Lecture 3

Protein Structure Prediction

Marc A. Marti-Renom
Assistant Adjunct Professor
Department of Biopharmaceutical Sciences

August 22, 2003

Summary

- Protein Structure Prediction and why is it useful?
- Methods in Protein Structure Prediction
- Comparative Modeling
  - ✓ Steps in CM (overview + resources)
  - ✓ Accuracy/Applications of comparative models
  - ✓ Case example in MODELLER
  - ✓ CM and Structural Genomics

Summary

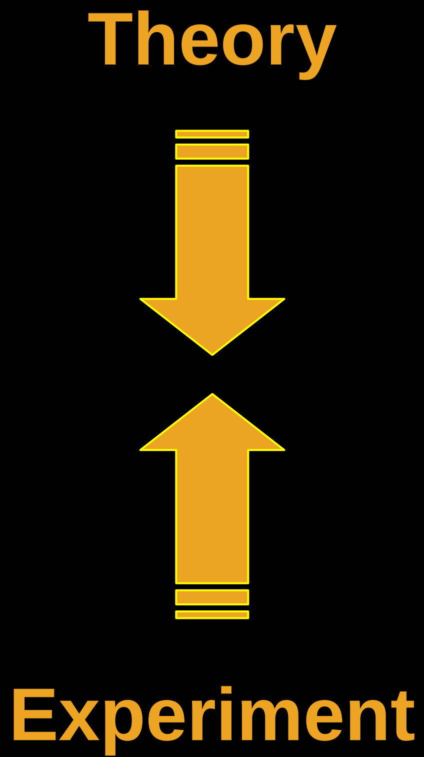
- Protein Structure Prediction and why is it useful?
- Methods in Protein Structure Prediction
- Comparative Modeling
  - ✓ Steps in CM (overview + resources)
  - ✓ Accuracy/Applications of comparative models
  - ✓ Case example in MODELLER
  - ✓ CM and Structural Genomics

Why protein structure prediction?

	Y 2003	Y 2005
Sequences	1,000,000	millions
Structures	28,000	50,000

Why protein structure prediction?

	Y 2003
Sequences	1,000,000
Structures	300,000



Theory

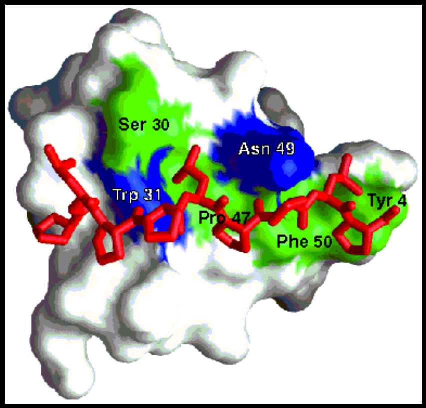
Experiment

http://salilab.org/modbase/

Function *via* Structure

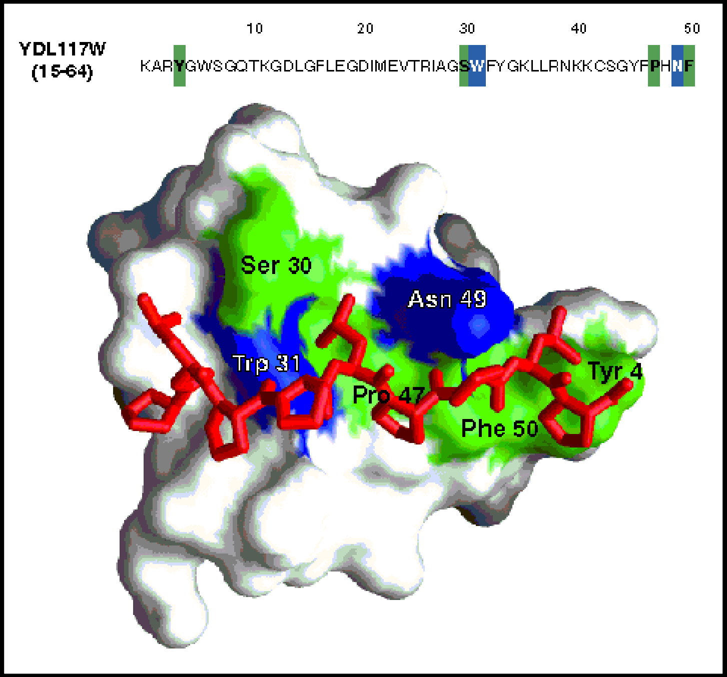
Sequence → **Structure** → Function

ASILPKRLFGNC



Why is it useful to know the structure of a protein, not only its sequence?

- The biochemical function (activity) of a protein is defined by its interactions with other molecules.
- The biological function is in large part a consequence of these interactions.
- The 3D structure is more informative than sequence because interactions are determined by residues that are close in space but are frequently distant in sequence.

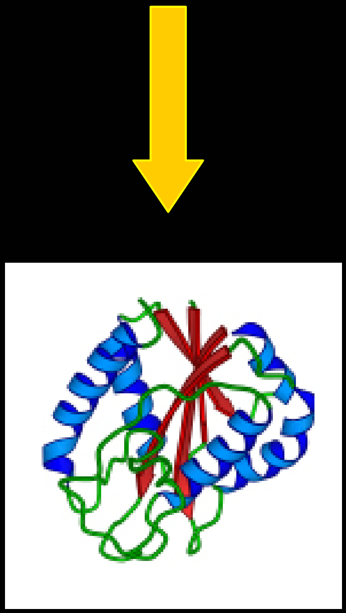


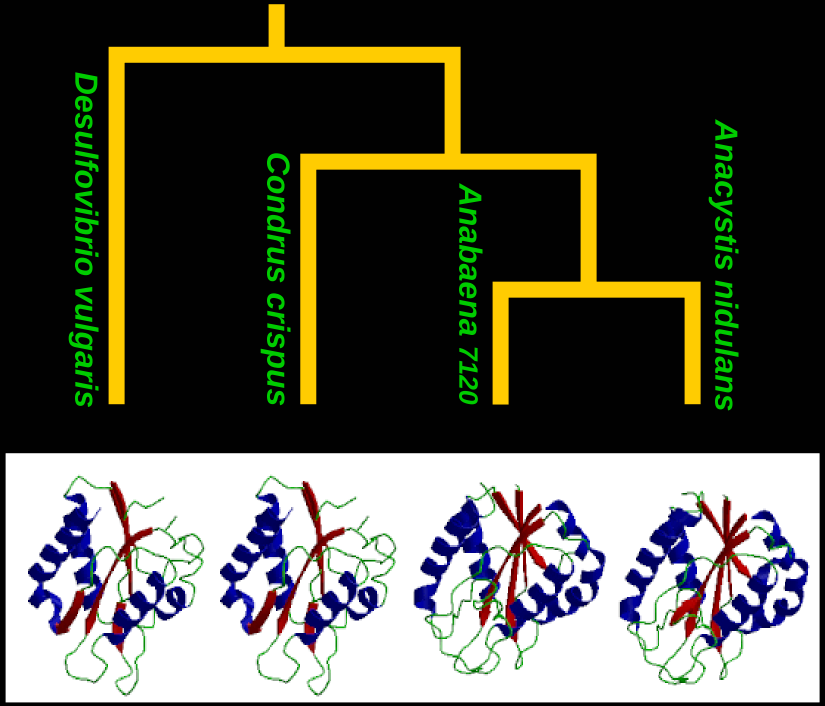
In addition, since evolution tends to conserve function and function depends more directly on structure than on sequence, structure is more conserved in evolution than sequence

The net result is that patterns in space are frequently more recognizable than patterns in sequence.

Principles of Protein Structure

GTCHIKAYTRLIMVG





Folding

Ab initio prediction

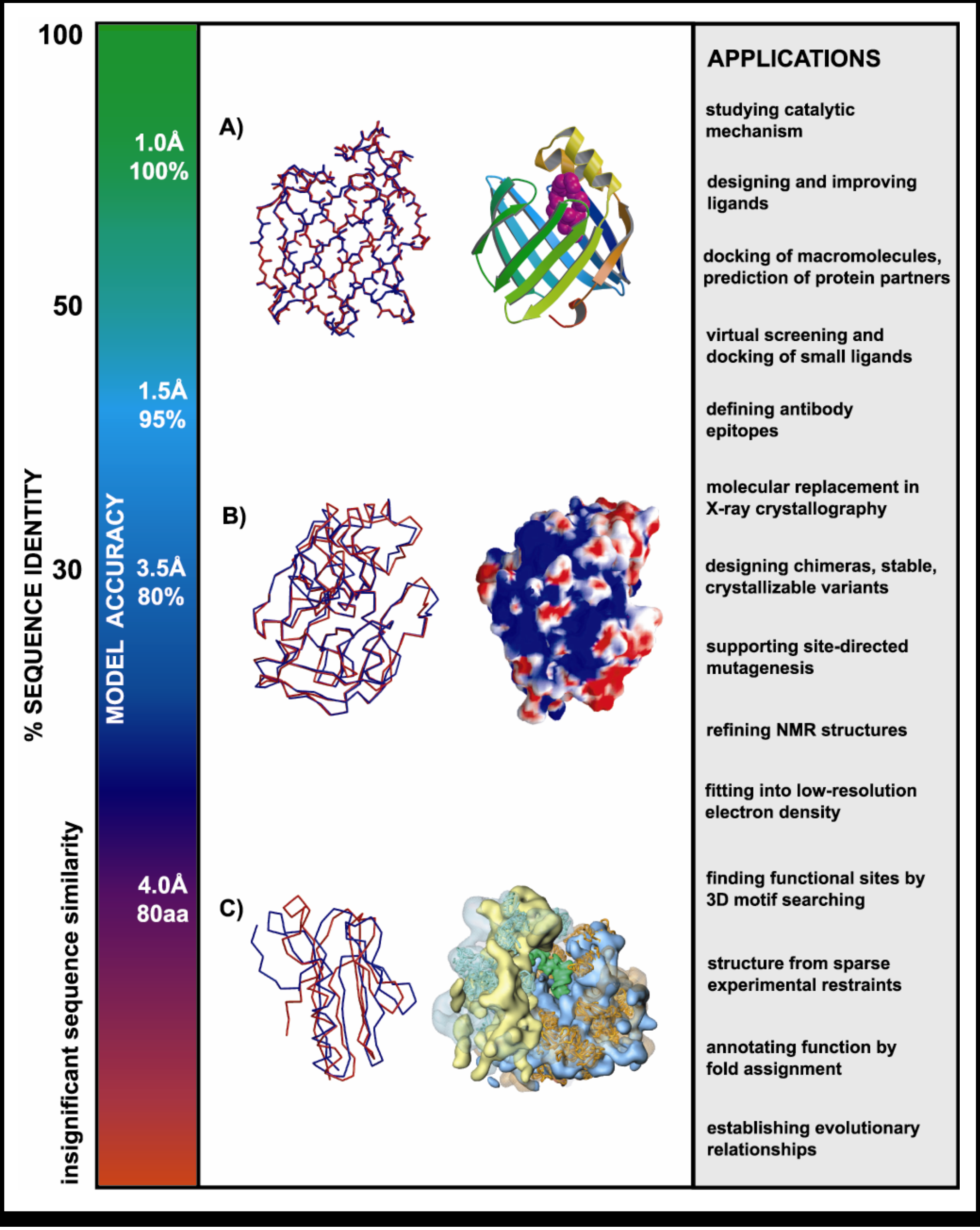
Evolution

Threading
Comparative Modeling

Summary

- Protein Structure Prediction and why is it useful?
- Methods in Protein Structure Prediction
- Comparative Modeling
    - ✓ Steps in CM (overview + resources)
    - ✓ Accuracy/Applications of comparative models
    - ✓ Case example in MODELLER
    - ✓ CM and Structural Genomics

Resolution ←→ Methods



APPLICATIONS

- studying catalytic mechanism
- designing and improving ligands
- docking of macromolecules, prediction of protein partners
- virtual screening 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
- finding functional sites by 3D motif searching
- structure from sparse experimental restraints
- annotating function by fold assignment
- establishing evolutionary relationships

A. Šali & J. Kuriyan.
TIBS **22**, M20, 1999.

Methods for Protein Structure Prediction

- *Ab Initio*
  - *ROSETTA*

    [http://depts.washington.edu/bakerpg/]
- *Threading – Fold assignment*
  - *THREADER*

    [http://www.hgmp.mrc.ac.uk/Registered/Option/threader.html]
- *Comparative Modeling*
  - *MODELLER*

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

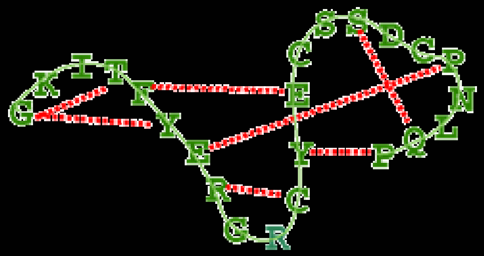
Summary

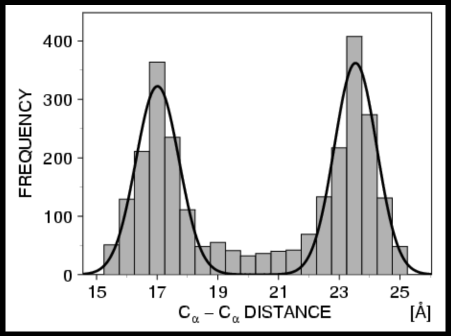
- Protein Structure Prediction and why is it useful?
- Methods in Protein Structure Prediction
- Comparative Modeling
    - ✓ Steps in CM (overview + resources)
    - ✓ Accuracy/Applications of comparative models
    - ✓ Case example in MODELLER
    - ✓ CM and Structural Genomics

Comparative Modeling by Satisfaction of Spatial Restraints (MODELLER)

```
3D  GKITFYERGFQGHCYESDC-NLQP…
SEQ GKITFYERG---RCYESDCPNLQP…
```

1. Extract spatial restraints





2. Satisfy spatial restraints

$$F(\mathbf{R}) = \prod_i p_i(f_i/I)$$

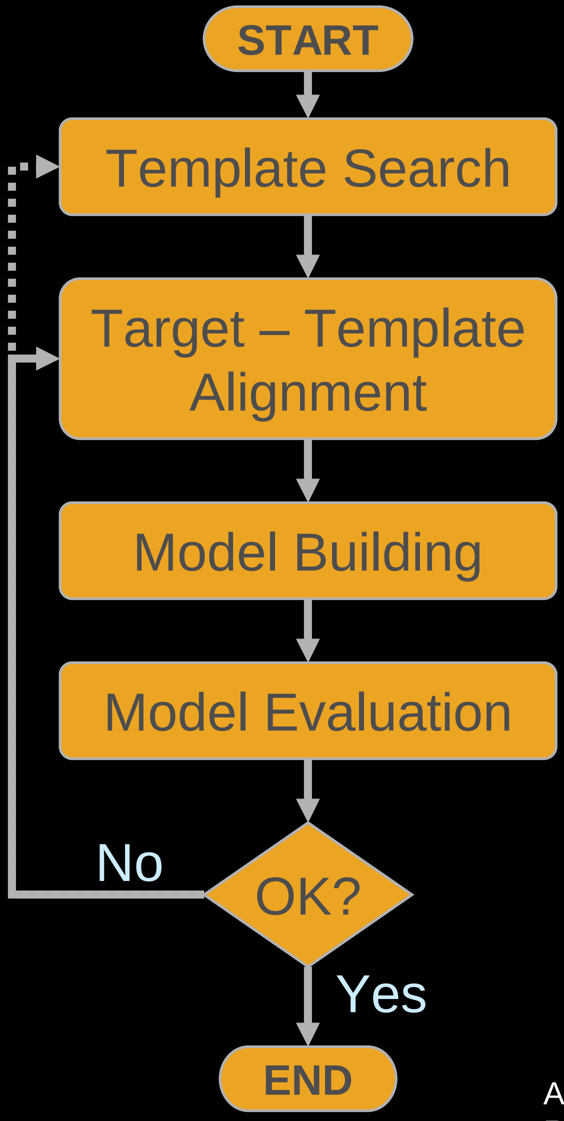
A. Šali & T. Blundell. *J. Mol. Biol.* **234**, 779, 1993.
J.P. Overington & A. Šali. *Prot. Sci*. **3**, 1582, 1994.
A. Fiser, R. Do & A. Šali, *Prot. Sci*., 9, 1753, 2000.

http://www.salilab.org/

Summary

- Protein Structure Prediction and why is it useful?
- Methods in Protein Structure Prediction
- Comparative Modeling
  - ✓ Steps in CM (overview + resources)
  - ✓ Accuracy/Applications of comparative models
  - ✓ Case example in MODELLER
  - ✓ CM and Structural Genomics

Steps in Comparative Protein Structure Modeling



START

Template Search

Target – Template Alignment

Model Building

Model Evaluation

OK?

No

Yes

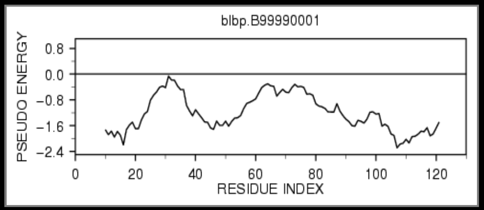
END

TARGET

TEMPLATE

```
ASILPKRLFGNCEQTSDEGLK
IERTPLVPHISAQNVCLKIDD
VPERLIPERASFQWMNDK
```

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



A. Šali, *Curr. Opin. Biotech.* 6, 437, 1995.
R. Sánchez & A. Šali, *Curr. Opin. Str. Biol.* 7, 206, 1997.
M. A. Martí-Renom *et al. Ann. Rev. Biophys. Biomolec. Struct.*, 29, 291, 2000.

Template Search Methods

- Sequence similarity searches
  - BLAST [http://www.ncbi.nlm.nih.gov/BLAST/]
  - FastA program [http://www.ebi.ac.uk/fasta33/]
- Profile and iterative methods
  - HMMs [http://www.cse.ucsc.edu/research/compbio/HMM-apps/]
  - PSI-BLAST [http://www.ncbi.nlm.nih.gov/BLAST/]
- Structure based threading
  - THREADER [http://bioinf.cs.ucl.ac.uk/threader/]
  - PROFIT [http://www.came.sbg.ac.at/]

Target – Template Alignment Methods

- Dynamic Programming Pairwise Alignment
  - ALIGN [http://www.salilab.org/modeller/]
- Multiple Alignments,
  - Psi-Blast [http://www.ncbi.nlm.nih.gov/BLAST/]
  - HMM [http://www.cse.ucsc.edu/research/compbio/HMM-apps/]
  - ALIGN4D [http://www.salilab.org/modeller/]
  - CLUSTALW [http://www.ebi.ac.uk/clustalw/]
- Structure based approaches
  - Threading [http://bioinf.cs.ucl.ac.uk/threader/]

Model Building Methods

- Rigid Body Assembly
    - COMPOSER [http://www-cryst.bioc.cam.ac.uk/]
- Segment Matching
    - SEGMOD
- Satisfaction of Spatial Restraints
    - MODELLER [http://www.salilab.org/modeller/]

Model Evaluation methods

- Stereochemistry
  - PROCHECK/ WHAT-IF [http://www.biochem.ucl.ac.uk/~roman/procheck/procheck.html]
- Environment
  - VERIFY3D [http://www.doe-mbi.ucla.edu/Services/Verify_3D/]
- Statistical potentials based methods
  - PROSAII [http://www.came.sbg.ac.at/]
  - ANOLEA [http://protein.bio.puc.cl/cardex/servers/index.html]

http://www.salilab.org/bioinformatics_resources.shtml

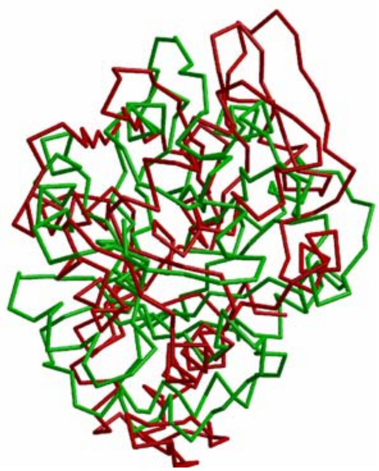
Summary

- Protein Structure Prediction and why is it useful?
- Methods in Protein Structure Prediction
- Comparative Modeling
  - ✓ Steps in CM (overview + some details)
  - ✓ Accuracy/Applications of comparative models
  - ✓ Case example in MODELLER
  - ✓ CM and Structural Genomics

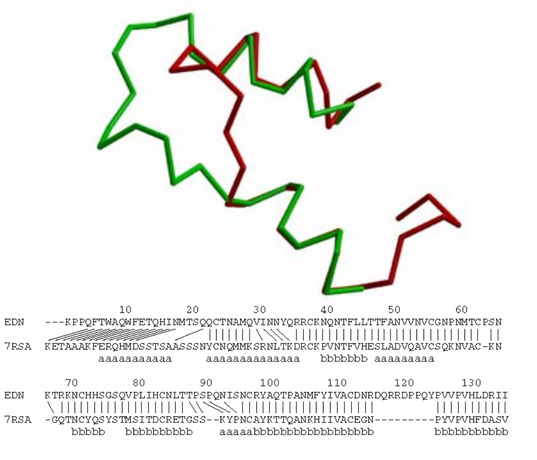
Typical Errors in Comparative Models

MODEL
X-RAY
TEMPLATE

Incorrect template



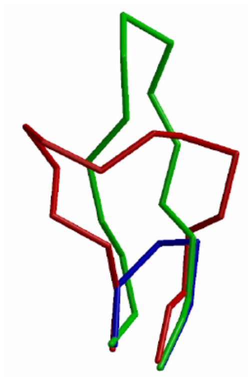
Misalignment



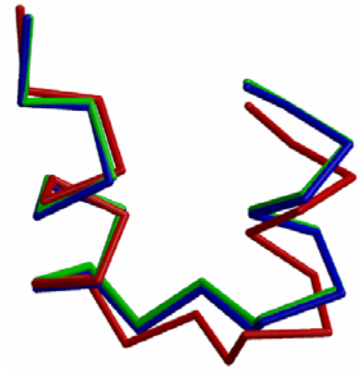
```
                 10        20        30        40        50        60
EDN   ---KPPQFTWAQWFETQHINMTSQQCTNAMQVINNYQRRCKNQNTFLLTTFANVVNVCGNPNMTCPSN
       ////////////////// /  |||||||\ \\\\||||||||||||||||||||||||||| ||
7RSA  KETAAAKFERQHMDSSTSAASSSNYCNQMMKSRNLTKDRCKPVNTFVHESLADVQAVCSQKNVAC-KN
              aaaaaaaaaaa    aaaaaaaaaaaaaa   bbbbbbb aaaaaaaaa

         70        80        90      100       110       120       130
EDN   KTRKNCHHSGSQVPLIHCNLTTPSPQNISNCRYAQTPANMFYIVACDNRDQRRDPPQYPVVPVHLDRII
      \ |||||||||||||||||| \\ \\ \||||||||||||||||||||          |||||||||||
7RSA  -GQTNCYQSYSTMSITDCRETGSS--KYPNCAYKTTQANKHIIVACEGN---------PYVPVHFDASV
         bbbbb    bbbbbbbbbb    aaaaabbbbbbbbbbbbbbbbbb         bbbbbbbbbbb
```

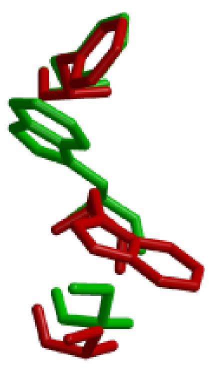
Region without a template



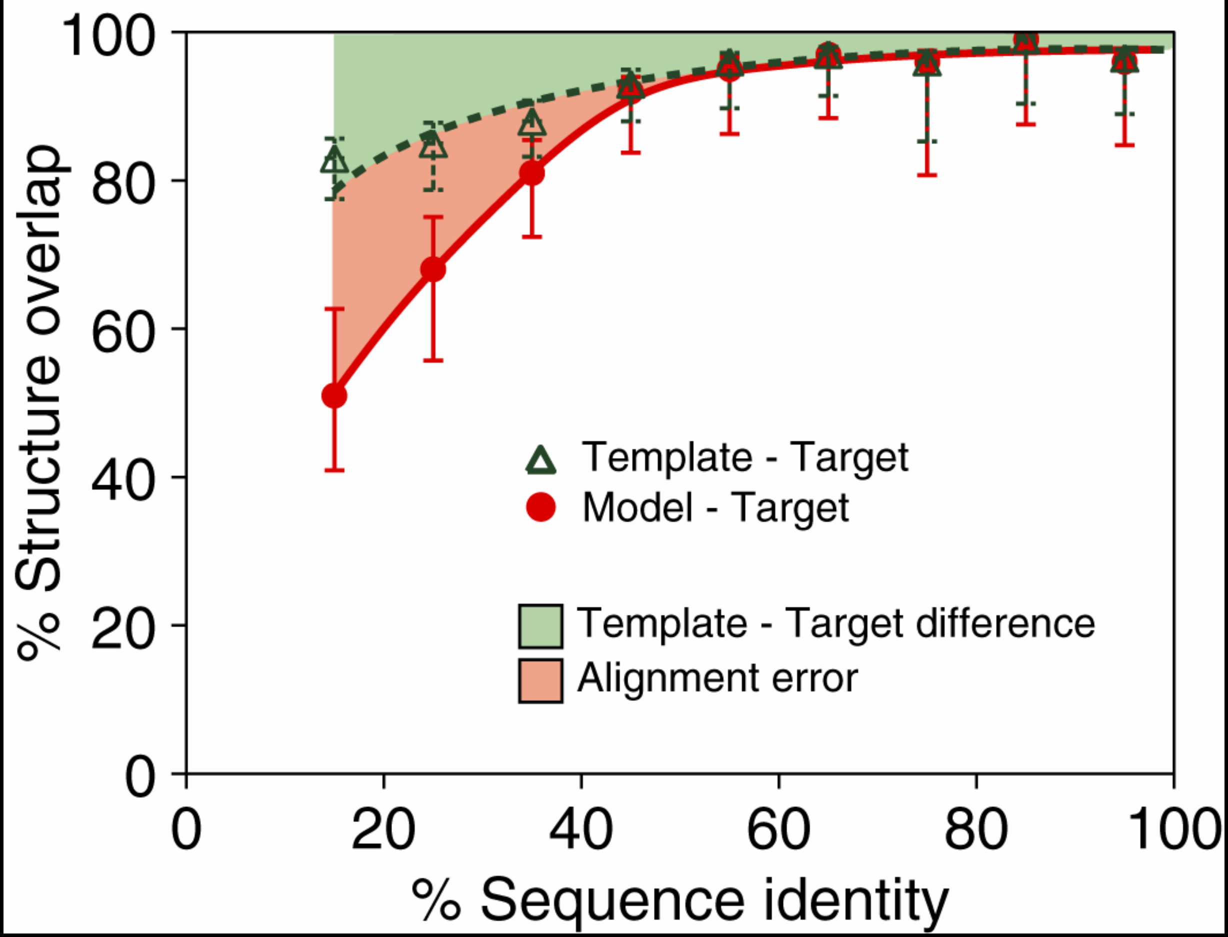
Distortion in correctly aligned regions



Sidechain packing



Model Accuracy as a Function of Target-Template Sequence Identity



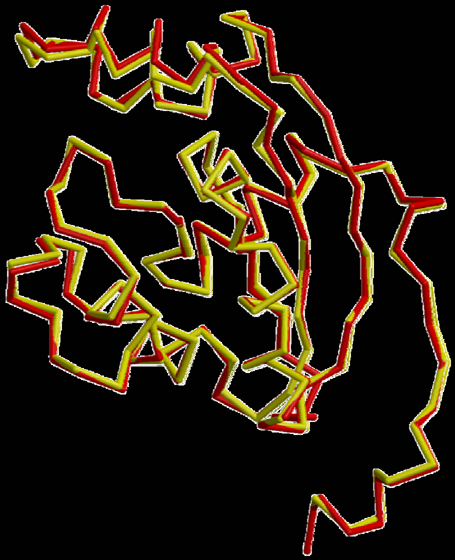
Sánchez, R., Šali, A. *Proc Natl Acad Sci U S A.* 95 pp13597-602. (1998).

Model Accuracy

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

HIGH ACCURACY

NM23
Seq id 77%
Cα equiv 147/148
RMSD 0.41Å

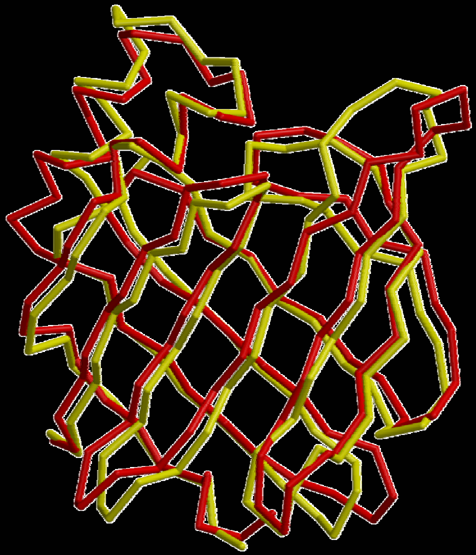


Sidechains

X-RAY **MODEL**

MEDIUM ACCURACY

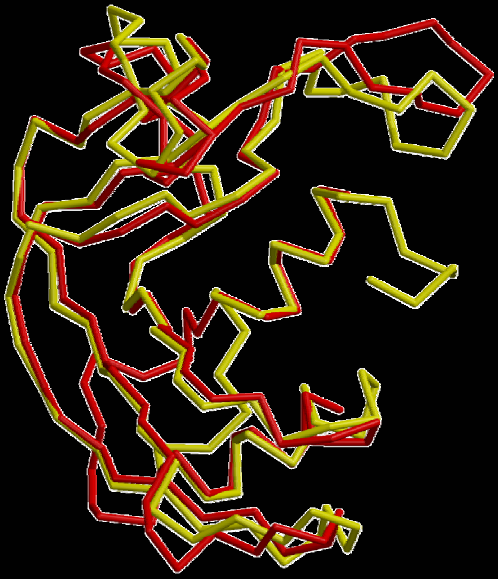
CRABP
Seq id 41%
Cα equiv 122/137
RMSD 1.34Å



Sidechains
Core backbone
Loops

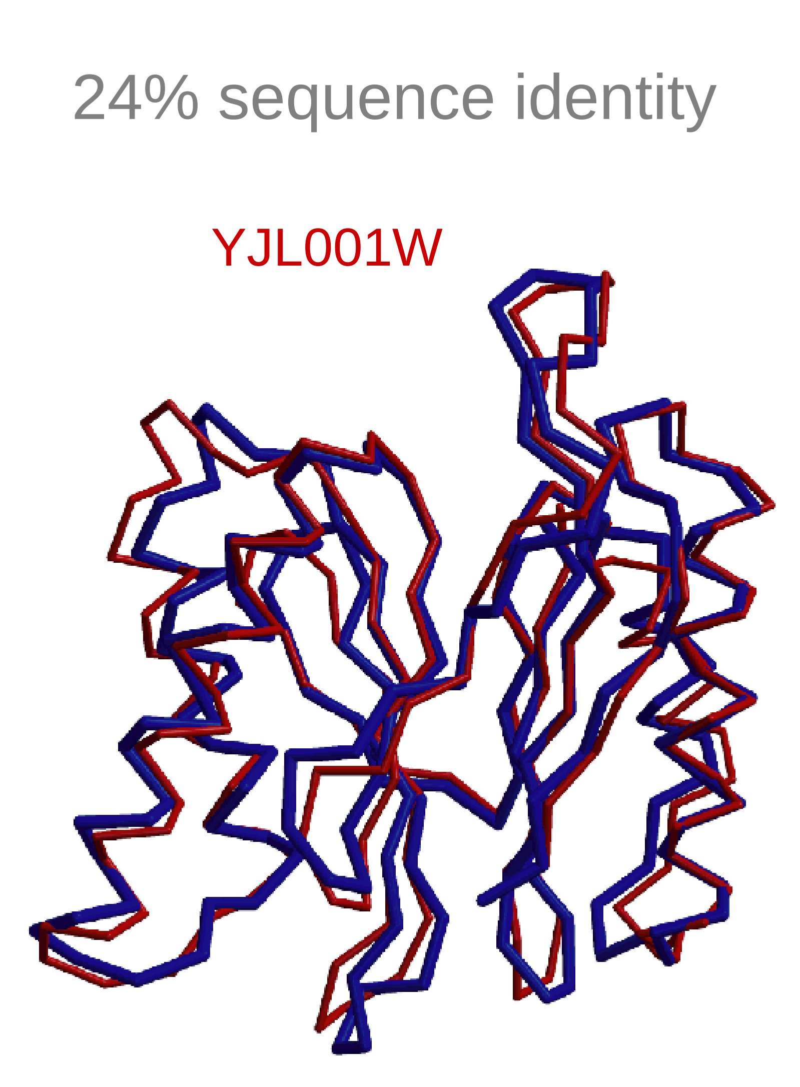
LOW ACCURACY

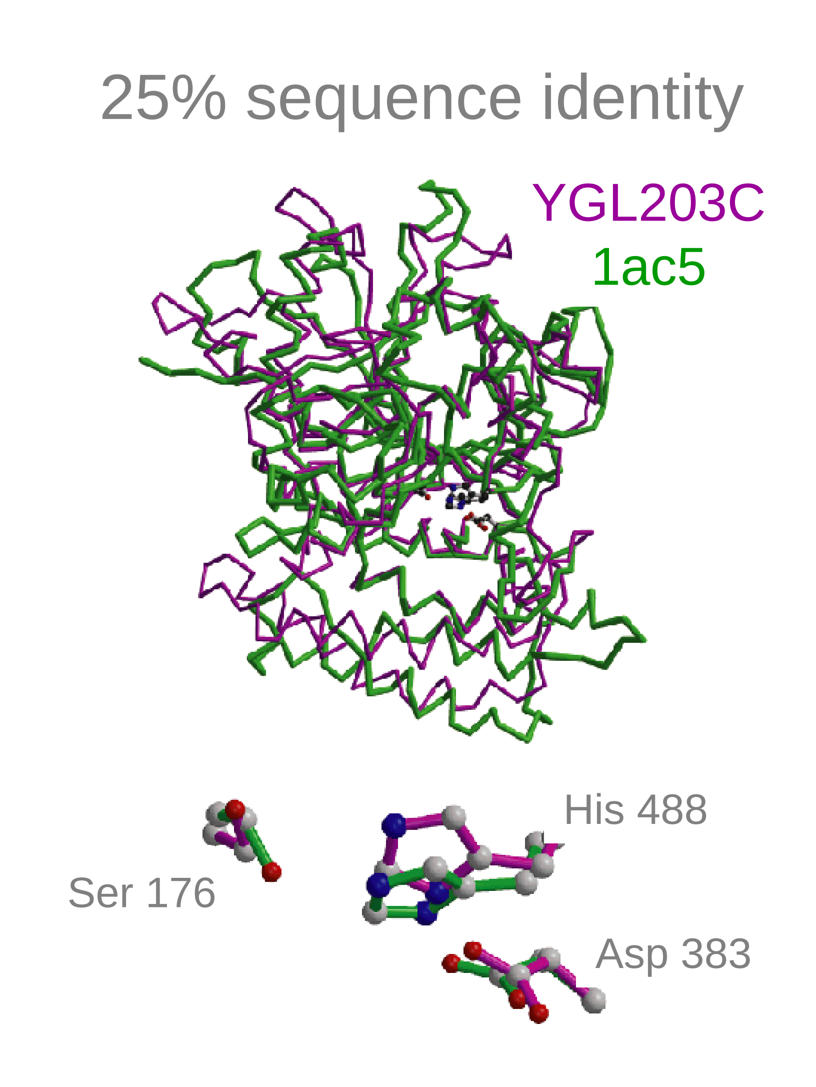
EDN
Seq id 33%
Cα equiv 90/134
RMSD 1.17Å



Sidechains
Core backbone
Loops
Alignment
Fold assignment

Some Models Can Be Surprisingly Accurate
(in Some Core or Active Site Regions)



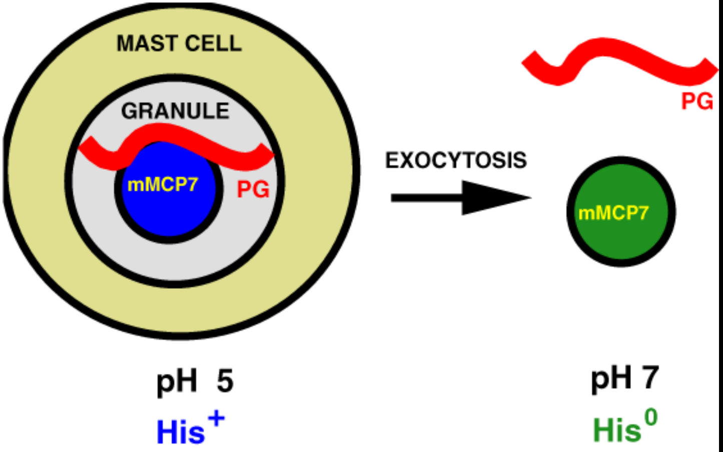


BayGenomics

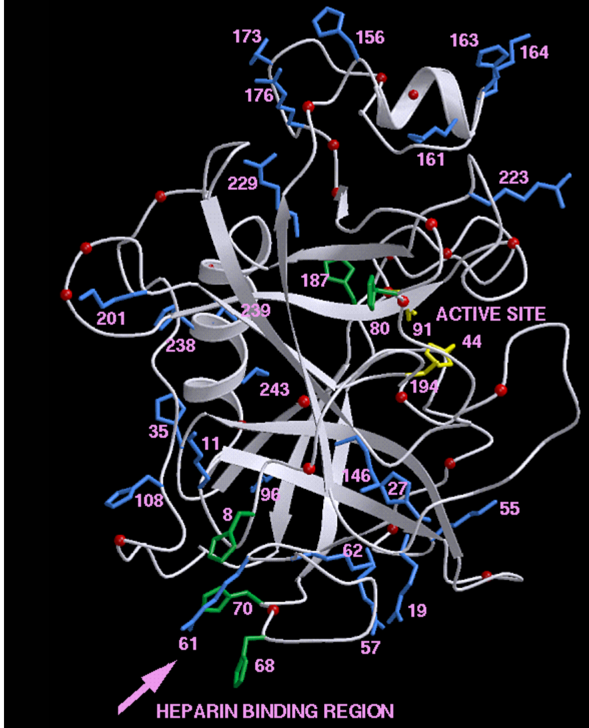
Do mast cell proteases bind proteoglycans? Where? When?

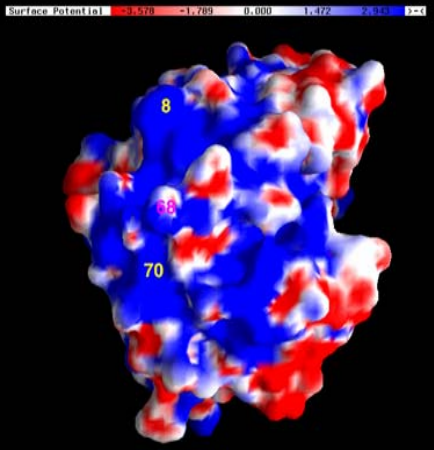
Predicting features of a model that are not present in the template

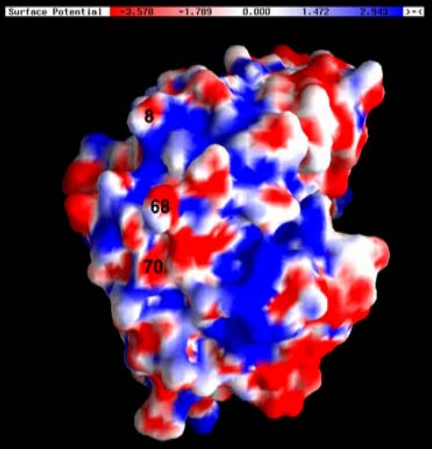
1. mMCPs bind negatively charged proteoglycans through electrostatic interactions?
2. Comparative models used to find clusters of positively charged surface residues.
3. Tested by site-directed mutagenesis.



Huang *et al*. *J. Clin. Immunol*. **18**,169,1998.
Matsumoto *et al*. *J.Biol.Chem*. **270**,19524,1995.
Šali *et al*. *J. Biol. Chem*. **268**, 9023, 1993.

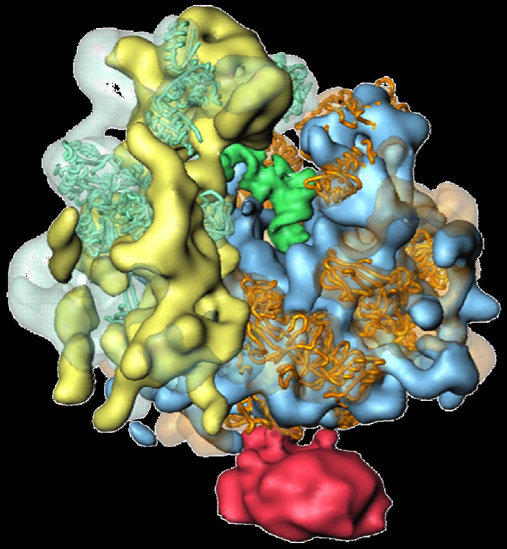


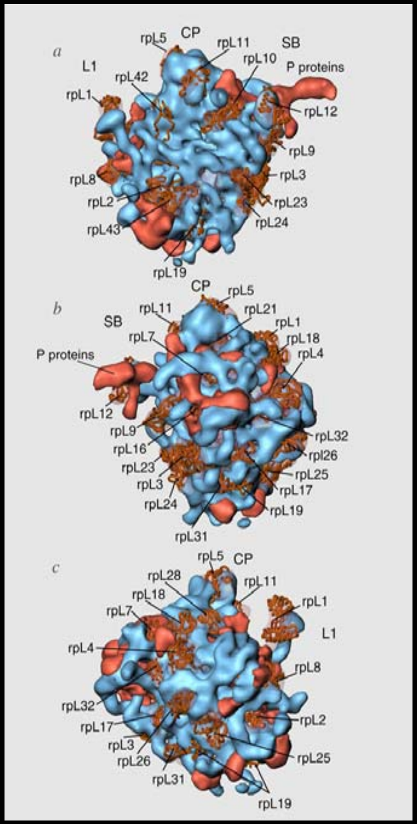




Some Models Can Be Used in Docking to Density Maps

(Yeast Ribosomal 40S subunit)





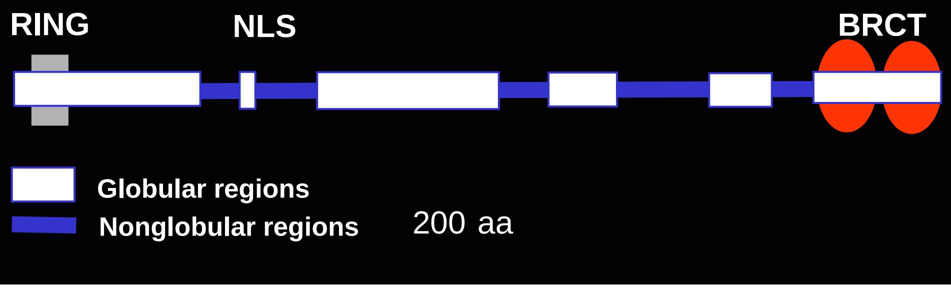
Docking of comparative models into the cryo-EM map.

Spahn *et al.* 2001 Cell **107**:373-386

Small 30S subunit from *Thermus thermophilus*
Large 50S subunit from *Haloarcula marismortui*

Human BRCA1 and its two BRCT domains

(structural analysis of missense mutations SNPs)

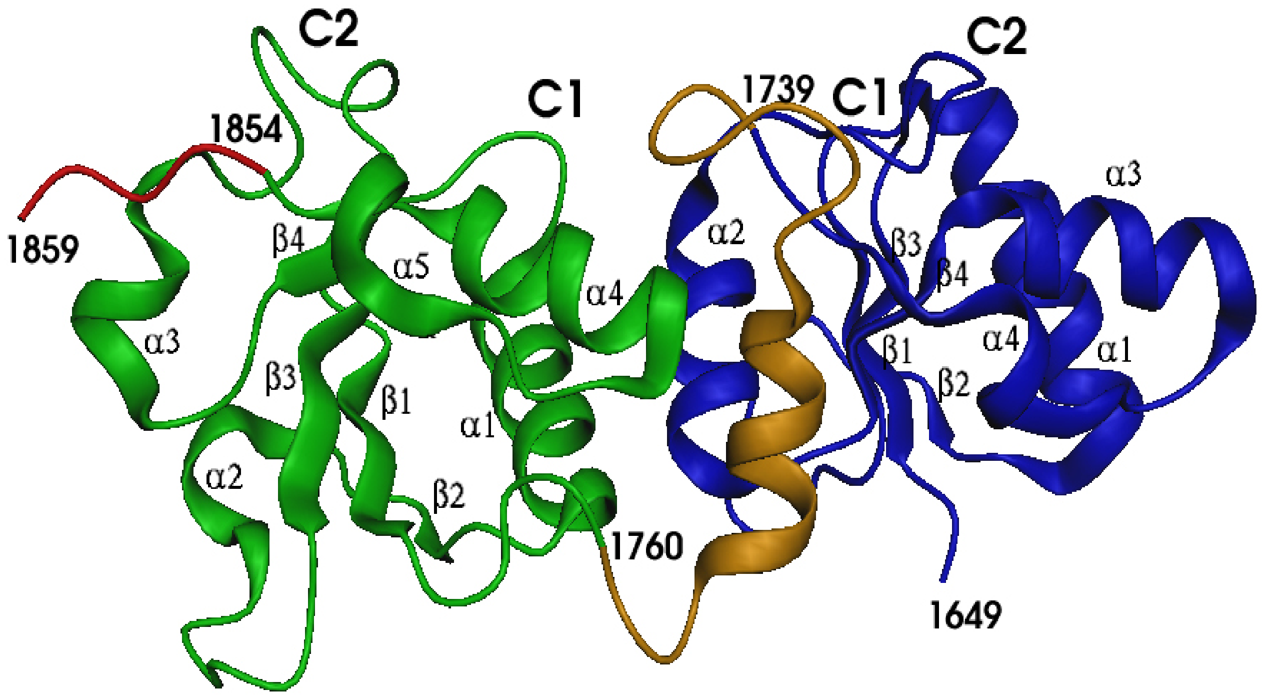


□ Globular regions

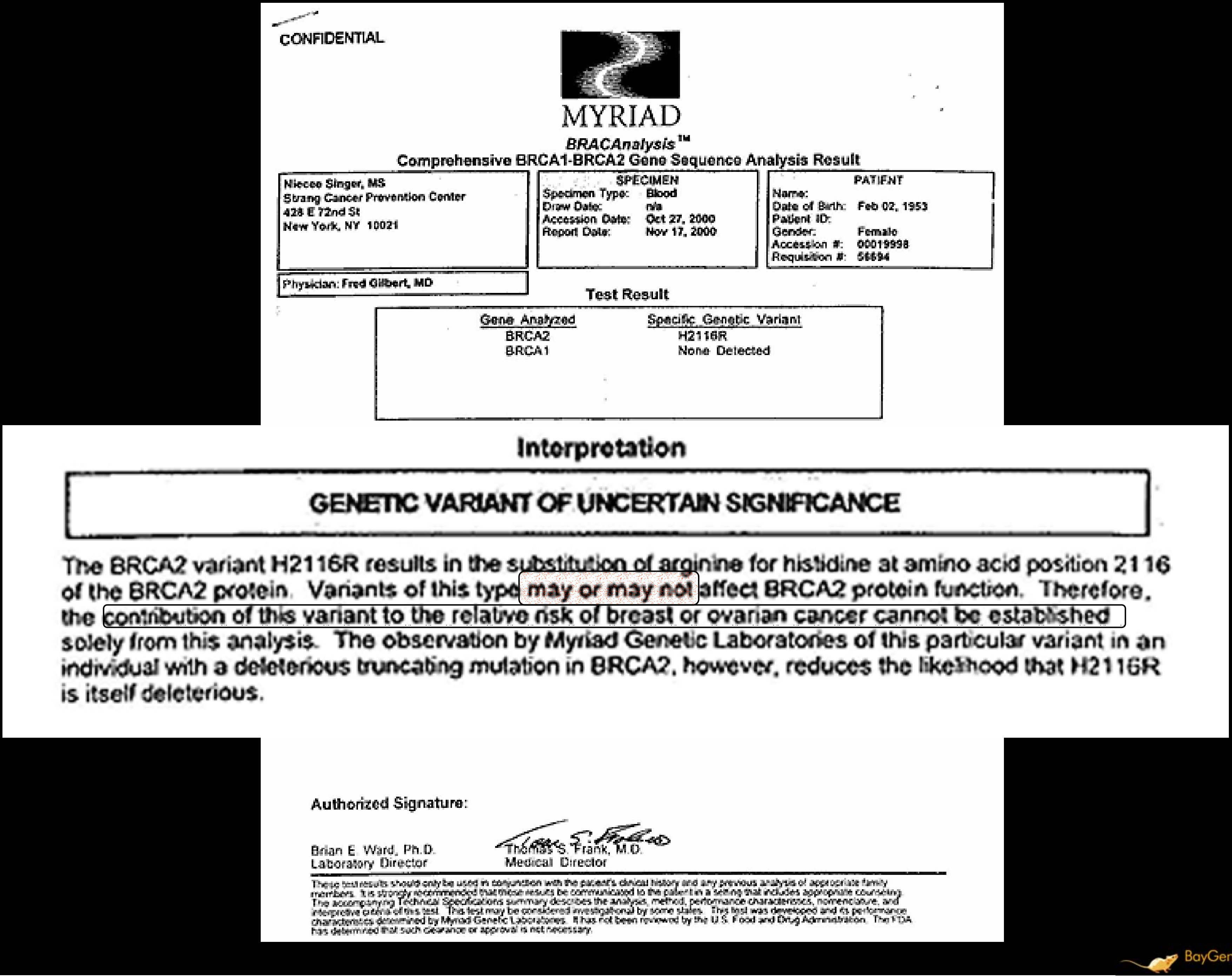
▬ Nonglobular regions

200 aa

BRCA1 BRCT repeats, 1jnx



Williams, Green, Glover. *Nat.Struct.Biol.* 8, 838, 2001



CONFIDENTIAL

MYRIAD

BRACAnalysis™

Comprehensive BRCA1-BRCA2 Gene Sequence Analysis Result

Niecee Singer, MS
Strang Cancer Prevention Center
428 E 72nd St
New York, NY 10021

Physician: Fred Gilbert, MD

SPECIMEN	
Specimen Type:	Blood
Draw Date:	n/a
Accession Date:	Oct 27, 2000
Report Date:	Nov 17, 2000

PATIENT	
Name:	
Date of Birth:	Feb 02, 1953
Patient ID:	
Gender:	Female
Accession #:	00019998
Requisition #:	56694

Test Result

Gene Analyzed	Specific Genetic Variant
BRCA2	H2116R
BRCA1	None Detected

Interpretation

GENETIC VARIANT OF UNCERTAIN SIGNIFICANCE

The BRCA2 variant H2116R results in the substitution of arginine for histidine at amino acid position 2116 of the BRCA2 protein. Variants of this type may or may not affect BRCA2 protein function. Therefore, the contribution of this variant to the relative risk of breast or ovarian cancer cannot be established solely from this analysis. The observation by Myriad Genetic Laboratories of this particular variant in an individual with a deleterious truncating mutation in BRCA2, however, reduces the likelihood that H2116R is itself deleterious.

Authorized Signature:

Brian E. Ward, Ph.D.
Laboratory Director

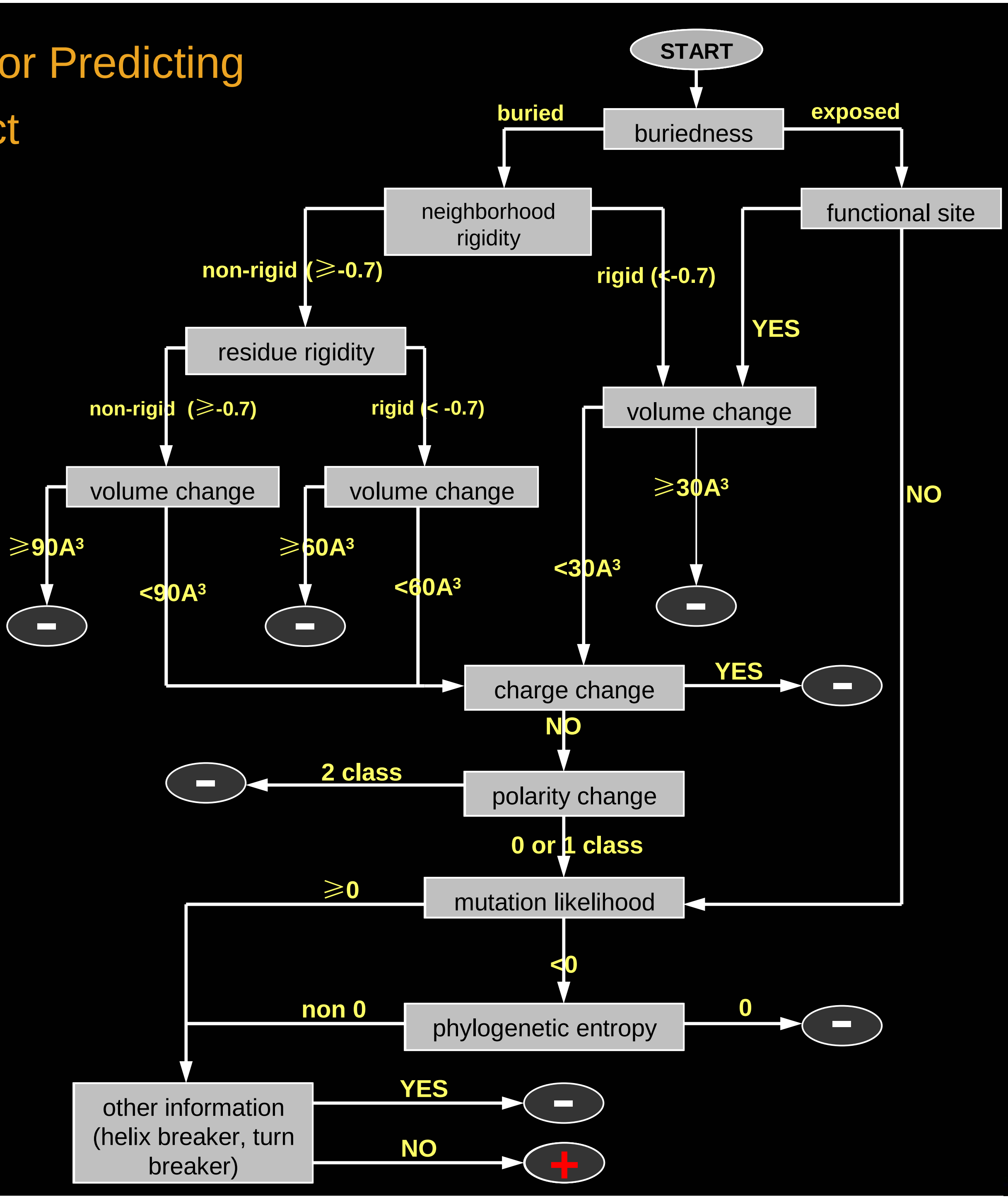
Thomas S. Frank, M.D.
Medical Director

These test results should only be used in conjunction with the patient's clinical history and any previous analysis of appropriate family members. It is strongly recommended that these results be communicated to the patient in a setting that includes appropriate counseling. The accompanying Technical Specifications summary describes the analysis, method, performance characteristics, nomenclature, and interpretive criteria of this test. This test may be considered investigational by some states. This test was developed and its performance characteristics determined by Myriad Genetic Laboratories. It has not been reviewed by the U.S. Food and Drug Administration. The FDA has determined that such clearance or approval is not necessary.

Missense Mutations in BRCT Domains by Function

	cancer associated	not cancer associated	?
no transcription activation	C1697R R1699W A1708E S1715R P1749R M1775R		M1652K L1657P E1660G H1686Q R1699Q K1702E Y1703H F1704S L1705P S1715N S1722F F1734L G1738E G1743R A1752P F1761I F1761S M1775E M1775K L1780P I1807S V1833E A1843T
transcription activation		M1652I A1669S	V1665M D1692N G1706A D1733G M1775V P1806A
?			M1652T V1653M L1664P T1685A T1685I M1689R D1692Y F1695L V1696L R1699L G1706E W1718C W1718S T1720A W1730S F1734S E1735K V1736A G1738R D1739E D1739G D1739Y V1741G H1746N R1751P R1751Q R1758G L1764P I1766S P1771L T1773S P1776S D1778N D1778G D1778H M1783T C1787S G1788D G1788V G1803A V1804D V1808A V1809A V1809F V1810G Q1811R P1812S N1819S A1823T V1833M W1837R W1837G S1841N A1843P T1852S P1856T P1859R

“Decision” Tree for Predicting Functional Impact of Genetic Variants



START

buriedness

buried

exposed

neighborhood rigidity

functional site

non-rigid ($\geq$-0.7)

rigid (<-0.7)

YES

NO

residue rigidity

non-rigid ($\geq$-0.7)

rigid (< -0.7)

volume change

volume change

volume change

$\geq 90A^3$

$<90A^3$

$\geq 60A^3$

$<60A^3$

$\geq 30A^3$

$<30A^3$

−

−

−

charge change

YES

−

NO

polarity change

2 class

−

0 or 1 class

mutation likelihood

$\geq$0

<0

phylogenetic entropy

non 0

0

−

other information (helix breaker, turn breaker)

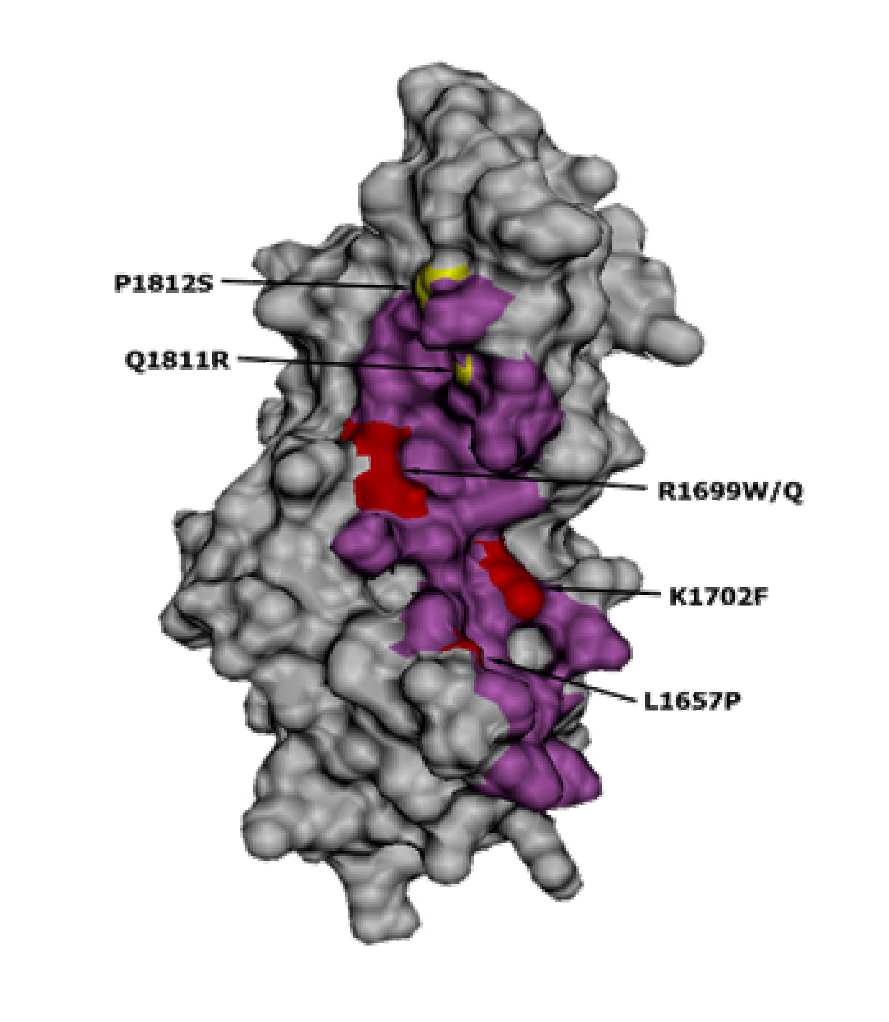
YES

−

NO

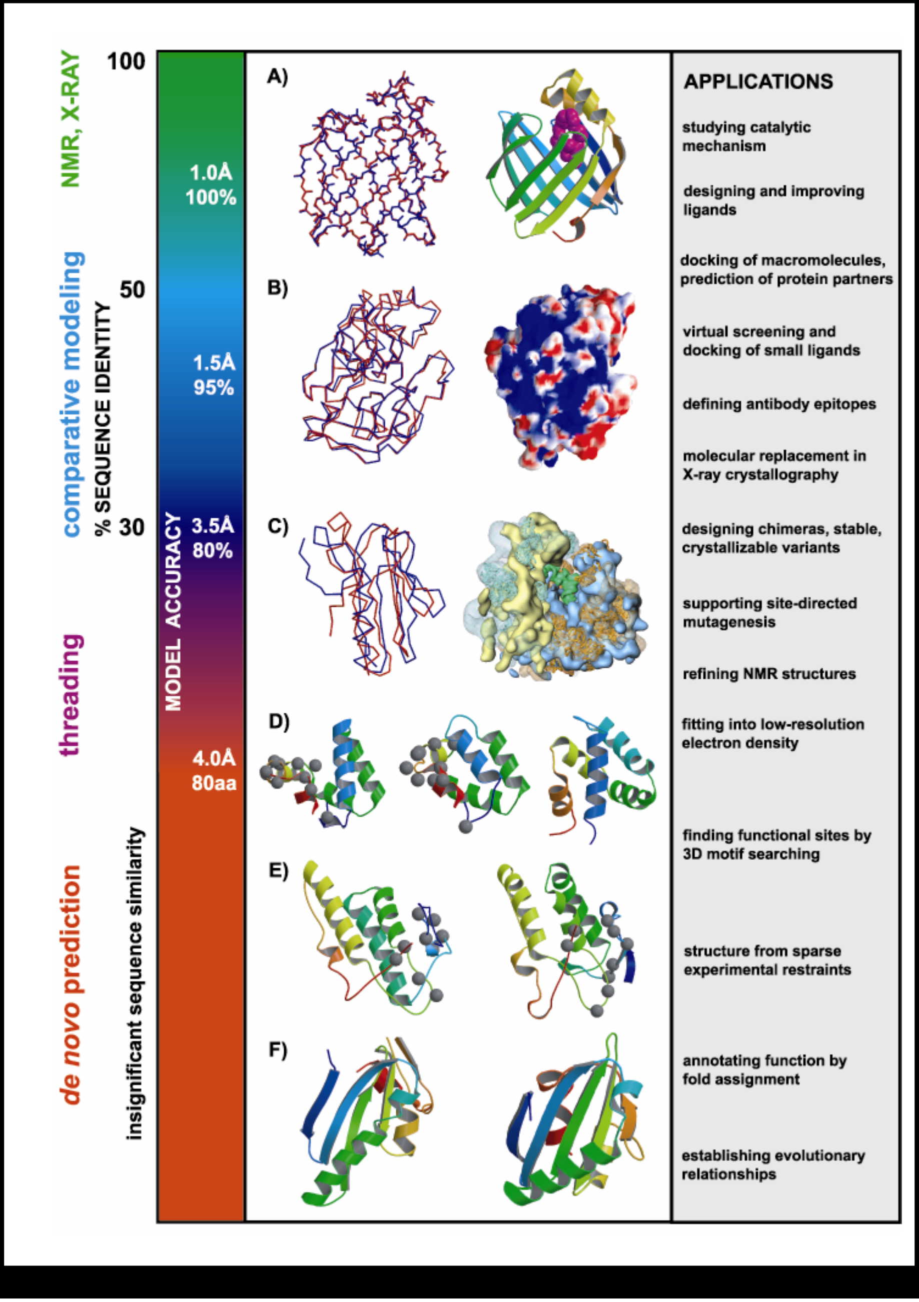
+

Putative Binding Site on BRCA1



RMSMV**VSGL**T**P**EE**FM**LV**Y**KFARKHHIT**L**T**N**LITEETTHVVMKTDAEFVC**ERTLK**Y**F**LGIAGGKWVVSYFWVTQSIKERKM
LNEHDFEVRGDVVNGRNHQGPKRARESQDRKIFRGLEICCYGPF**TNM**PTDQLEWMVQLCGASVVKELSSFTLGTGVHPIV
VV**QPD**AWTEDNGFHAIGQMCEAPVVT**RE**WV**L**DSVALYQCQELDTYLIPQIP

Applications of Comparative Models


NMR, X-RAY
comparative modeling
threading
de novo prediction
% SEQUENCE IDENTITY
100
50
30
insignificant sequence similarity
MODEL ACCURACY
1.0Å
100%
1.5Å
95%
3.5Å
80%
4.0Å
80aa
A)
B)
C)
D)
E)
F)
APPLICATIONS
studying catalytic mechanism
designing and improving ligands
docking of macromolecules, prediction of protein partners
virtual screening 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
finding functional sites by 3D motif searching
structure from sparse experimental restraints
annotating function by fold assignment
establishing evolutionary relationships

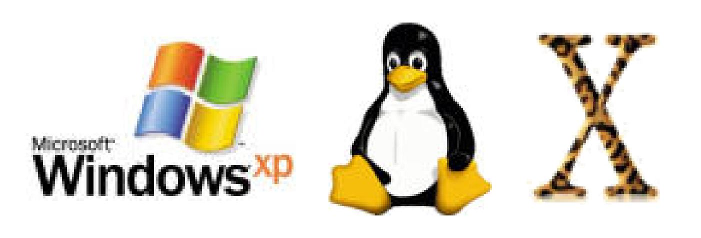
Baker & Sali
Science **294**, 93-96, 2001

Summary

- Protein Structure Prediction and why is it useful?
- Methods in Protein Structure Prediction
- Comparative Modeling
    - ✓ Steps in CM (overview + resources)
    - ✓ Accuracy/Applications of comparative models
    - ✓ Case example in MODELLER
    - ✓ CM and Structural Genomics

Obtaining MODELLER and related information

- MODELLER (6v2) 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) ☹☹
- Script is written in TOP language ☹☹☹
- TOP language is simple ☺☺☺☺

Using MODELLER

- INPUT:
  - Target Sequence (FASTA/PIR format)
  - Template Structure (PDB format)
  - TOP command file
- OUTPUT:
  - Target-Template Alignment
  - Model in PDB format
  - Other data

Example 1: Modeling of BLBP

Input

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

```
>P1;blbp

sequence:blbp::::::::

VDAFCATWKLTDSQNFDEYMKALGVGFATRQVGNVTKPTVIISQEGGKVVIRTQCTFKNTEINFQLGEEFEETSI
DDRNCKSVVRLDGDKLIHVQKWDGKETNCTREIKDGKMVVTLTFGDIVAVRCYEKA*
```

- PSI-BLAST template search: Template: PDB file 1HMS:_

Example 1: Modeling of BLBP

STEP 1: Align blbp and 1hms sequences

TOP script for target-template alignment

```
READ_MODEL FILE = '1hms.pdb'

SEQUENCE_TO_ALI ALIGN_CODES = '1hms'

READ_ALIGNMENT FILE = 'blbp.seq', ALIGN_CODES = 'blbp', ADD_SEQUENCE = on

ALIGN

WRITE_ALIGNMENT FILE='blbp-1hms.ali', ALIGNMENT_FORMAT = 'PIR'

WRITE_ALIGNMENT FILE='blbp-1hms.pap', ALIGNMENT_FORMAT = 'PAP'
```

Run by typing `mod align.top` directory where you have the TOP file.
MODELLER will produce a `align.log` file

BayGenomics

Example 1: Modeling of BLBP

STEP 1: Align blbp and 1hms sequences

TOP script for target-template alignment

```
READ_MODEL FILE = '1hms.pdb'

SEQUENCE_TO_ALI ALIGN_CODES = '1hms'

READ_ALIGNMENT FILE = 'blbp.seq', ALIGN_CODES = 'blbp', ADD_SEQUENCE = on

ALIGN

WRITE_ALIGNMENT FILE='blbp-1hms.ali', ALIGNMENT_FORMAT = 'PIR'

WRITE_ALIGNMENT FILE='blbp-1hms.pap', ALIGNMENT_FORMAT = 'PAP'
```

Run by typing `mod align.top` directory where you have the TOP file.
MODELLER will produce a `align.log` file

Example 1: Modeling of BLBP

STEP 1: Align blbp and 1hms sequences

TOP script for target-template alignment

```
READ_MODEL FILE = '1hms.pdb'

SEQUENCE_TO_ALI ALIGN_CODES = '1hms'

READ_ALIGNMENT FILE = 'blbp.seq', ALIGN_CODES = 'blbp', ADD_SEQUENCE = on

ALIGN

WRITE_ALIGNMENT FILE 'blbp-1hms.ali', ALIGNMENT_FORMAT = 'PIR'

WRITE_ALIGNMENT FILE 'blbp-1hms.pap', ALIGNMENT_FORMAT = 'PAP'
```

Run by typing `mod align.top` directory where you have the TOP file.
MODELLER will produce a `align.log` file

Example 1: Modeling of BLBP

STEP 1: Align blbp and 1hms sequences

TOP script for target-template alignment

```
READ_MODEL FILE = '1hms.pdb'

SEQUENCE_TO_ALI ALIGN_CODES = '1hms'

READ_ALIGNMENT FILE = 'blbp.seq', ALIGN_CODES = 'blbp', ADD_SEQUENCE = on

ALIGN

WRITE_ALIGNMENT FILE='blbp-1hms.ali', ALIGNMENT_FORMAT = 'PIR'

WRITE_ALIGNMENT FILE='blbp-1hms.pap', ALIGNMENT_FORMAT = 'PAP'
```

Run by typing `mod align.top` directory where you have the TOP file. MODELLER will produce a `align.log` file

Example 1: 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*
```

Example 1: 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*
```

Example 1: Modeling of BLBP

STEP 1: Align blbp and 1hms sequences

Output

```
 _aln.pos         10        20        30        40        50        60
1hms      VDAFLGTWKLVDSKNFDDYMKSLGVGFATRQVASMTKPTTIIEKNGDILTLKTHSTFKNT
blbp      VDAFCATWKLTDSQNFDEYMKALGVGFATRQVGNVTKPTVIISQEGGKVVIRTQCTFKNT
 _consrvd ****  **** ** *** *** **********   **** **   *      *  *****

 _aln.pos         70        80        90       100       110       120
1hms      EISFKLGVEFDETTADDRKVKSIVTLDGGKLVHLQKWDGQETTLVRELIDGKLILTLTHG
blbp      EINFQLGEEFEETSIDDRNCKSVVRLDGDKLIHVQKWDGKETNCTREIKDGKMVVTLTFG
 _consrvd ** * ** ** **  ***  ** * *** ** * ***** **   **  ***   *** *

 _aln.pos        130
1hms      TAVCTRTYEKE
blbp      DIVAVRCYEKA
 _consrvd   *  * ***
```

BayGenomics

Example 1: Modeling of BLBP

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

TOP script for model building

```
INCLUDE
SET ALNFILE = 'blbp-1hms.ali'
SET KNOWNS = '1hms'
SET SEQUENCE = 'blbp'
SET STARTING_MODEL = 1
SET ENDING_MODEL = 1
CALL ROUTINE = 'model'
```

Run by typing `mod model.top`.

Check file `model.log`

Example 1: Modeling of BLBP

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

TOP script for model building

```
INCLUDE
SET ALNFILE = 'blbp-1hms.ali'
SET KNOWNS = '1hms'
SET SEQUENCE = 'blbp'
SET STARTING_MODEL = 1
SET ENDING_MODEL = 1
CALL ROUTINE = 'model'
```

Run by typing `mod model.top`.

Check file `model.log`

Example 1: Modeling of BLBP

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

TOP script for model building

```
INCLUDE
SET ALNFILE = 'blbp-1hms.ali'
SET KNOWNS = '1hms'
SET SEQUENCE = 'blbp'
SET STARTING_MODEL = 1
SET ENDING_MODEL = 1
CALL ROUTINE = 'model'
```

Run by typing `mod model.top`.
Check file `model.log`

Example 1: Modeling of BLBP

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

Output coordinates file

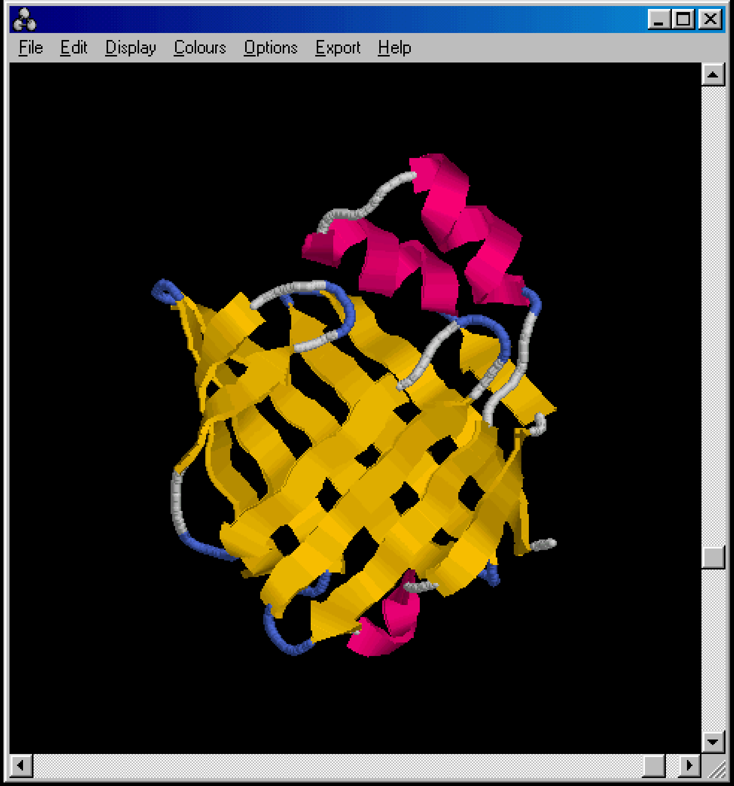
Model file → `blbp.B99990001`

- PDB file
- Can be viewed with Chimera

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

Rasmol

[http://www.bernstein-plus-sons.com/software/rasmol/]

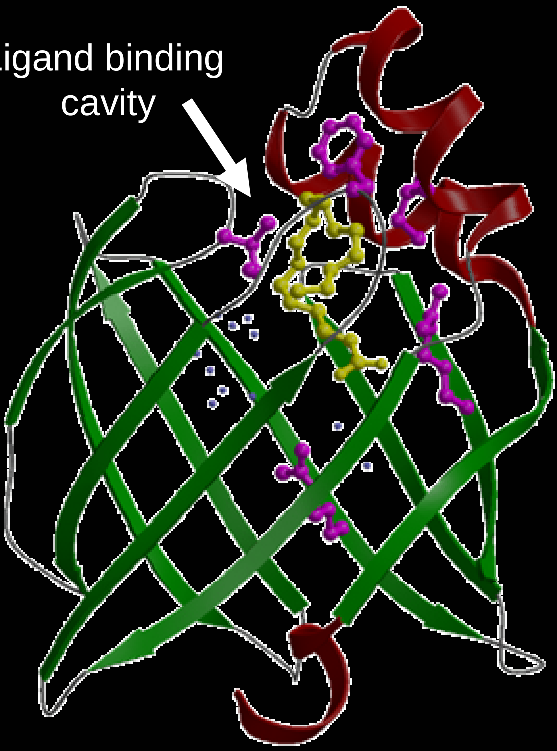


What is the physiological ligand of Brain Lipid-Binding Protein?

Predicting features of a model that are not present in the template

BLBP/oleic acid

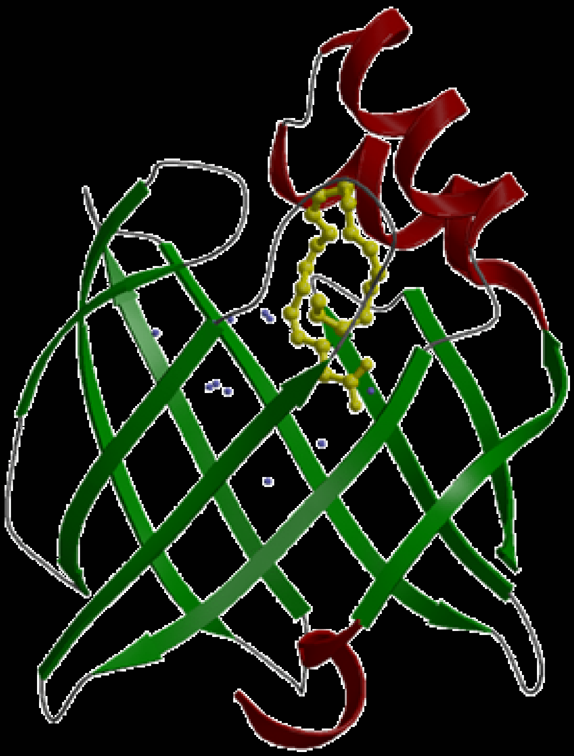
*Cavity is **not** filled*



Ligand binding cavity

BLBP/Docosahexaenoic acid

*Cavity **is** filled*



1. BLBP binds fatty acids.
2. Build a 3D model.
3. Find the fatty acid that fits most snuggly into the ligand binding cavity.

L. Xu, R. Sánchez, A. Šali, N. Heintz, *J. Biol. Chem.* **271**, 24711, 1996.

Summary

- Protein Structure Prediction and why is it useful?
- Methods in Protein Structure Prediction
- Comparative Modeling
    - ✓ Steps in CM (overview + some details)
    - ✓ Accuracy of comparative models
    - ✓ Case example in MODELLER
    - ✓ CM and Structural Genomics

Structural Genomics

- **Definition:**
  - The aim of structural genomics is to put every protein sequence within a modeling distance of a known protein structure.

- **Size of the problem:**
  - There are a few thousand domain fold families.
  - There are ~16,000 sequence families (30% sequence id).

- **Solution:**
  - Determine many protein structures.
  - Increase modeling distance.

Šali. *Nat. Struct. Biol*. **5**, 1029, 1998.
Šali & Kuriyan. *TIBS* **22**, M20, 1999.

Burley *et al*. *Nat. Genet*. **23**, 151, 1999.
Sanchez *et al. Nat. Str. Biol.* 7, 986, 2000

How can Comparative Modeling be used in Structural Genomics?

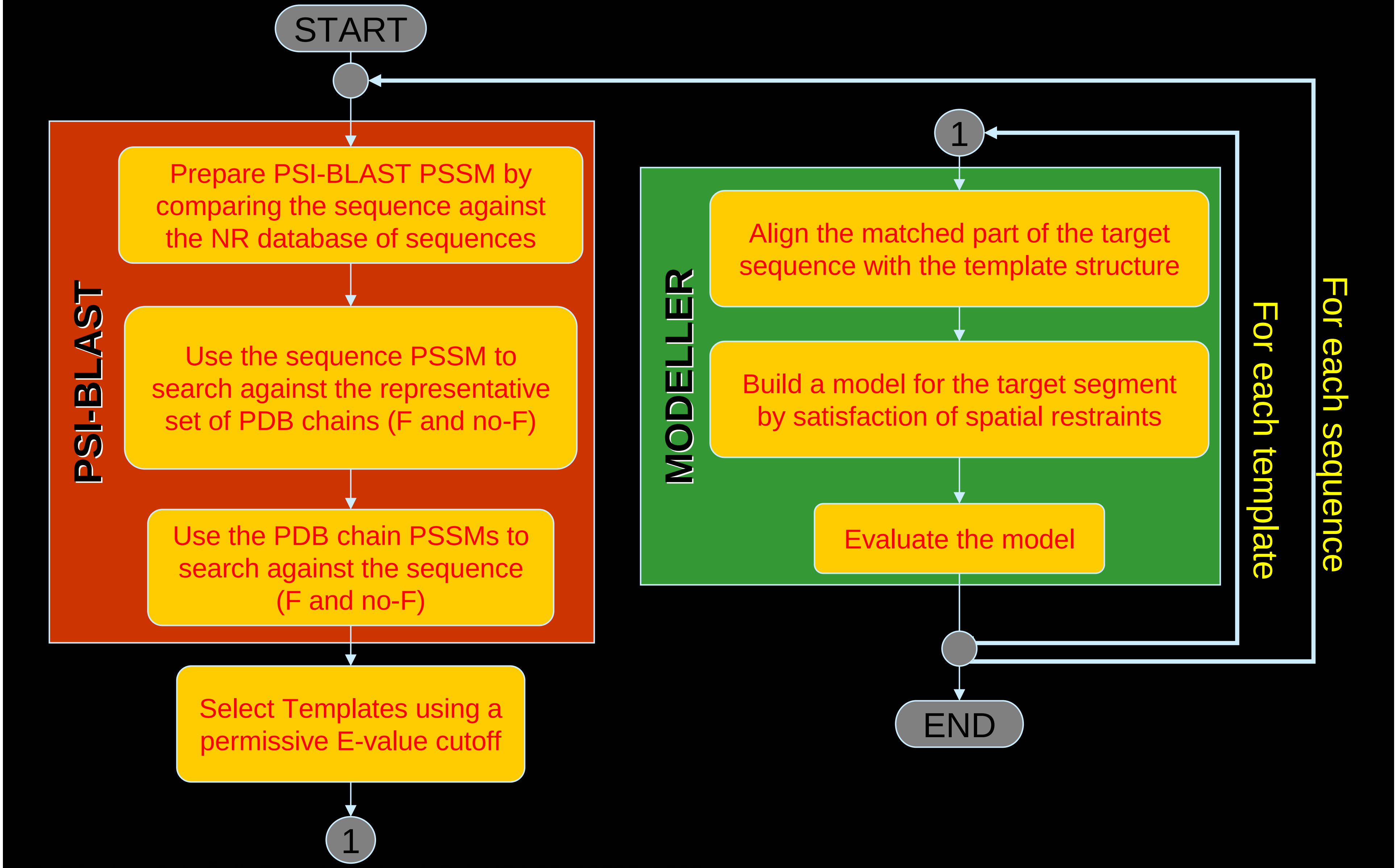
- Target Selection

  How many structures need to be solved?
  Which structures should we solve first?

- Target Amplification

  How much of the sequence space is covered by:
  a new structure
  all structures

MODPIPE: Large-Scale Comparative Protein Structure Modeling



Prepare PSI-BLAST PSSM by comparing the sequence against the NR database of sequences

Use the sequence PSSM to search against the representative set of PDB chains (F and no-F)

Use the PDB chain PSSMs to search against the sequence (F and no-F)

Select Templates using a permissive E-value cutoff

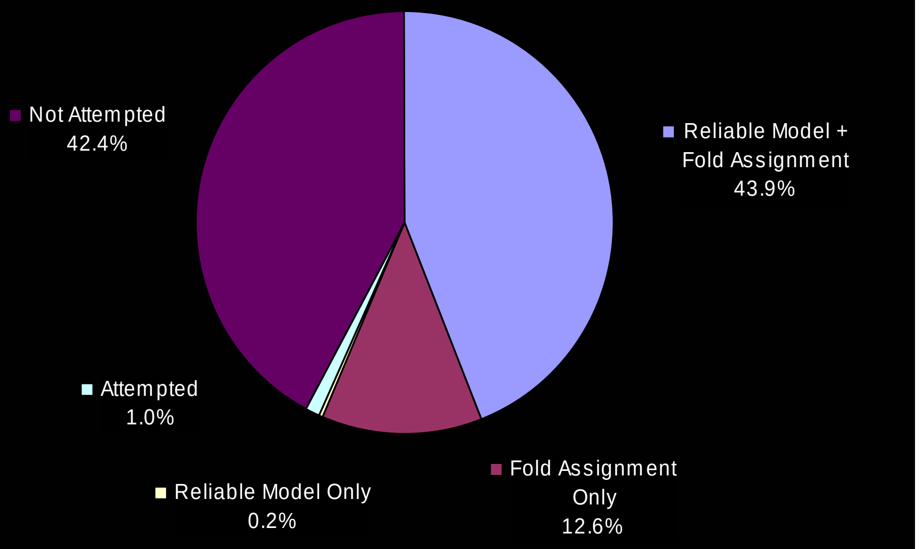
Align the matched part of the target sequence with the template structure

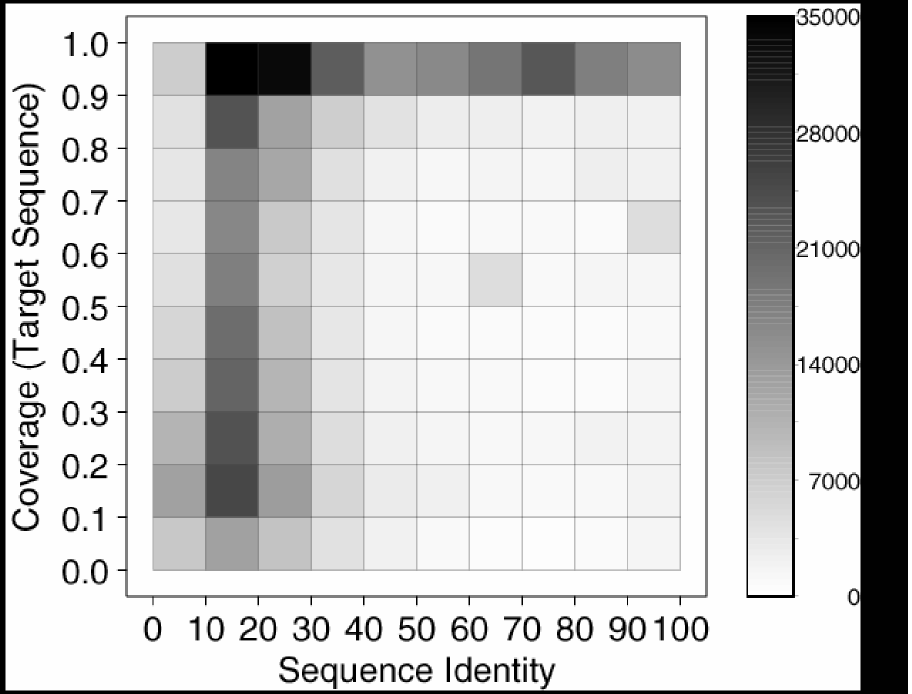
Build a model for the target segment by satisfaction of spatial restraints

Evaluate the model

BayGenomics

Modeling Coverage Of The Sequence Space



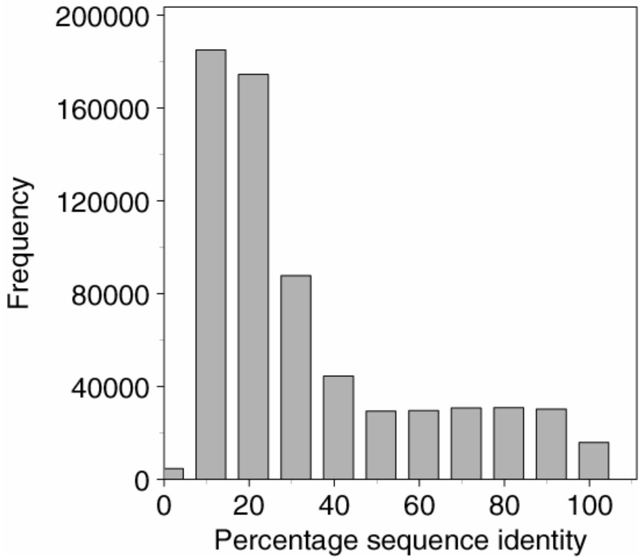


Fold assignment: PSI-BLAST E-value ≤ $1e^{-4}$

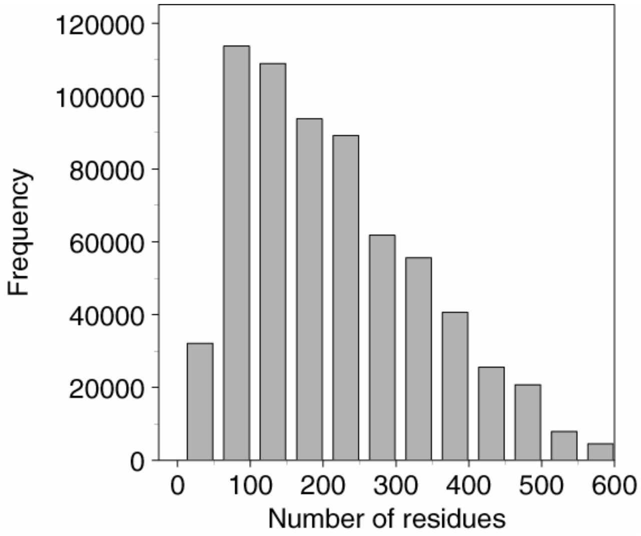
Reliable Model: Model Score ≥ 0.7

Comparative Models for TrEMBL Sequences

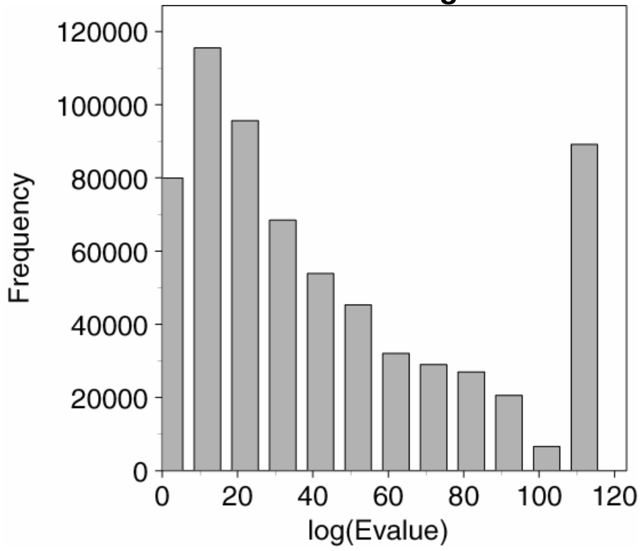
Target-Template Sequence Identity



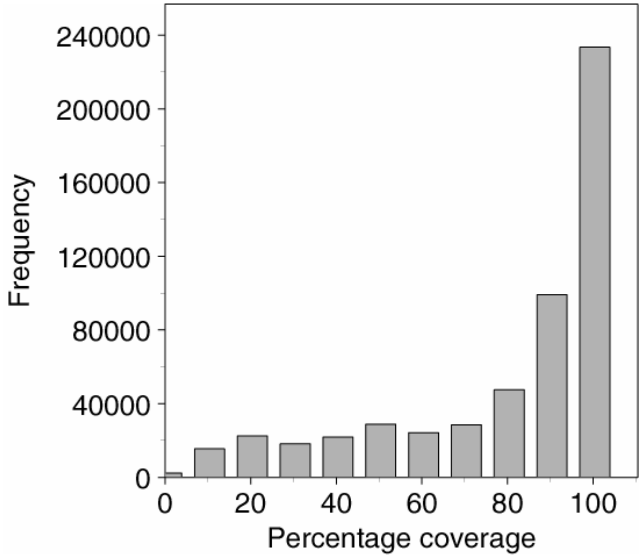
Model Length



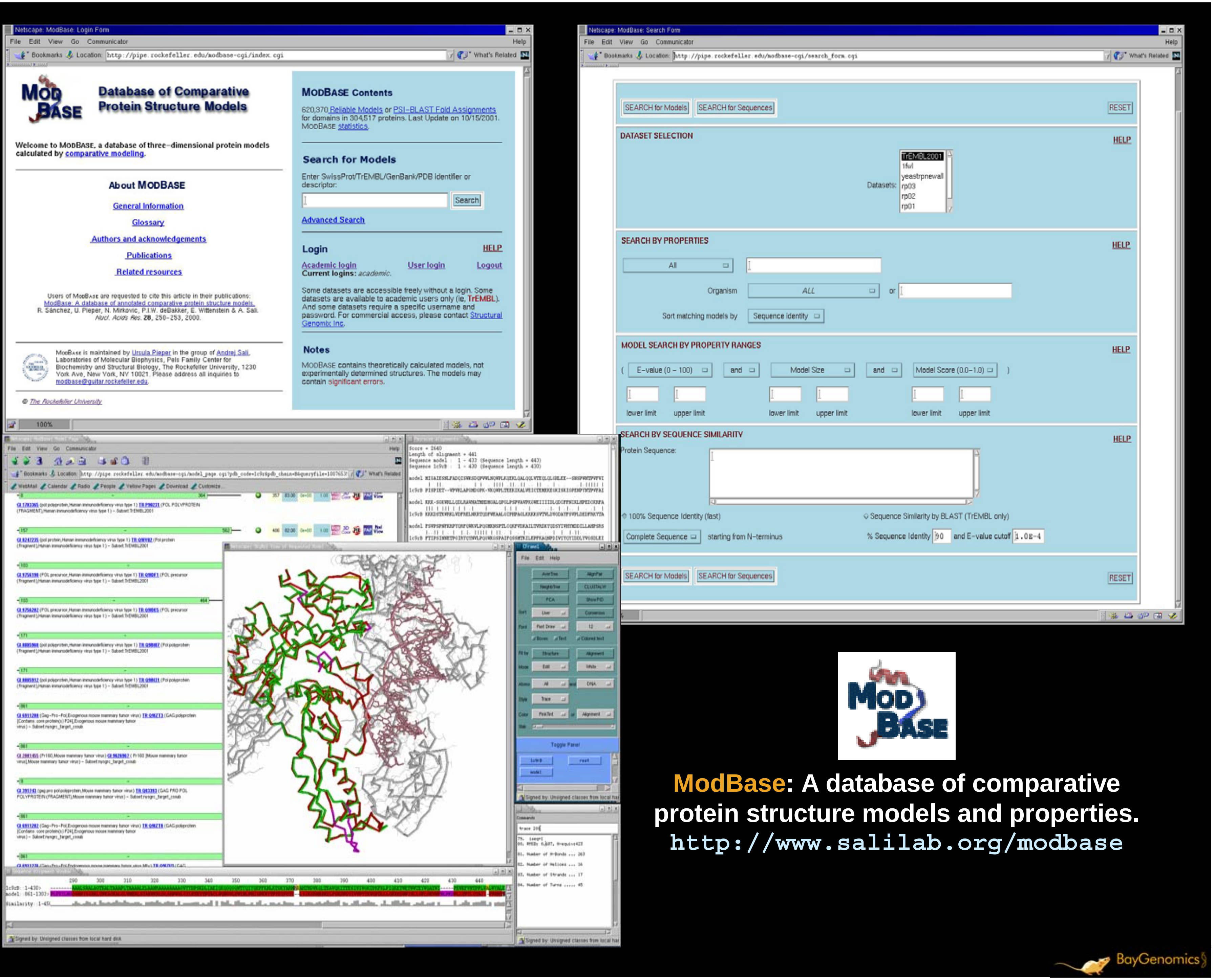
E-values of Alignment



Coverage of SCOP domains



BayGenomics



Netscape: ModBase: Login Form

http://pipe.rockefeller.edu/modbase-cgi/index.cgi

Database of Comparative Protein Structure Models

Welcome to MODBASE, a database of three-dimensional protein models calculated by comparative modeling.

About MODBASE

General Information

Glossary

Authors and acknowledgements

Publications

Related resources

Users of MODBASE are requested to cite this article in their publications:
ModBase: A database of annotated comparative protein structure models.
R. Sánchez, U. Pieper, N. Mirkovic, P.I.W. deBakker, E. Wittenstein & A. Sali.
Nucl. Acids Res. **28**, 250-253, 2000.

MODBASE is maintained by Ursula Pieper in the group of Andrej Sali, Laboratories of Molecular Biophysics, Pels Family Center for Biochemistry and Structural Biology, The Rockefeller University, 1230 York Ave, New York, NY 10021. Please address all inquiries to modbase@guitar.rockefeller.edu.

© *The Rockefeller University*

MODBASE Contents

620,370 Reliable Models or PSI-BLAST Fold Assignments for domains in 304,517 proteins. Last Update on 10/15/2001. MODBASE statistics.

Search for Models

Enter SwissProt/TrEMBL/GenBank/PDB identifier or descriptor:

[ ] Search

Advanced Search

Login HELP

Academic login User login Logout
Current logins: *academic.*

Some datasets are accessible freely without a login. Some datasets are available to academic users only (ie, TrEMBL). And some datasets require a specific username and password. For commercial access, please contact Structural Genomix Inc.

Notes

MODBASE contains theoretically calculated models, not experimentally determined structures. The models may contain significant errors.

Netscape: ModBase: Search Form

http://pipe.rockefeller.edu/modbase-cgi/search_form.cgi

SEARCH for Models SEARCH for Sequences RESET

DATASET SELECTION HELP

Datasets: TrEMBL2001, 1fwl, yeastrpnewall, rp03, rp02, rp01

SEARCH BY PROPERTIES HELP

All [ ]

Organism ALL or [ ]

Sort matching models by Sequence identity

MODEL SEARCH BY PROPERTY RANGES HELP

(E-value (0 - 100) and Model Size and Model Score (0.0-1.0))

lower limit upper limit lower limit upper limit lower limit upper limit

SEARCH BY SEQUENCE SIMILARITY HELP

Protein Sequence:

○ 100% Sequence Identity (fast) ○ Sequence Similarity by BLAST (TrEMBL only)

Complete Sequence starting from N-terminus % Sequence Identity 90 and E-value cutoff 1.0E-4

SEARCH for Models SEARCH for Sequences RESET

```
Score = 2640
Length of alignment = 441
Sequence model  :  1 - 433 (Sequence length = 443)
Sequence 1c9rB  :  1 - 430 (Sequence length = 430)
```

```
trace 20
81. Number of H-Bonds ... 263
82. Number of Helices ... 16
83. Number of Strands ... 17
84. Number of Turns ..... 45
```

File Edit View Go Communicator Help
Bookmarks Location What's Related
100%
Signed by: Unsigned classes from local hard disk

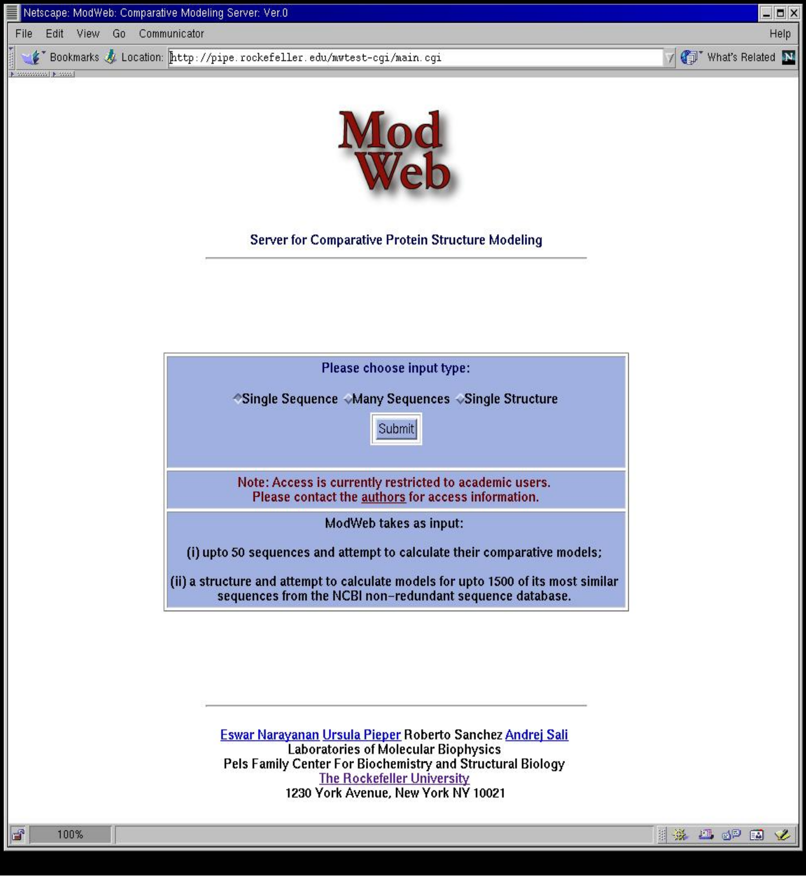
ModBase: A database of comparative protein structure models and properties.

http://www.salilab.org/modbase

Mod Web

A Server for Comparative Protein Structure Modeling

http://www.salilab.org/modweb


Netscape: ModWeb: Comparative Modeling Server: Ver.0
File Edit View Go Communicator Help
Bookmarks Location: http://pipe.rockefeller.edu/mwtest-cgi/main.cgi
What's Related
Submit
100%

Mod Web

Server for Comparative Protein Structure Modeling

Please choose input type:

◇Single Sequence ◇Many Sequences ◇Single Structure

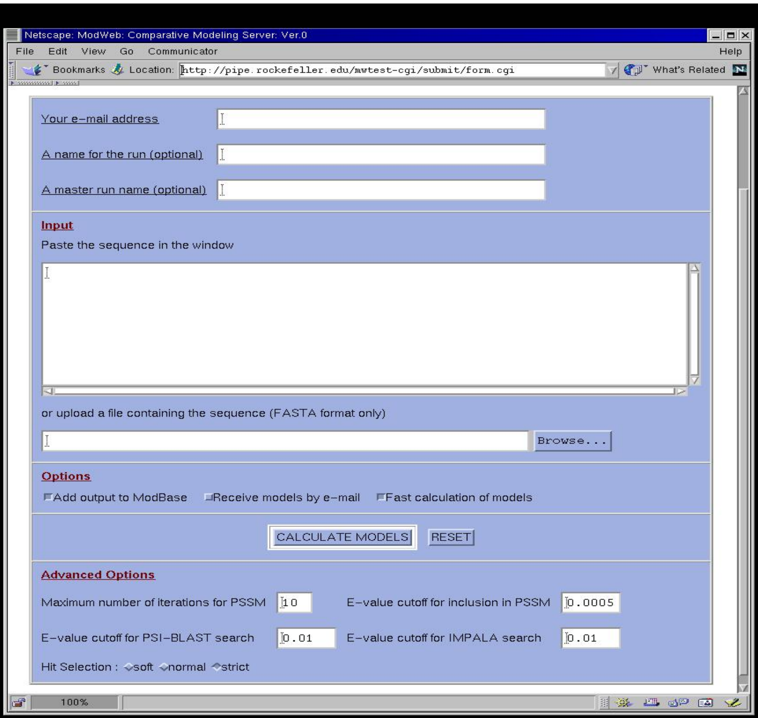
Note: Access is currently restricted to academic users.
Please contact the authors for access information.

ModWeb takes as input:

(i) upto 50 sequences and attempt to calculate their comparative models;

(ii) a structure and attempt to calculate models for upto 1500 of its most similar sequences from the NCBI non-redundant sequence database.

Eswar Narayanan Ursula Pieper Roberto Sanchez Andrej Sali
Laboratories of Molecular Biophysics
Pels Family Center For Biochemistry and Structural Biology
The Rockefeller University
1230 York Avenue, New York NY 10021


Netscape: ModWeb: Comparative Modeling Server: Ver.0
File Edit View Go Communicator Help
Bookmarks Location: http://pipe.rockefeller.edu/mwtest-cgi/submit/form.cgi
What's Related
Browse...
CALCULATE MODELS
RESET
100%

Your e-mail address

A name for the run (optional)

A master run name (optional)

Input

Paste the sequence in the window

or upload a file containing the sequence (FASTA format only)

Options

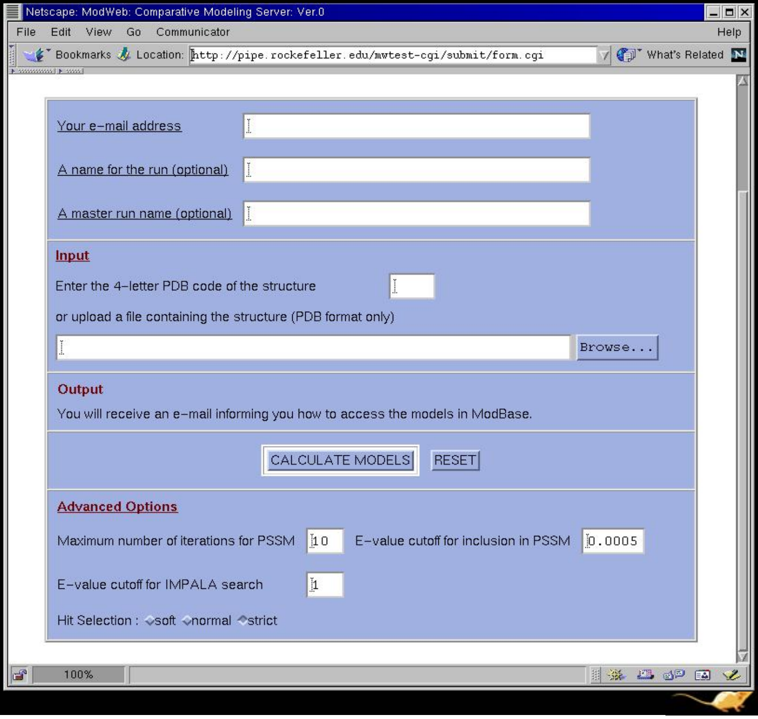
☑ Add output to ModBase ☐ Receive models by e-mail ☑ Fast calculation of models

Advanced Options

Maximum number of iterations for PSSM 10 E-value cutoff for inclusion in PSSM 0.0005

E-value cutoff for PSI-BLAST search 0.01 E-value cutoff for IMPALA search 0.01

Hit Selection : ◇soft ◇normal ◆strict


Netscape: ModWeb: Comparative Modeling Server: Ver.0
File Edit View Go Communicator Help
Bookmarks Location: http://pipe.rockefeller.edu/mwtest-cgi/submit/form.cgi
What's Related
Browse...
CALCULATE MODELS
RESET
100%

Your e-mail address

A name for the run (optional)

A master run name (optional)

Input

Enter the 4-letter PDB code of the structure

or upload a file containing the structure (PDB format only)

Output

You will receive an e-mail informing you how to access the models in ModBase.

Advanced Options

Maximum number of iterations for PSSM 10 E-value cutoff for inclusion in PSSM 0.0005

E-value cutoff for IMPALA search 1

Hit Selection : ◇soft ◇normal ◆strict

BayGenomics

Conclusions

- ✓ Comparative models help to understand protein’s function:
    - ✓ Detecting remote structural (functional?) relationships.
    - ✓ Revealing features that are not present in the templates.
    - ✓ Revealing features that are not recognizable from the sequence.

- ✓ Currently, useful 3D models can be obtained for domains in approximately 50% of the proteins (30% of domains), because of the improved **methods** and because of the many **known protein structures** and **sequences**.

- ✓ We will be able to calculate useful models for most globular domains soon after the completion of the genome projects, because of **structural genomics**.

Acknowledgments

Andrej Sali

Frank Alber
Fred Davis
Damien Devos
Narayanan Eswar
Rachel Karchin
Libusha Kelly
Michael F. Kim
Dmitry Korkin
M. S. Madhusudhan
Nebosja Mirkovic
Ursula Pieper
Andrea Rossi
Min-yi Shen
Maya Topf

http://www.salilab.org