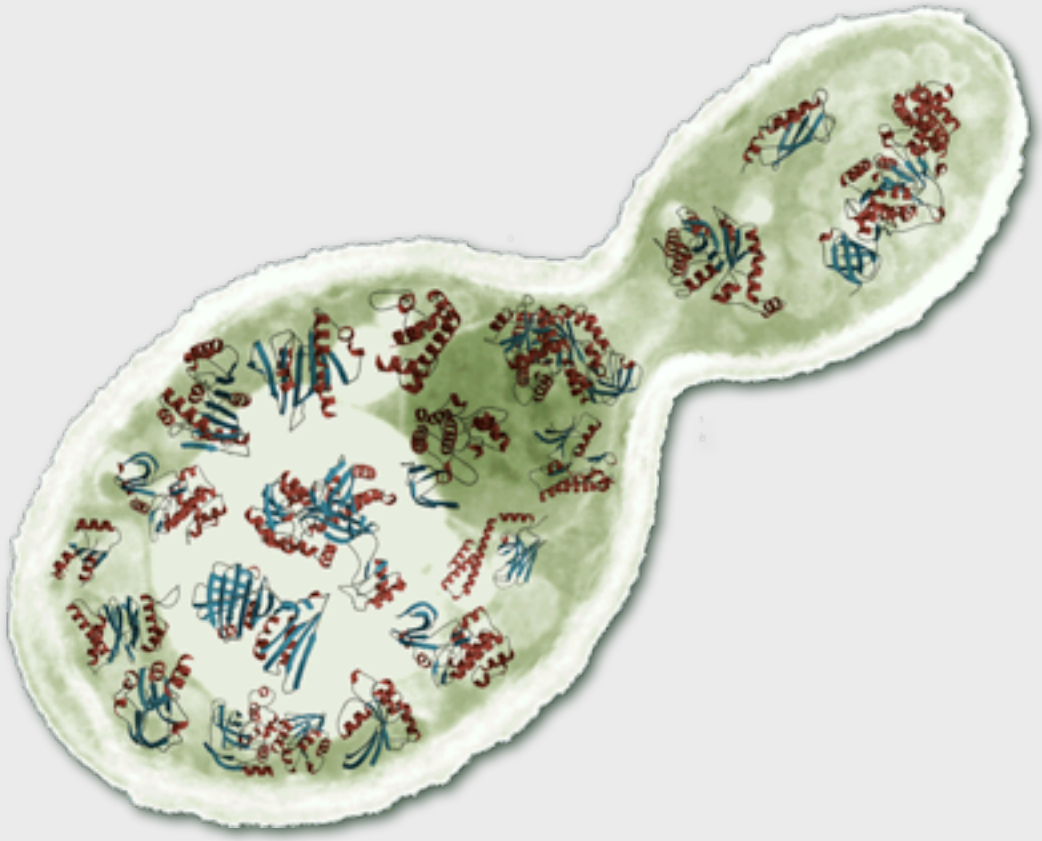
Comparative Protein Structure Prediction



Program

Intro to comparative protein structure prediction

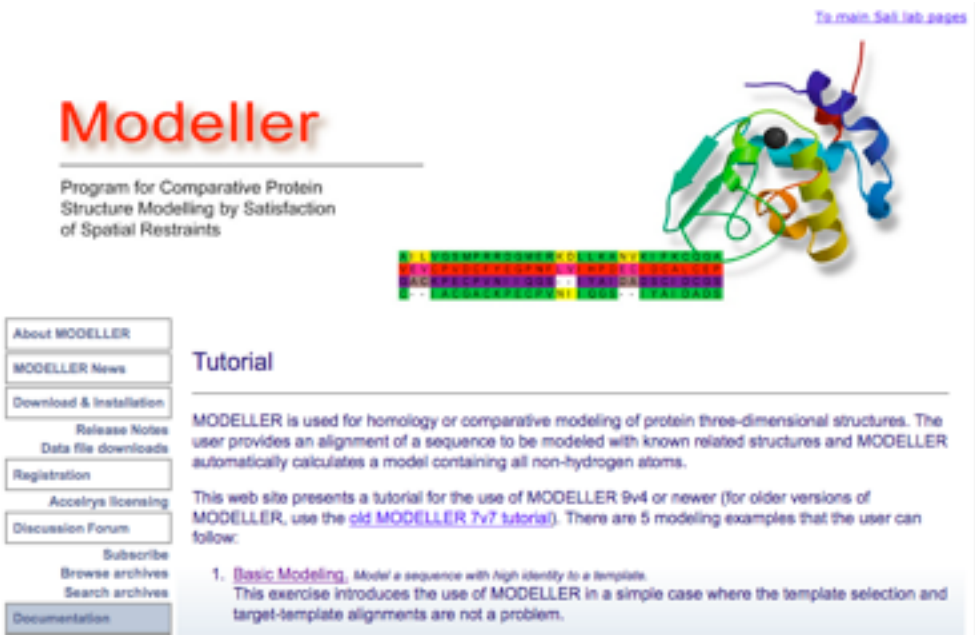
Template Search

Target – Template Alignment

Model Building

Model Evaluation

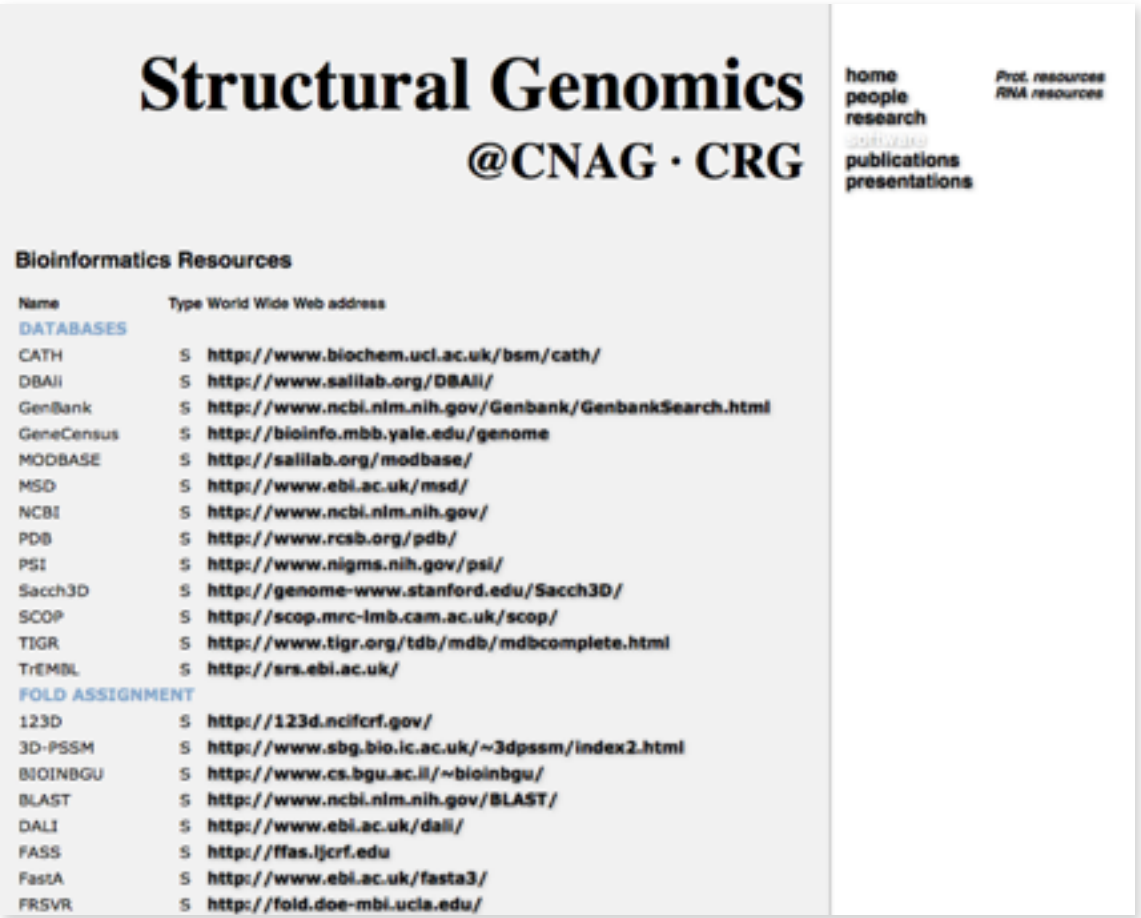
http://www.salilab.org/modeller/tutorial/



Objective

TO LEARN HOW-TO MODEL A 3D-STRUCTURE FROM A SEQUENCE AND A KNOWN STRUCTURE

DISCLAIMER!



Structural Genomics
@CNAG · CRG

home
people
research
publications
presentations

Prot. resources
RNA resources

Bioinformatics Resources

Name	Type	World Wide Web address
DATABASES		
CATH	S	http://www.biochem.ucl.ac.uk/bsm/cath/
DBAli	S	http://www.salilab.org/DBAli/
GenBank	S	http://www.ncbi.nlm.nih.gov/Genbank/GenbankSearch.html
GeneCensus	S	http://bioinfo.mbb.yale.edu/genome
MODBASE	S	http://salilab.org/modbase/
MSD	S	http://www.ebi.ac.uk/msd/
NCBI	S	http://www.ncbi.nlm.nih.gov/
PDB	S	http://www.rcsb.org/pdb/
PSI	S	http://www.nigms.nih.gov/psi/
Sacch3D	S	http://genome-www.stanford.edu/Sacch3D/
SCOP	S	http://scop.mrc-lmb.cam.ac.uk/scop/
TIGR	S	http://www.tigr.org/tdb/mdb/mdbcomplete.html
TrEMBL	S	http://srs.ebi.ac.uk/
FOLD ASSIGNMENT		
123D	S	http://123d.ncifcrf.gov/
3D-PSSM	S	http://www.sbg.bio.ic.ac.uk/~3dpssm/index2.html
BIOINBGU	S	http://www.cs.bgu.ac.il/~bioinbgu/
BLAST	S	http://www.ncbi.nlm.nih.gov/BLAST/
DALI	S	http://www.ebi.ac.uk/dali/
FASS	S	http://ffas.ljcrf.edu
FastA	S	http://www.ebi.ac.uk/fasta3/
FRSVR	S	http://fold.doe-mbi.ucla.edu/

http://sgt.cnag.cat/www/software/?prot_resources

Nomenclature

Homology: Sharing a common ancestor, may have similar or dissimilar functions

Similarity: Score that quantifies the degree of relationship between two sequences.

Identity: Fraction of identical aminoacids between two aligned sequences (case of similarity).

Target: Sequence corresponding to the protein to be modeled.

Template: 3D structure/s to be used during protein structure prediction.

Model: Predicted 3D structure of the target sequence.

Nomenclature

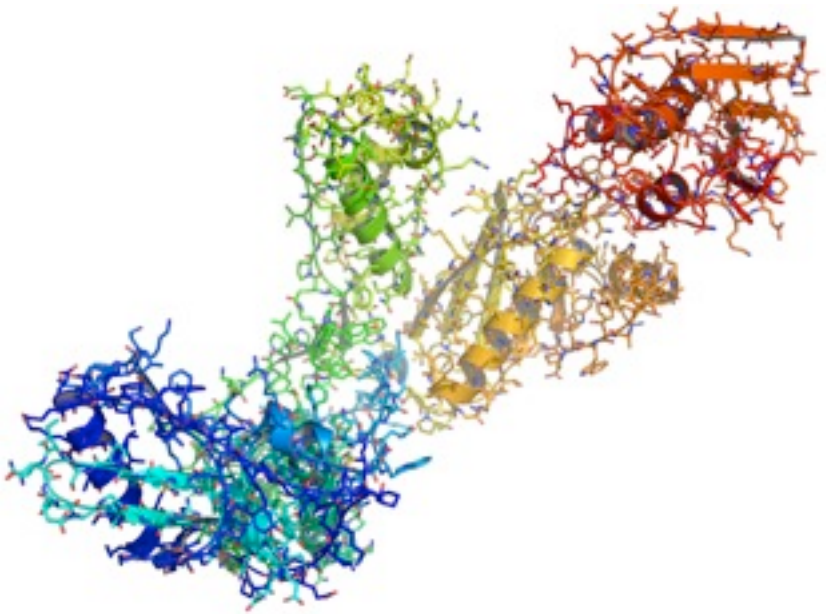
Fold: Three dimensional conformation of a protein sequence (usually at domain level).

Domain: Structurally globular part of a protein, which may independently fold.

Secondary Structure: Regular sub-domain structures composed by alpha-helices, beta-sheets and coils (or loops).

Backbone: Protein structure skeleton composed by the carbon, nitrogen and oxygen atoms.

Side-Chain: Specific atoms identifying each of the 20 residues types.



General References

Protein Structure Prediction:

Marti-Renom et al. Annu. Rev. Biophys. Biomol. Struct. 29, 291-325, 2000.
Baker & Sali. Science 294, 93-96, 2001.

Comparative Modeling:

Madhusudhan et al. The Proteomics Protocols Handbook. Ed. Walker. Humana Press Inc., Totowa, NJ. 831-860, 2005.

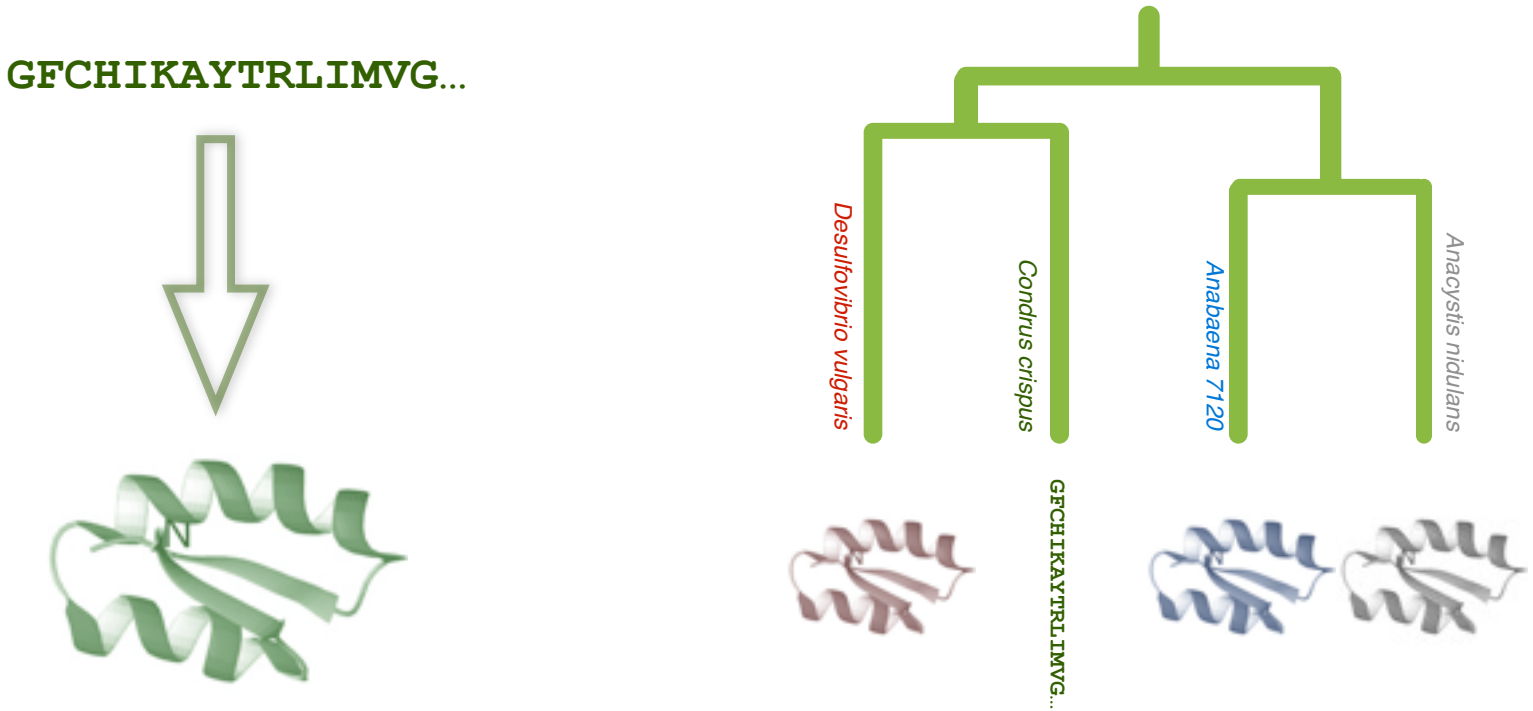
MODELLER:

Sali & Blundell. J. Mol. Biol. 234, 779-815, 1993.
Eswar, M. A. et al. Comparative Protein Structure Modeling With MODELLER. Current Protocols in Bioinformatics, John Wiley & Sons, Inc., Supplement 15, 5.6.1-5.6.30, 2006.

Structural Genomics:

Sali. Nat. Struct. Biol. 5, 1029, 1998.
Burley et al. Nat. Genet. 23, 151, 1999.
Sali & Kuriyan. TIBS 22, M20, 1999.
Sanchez et al. Nat. Str. Biol. 7, 986, 2000.
Baker & Sali. Science 294, 93-96, 2001.
Terwilliger et al. Annu. Rev. Biophys. 38, 371-383, 2009.

Principles of protein structure



Folding (physics)

Ab initio prediction

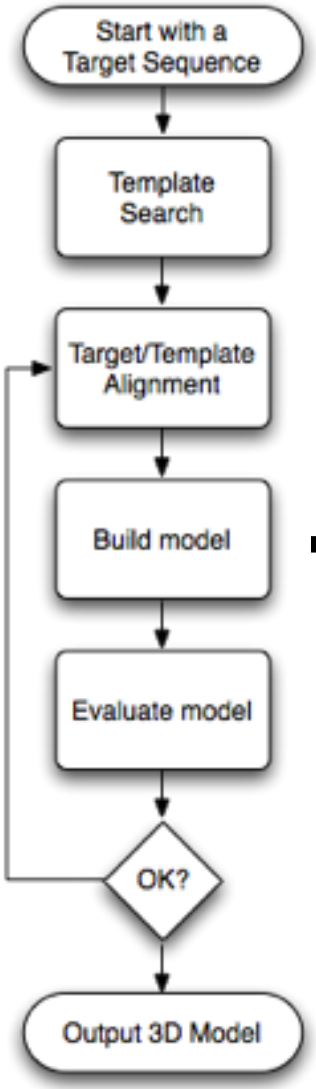
Evolution (rules)

Threading

Comparative Modeling

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

Comparative modeling by satisfaction of spatial restraints
MODELLER

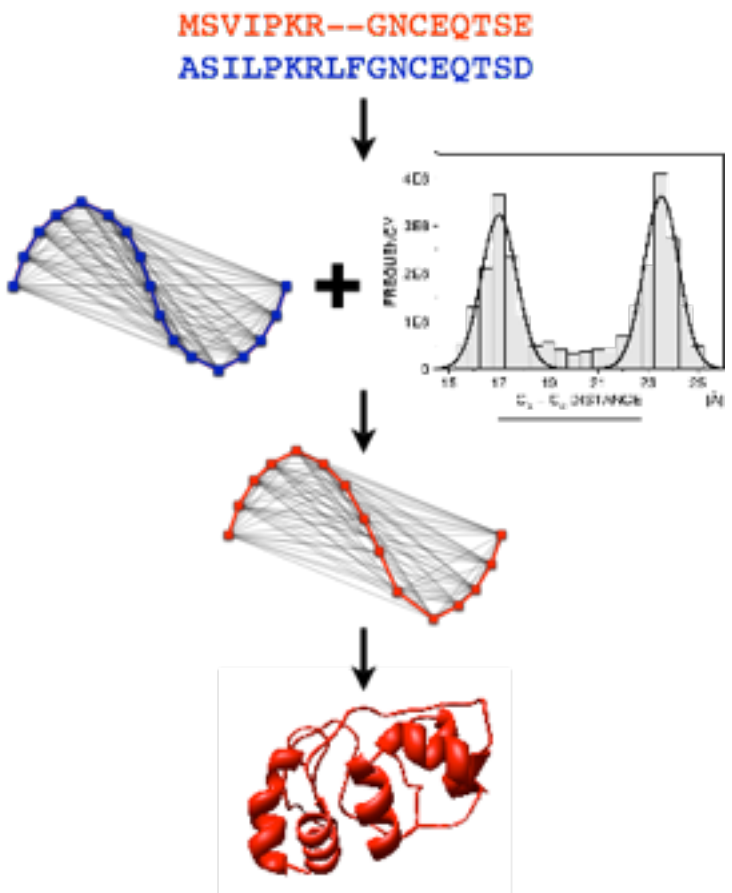


Given an alignment...

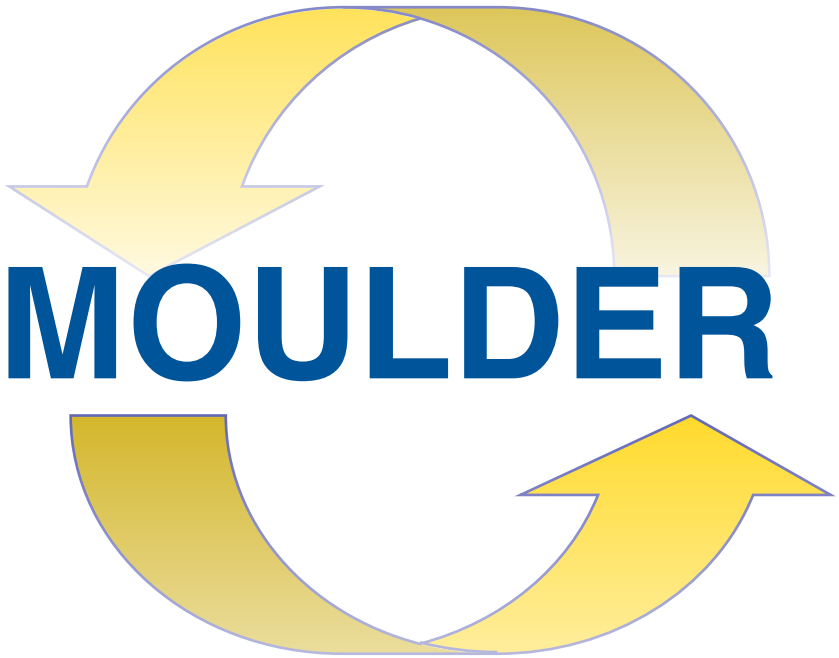
extract spatial features from the template(s) and statistics from known structures

apply these features as restraints on your target sequence

optimize to find the best solution for the restraints to produce your 3D model

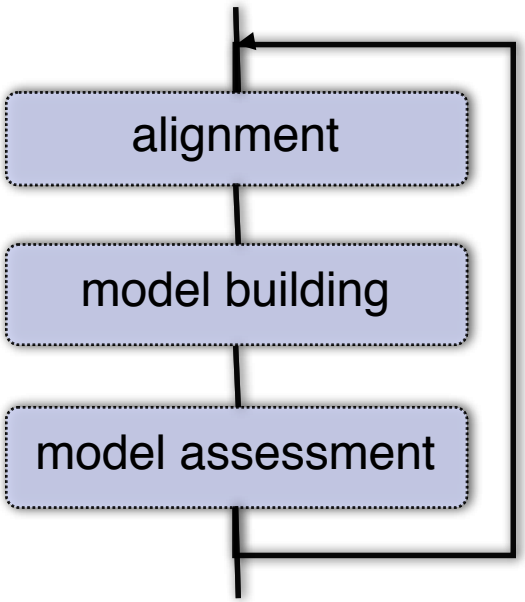


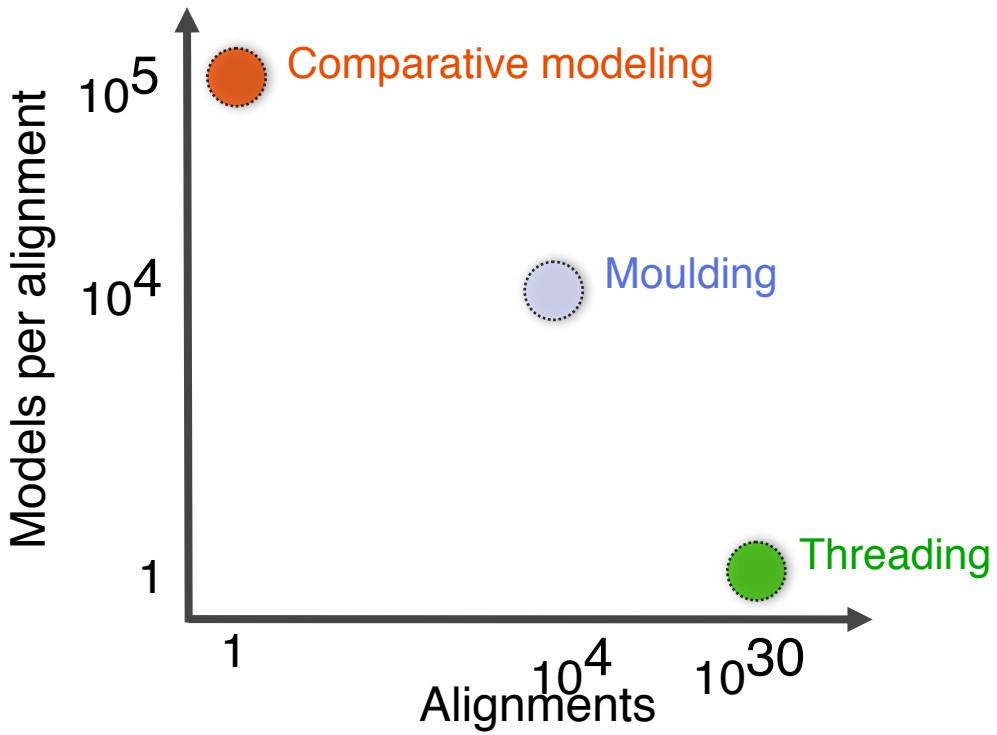
A. Šali & T. Blundell. J. Mol. Biol. 234, 779, 1993.
J.P. Overington & A. Šali. Prot. Sci. 3, 1582, 1994.
A. Fiser, R. Do & A. Šali, Prot. Sci., 9, 1753, 2000.



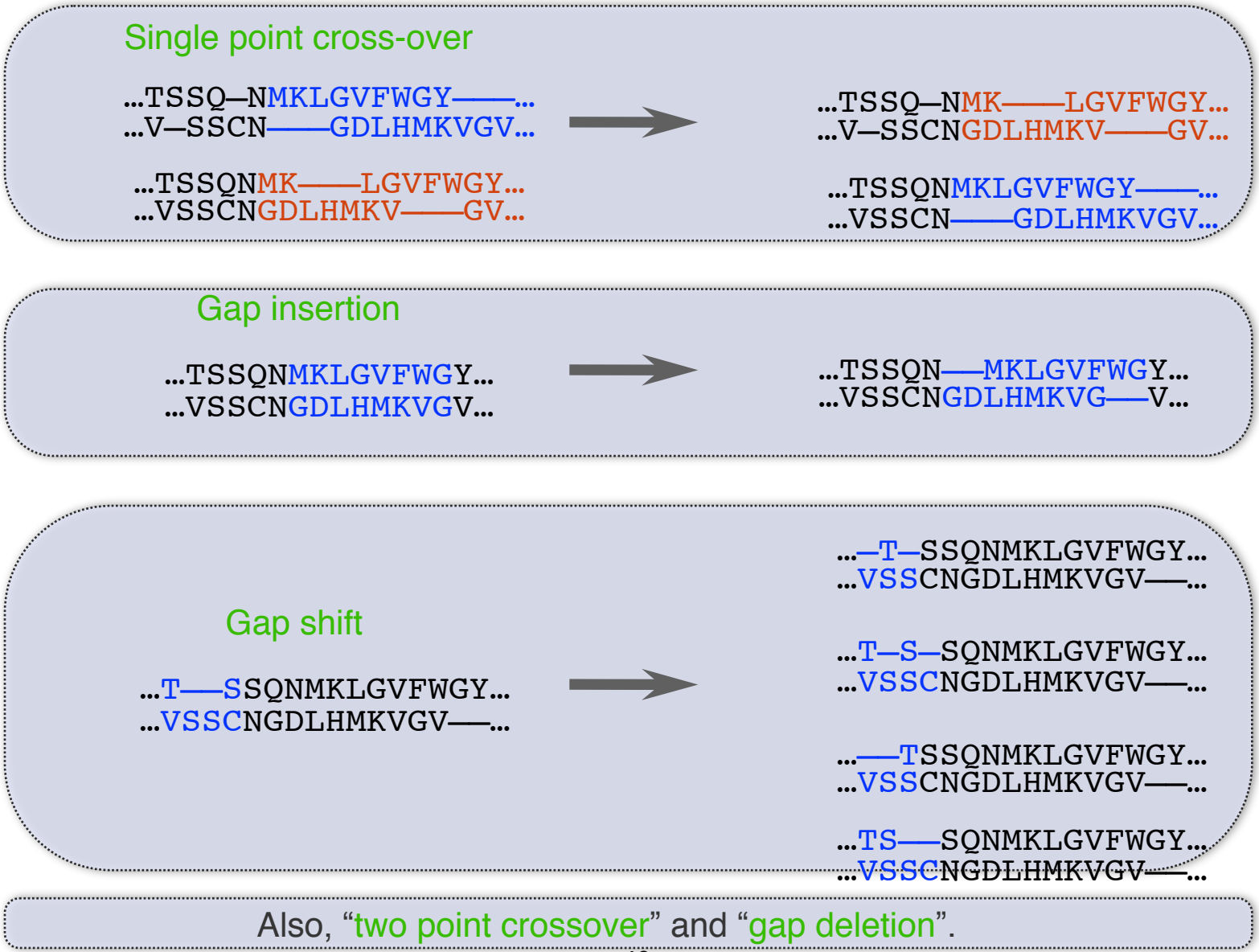
John, Sali (2003). NAR pp31 3982

Moulding: iterative alignment, model building, model assessment





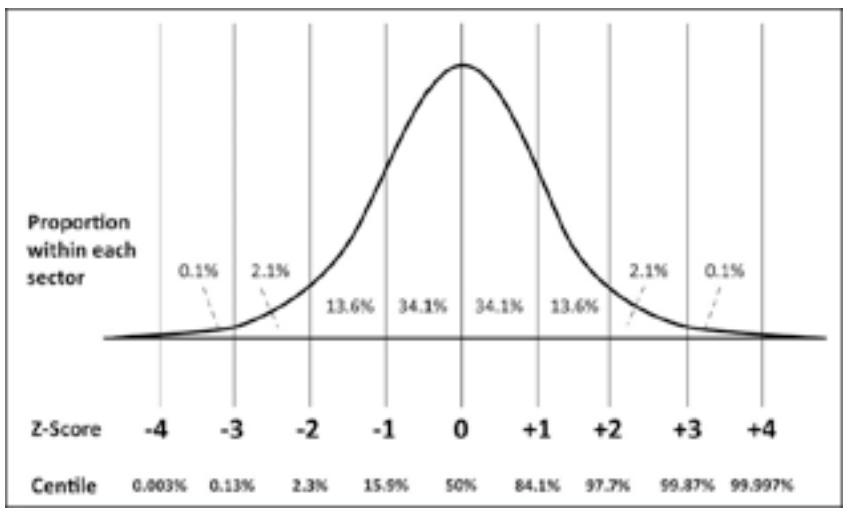
Genetic algorithm operators



Composite model assessment score

- Weighted linear combination of several structural properties:

$$Z = 0.17\,Z(P_P) + 0.02\,Z(P_S) + 0.10\,Z(S_C) + 0.26\,Z(H_a) + 0.45\,(A_S)$$



John, Sali (2003). NAR pp31 3982

Application to a difficult modeling case

1BOV-1LTS

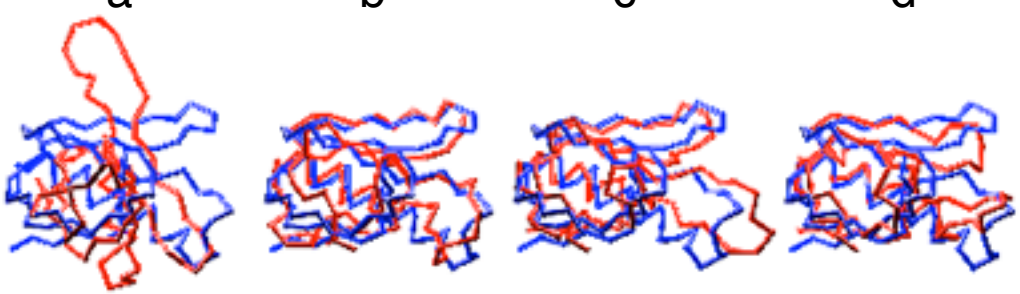
Best Z

Sequence identity 4.4%

Initial model Cα RMSD 10.1Å

Final model Cα RMSD 3.6Å

a b c d



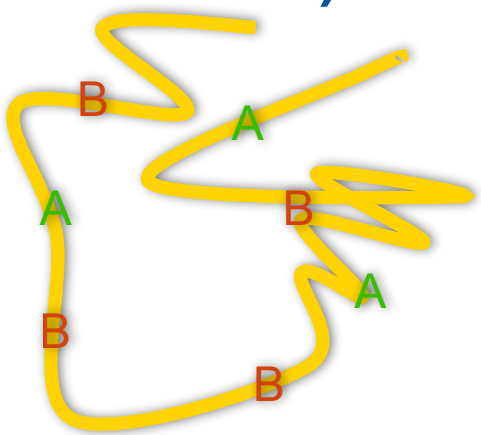
Benchmark with the “very difficult” test set

D. Fischer threading test set of 68 structural pairs (a subset of 19)

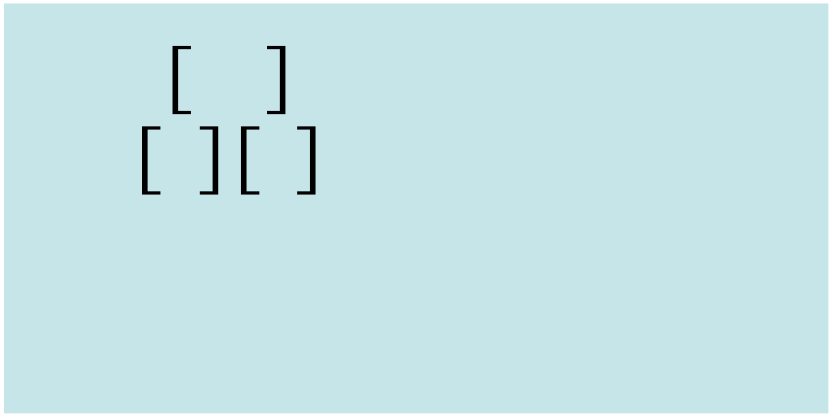
Target -template	Sequence identity [%]	Coverage [% aa]	Initial prediction		Final prediction		Best prediction	
			Cα RMSD [Å]	CE overlap [%]	Cα RMSD [Å]	CE overlap [%]	Cα RMSD [Å]	CE overlap [%]
1ATR-1ATN	13.8	94.3	19.2	20.2	18.8	20.2	17.1	24.6
1BOV-1LTS	4.4	83.5	10.1	29.4	3.6	79.4	3.1	92.6
1CAU-1CAU	18.8	96.7	11.7	15.6	10.0	27.4	7.6	47.4
1COL-1CPC	11.2	81.4	8.6	44.0	5.6	58.6	4.8	59.3
1LFB-1HOM	17.6	75.0	1.2	100.0	1.2	100.0	1.1	100.0
1NSB-2SIM	10.1	89.2	13.2	20.2	13.2	20.1	12.3	26.8
1RNH-1HRH	26.6	91.2	13.0	21.2	4.8	35.4	3.5	57.5
1YCC-2MTA	14.5	55.1	3.4	72.4	5.3	58.4	3.1	75.0
2AYH-1SAC	8.8	78.4	5.8	33.8	5.5	48.0	4.8	64.9
2CCY-1BBH	21.3	97.0	4.1	52.4	3.1	73.0	2.6	77.0
2PLV-1BBT	20.2	91.4	7.3	58.9	7.3	58.9	6.2	60.7
2POR-2OMF	13.2	97.3	18.3	11.3	11.4	14.7	10.5	25.9
2RHE-1CID	21.2	61.6	9.2	33.7	7.5	51.1	4.4	71.1
2RHE-3HLA	2.4	96.0	8.1	16.5	7.6	9.4	6.7	43.5
3ADK-1GKY	19.5	100.0	13.8	26.6	11.5	37.7	7.7	48.1
3HHR-1TEN	18.4	98.9	7.3	60.9	6.0	66.7	4.9	79.3
4FGF-81IB	14.1	98.6	11.3	24.0	9.3	30.6	5.4	41.2
6XIA-3RUB	8.7	44.1	10.5	14.5	10.1	11.0	9.0	34.3
9RNT-2SAR	13.1	88.5	5.8	41.7	5.1	51.2	4.8	69.0
AVERAGE	14.2	85.2	9.6	36.7	7.7	44.8	6.3	57.8

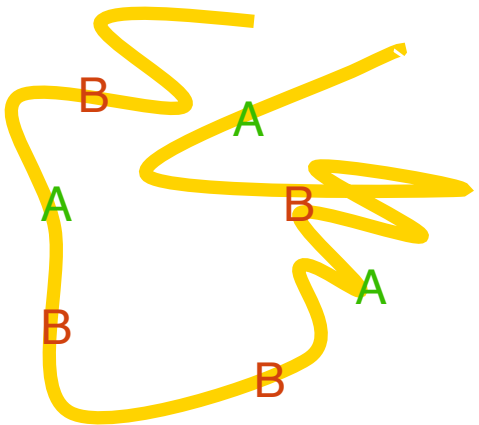
Model Assessment

(Potentials of Mean Force)



Statistical Potential (inspiration)





From statistical physics, we know that energy difference between two states (ΔE) and the ratio of their occupancies (N_1:N_2) are related [9]:

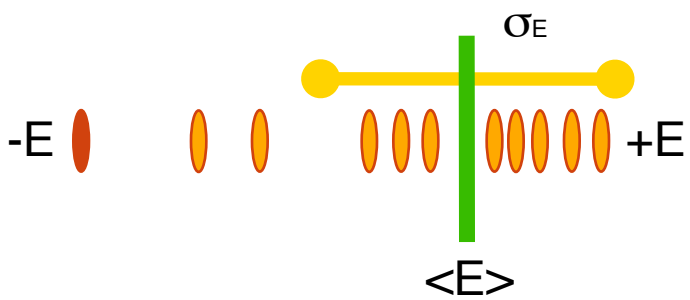
$$\Delta E = -kT \ln \left(\frac{N_1}{N_2} \right) \qquad (1)$$

in which T is the absolute temperature and k is the Boltzmann's constant. As we are interested in an interaction energy between two amino acid side chains, it would seem natural to define N_1 as the number of interactions between these two residues types in a group of real protein structures, a number which is readily available from simple database analysis. But this number must be compared with the number of interactions in some other system, N_2, to obtain the energy difference between them.

*Tanaka and Sheraga (1975) PNAS, **72** pp3802*
Sippl, (1990) J.Mo.Biol. 213 pp859
*Godzik, (1996) Structure **15** pp363*

Significance of an alignment (score)

"Energy" Z-score the model with respect the energy of random models (or rest of decoys).



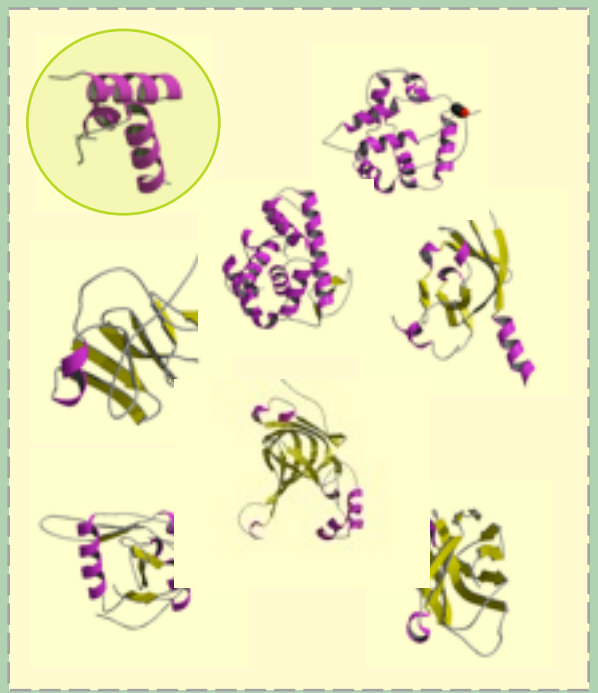
$$\left(\langle \quad \rangle \quad \right)$$

ProSa

http://www.came.sbg.ac.at/prosa.php

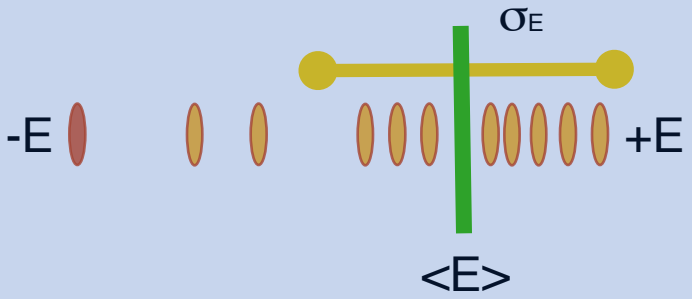
Deriving

Structural space



Scoring

[ ]
[ ][ ]



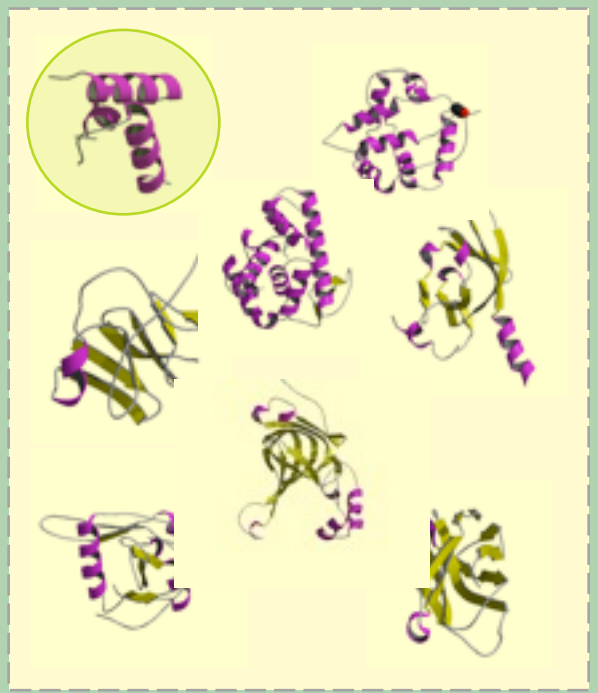
$$Zscore = \frac{\left(\langle E \rangle - E_m\right)}{\sigma_E}$$

DFIRE/DFIRE2

http://sparks.informatics.iupui.edu/yueyang/DFIRE/dDFIRE-service

Deriving

Structural space



Scoring

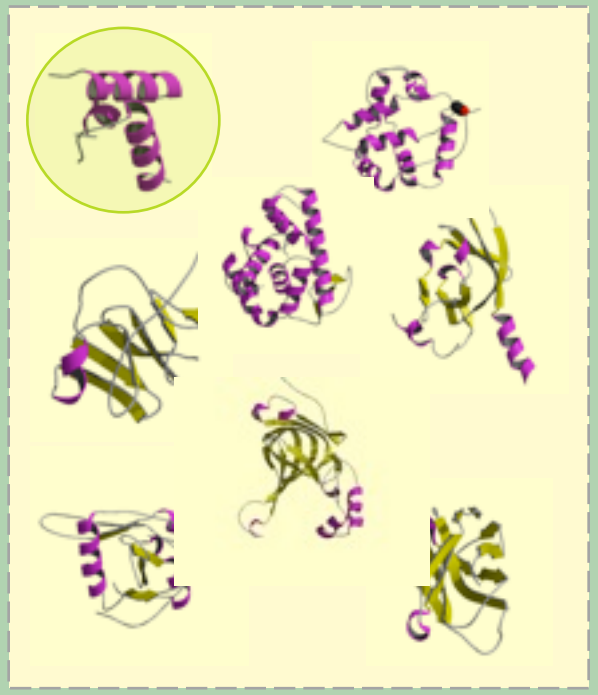
Pseudo-Energy with respect a ideal gas-phase reference state

DOPE (MODELLER)

http://www.salilab.org/modeller/

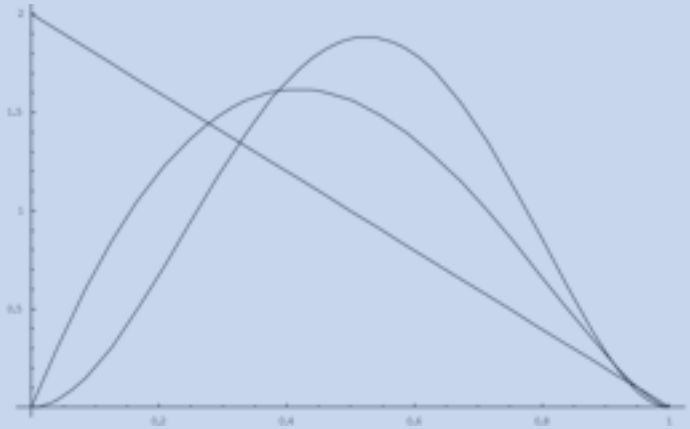
Deriving

Structural space



Scoring

Pseudo-Energy with respect a ideal spherical protein as a reference state

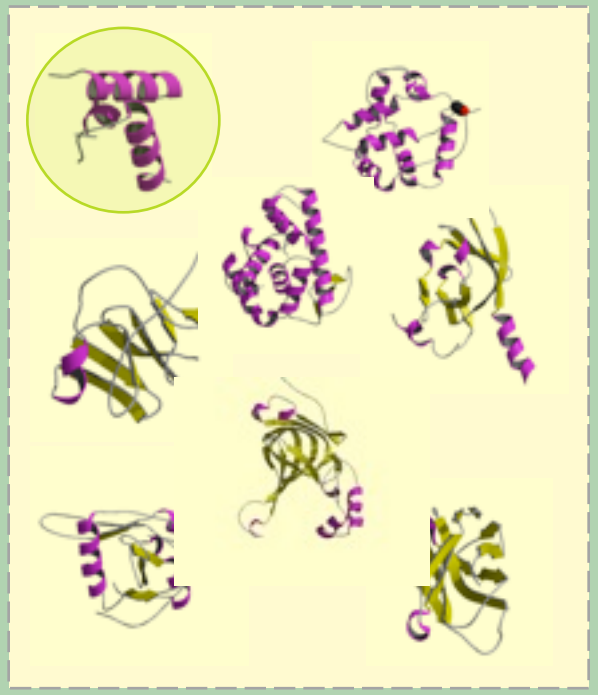


Verify3D

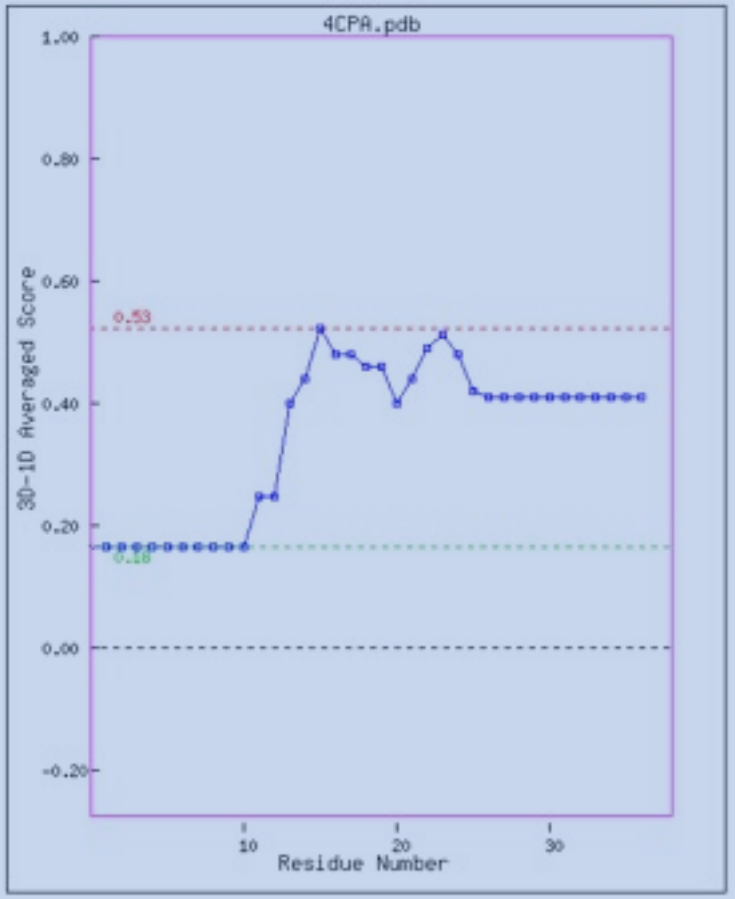
http://nihserver.mbi.ucla.edu/Verify_3D/

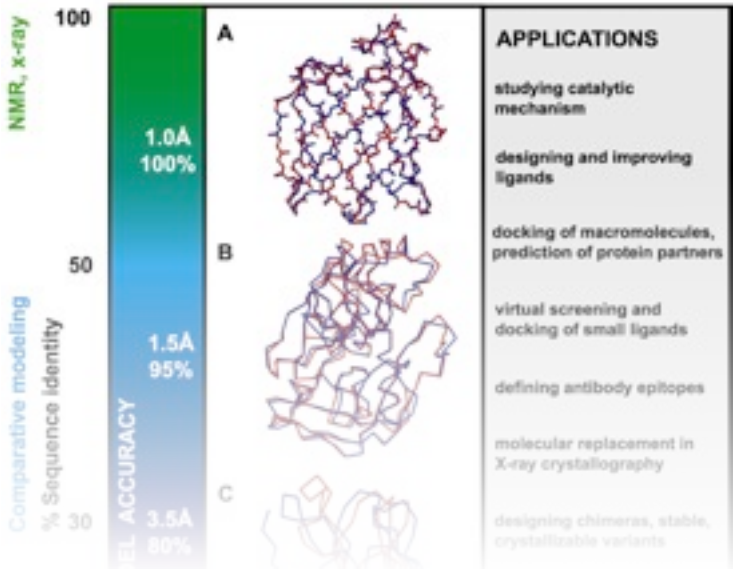
Deriving

Structural space



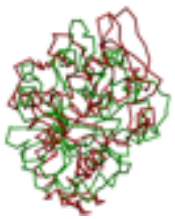
Scoring





Accuracy and applicability of comparative models

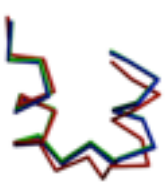
Comparative modeling by satisfaction of spatial restraints Types of errors and their impact



Wrong fold

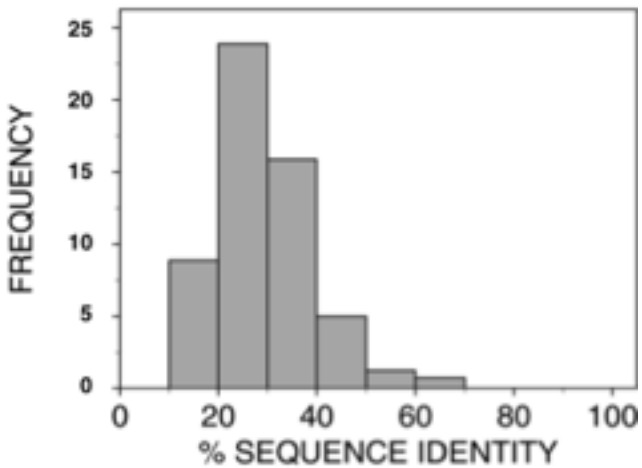
Miss alignments

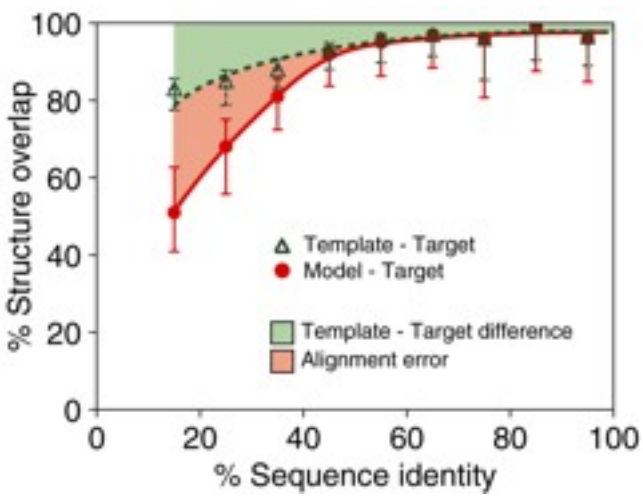
Loop regions



Rigid body distortions

Side-chain packing

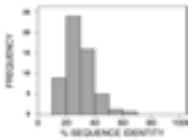




Template - Model - Real Structure

Marti-Renom etal. Ann Rev Biophys Biomol Struct (2000) 29, 291

Model Accuracy



HIGH ACCURACY (2/10) **MEDIUM ACCURACY (3/10)** **LOW ACCURACY (5/10)**

NM23 Seq id 77%

Cα equiv 147/148
RMSD 0.41Å

Sidechains
Core backbone
Loops

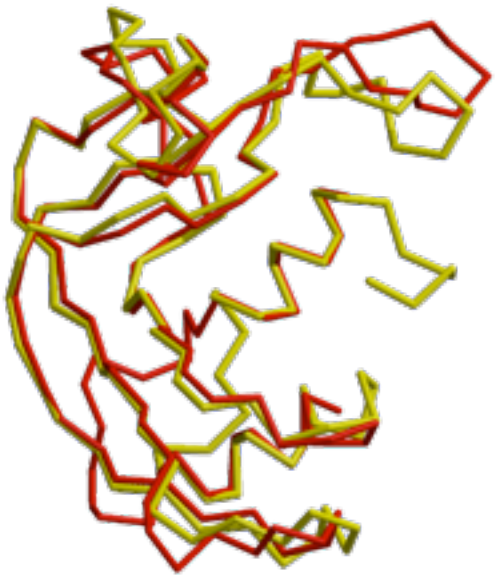
CRABP Seq id 41%

Cα equiv 122/137
RMSD 1.34Å

Sidechains
Core backbone
Loops
Alignment

EDN Seq id 33%

Cα equiv 90/134
RMSD 1.17Å

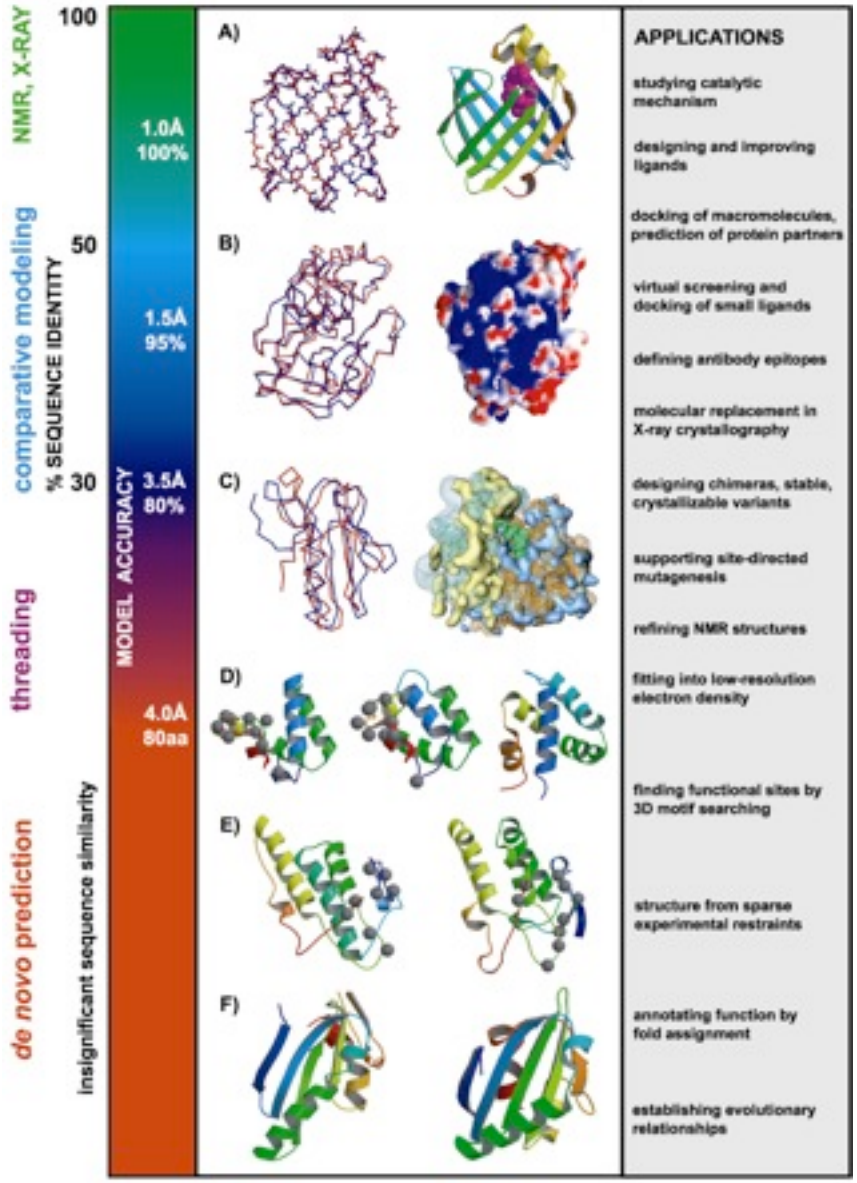


Sidechains
Core backbone
Loops
Alignment
Fold assignment

X-RAY / MODEL

Marti-Renom et al. Annu.Rev.Biophys.Biomol.Struct. 29, 291-325, 2000.

Utility of protein structure models, despite errors



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

Comparative Protein Structure Prediction
MODELLER tutorial

```
$>mod9v11 model.py
```

Obtaining MODELLER and related information

- MODELLER (9v11) web page
- http://www.salilab.org/modeller/
  - Download Software (Linux/Windows/Mac)
  - HTML Manual
  - **Join Mailing List**



Using MODELLER

- No GUI! ☹
- Controlled by command file ☹☹
- Script is written in PYTHON language ☺
- You may know Python language is simple ☺☺

“MINIMAL” MODELLER

- INPUT:
  - **Target** Sequence (FASTA/PIR format)
  - **Template** Structure (PDB format)
  - **Python** script file
- OUTPUT:
  - Target-Template **Alignment**
  - **Model** in PDB format
  - Additional data (in .log)

Modeling of BLBP
Input

- Target: Brain lipid-binding protein (BLBP)
- BLBP sequence in PIR (MODELLER) format:

blbp.seq

```
>P1;blbp
sequence:blbp:::::::
VDAFCATWKLTDSQNFDEYMKALGVGFATRQVGNVTKPTVIISQEGGKVVIRTQCTFKNTEINFQLGEEFEETSIDDRNCKSVVRLDGD
KLIHVQKWDGKETNCTREIKDGKMVVTLTFGDIVAVRCYEKA*
```

Modeling of BLBP

STEP 1: Align blbp and 1hms sequences

Python script for target-template alignment

align.py

```
# Example for: alignment.align()

# This will read two sequences, align them, and write the alignment
# to a file:


log.verbose()
env = environ()


aln = alignment(env)
mdl = model(env, file='1hms')
aln.append_model(mdl, align_codes='1hms')
aln.append(file='blbp.seq', align_codes=('blbp'))

# The as1.sim.mat similarity matrix is used by default:
aln.align(gap_penalties_1d=(-600, -400))
aln.write(file='blbp-1hms.ali', alignment_format='PIR')
aln.write(file='blbp-1hms.pap', alignment_format='PAP')
```

Run by typing `mod9v11 align.py` in the directory where you have the python file.
MODELLER will produce a `align.log` file

Modeling of BLBP

STEP 1: Align blbp and 1hms sequences

Python script for target-template alignment

```
# Example for: alignment.align()

# This will read two sequences, align them, and write the alignment
# to a file:


log.verbose()
env = environ()


aln = alignment(env)
mdl = model(env, file='1hms')
aln.append_model(mdl, align_codes='1hms')
aln.append(file='blbp.seq', align_codes=('blbp'))

# The as1.sim.mat similarity matrix is used by default:
aln.align(gap_penalties_1d=(-600, -400))
aln.write(file='blbp-1hms.ali', alignment_format='PIR')
aln.write(file='blbp-1hms.pap', alignment_format='PAP')
```

Run by typing `mod9v11 align.py` in the directory where you have the python file.
MODELLER will produce a `align.log` file

Modeling of BLBP
STEP 1: Align blbp and 1hms sequences
Python script for target-template alignment

```
# Example for: alignment.align()

# This will read two sequences, align them, and write the alignment
# to a file:


log.verbose()
env = environ()


aln = alignment(env)
mdl = model(env, file='1hms')
aln.append_model(mdl, align_codes='1hms')
aln.append(file='blbp.seq', align_codes=('blbp'))

# The as1.sim.mat similarity matrix is used by default:
aln.align(gap_penalties_1d=(-600, -400))
aln.write(file='blbp-1hms.ali', alignment_format='PIR')
aln.write(file='blbp-1hms.pap', alignment_format='PAP')
```

Run by typing `mod9v11 align.py` in the directory where you have the python file.
MODELLER will produce a `align.log` file

Modeling of BLBP

STEP 1: Align blbp and 1hms sequences

Python script for target-template alignment

```
# Example for: alignment.align()

# This will read two sequences, align them, and write the alignment
# to a file:


log.verbose()
env = environ()


aln = alignment(env)
mdl = model(env, file='1hms')
aln.append_model(mdl, align_codes='1hms')
aln.append(file='blbp.seq', align_codes=('blbp'))

# The as1.sim.mat similarity matrix is used by default:
aln.align(gap_penalties_1d=(-600, -400))
aln.write(file='blbp-1hms.ali', alignment_format='PIR')
aln.write(file='blbp-1hms.pap', alignment_format='PAP')
```

Run by typing `mod9v11 align.py` in the directory where you have the python file.
MODELLER will produce a `align.log` file

Modeling of BLBP
STEP 1: Align blbp and 1hms sequences
Output

blbp-1hms.ali

```
>P1;1hms
structureX:1hms:    1 :A: 131 :A:undefined:undefined:-1.00:-1.00
VDAFLGTWKLVDSKNFDDYMKSLGVGFATRQVASMTKPTTIIEKNGDILTLKTHSTFKNTEISFKLGVEFDETTA
DDRKVKSIVTLDGGKLVHLQKWDGQETTLVRELIDGKLILTLTHGTAVCTRTYEKE*
>P1;blbp
sequence:blbp:      : :      : : : : 0.00: 0.00
VDAFCATWKLTDSQNFDEYMKALGVGFATRQVGNVTKPTVIISQEGGKVVIRTQCTFKNTEINFQLGEEFEETSI
DDRNCKSVVRLDGDKLIHVQKWDGKETNCTREIKDGKMVVTLTFGDIVAVRCYEKA*
```

Modeling of BLBP
STEP 1: Align blbp and 1hms sequences
Output

```
>P1;1hms
structureX:1hms:    1 : : 131 : :undefined:undefined:-1.00:-1.00
VDAFLGTWKLVDSKNFDDYMKSLGVGFATRQVASMTKPTTIIEKNGDILTLKTHSTFKNTEISFKLGVEFDETTA
DDRKVKSIVTLDGGKLVHLQKWDGQETTLVRELIDGKLILTLTHGTAVCTRTYEKE*
>P1;blbp
sequence:blbp:      : :      : : : : 0.00: 0.00
VDAFCATWKLTDSQNFDEYMKALGVGFATRQVGNVTKPTVIISQEGGKVVIRTQCTFKNTEINFQLGEEFEETSI
DDRNCKSVVRLDGDKLIHVQKWDGKETNCTREIKDGKMVVTLTFGDIVAVRCYEKA*
```

Modeling of BLBP
STEP 1: Align blbp and 1hms sequences
Output

blbp-1hms.pap

```
 _aln.pos          10         20         30         40         50         60
1hms       VDAFLGTWKLVDSKNFDDYMKSLGVGFATRQVASMTKPTTIIEKNGDILTLKTHSTFKNTEISFKLGV
blbp       VDAFCATWKLTDSQNFDEYMKALGVGFATRQVGNVTKPTVIISQEGGKVVIRTQCTFKNTEINFQLGE
 _consrvd ****   **** ** *** *** **********    **** **    *       *  ******* * **


 _aln.p    70         80         90         100        110        120        130
1hms       EFDETTADDRKVKSIVTLDGGKLVHLQKWDGQETTLVRELIDGKLILTLTHGTAVCTRTYEKE
blbp       EFEETSIDDRNCKSVVRLDGDKLIHVQKWDGKETNCTREIKDGKMVVTLTFGDIVAVRCYEKA
 _consrvd ** **   ***   ** * *** ** * ***** **    **   ***    *** *   *   * ***
```

Modeling of BLBP

STEP 2: Model the blbp structure using the alignment from step 1.

Python script for model building

```
# 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

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


a = automodel(env,
              alnfile  = 'blbp-1hms.ali',     # alignment filename
              knowns   = '1hms',              # codes of the templates
              sequence = 'blbp')              # 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
```

Run by typing `mod9v11 model.py` in the directory where you have the python file.
MODELLER will produce a `model.log` file

Modeling of BLBP

STEP 2: Model the blbp structure using the alignment from step 1.

Python script for model building

```
# 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

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


a = automodel(env,
              alnfile  = 'blbp-1hms.ali',     # alignment filename
              knowns   = '1hms',              # codes of the templates
              sequence = 'blbp')              # 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
```

Run by typing `mod9v11 model.py` in the directory where you have the python file.
MODELLER will produce a `model.log` file

Modeling of BLBP

STEP 2: Model the blbp structure using the alignment from step 1.

Python script for model building

```
# 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

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


a = automodel(env,
              alnfile  = 'blbp-1hms.ali',     # alignment filename
              knowns   = '1hms',              # codes of the templates
              sequence = 'blbp')              # 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
```

Run by typing `mod9v11 model.py` in the directory where you have the python file.
MODELLER will produce a `model.log` file

Modeling of BLBP

STEP 2: Model the blbp structure using the alignment from step 1.

Python script for model building

PDB file

Can be viewed with Chimera

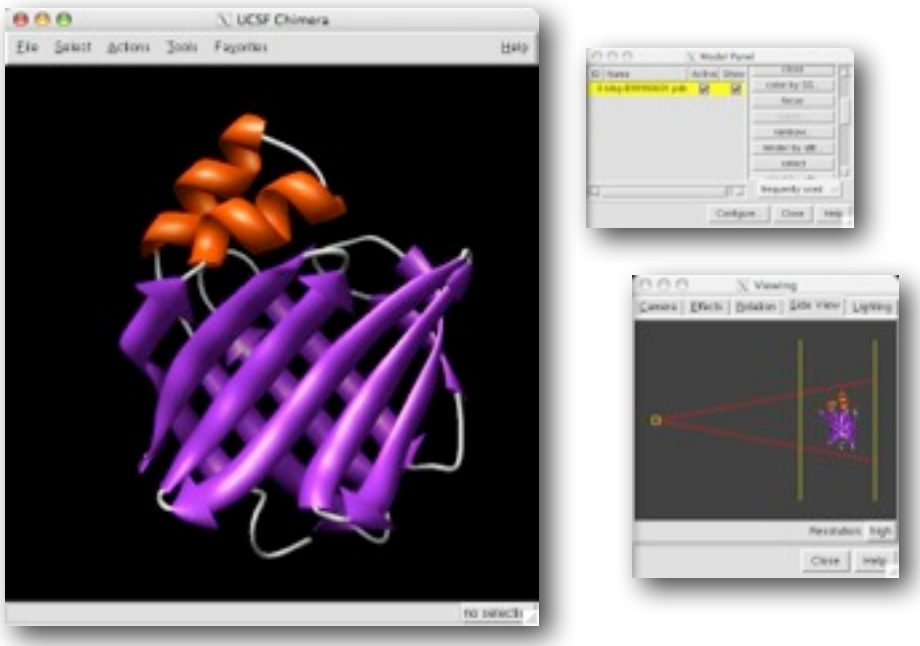
http://www.cgl.ucsf.edu/chimera/

PyMol

http://www.pymol.org

Rasmol

http://www.openrasmol.org

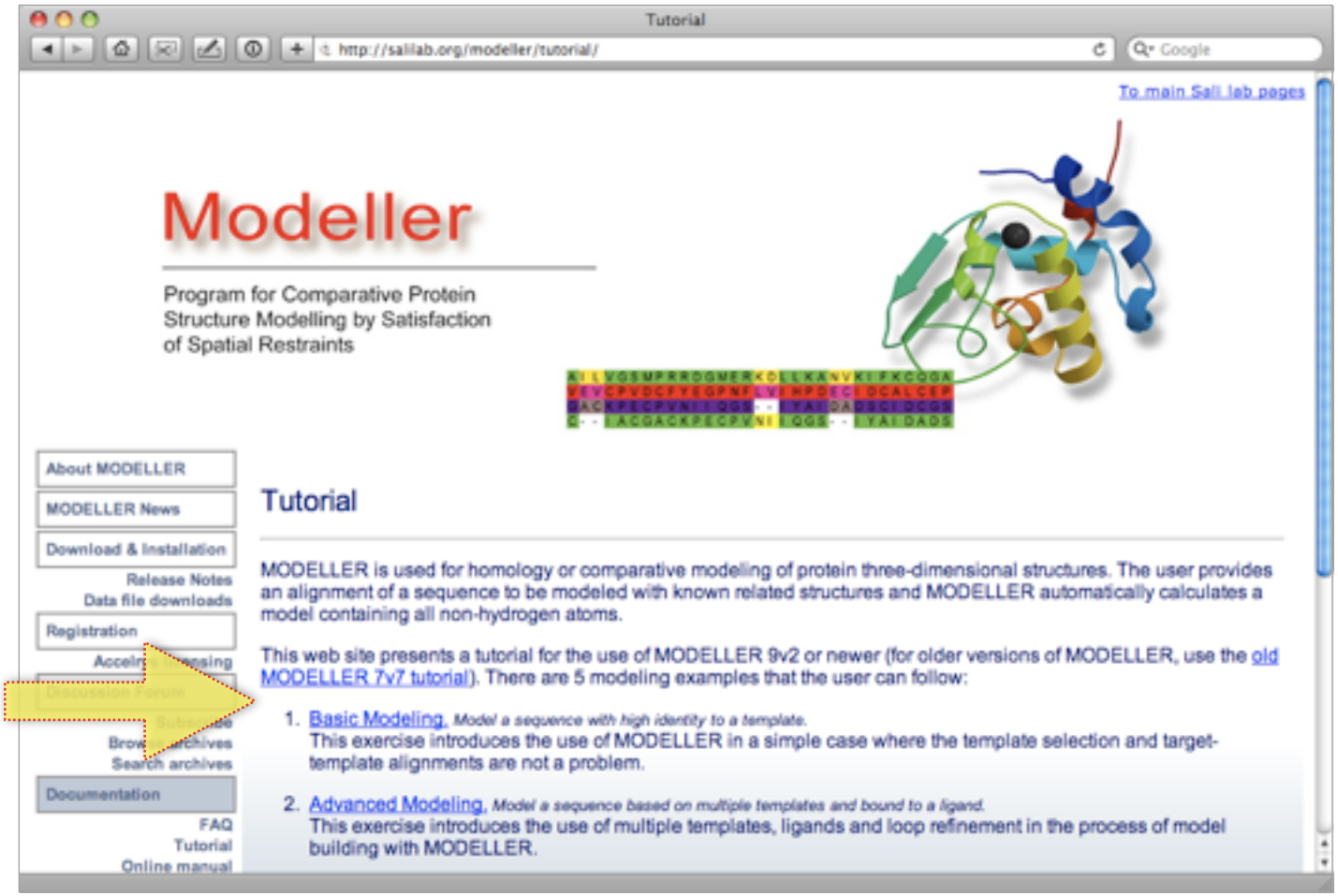


- Model file →

  `blbp.B99990001.pdb`

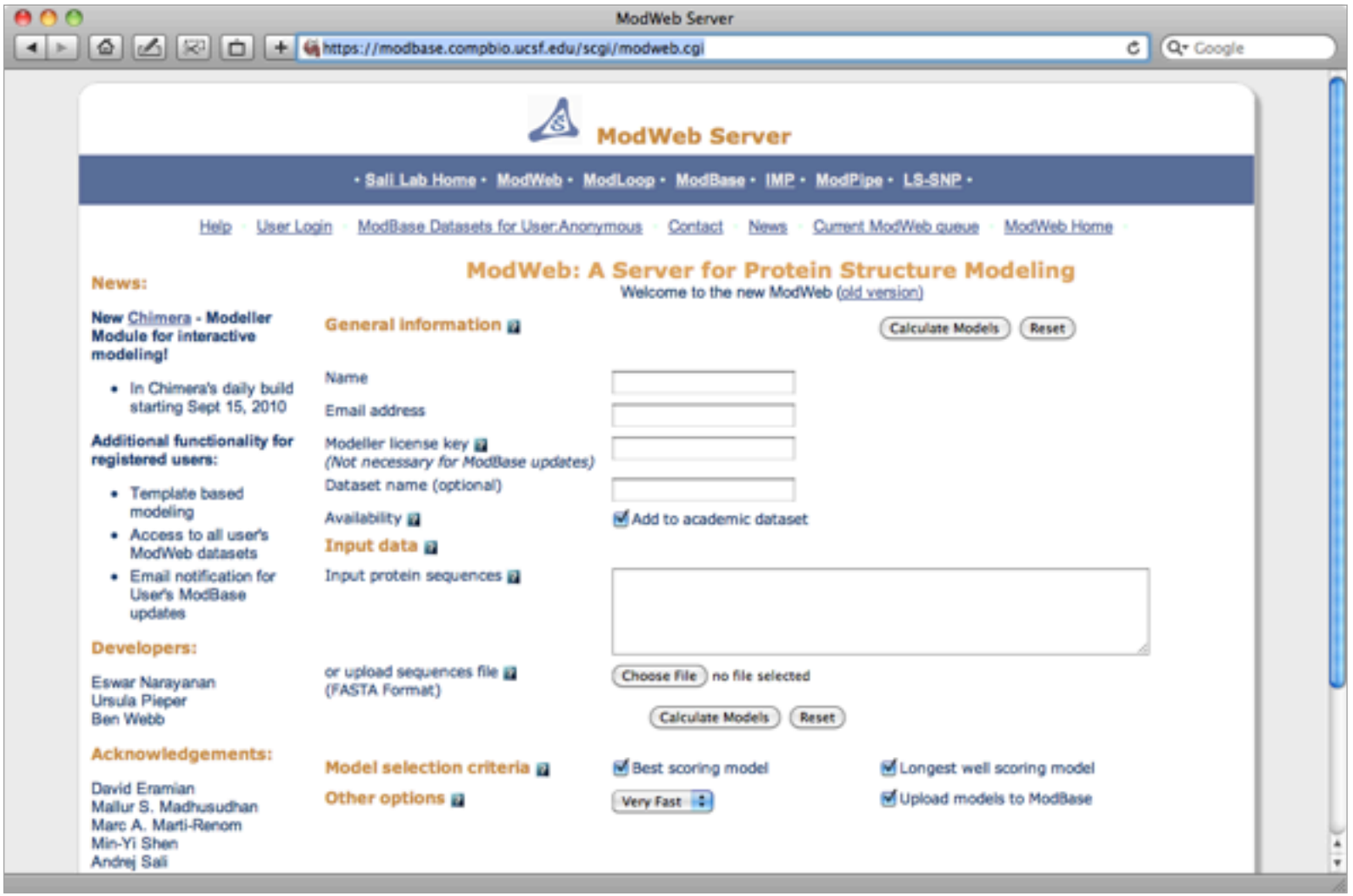
In 15 days!

http://www.salilab.org/modeller/tutorial/



FULLY AUTOMATIC MODWEB

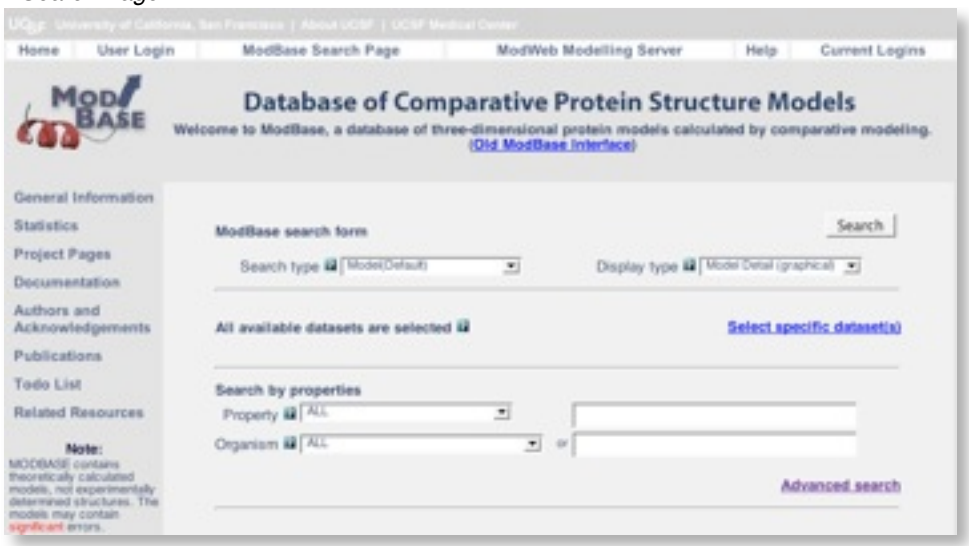
http://salilab.org/modweb



MODBASE

http://salilab.org/modbase

Search Page



Model Details

Sequence Overview

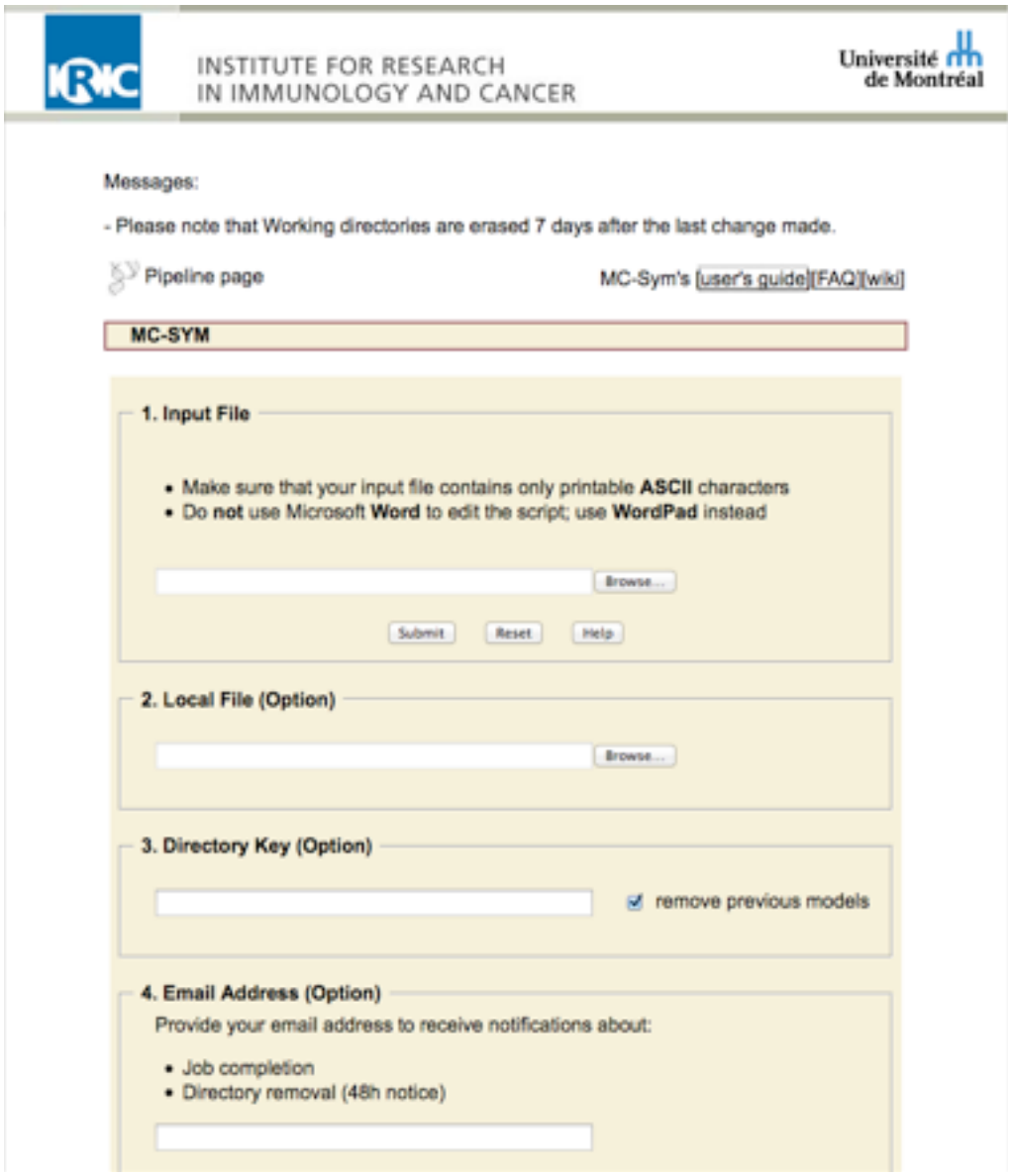
Model Overview

- *Pieper et al. (2004) Nucleic Acids Research 32, D217-D222*

Modeling RNA

MC-FOLD / MC-SYM

http://www.major.iric.ca/MajorLabEn/MC-Tools.html
http://www.major.iric.ca/MC-Sym/



- Coarse-grained model
- Fragment-based approach
- Fully automated

- Slow
- Sampling bottleneck
- Relative poor accuracy

ROSIE

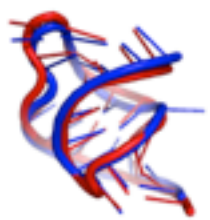
http://rosie.rosettacommons.org/rna_denovo

Welcome to ROSIE

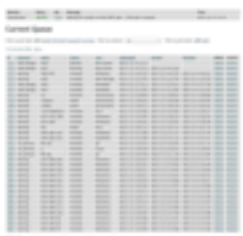
Rosetta Online Server that Includes Everyone

Welcome | Queue | Documentation | Support | Login | Create an account

Rosetta RNA_Denovo Protocol



[Submit RNA_Denovo task]



[RNA Denovo Queue]

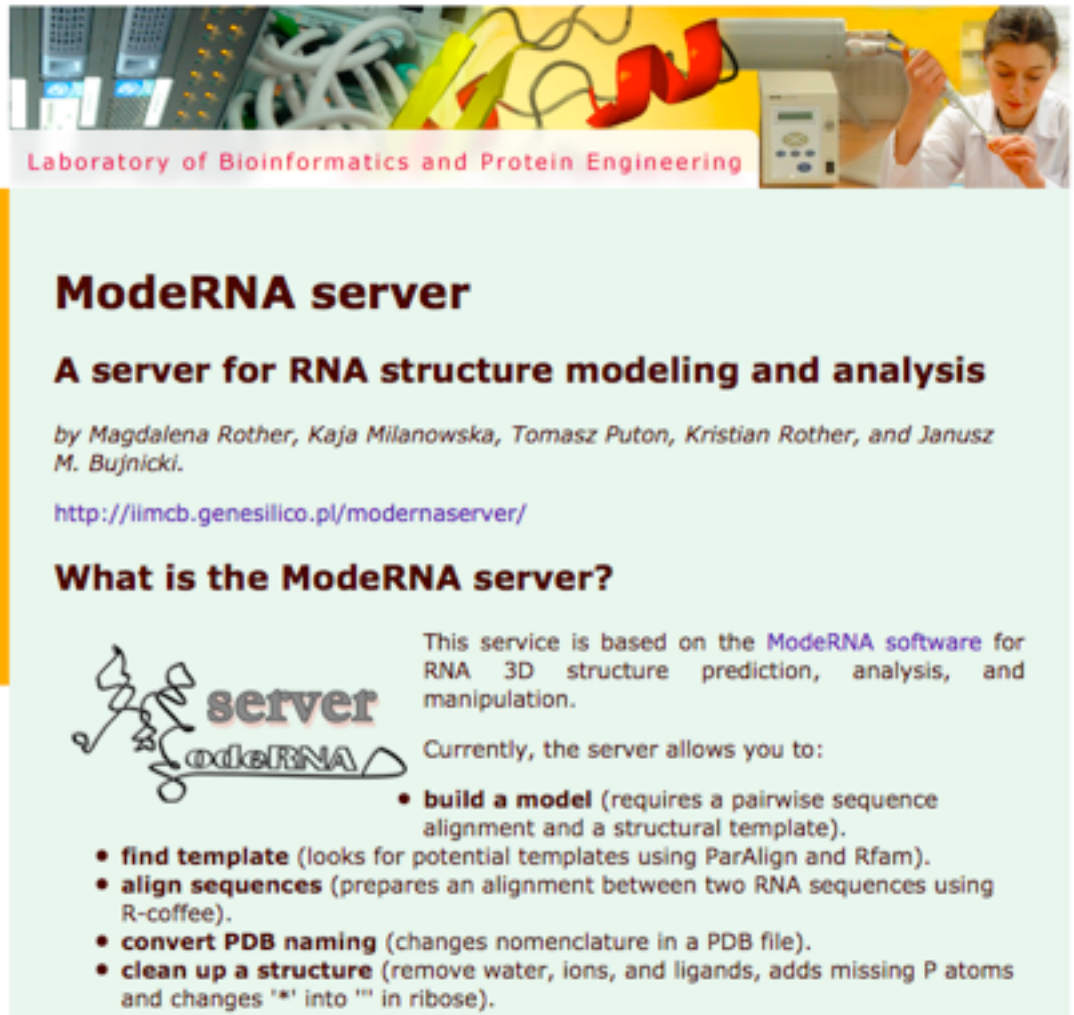
[RNA Denovo Documentation]

Please cite the following article when referring to results from our RNA_Denovo server:

- Das, R., Karanicolas, J., Baker., D. "Atomic accuracy in predicting and designing noncanonical RNA structure". *Nature Methods* 7:291–294. (2010). Online | PDF

MODERNA

http://iimcb.genesilico.pl/modernaserver/



- Comparative RNA modeling
- Able to handle 115 nucleotide modifications