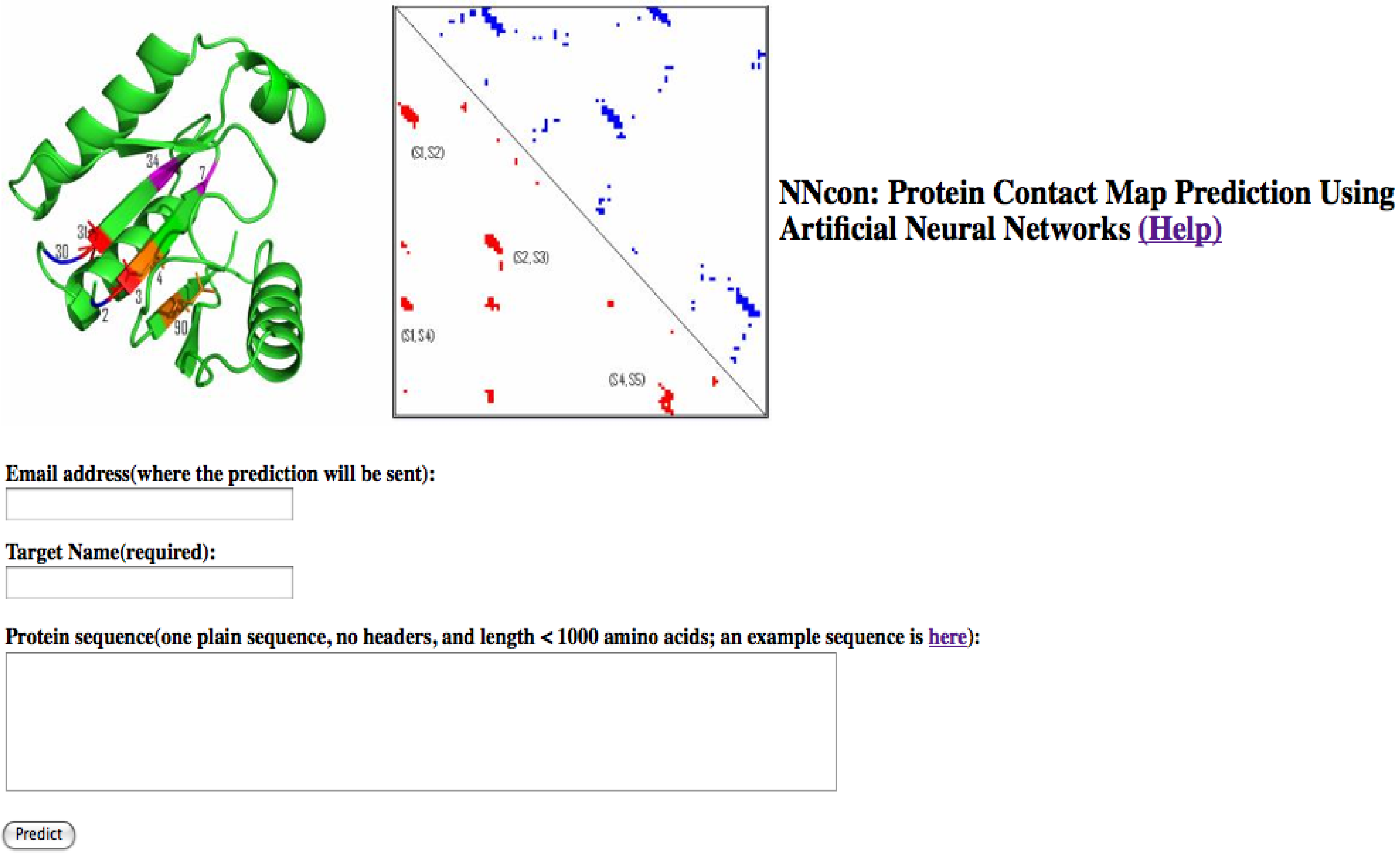
Distance Threshold = 8Aº

Cheng, Randall, Sweredoski, Baldi. *Nucleic Acid Research*, 2005

Contact Prediction

- SVMcon: http://casp.rnet.missouri.edu/svmcon.html
- NNcon:

  http://casp.rnet.missouri.edu/nncon.html
- SCRATCH: http://scratch.proteomics.ics.uci.edu/
- SAM: http://compbio.soe.ucsc.edu/HMM-apps/HMM-applications.html



NNcon: Protein Contact Map Prediction Using Artificial Neural Networks (Help)

Email address(where the prediction will be sent):

Target Name(required):

Protein sequence(one plain sequence, no headers, and length < 1000 amino acids; an example sequence is here):

Tegge, Wang, Eickholt, Cheng, Nucleic Acids Research, 2009

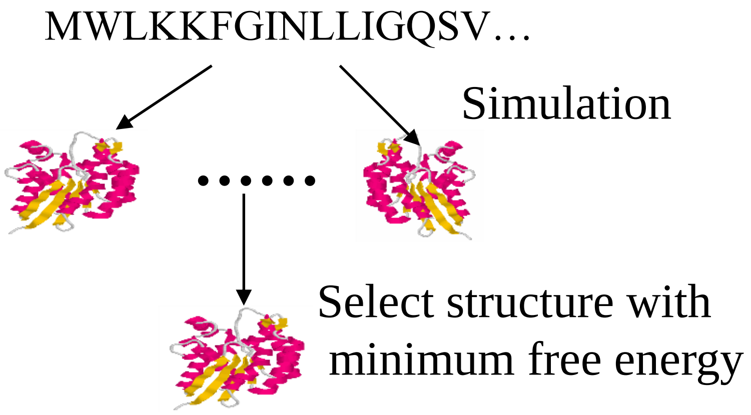
Two Methodologies for 3D Structure Prediction

- AB Initio Method (physical-chemical principles / molecular dynamics, knowledge-based approaches)
- Template-Based Method (knowledge-based approaches)

Two Approaches

- **Ab Initio Structure Prediction**

  Physical force field – protein folding
  Contact map - reconstruction



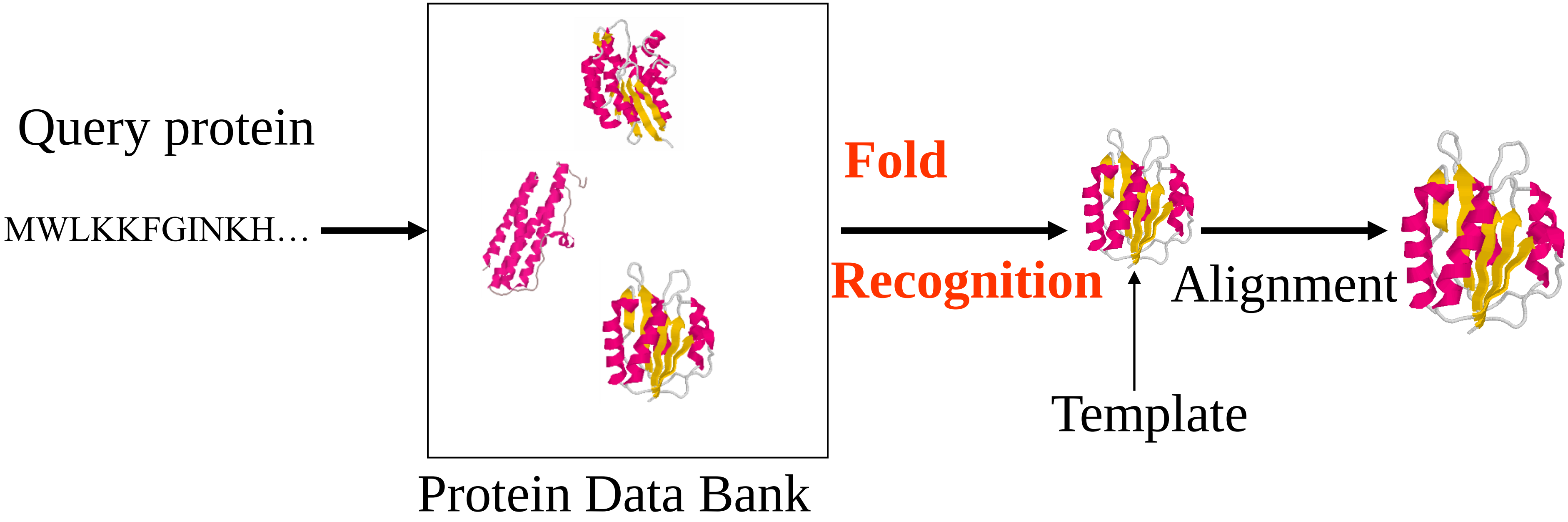
MWLKKFGINLLIGQSV…

Simulation

……

Select structure with
minimum free energy

- **Template-Based Structure Prediction**



Query protein

MWLKKFGINKH…

Fold

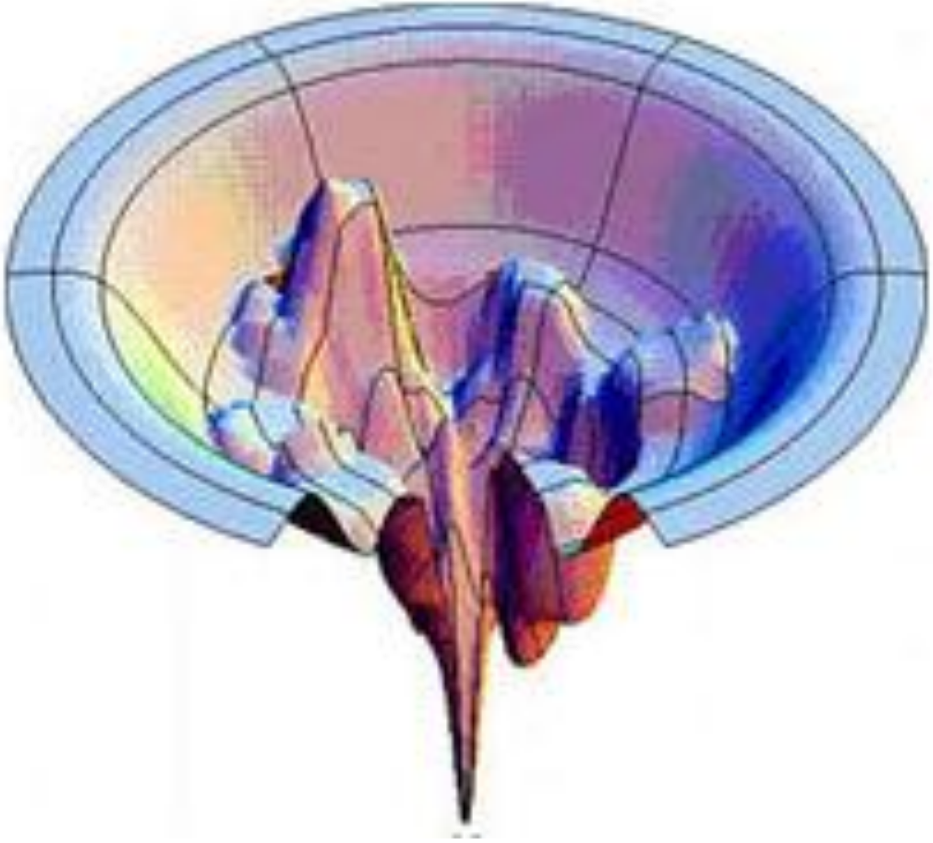
Recognition

Alignment

Template

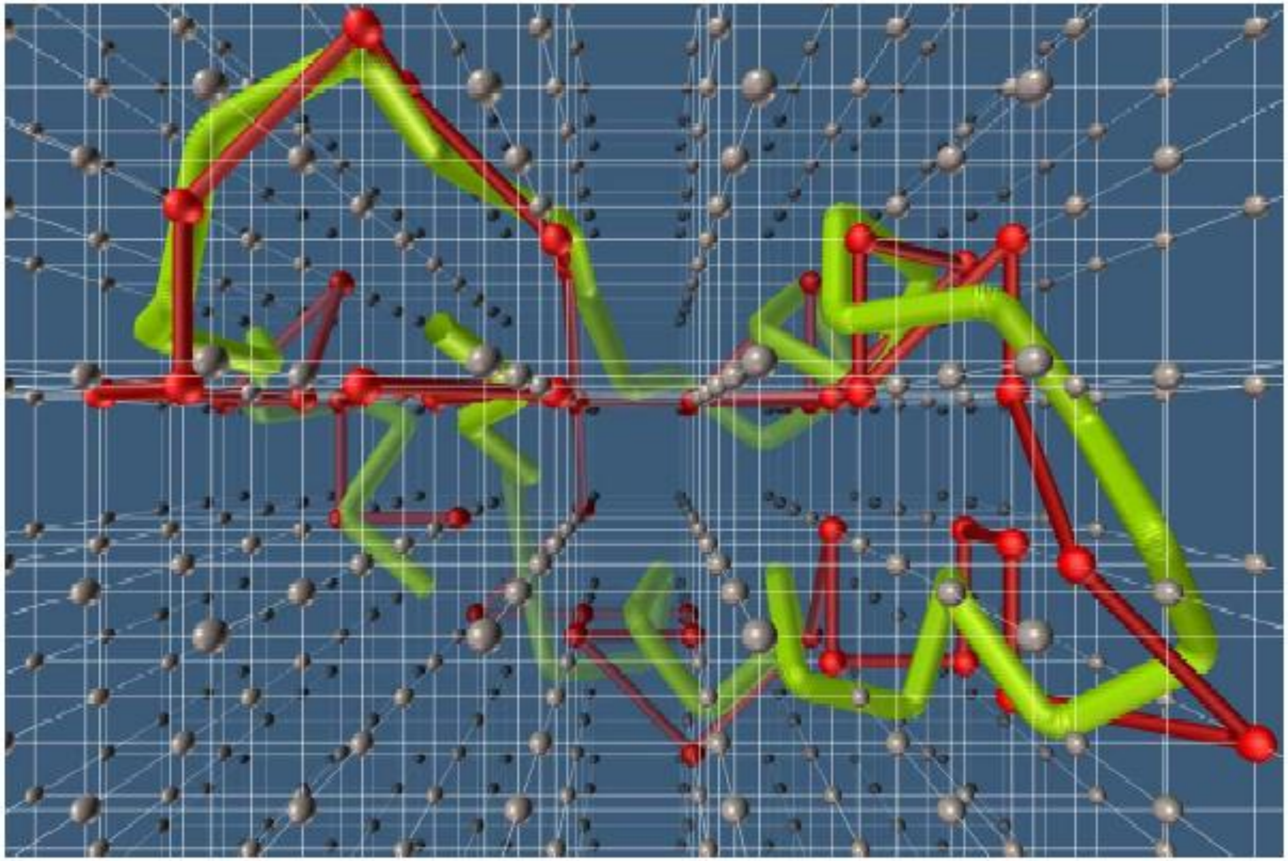
Protein Data Bank

Protein Energy Landscape



C. Park, 2005

Markov Chain Monte Carlo Simulation



picture from http://www.scripps.edu/pub/olson-web/people/sanner/html/lat_gallery.html

Template-Based Structure Prediction

1. Template identification
2. Query-template alignment
3. Model generation
4. Model evaluation
5. Model refinement

Notes: if template is easy to identify, it is often called **comparative Modeling or homology** modeling. If template is hard to identify, it is often called **fold recognition**.

TARGET

TEMPLATE

ASILPKRLFGNCEQTSDEGLK
IERTPLVPHISAQNVCLKIDD
VPERLIPERASFQWMNDK

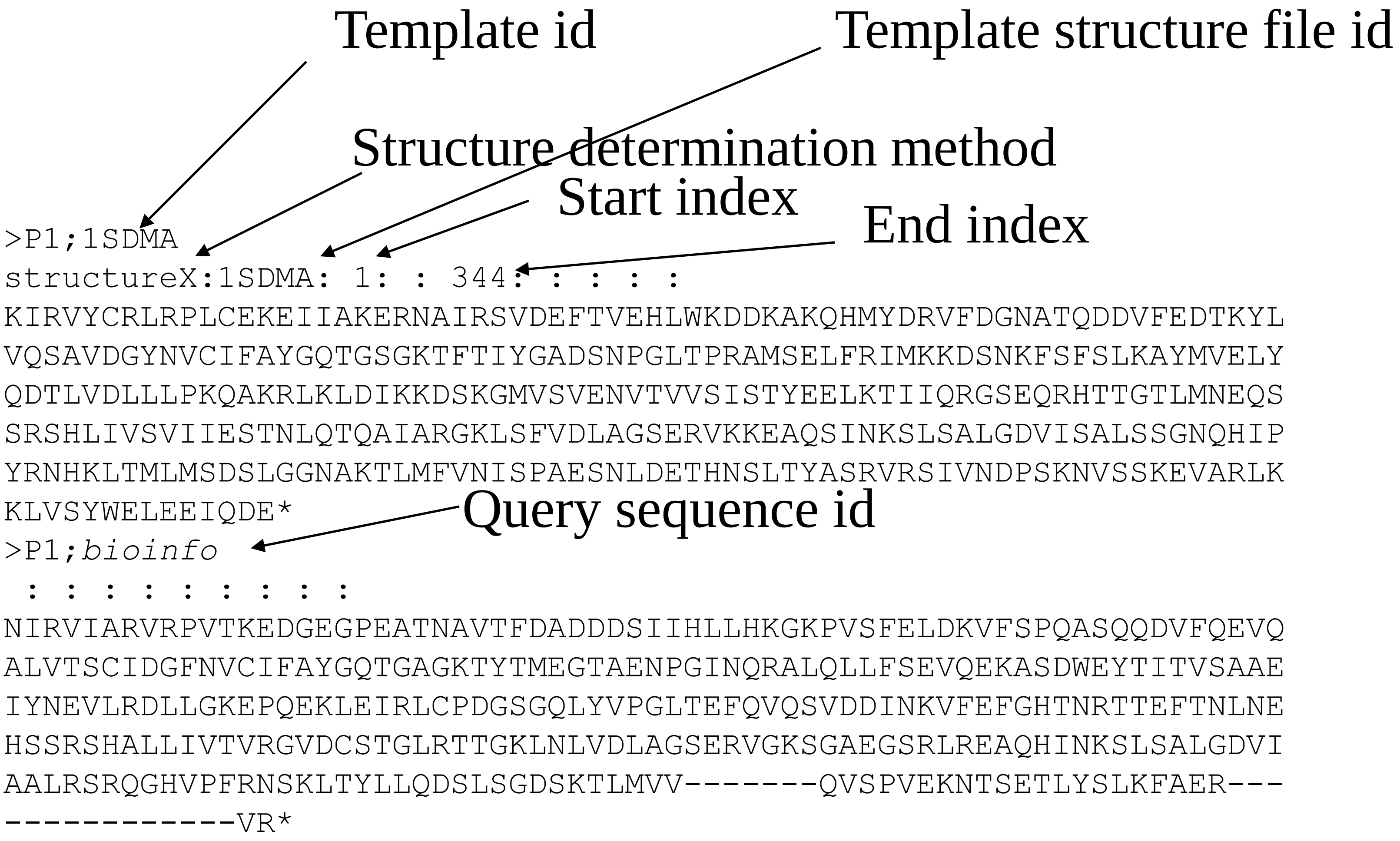
```
ASILPKRLFGNCEQTSDEGLKIERTPLVPHISAQNVCLKIDDVPERLIPE
MSVIPKRLYGNCEQTSEEAIRIEDSPIV---TADLVCLKIDEIPERLVGE
```

A. Fisher, 2005

Modeller

- Need an alignment file between query and template sequence in the PIR format
- Need the structure (atom coordinates) file of template protein
- You need to write a simple script (Python for version 8.2) to tell how to generate the model and where to find the alignment file and template structure file.
- Run Modeller on the script. Modeller will automatically copy coordinates and make necessary adjustments to generate a model.
- See project step 5-8 for more details.

An PIR Alignment Example



Template id

Template structure file id

Structure determination method

Start index

End index

Query sequence id

```
>P1;1SDMA
structureX:1SDMA: 1: : 344: : : : :
KIRVYCRLRPLCEKEIIAKERNAIRSVDEFTVEHLWKDDKAKQHMYDRVFDGNATQDDVFEDTKYL
VQSAVDGYNVCIFAYGQTGSGKTFTIYGADSNPGLTPRAMSELFRIMKKDSNKFSFSLKAYMVELY
QDTLVDLLLPKQAKRLKLDIKKDSKGMVSVENVTVVSISTYEELKTIIQRGSEQRHTTGTLMNEQS
SRSHLIVSVIIESTNLQTQAIARGKLSFVDLAGSERVKKEAQSINKSLSALGDVISALSSGNQHIP
YRNHKLTMLMSDSLGGNAKTLMFVNISPAESNLDETHNSLTYASRVRSIVNDPSKNVSSKEVARLK
KLVSYWELEEIQDE*
>P1;bioinfo
 : : : : : : : : :
NIRVIARVRPVTKEDGEGPEATNAVTFDADDDSIIHLLHKGKPVSFELDKVFSPQASQQDVFQEVQ
ALVTSCIDGFNVCIFAYGQTGAGKTYTMEGTAENPGINQRALQLLFSEVQEKASDWEYTITVSAAE
IYNEVLRDLLGKEPQEKLEIRLCPDGSGQLYVPGLTEFQVQSVDDINKVFEFGHTNRTTEFTNLNE
HSSRSHALLIVTVRGVDCSTGLRTTGKLNLVDLAGSERVGKSGAEGSRLREAQHINKSLSALGDVI
AALRSRQGHVPFRNSKLTYLLQDSLSGDSKTLMVV-------QVSPVEKNTSETLYSLKFAER---
------------VR*
```

Structure File Example (1SDMA.atm)

```
ATOM      1  N   LYS     1      -3.978  26.298 113.043  1.00 31.75           N
ATOM      2  CA  LYS     1      -4.532  25.067 113.678  1.00 31.58           C
ATOM      3  C   LYS     1      -5.805  25.389 114.448  1.00 30.38           C
ATOM      4  O   LYS     1      -6.887  24.945 114.072  1.00 32.68           O
ATOM      5  CB  LYS     1      -3.507  24.446 114.631  1.00 34.97           C
ATOM      6  CG  LYS     1      -3.743  22.970 114.942  1.00 36.49           C
ATOM      7  CD  LYS     1      -3.886  22.172 113.644  1.00 39.52           C
ATOM      8  CE  LYS     1      -3.318  20.766 113.761  1.00 41.58           C
ATOM      9  NZ  LYS     1      -1.817  20.761 113.756  1.00 43.48           N
ATOM     10  N   ILE     2      -5.687  26.161 115.522  1.00 26.16           N
ATOM     11  CA  ILE     2      -6.867  26.500 116.302  1.00 22.75           C
ATOM     12  C   ILE     2      -7.887  27.226 115.439  1.00 21.35           C
ATOM     13  O   ILE     2      -7.565  28.200 114.770  1.00 20.95           O
ATOM     14  CB  ILE     2      -6.513  27.377 117.523  1.00 21.68           C
ATOM     15  CG1 ILE     2      -5.701  26.563 118.526  1.00 21.13           C
ATOM     16  CG2 ILE     2      -7.782  27.875 118.200  1.00 18.96           C
ATOM     17  CD1 ILE     2      -5.368  27.325 119.787  1.00 21.39           C
ATOM     18  N   ARG     3      -9.120  26.737 115.461  1.00 22.04           N
ATOM     19  CA  ARG     3     -10.214  27.327 114.693  1.00 23.95           C
ATOM     20  C   ARG     3     -10.783  28.563 115.400  1.00 22.82           C
ATOM     21  O   ARG     3     -10.771  28.645 116.629  1.00 22.62           O
ATOM     22  CB  ARG     3     -11.327  26.290 114.510  1.00 26.34           C
ATOM     23  CG  ARG     3     -11.351  25.586 113.161  1.00 30.68           C
ATOM     24  CD  ARG     3     -10.004  25.034 112.771  1.00 35.43           C
ATOM     25  NE  ARG     3     -10.104  24.072 111.672  1.00 43.37           N
ATOM     26  CZ  ARG     3     -10.575  24.350 110.458  1.00 46.04           C
ATOM     27  NH1 ARG     3     -10.997  25.572 110.168  1.00 48.68           N
ATOM     28  NH2 ARG     3     -10.627  23.400 109.532  1.00 48.37           N
ATOM     29  N   VAL     4     -11.278  29.524 114.630  1.00 20.49           N
ATOM     30  CA  VAL     4     -11.853  30.724 115.225  1.00 17.59           C
ATOM     31  C   VAL     4     -13.082  31.211 114.471  1.00 18.31           C
ATOM     32  O   VAL     4     -13.030  31.446 113.264  1.00 16.37           O
ATOM     33  CB  VAL     4     -10.834  31.872 115.272  1.00 19.94           C
ATOM     34  CG1 VAL     4     -11.512  33.168 115.759  1.00 15.64           C
ATOM     35  CG2 VAL     4      -9.668  31.489 116.168  1.00 15.45           C
```

Modeller Python Script

(bioinfo.py)

```
# Homology modelling by the automodel class

from modeller.automodel import *    # Load the automodel class

log.verbose()    # request verbose output
env = environ()  # create a new MODELLER environment to build this model in

# directories for input atom files
env.io.atom_files_directory = './:../atom_files'

a = automodel(env,
              alnfile  = 'bioinfo.pir',    # alignment filename
               knowns   = '1SDMA',              # codes of the templates
               sequence = 'bioinfo')            # code of the target
a.starting_model= 1                 # index of the first model
a.ending_model  = 1                 # index of the last model
                                    # (determines how many models to calculate)
a.make()                            # do the actual homology modelling
```

Where to find structure file

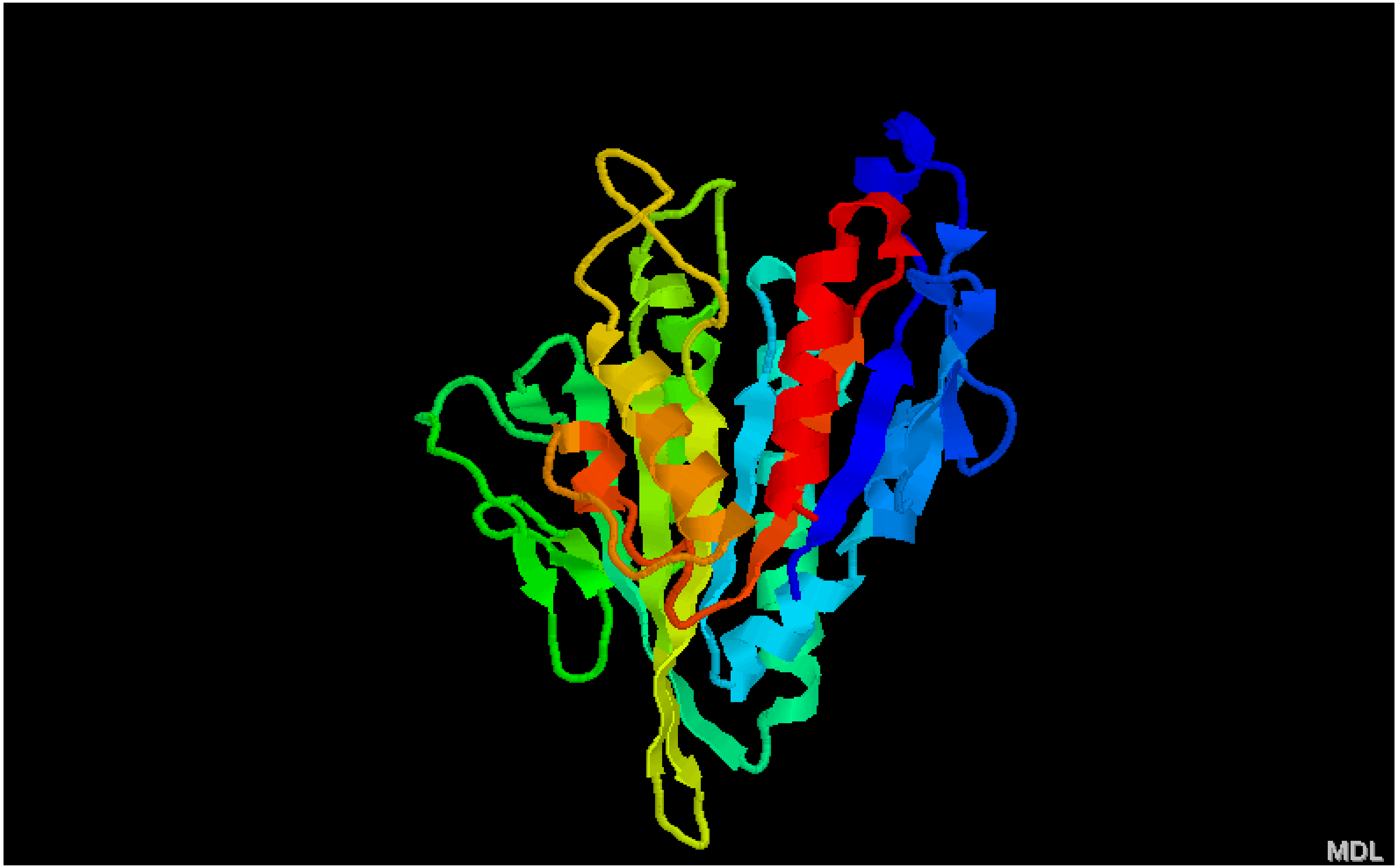
PIR alignment file name

Template structure file id

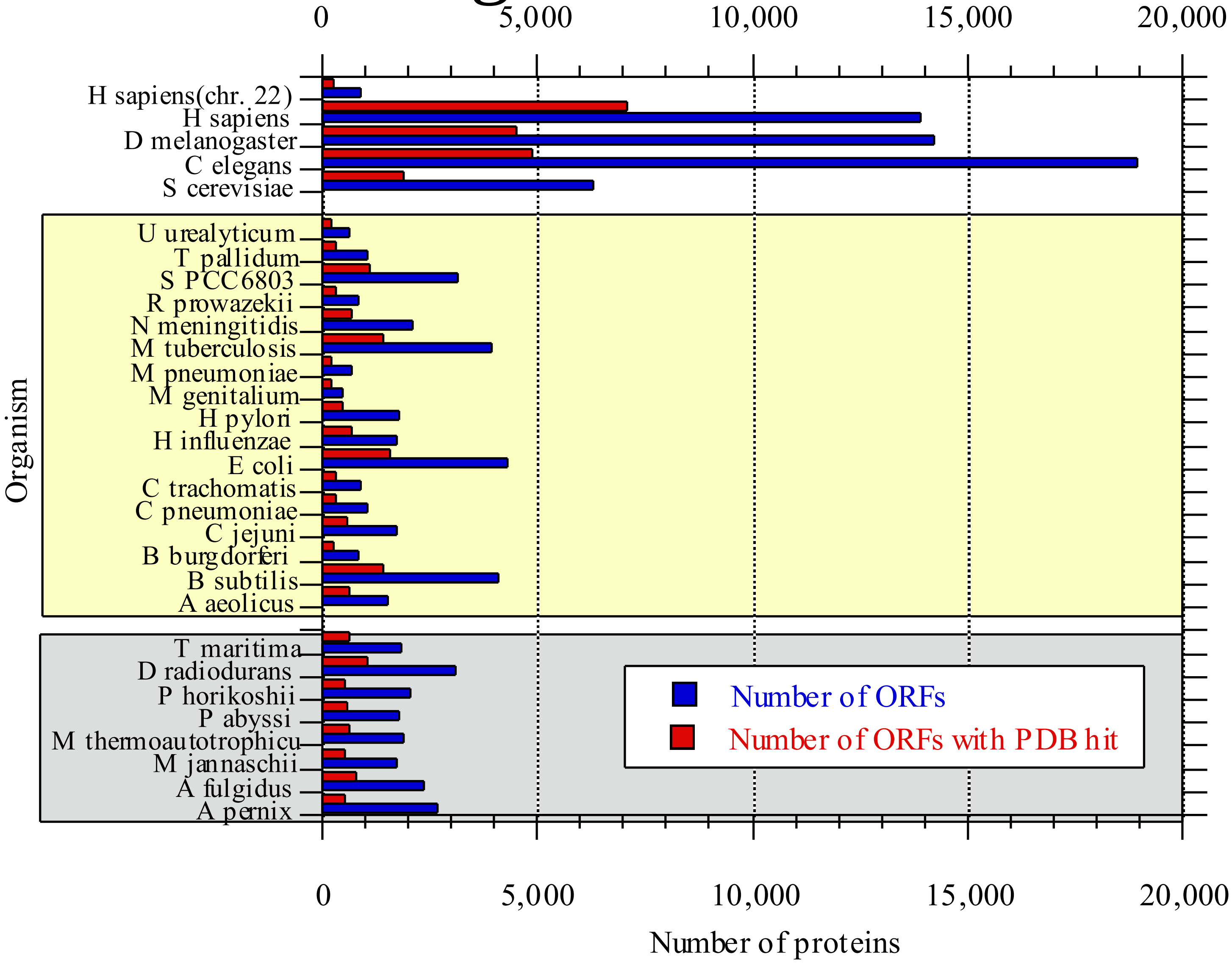
Query sequence id

Output Example

Command: mod8v2 bioinfo.py

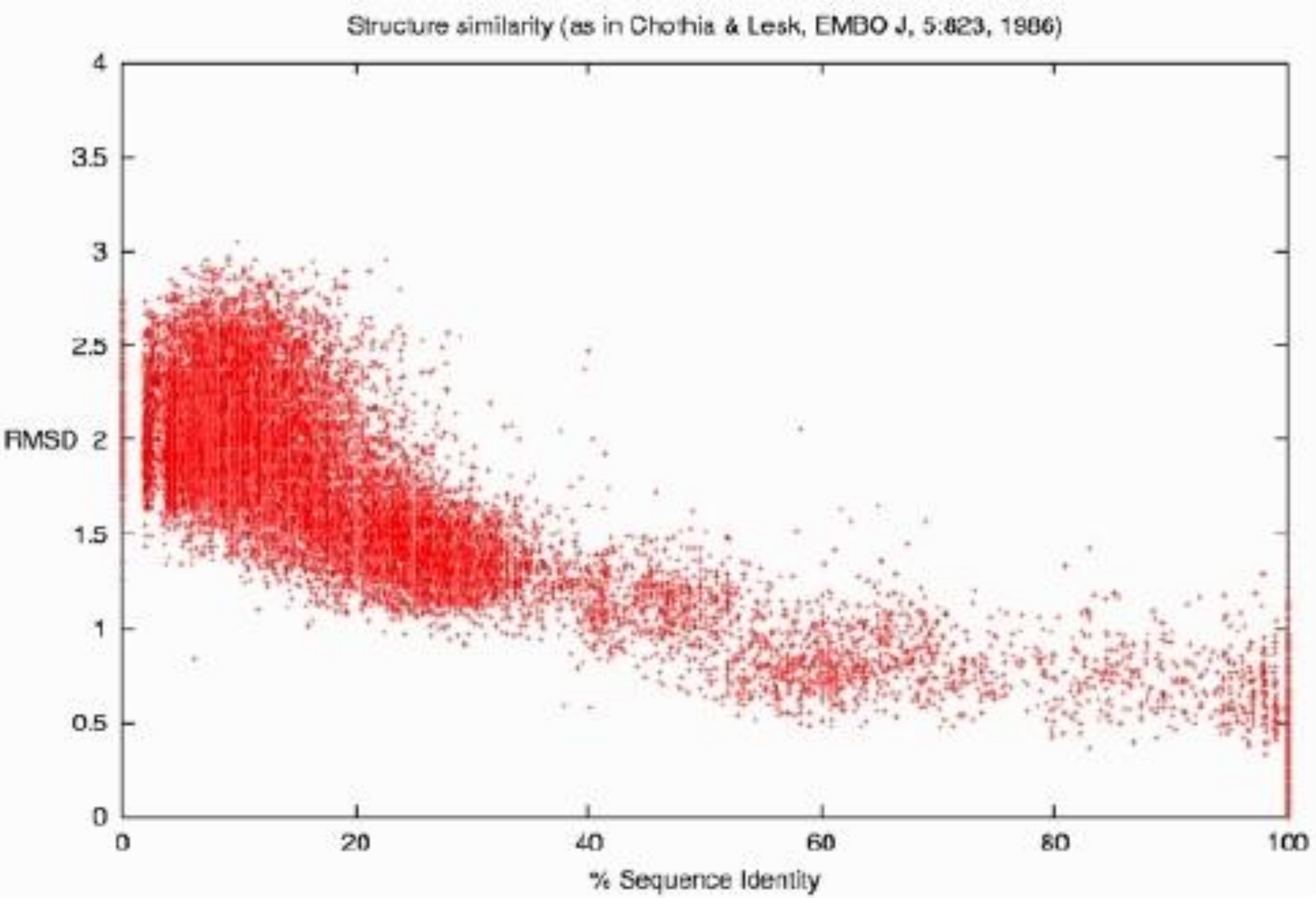


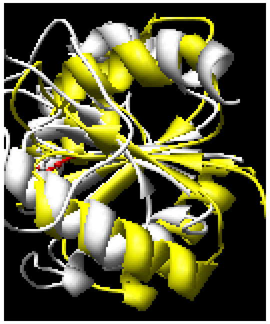
Homology modelling for entire genomes



B. Rost, 2005

Sequence Identity and Alignment Quality in Structure Prediction





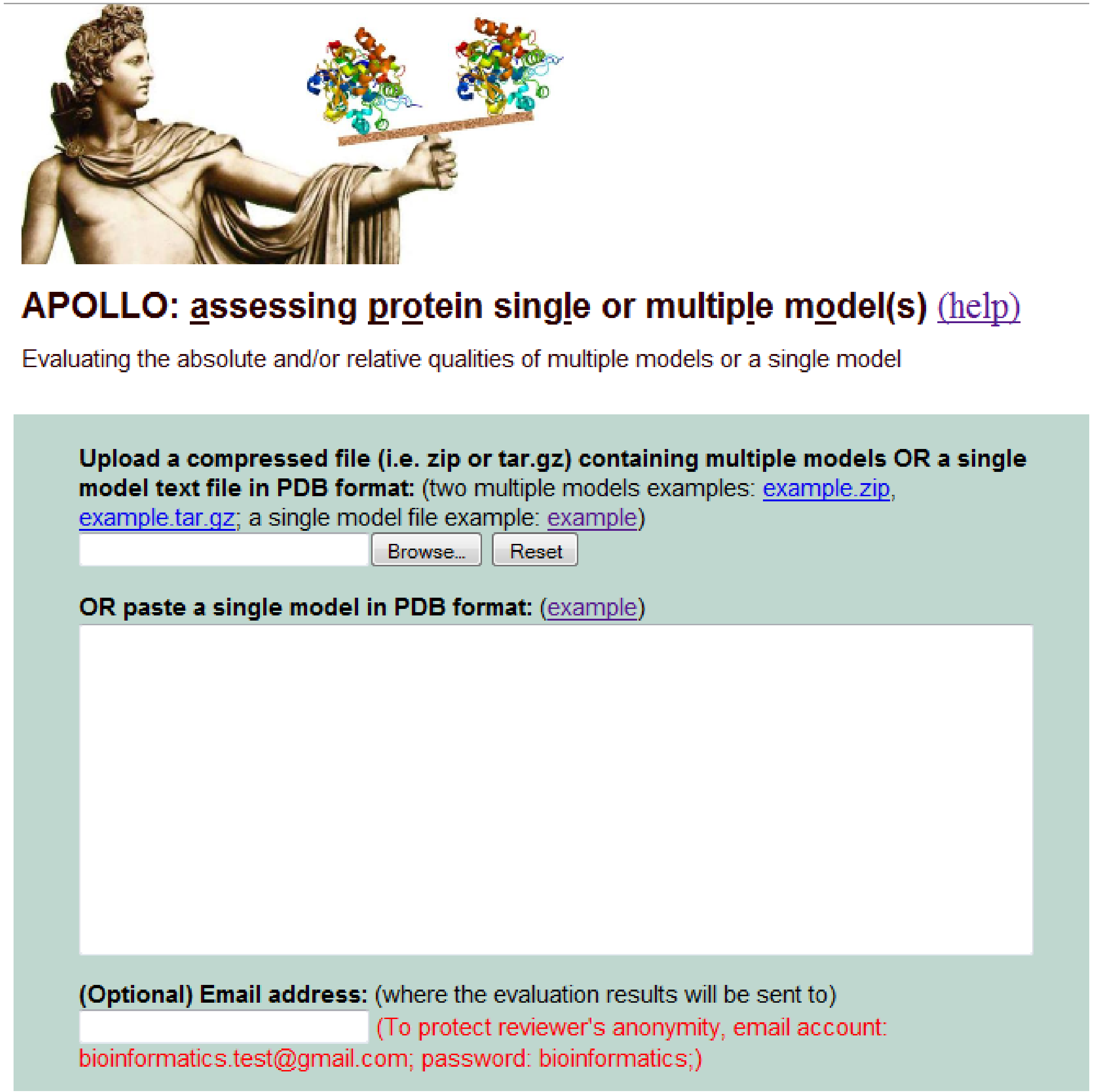
Superimpose
-> RMSD

%Sequence Identity: percent of identical residues in alignment
RMSD: square root of average distance between predicted structure and native structure.

3D Structure Prediction Tools

- MULTICOM (http://sysbio.rnet.missouri.edu/multicom_toolbox/index.html)
- I-TASSER (http://zhang.bioinformatics.ku.edu/I-TASSER/)
- HHpred (http://protevo.eb.tuebingen.mpg.de/toolkit/index.php?view=hhpred)
- Robetta (http://robetta.bakerlab.org/)
- 3D-Jury (http://bioinfo.pl/Meta/)
- FFAS (http://ffas.ljcrf.edu/ffas-cgi/cgi/ffas.pl)
- Pcons (http://pcons.net/)
- Sparks (http://phyyz4.med.buffalo.edu/hzhou/anonymous-fold-sp3.html)
- FUGUE (http://www-cryst.bioc.cam.ac.uk/%7Efugue/prfsearch.html)
- FOLDpro (http://mine5.ics.uci.edu:1026/foldpro.html)
- SAM (http://www.cse.ucsc.edu/research/compbio/sam.html)
- Phyre (http://www.sbg.bio.ic.ac.uk/~phyre/)
- 3D-PSSM (http://www.sbg.bio.ic.ac.uk/3dpssm/)
- mGenThreader (http://bioinf.cs.ucl.ac.uk/psipred/psiform.html)

Protein Model Quality Assessment



APOLLO: assessing protein single or multiple model(s) (help)

Evaluating the absolute and/or relative qualities of multiple models or a single model

Upload a compressed file (i.e. zip or tar.gz) containing multiple models OR a single model text file in PDB format: (two multiple models examples: example.zip, example.tar.gz; a single model file example: example)

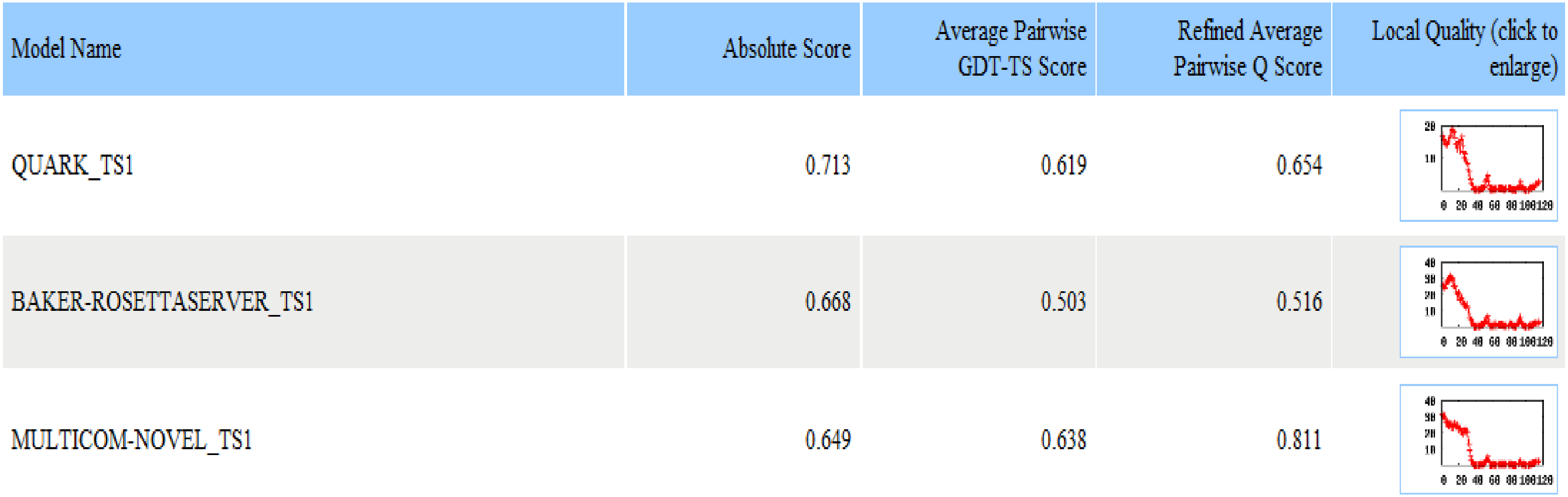
OR paste a single model in PDB format: (example)

(Optional) Email address: (where the evaluation results will be sent to)

(To protect reviewer's anonymity, email account: bioinformatics.test@gmail.com; password: bioinformatics;)

http://sysbio.rnet.missouri.edu/apollo/

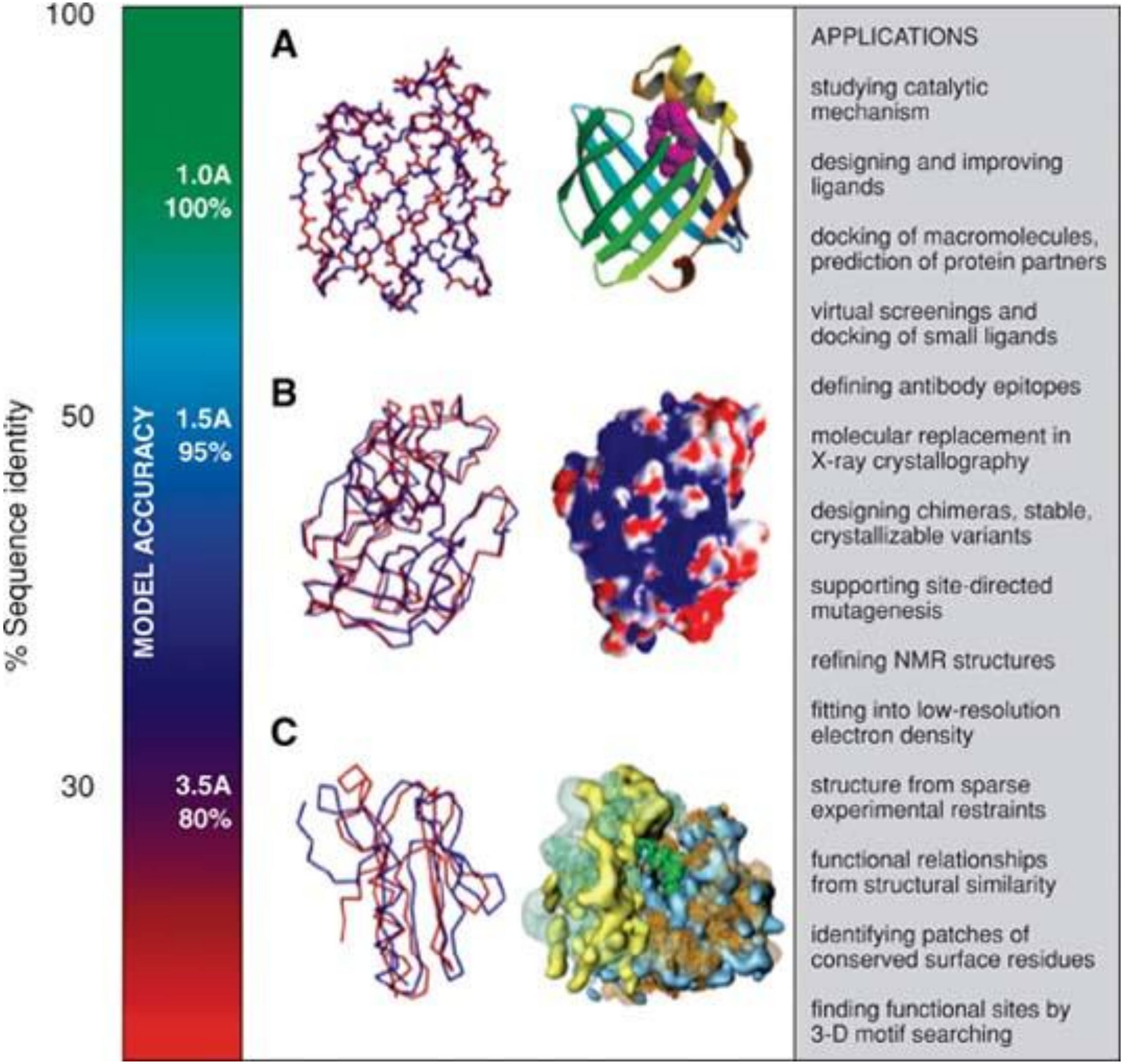
APOLLO Output



Model Name	Absolute Score	Average Pairwise GDT-TS Score	Refined Average Pairwise Q Score	Local Quality (click to enlarge)
QUARK_TS1	0.713	0.619	0.654	
BAKER-ROSETTASERVER_TS1	0.668	0.503	0.516	
MULTICOM-NOVEL_TS1	0.649	0.638	0.811	

Application of Structure Prediction

- Structure prediction is improving
- Template-based structure become more and more practical. Particularly, comparative / homology modeling is pretty accurate in many cases.
- Comparative modeling has been widely used in drug design.
- Protein structure prediction (both secondary and tertiary) has become an indispensable tool of investigating function of proteins and mechanisms of biological processes.



APPLICATIONS

- studying catalytic mechanism
- designing and improving ligands
- docking of macromolecules, prediction of protein partners
- virtual screenings and docking of small ligands
- defining antibody epitopes
- molecular replacement in X-ray crystallography
- designing chimeras, stable, crystallizable variants
- supporting site-directed mutagenesis
- refining NMR structures
- fitting into low-resolution electron density
- structure from sparse experimental restraints
- functional relationships from structural similarity
- identifying patches of conserved surface residues
- finding functional sites by 3-D motif searching

Baker and Sali (2000)

J. Pevsner, 2005

Project 2

- Select 5 soybean proteins
- Predict 3D structures
- Visualize the structures

(data:http://casp.rnet.missouri.edu /marc/muii_7005/SEQ_TFP_90_2500.txt)