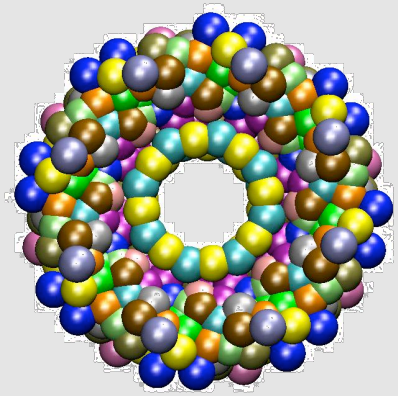
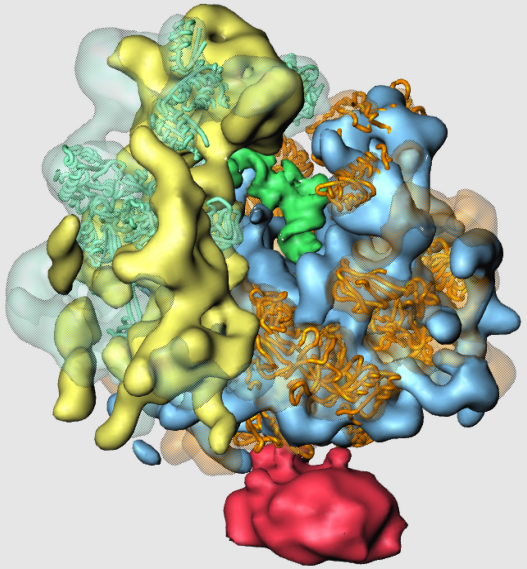
Modeling the Structures of Proteins and Macromolecular Assemblies



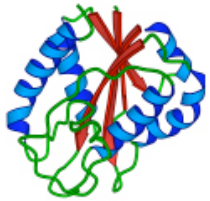


Marc A. Marti-Renom
The Sali Lab
http://salilab.org/

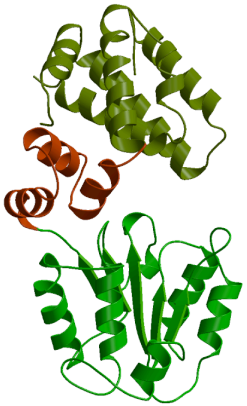
Depts. of Biopharmaceutical Sciences and Pharmaceutical Chemistry
California Institute for Quantitative Biomedical Research
University of California at San Francisco
UCSF

From domains to assemblies

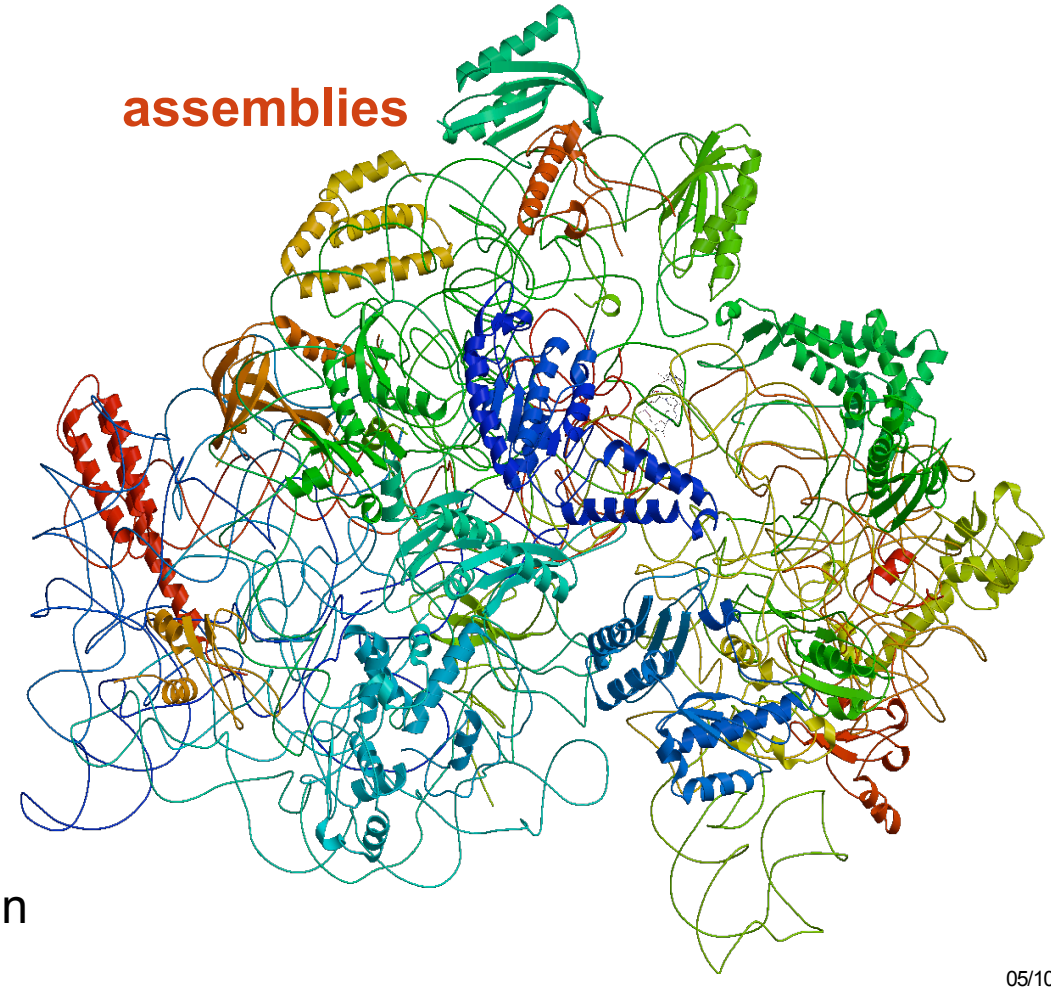
domains



proteins



assemblies



~2.5 domains in a protein
a few domain partners per domain

Sequence *versus* Structure

GDCAGDFKIWYFGRTLLVAGAKDEFGAIDAW…

RTLAWYAGHLVAGAKDEFGGDFKIWYFGAID…

DFLLVAGAKDEFGKIWYFGGIDAWRTAGDCA…

HLVAGARTLAFGAIDWYAKDEFGGGDFKIWY…

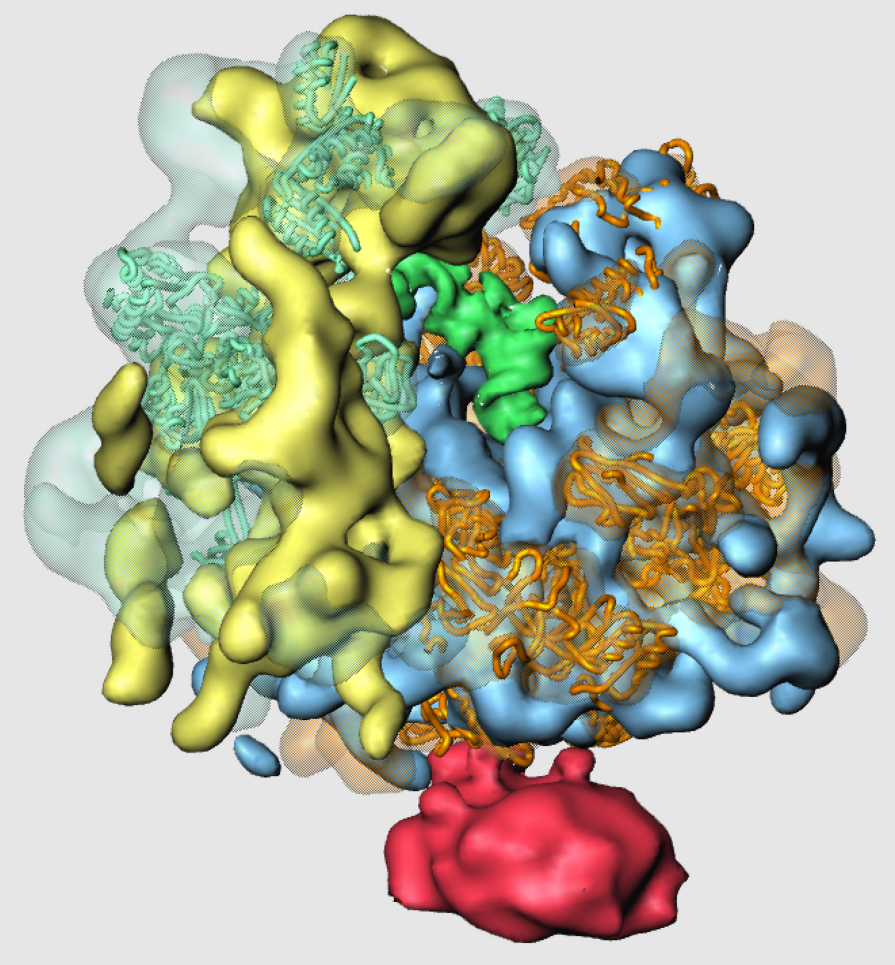
ARTHLVAGFGGGAIDWYFKIWYAKLAFGDED…

GCTAGCTTAAGGCCTTCATGATCTTCTGAG…

AGGGCTCCTTCATGATAGCTTAAGGCTTAA…

AGGCCTTCATGGGGTTAACATATCTTCTGA…

CCTTCATGCTAGCTTAAGGGATCTTAACCG…



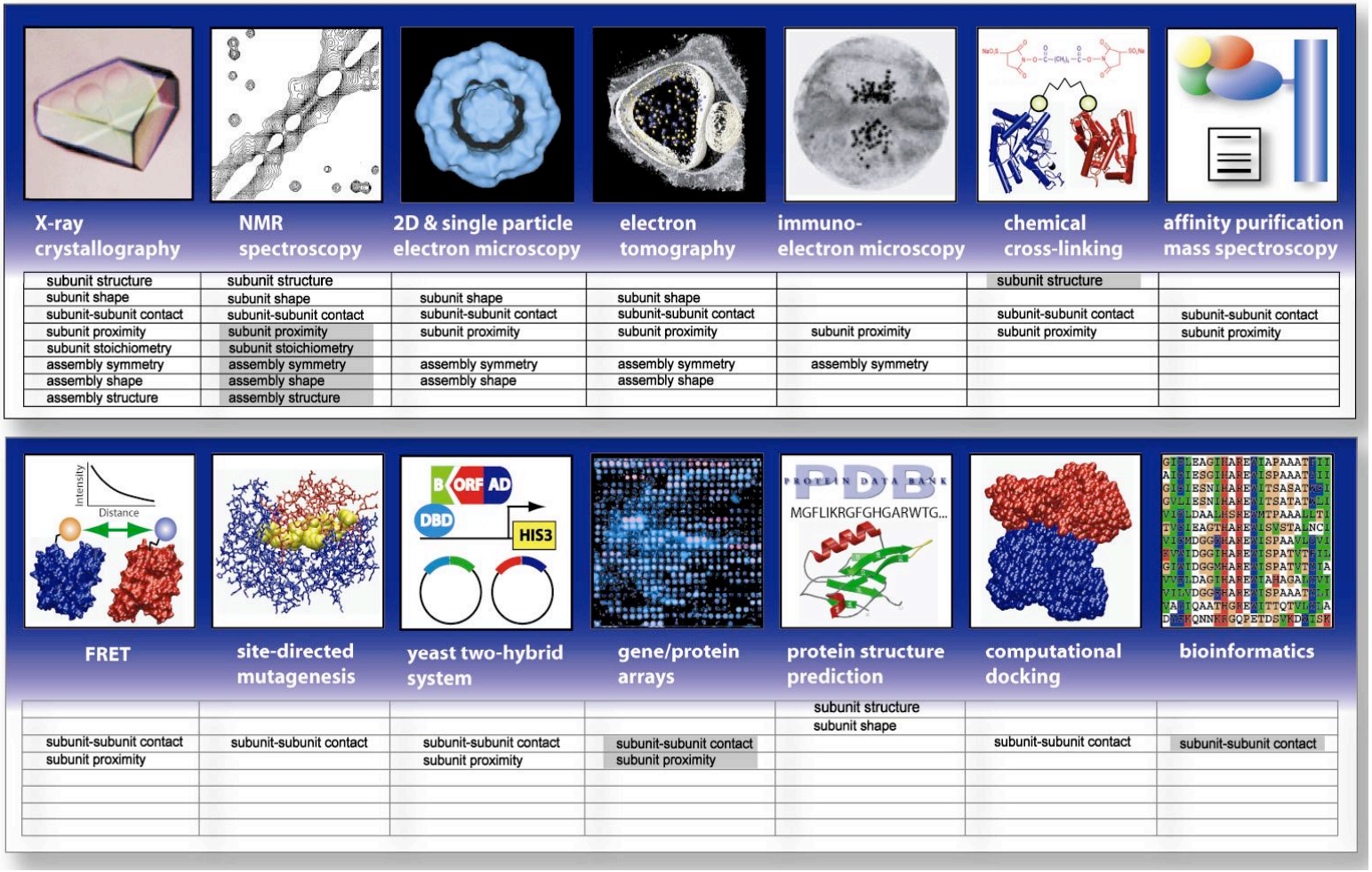
Determining the structures of proteins and assemblies

Use structural information from any

source: measurement, first principles, rules,

resolution: low or high resolution

to obtain the set of all models that are consistent with it.

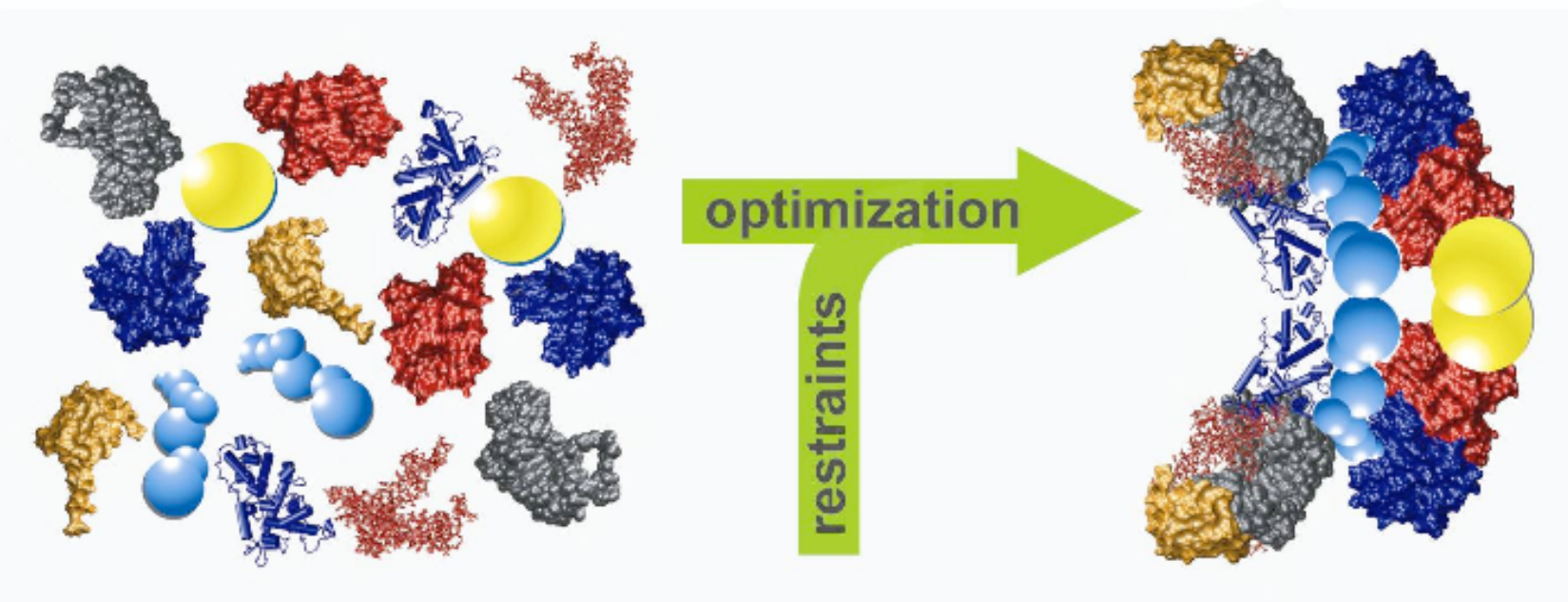


Sali, Earnest, Glaeser, Baumeister. From words to literature in structural proteomics. Nature 422, 216-225, 2003.

Modeling proteins and macromolecular assemblies by satisfaction of spatial restraints

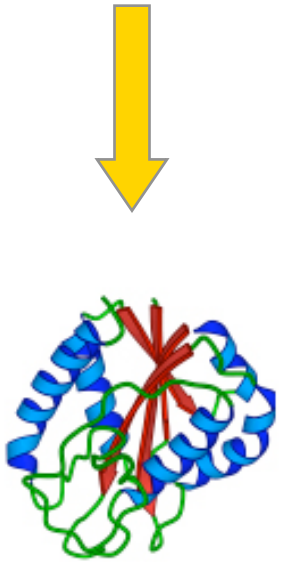
1) Representation of a system.
2) Scoring function (spatial restraints).
3) Optimization.

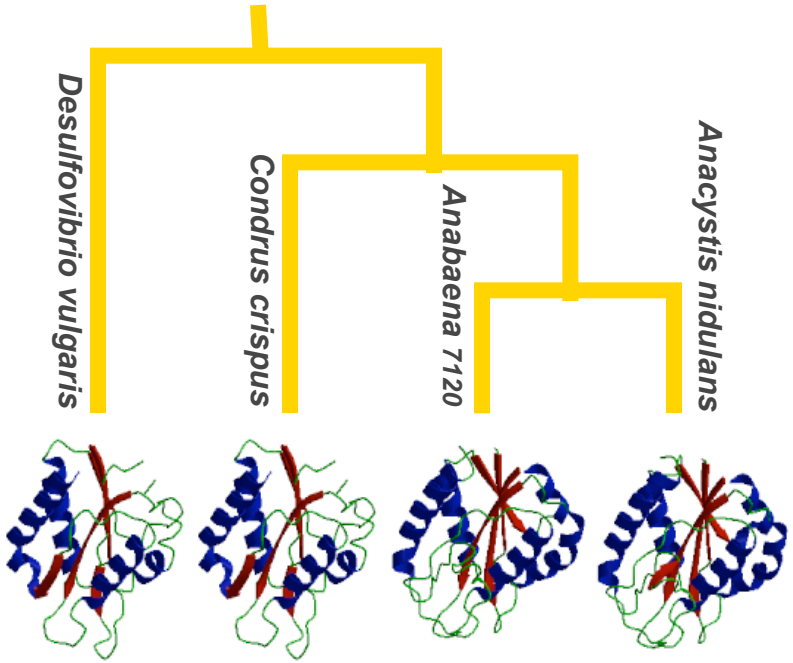
There is nothing but points and restraints on them.



Principles of protein structure

GFCHIKAYTRLIMVG…





Folding

(physics)

Ab initio prediction

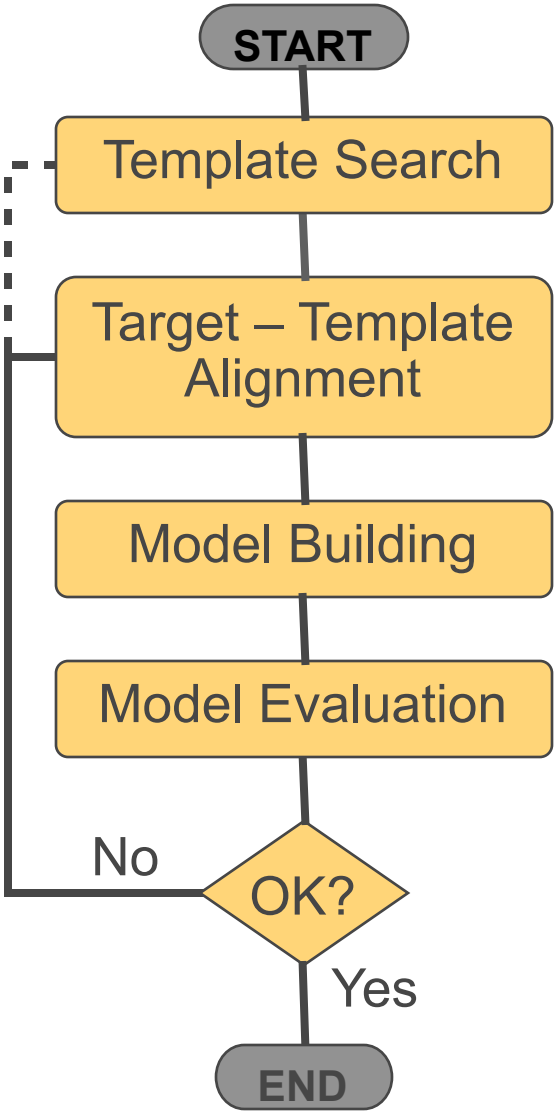
Evolution

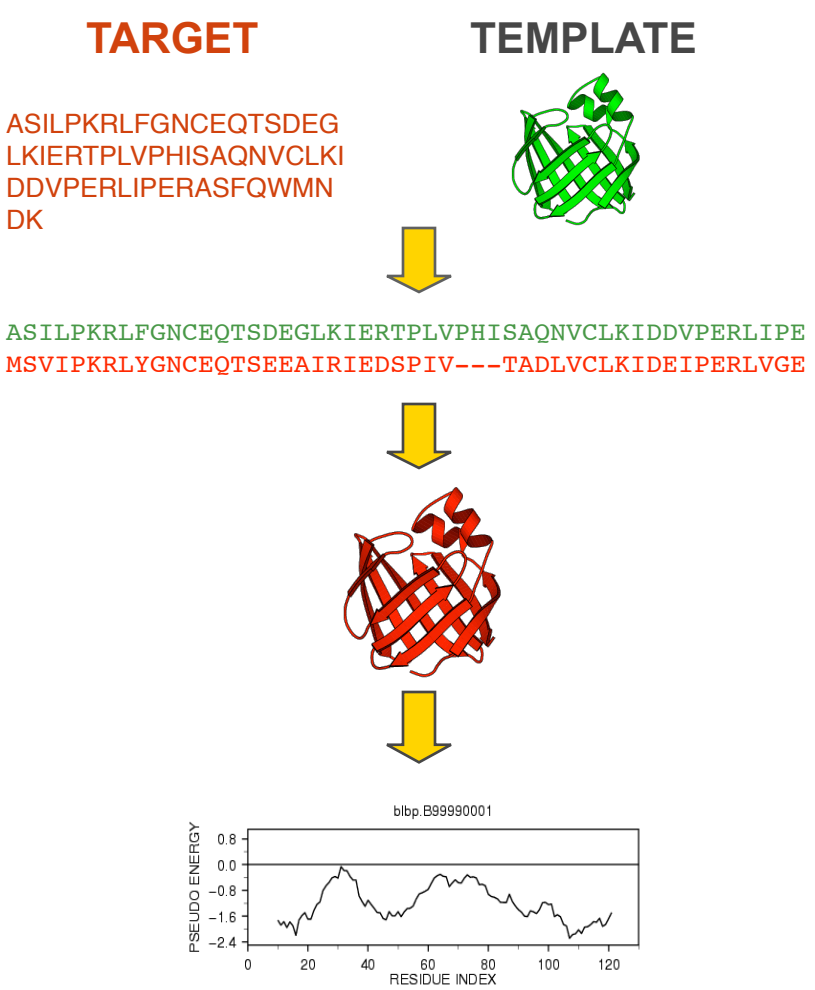
(rules)

Threading
Comparative Modeling

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

Steps in Comparative Protein Structure Modeling



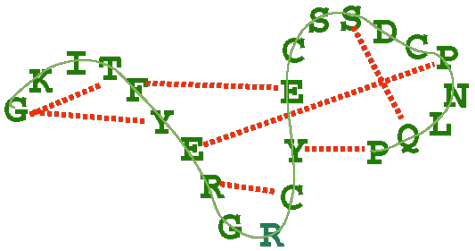


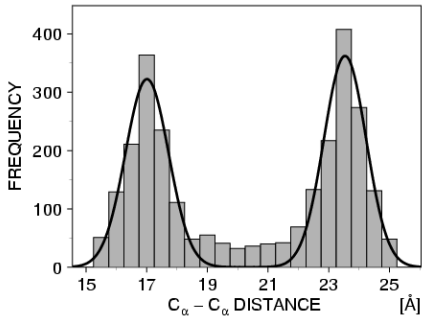
A. Šali, *Curr. Opin. Biotech.* 6, 437, 1995.
R. Sánchez & A. Šali, *Curr. Opin. Str. Biol.* 7, 206, 1997.
M. Marti *et al*. *Ann. Rev. Biophys. Biomolec. Struct*., 29, 291, 2000.
http://salilab.org/

Comparative modeling by satisfaction of spatial restraints
MODELLER

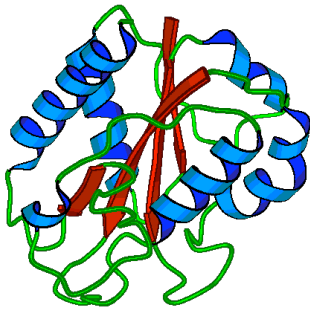
3D GKITFYERGFQGHCYESDC-NLQP…
SEQ GKITFYERG---RCYESDCPNLQP…

1. Extract spatial restraints





2. Satisfy spatial restraints



$$F(\mathbf{R}) = \prod_{i} p_i \, (f_i / I)$$

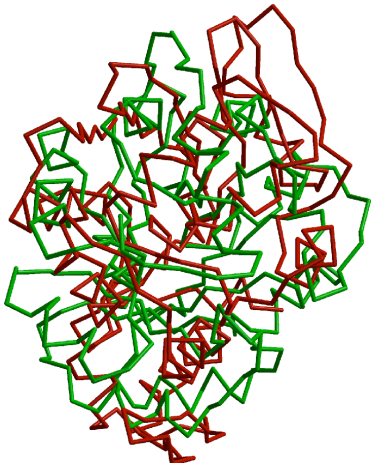
A. Šali & T. Blundell. *J. Mol. Biol.* **234**, 779, 1993.
J.P. Overington & A. Šali. *Prot. Sci*. **3**, 1582, 1994.
A. Fiser, R. Do & A. Šali, *Prot. Sci*., **9**, 1753, 2000.

http://salilab.org/

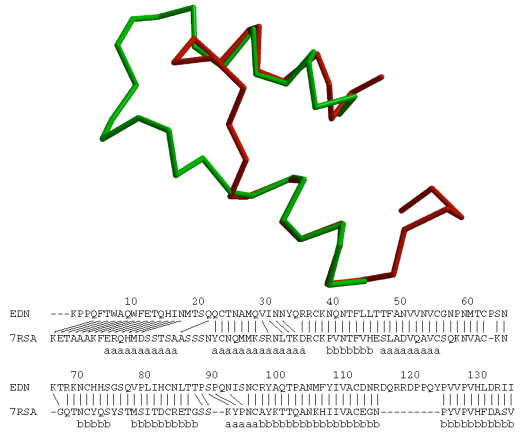
Typical errors in comparative models

MODEL
X-RAY
TEMPLATE

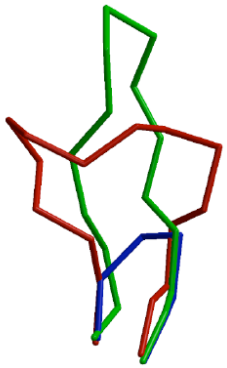
Incorrect template



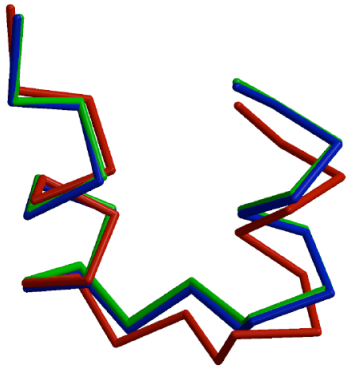
Misalignment



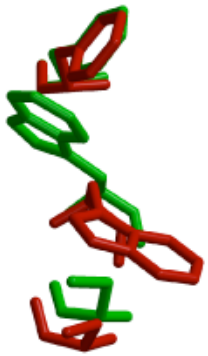
Region without a template



Distortion/shifts in aligned regions



Sidechain packing



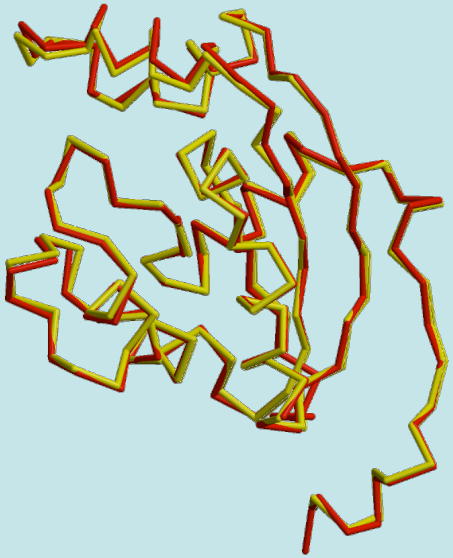
Marti-Renom et al. Annu.Rev.Biophys.Biomol.Struct. 29, 291-325, 2000.

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

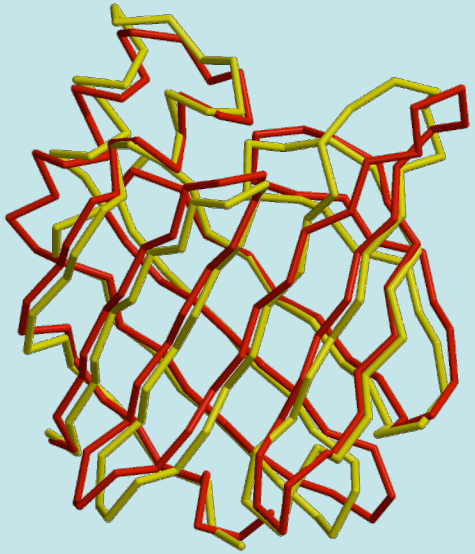


Sidechains
Core backbone
Loops

X-RAY / MODEL

MEDIUM ACCURACY

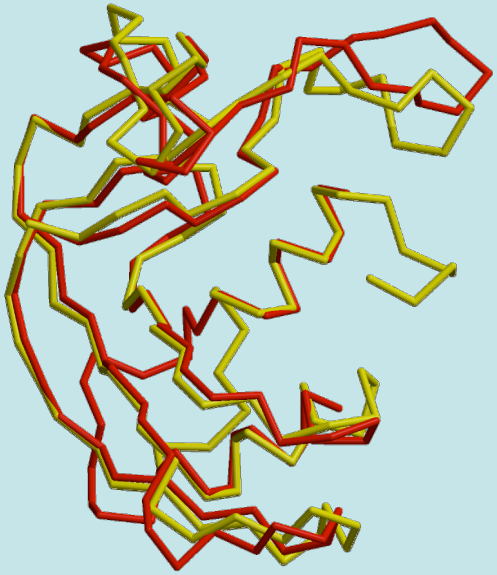
CRABP
Seq id 41%
Cα equiv 122/137
RMSD 1.34Å



Sidechains
Core backbone
Loops
Alignment

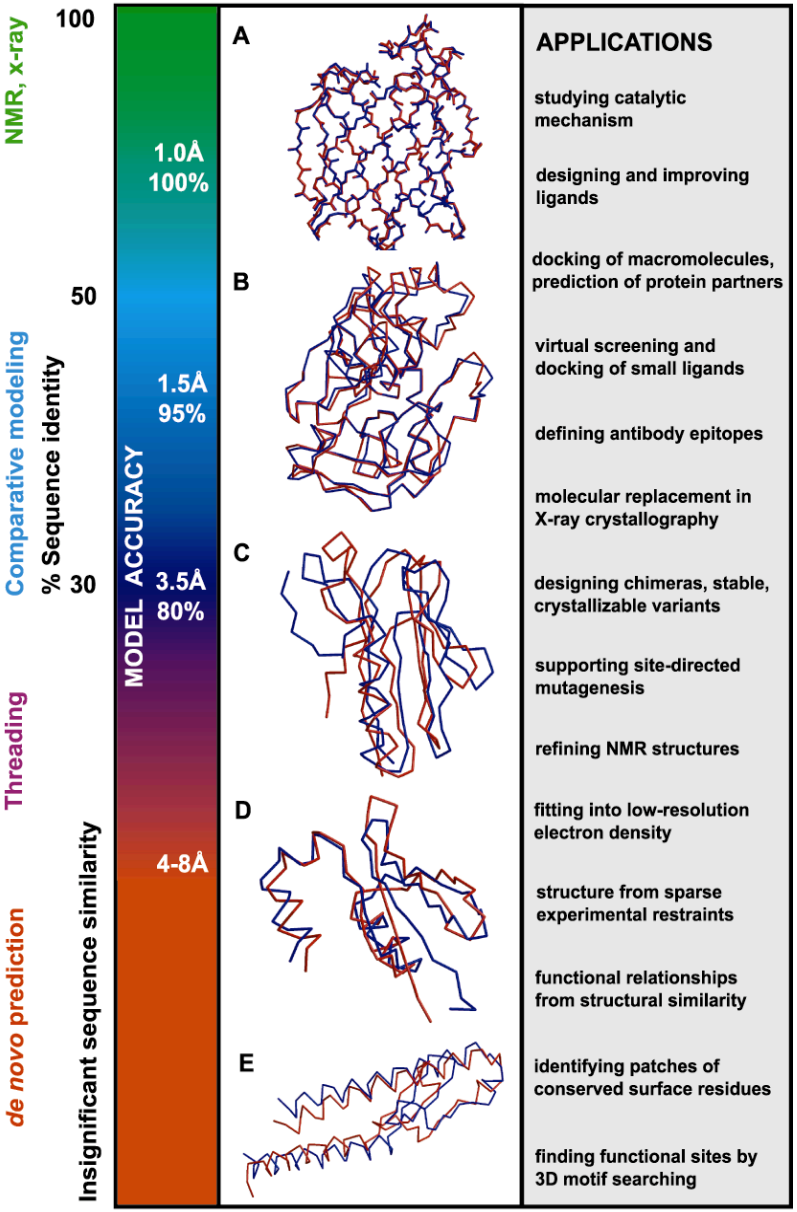
LOW ACCURACY

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



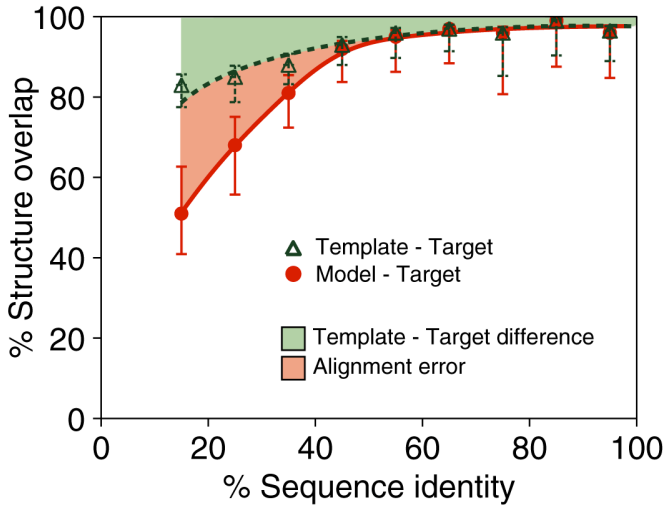
Sidechains
Core backbone
Loops
Alignment
Fold assignment

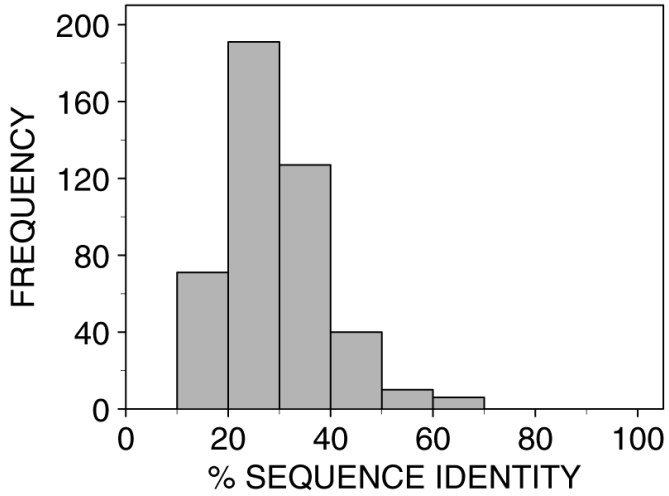
Utility of protein structure models, despite errors



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

Alignment errors are frequent and large





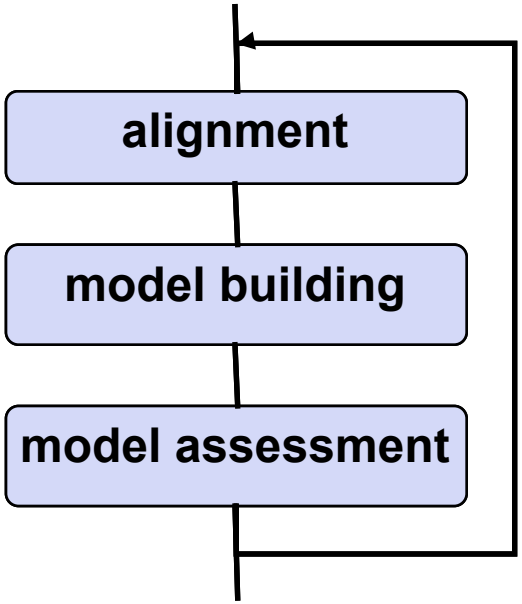
R. Sánchez & A. Šali, *Proc. Natl. Acad. Sci. USA* **95**, 13597, 1998.

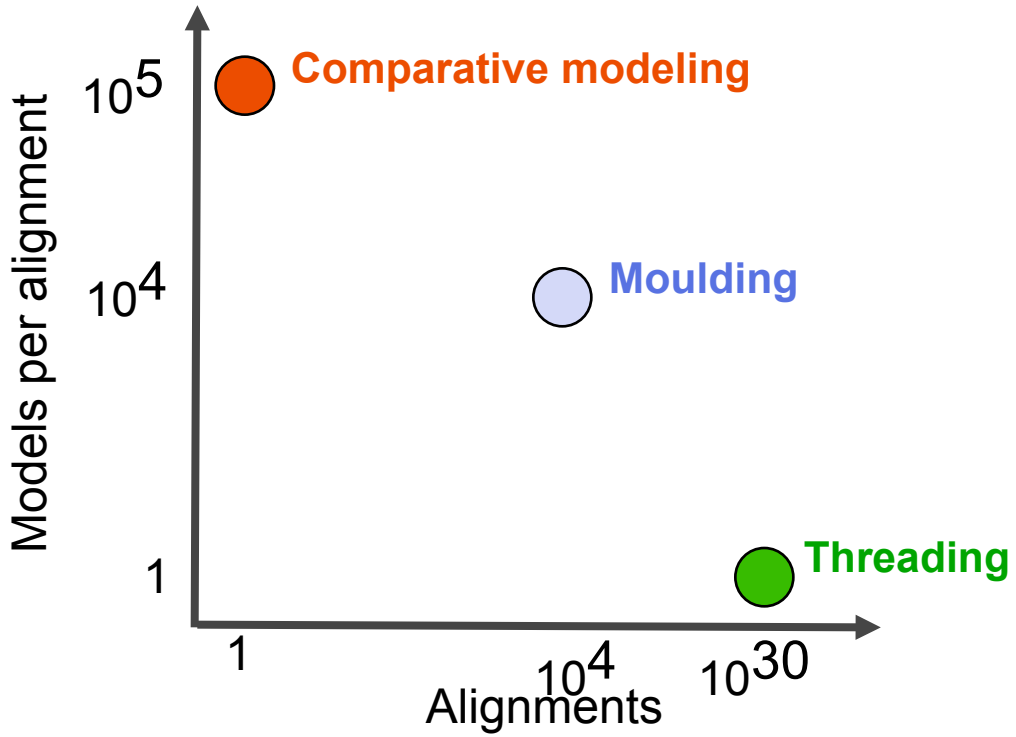
Minimizing errors in sequence-structure alignment

- Multiple sequence profiles.
- Complex gap penalty functions.
- Hidden Markov Models.
- Threading.

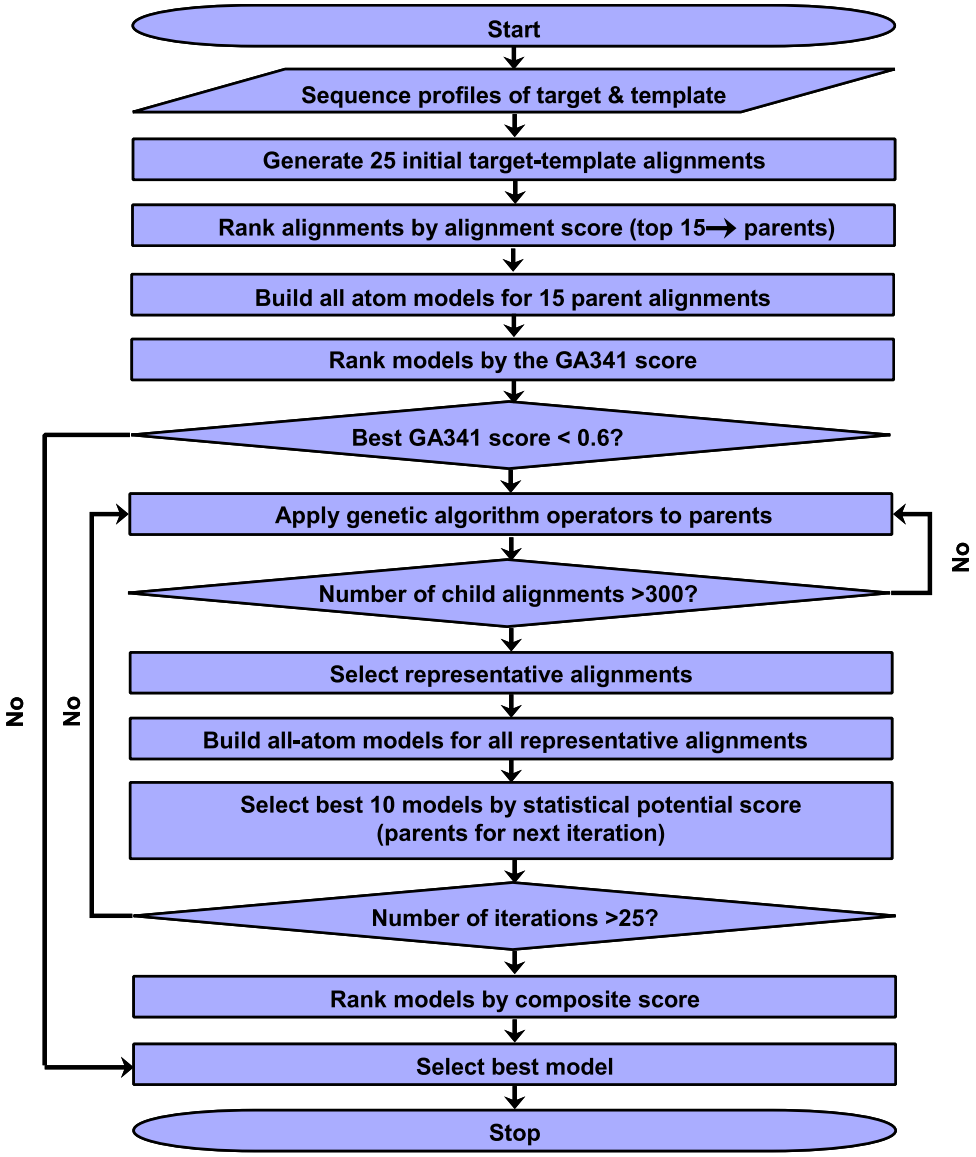
Moulding: iterative alignment, model building, model assessment

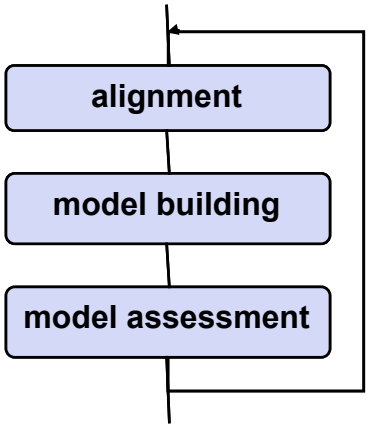
B. John, A. Sali. *Nucl. Acids Res.*, **31**, 1982-1992, 2003.





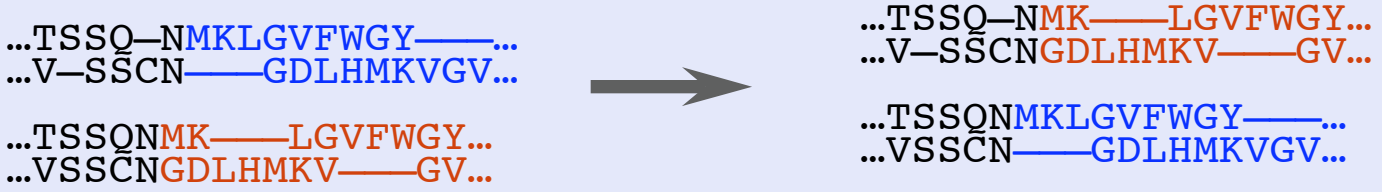
Moulding by a *Genetic Algorithm* approach





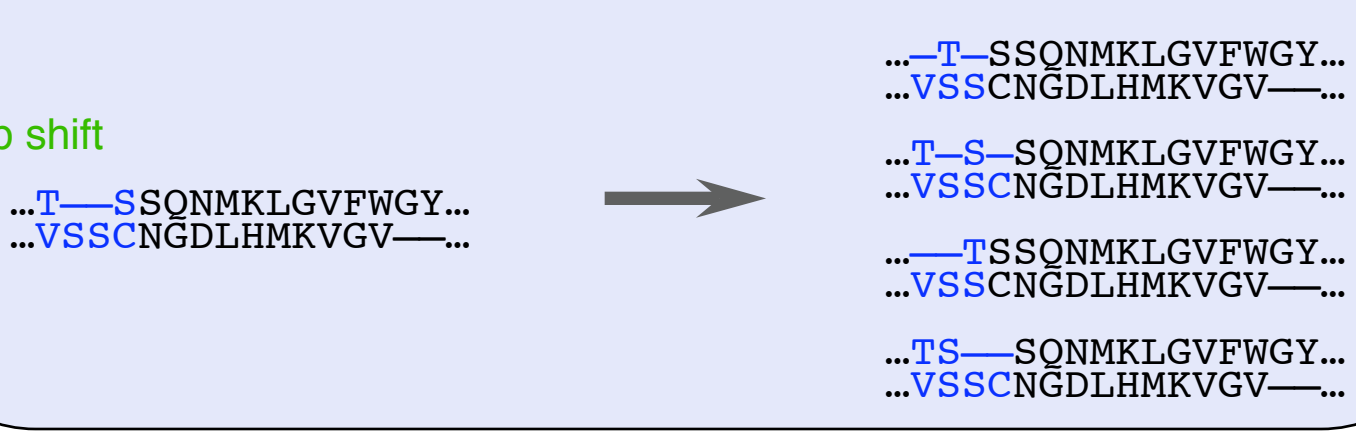
Genetic algorithm operators

Single point cross-over



Gap insertion

Gap shift



Also, “**two point crossover**” and “**gap deletion**”.

Composite model assessment score

Weighted linear combination of several scores:

- Pair (P_p) and surface (P_s) statistical potentials;
- Structural compactness (S_c);
- Harmonic average distance score (H_a);
- Alignment score (A_s).

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

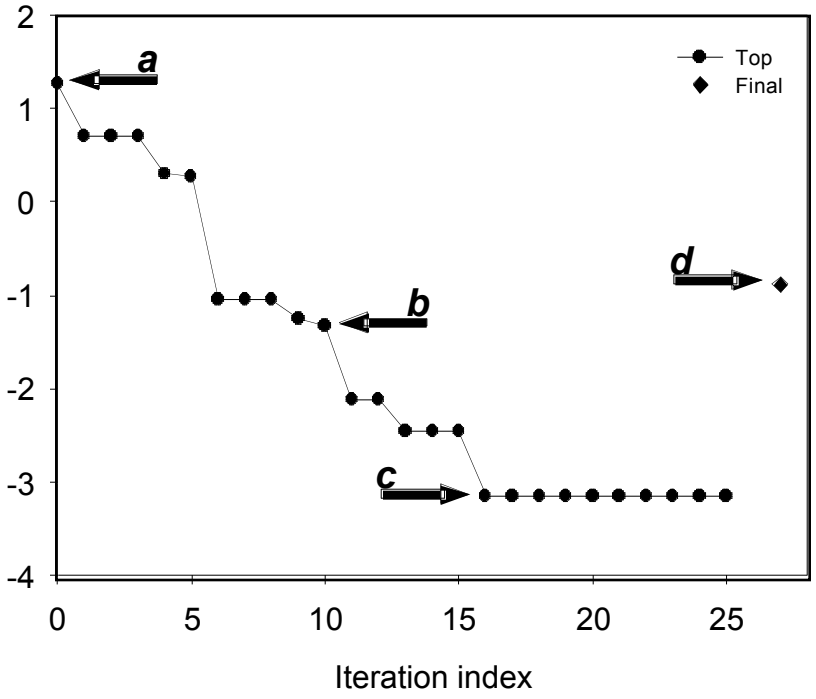
$Z(score) = (score - \mu)/\sigma$

μ … average score of all models

σ … standard deviation of the scores

Application to a difficult modeling case

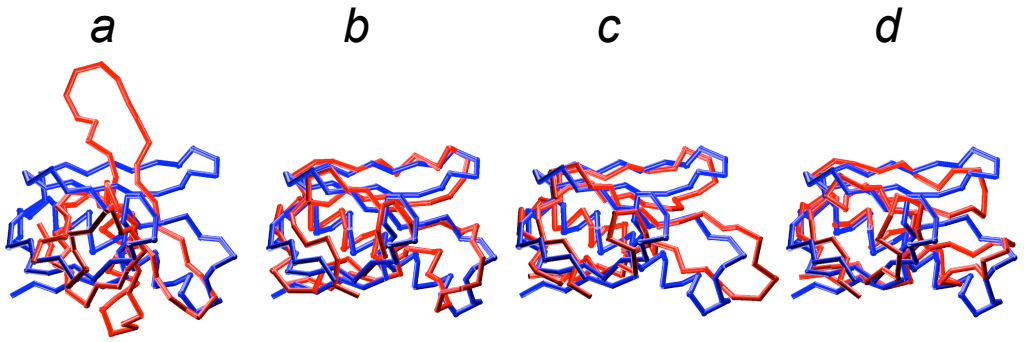
1BOV-1LTS



Sequence identity **4.4%**

Initial model Cα RMSD **10.1Å**

Final model Cα RMSD **3.6Å**



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

Alignment accuracy (CE overlap)

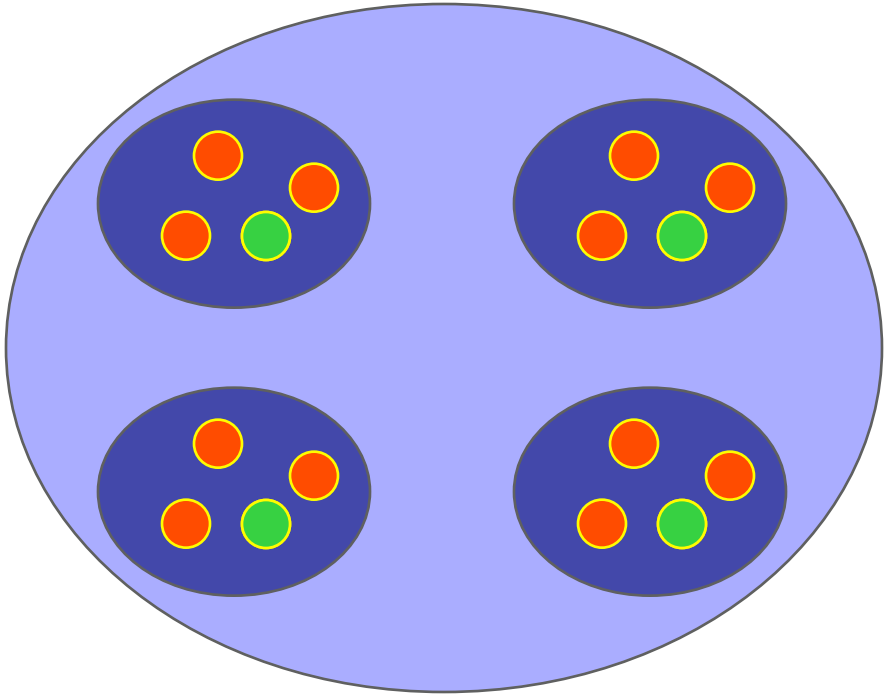
D. Fischer threading test set of 68 structural pairs (a subset of 19):

PSI-BLAST (sequence-profile alignment)	25%
SAM (Hidden Markov Models)	36%
MOULDER (iterative sequence-structure alignment)	45%

Structural Genomics

Sali. *Nat. Struct. Biol*. **5**, 1029, 1998.
Sali *et al. Nat. Struct. Biol*., **7**, 986, 2000.
Sali. *Nat. Struct. Biol.* **7**, 484, 2001.
Baker & Sali. *Science* **294,** 93, 2001.

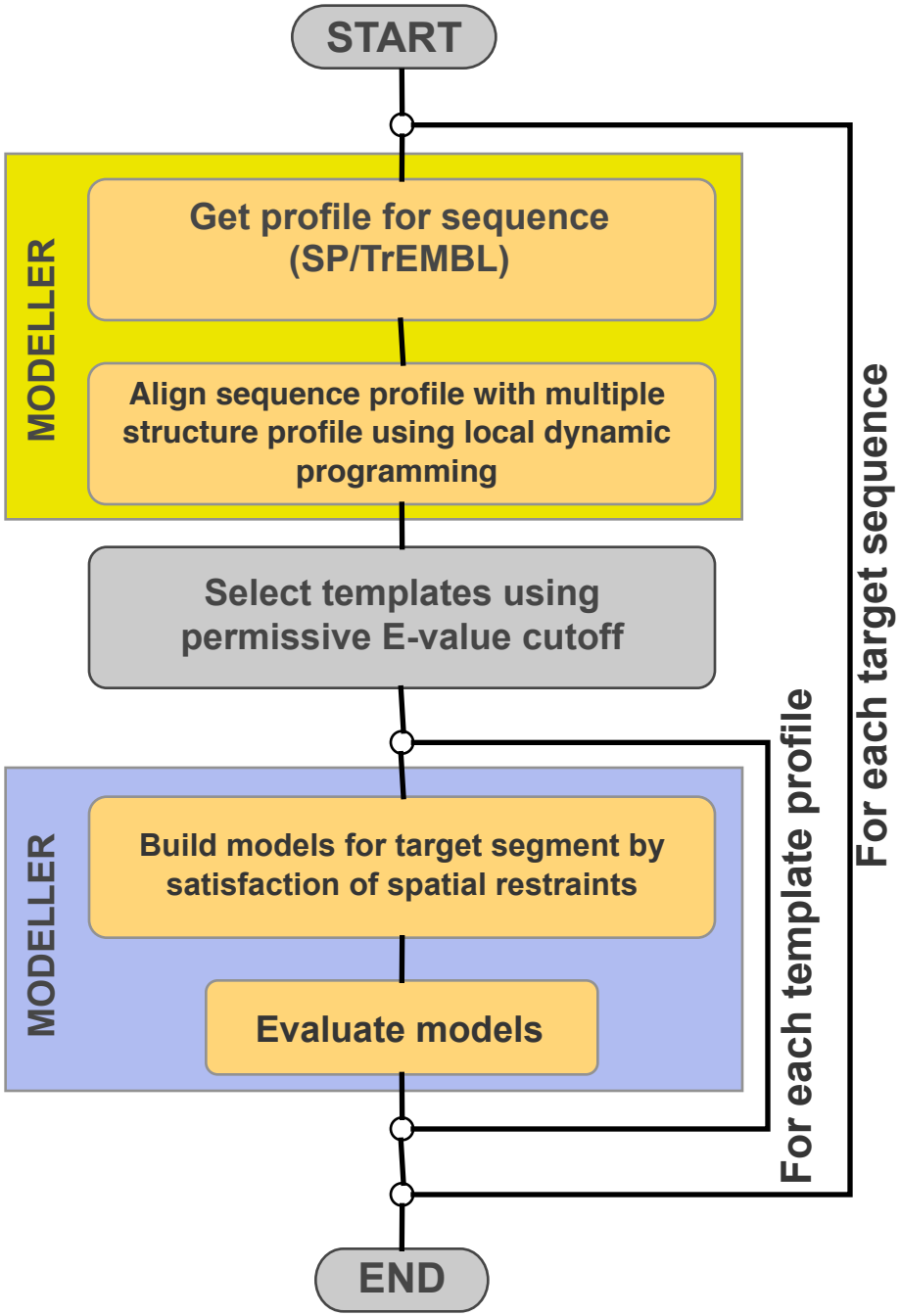
Characterize most protein sequences based on related known structures.



The number of “families” is much smaller than the number of proteins.

Any one of the members of a family is fine.

There are ~16,000 30% seq id families (90%)
(Vitkup *et al*. *Nat. Struct. Biol*. **8**, 559, 2001).



MODPIPE: Automated Large-Scale Comparative Modeling

R. Sánchez & A. Šali, *Proc. Natl. Acad. Sci. USA* 95, 13597, 1998.

Eswar *et al*. Nucl. Acids Res. 31, 3375–3380, 2003.

Pieper *et al*., *Nucl. Acids Res.* 32, 2004.

N. Eswar, M. Marti-Renom, M.S. Madhusudhan, B. John, A. Fiser, R. Sánchez, F. Melo, N. Mirkovic, B. Webb, M.-Y. Shen, A. Šali.

Synergy of crystallography and comparative modeling in structural genomics

Pieper *et al.*, *Nucl. Acids Res.* 32, 2004.

http://salilab.org/modbase/models_nysgxrc.html

NYSGXRC X-ray Structure			MODBASE Models			
PDB Code	Database Accession Number	Annotation	Total Sequences	Fold & Model	Fold	Model
1b54	P38197	Hypothetical UPF0001 protein YBL036C	151	132	2	17
1f89	P49954	Hypothetical 32.5 kDa protein YLR351C	553	488	55	10
1njr	Q04299	Hypothetical 32.1 kDa protein in ADH3-RCA1 intergenic region	4	1	0	3
1nkq	P53889	Hypothetical 28.8 kDa protein in PSD1-SKO1 intergenic region	379	207	172	0
1jzt	P40165	Hypothetical 27.5 kDa protein in SPX19-GCR2 intergenic region	1058	39	1006	13
1jr7	P76621	Hypothetical protein ygaT	11	10	0	1
1ku9	3025177	YF63_METJA hypothetical protein MJ1563	598	131	214	253

Comparative modeling of the TrEMBL database

Unique sequences processed: 1,182,126

Sequences with fold assignments or models: 659,495 (56%)

70% of models based on <30% sequence identity to template.

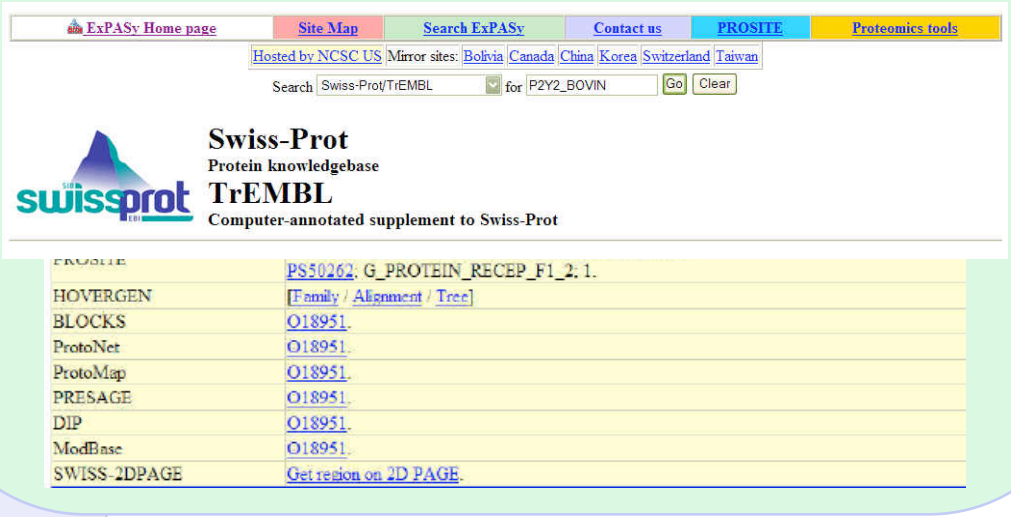
On average, only a domain per protein is modeled
(an “average” protein has 2.5 domains of 175 aa).

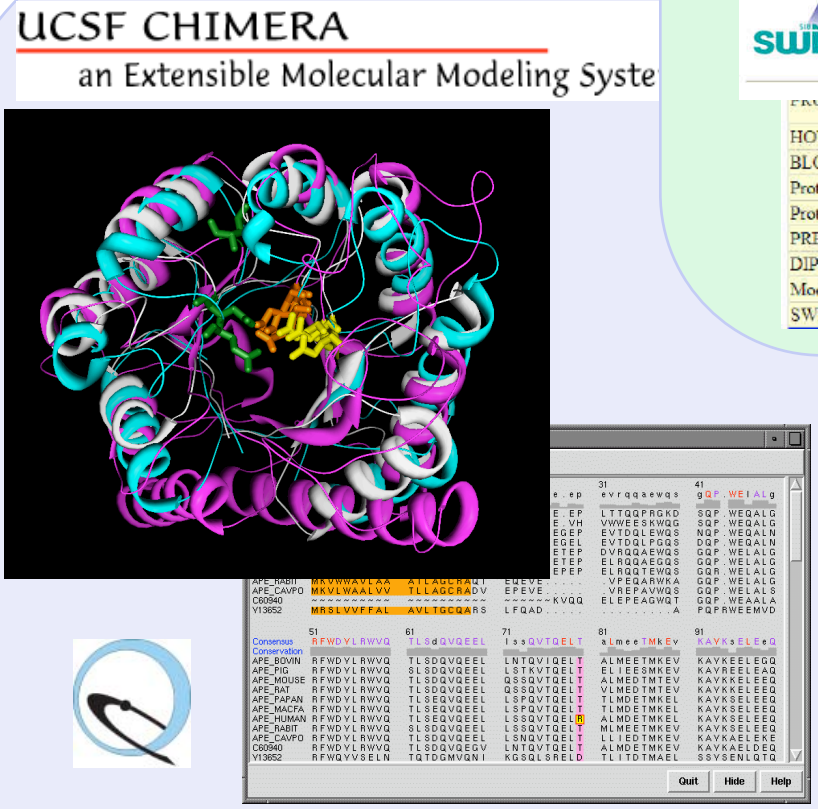
9/15/03 ~4 weeks on 660 Pentium CPUs

http://salilab.org/modbase

Pieper *et al.*, Nucl. Acids Res. 2004.

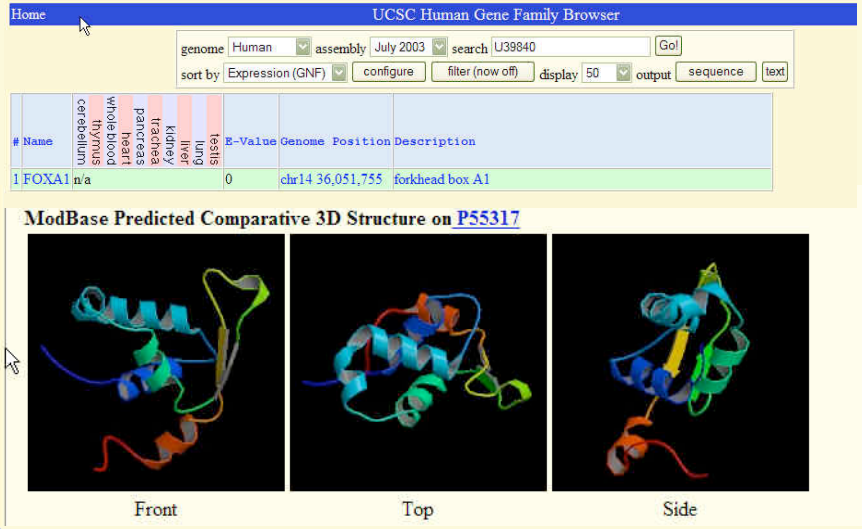
Major bidirectional resources involving ModBase



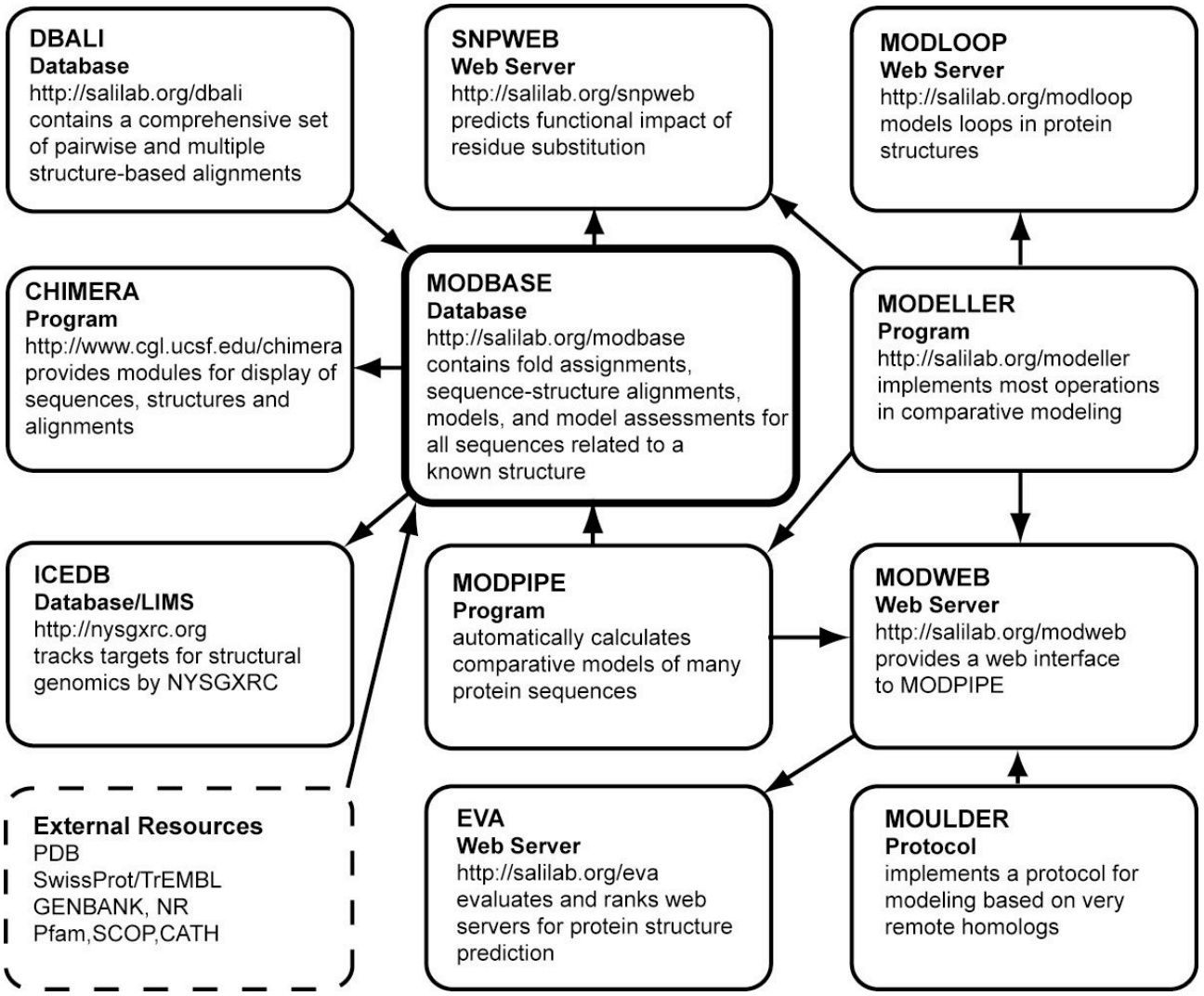


http://www.cgl.ucsf.edu/chimera/
Daniel Greenblatt, Conrad C. Huang, Thomas E. Ferrin

UCSC Human Gene Family Browser



MODBASE and associated resources
http://salilab.org/



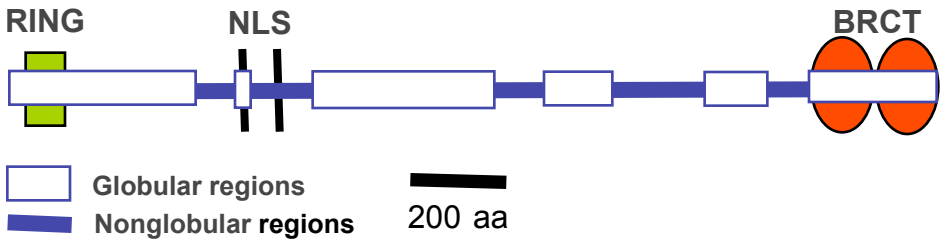
Structural analysis of missense mutations in human BRCA1 BRCT domains

Nebojsa Mirkovic, Marc A. Marti-Renom, Barbara L. Weber, Andrej Sali and Alvaro N.A. Monteiro

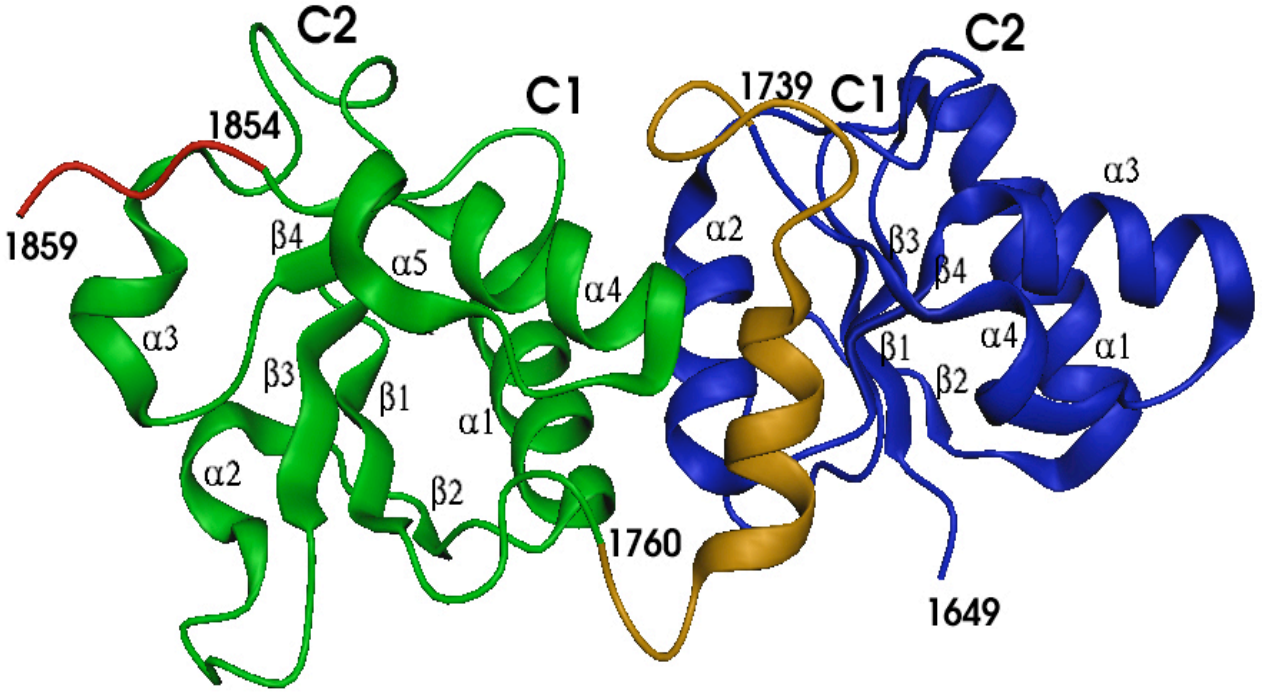
Cancer Research (June 2004). 64:3790-97

Cannot measure the functional impact of every possible SNP at all positions in each protein! Thus, prediction based on general principles of protein structure is needed.

Human BRCA1 and its two BRCT domains



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	**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
Physician: Fred Gilbert, MD		

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.

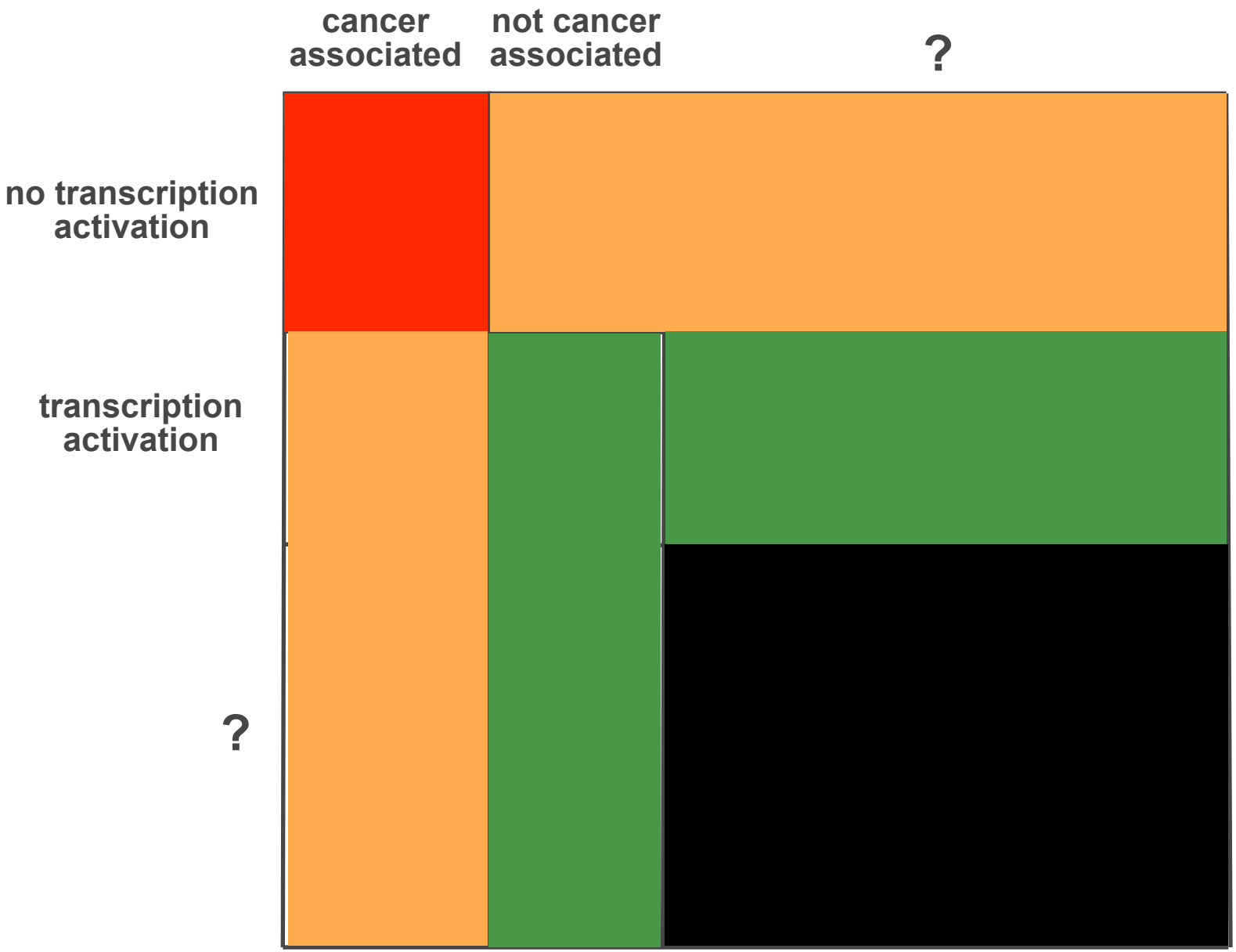
Authorized Signature:

Brian E. Ward, Ph.D.
Laboratory Director

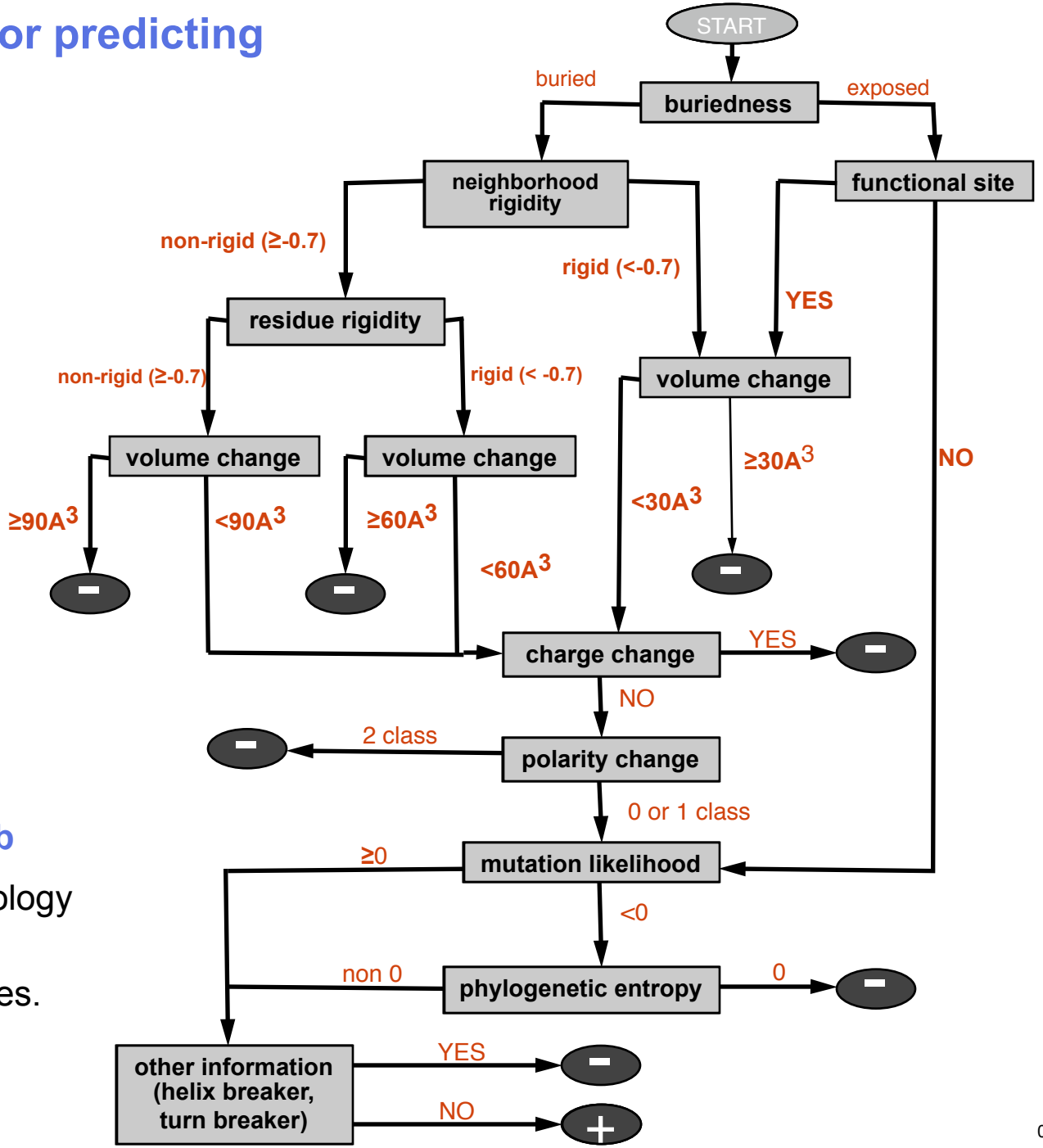
Thomas S. Frank, M.D.
Medical Director

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 analytes, 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



"Decision" tree for predicting functional impact of genetic variants

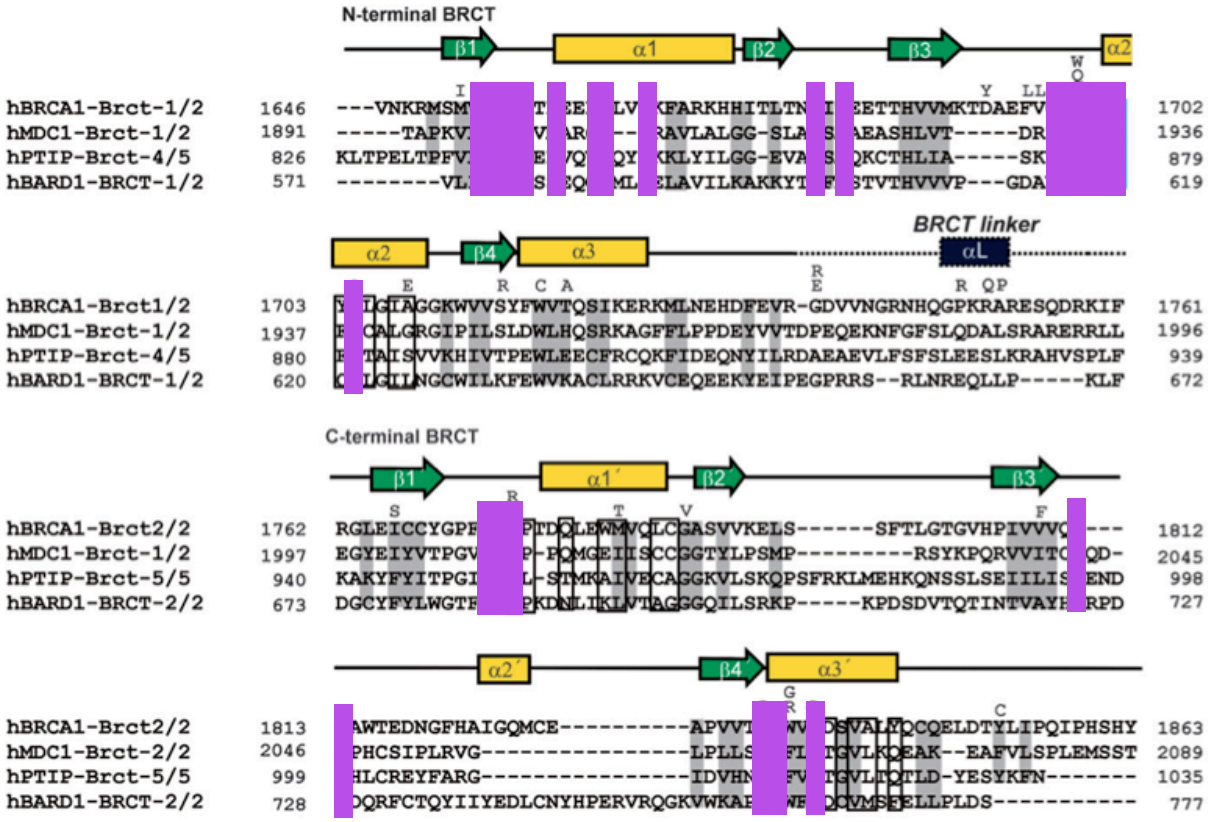


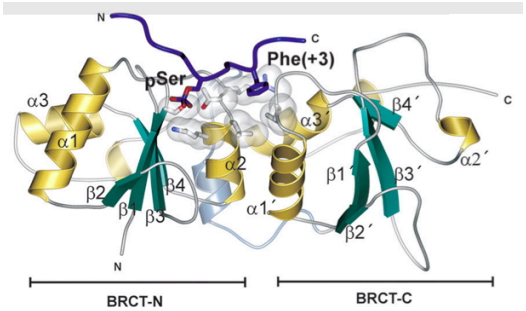
http://salilab.org/snpweb

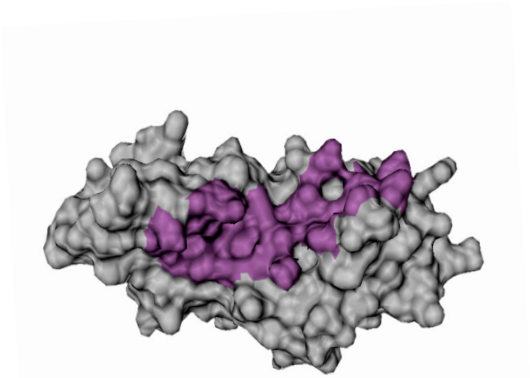
Mirkovic *et al*., Cancer Biology (2004) **64**:3790-97

Eswar *et al*. Nucl.Acids Res. **31**, 3375, 2003.

Putative binding site on BRCA1







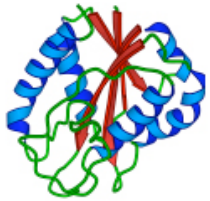
Putative binding site predicted in 2003 and accepted for publication on March 2004.

Williams *et al.* 2004 Nature Structure Biology. **June 2004 11**:519

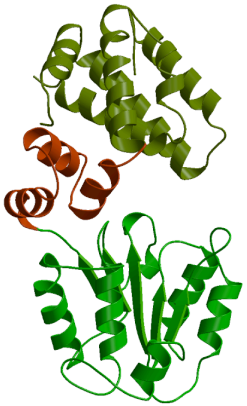
Mirkovic *et al.* 2004 Cancer Research. **June 2004 64**:3790

From domains to assemblies

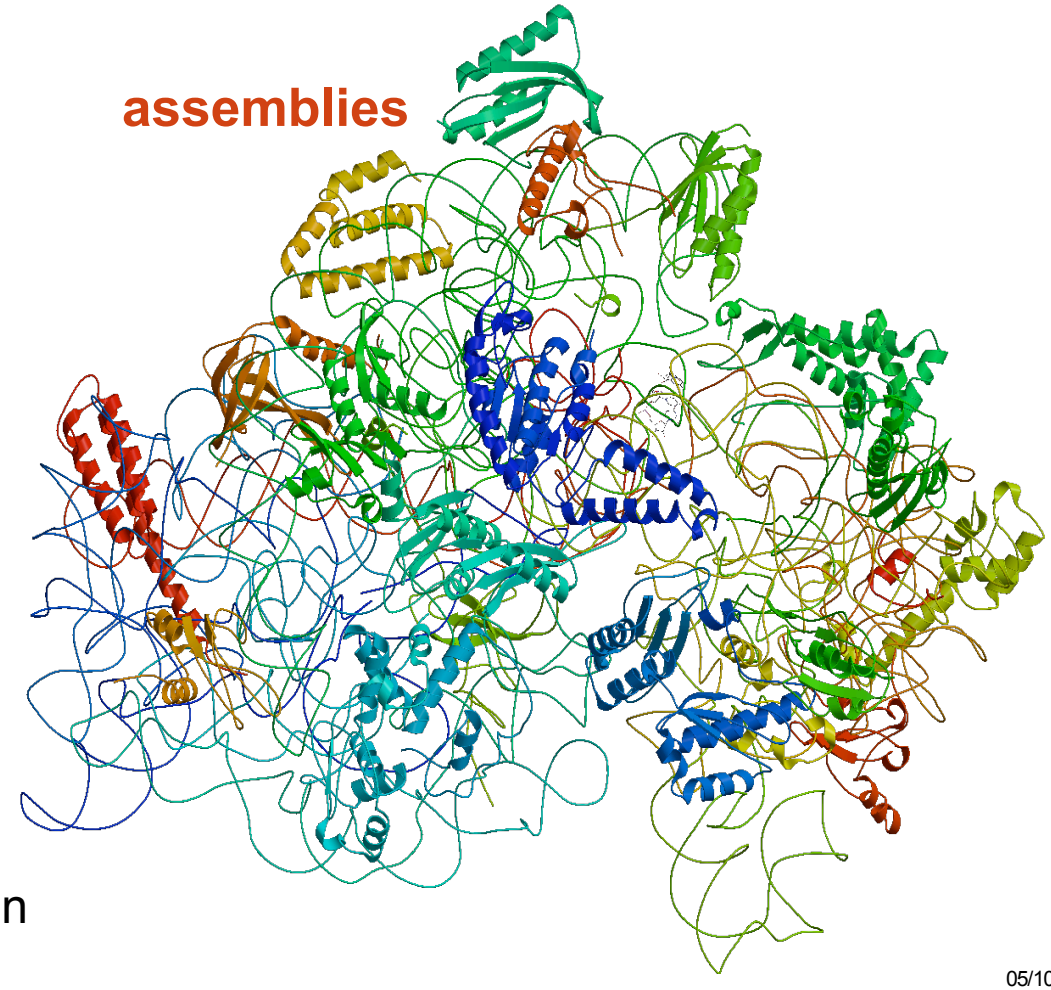
domains



proteins



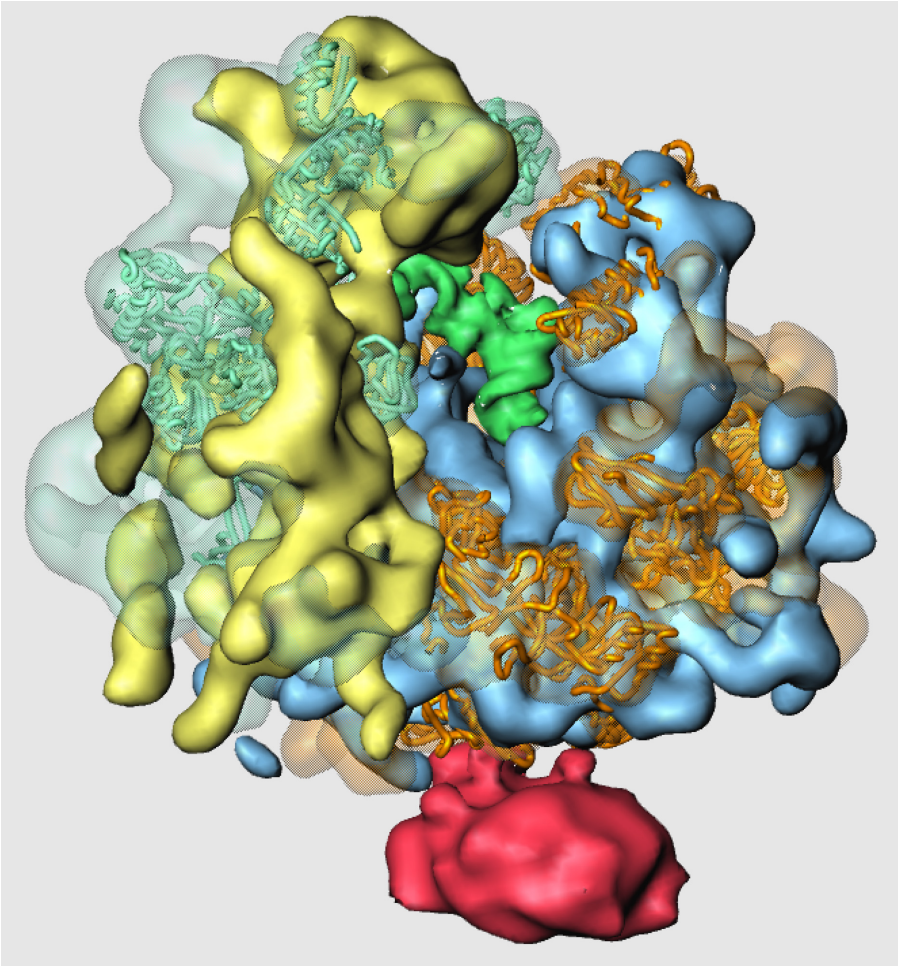
assemblies



A. Sali. NIH Workshop on Structural Proteomics of Biological Complexes. *Structure* **11**, 1043, 2003**.**

~2.5 domains in a protein
a few domain partners per domain

S. cerevisiae ribosome



Fitting of comparative models into 15Å cryo-electron density map.

43 proteins could be modeled on 20-56% seq.id. to a known structure.

The modeled fraction of the proteins ranges from 34-99%.

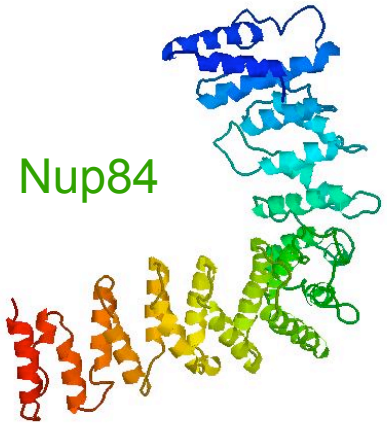
C. Spahn, R. Beckmann, N. Eswar, P. Penczek, A. Sali, G. Blobel, J. Frank.
Cell **107**, 361-372, 2001.

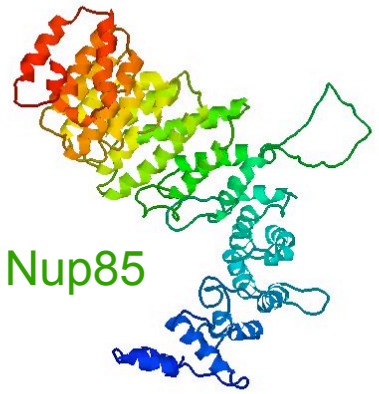
Common Evolutionary Origin of Coated Vesicles and Nuclear Pore Complexes

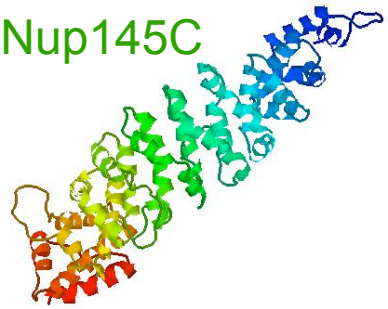
D. Devos*, S. Dokudovskaya, F. Alber*, M.A. Marti-Renom*,
A. Sali *, M. Rout, B. Chait

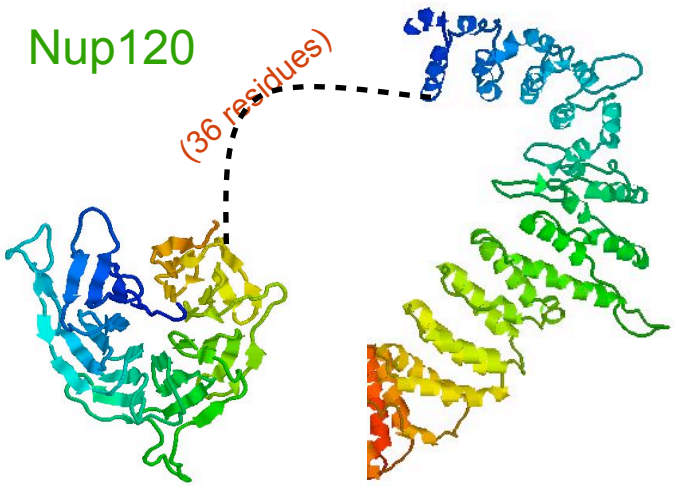
Rockefeller University, New York
*UCSF

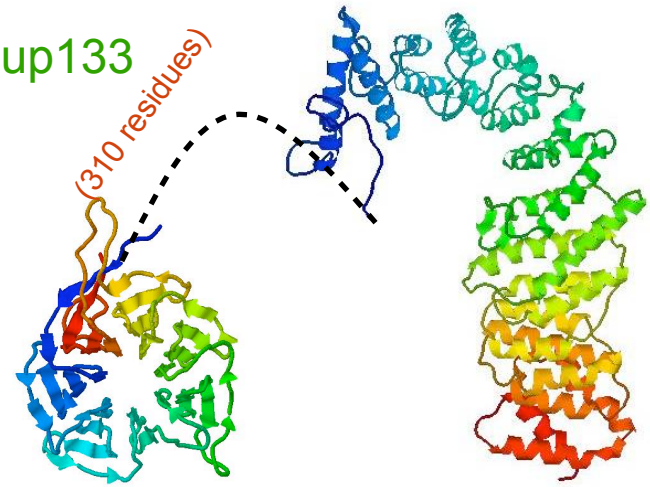
All Nucleoporins in the Nup84 Complex are Predicted to Contain β-Propeller and/or α-Solenoid Folds



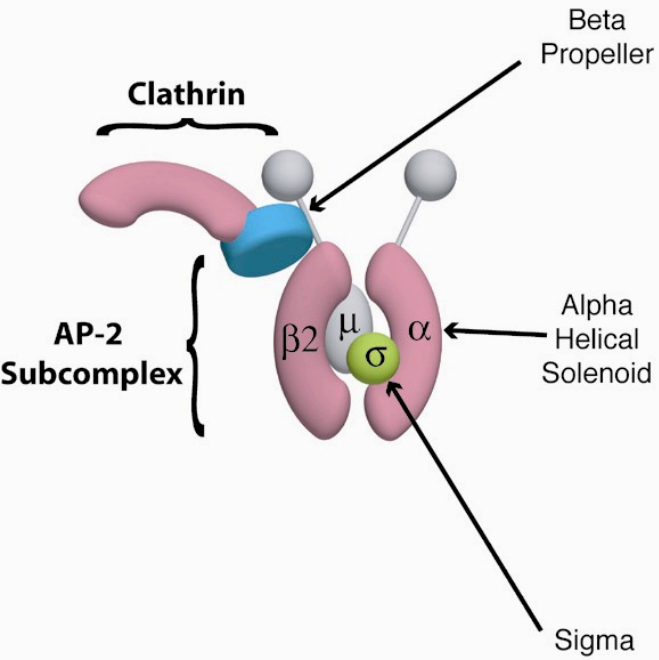


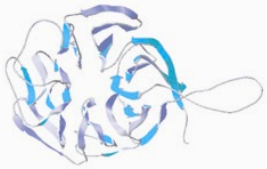


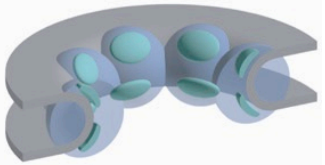




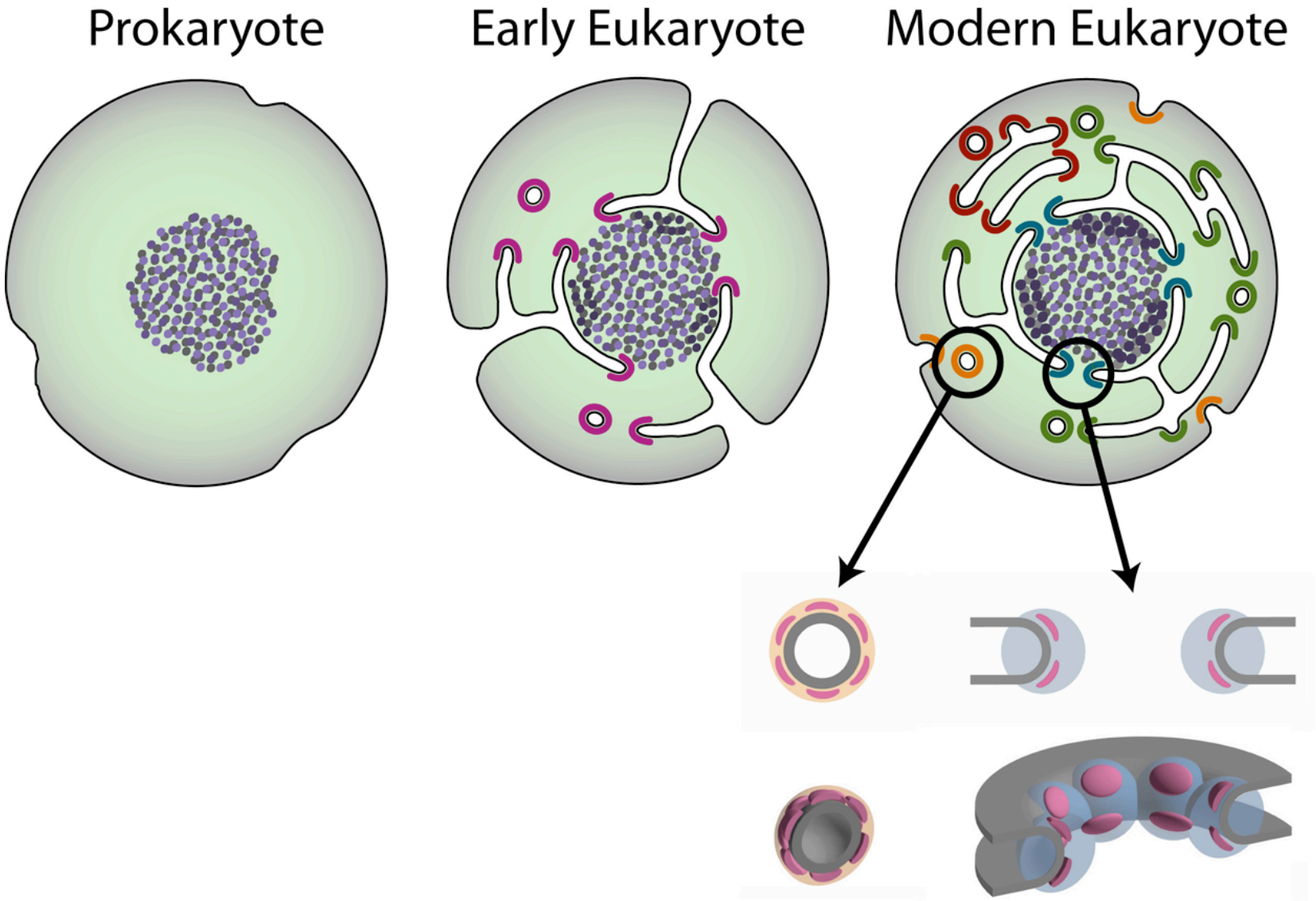
NPC and Coated Vesicles Share the β-Propeller and α-Solenoid Folds and Associate with Membranes



Fold	CVs	COPI	COPII	NPC
	Clathrin	α-subunit β'-subunit	Sec13 Sec31	Sec13 Seh1
	Clathrin α/γ/δ/ε-subunit β-subunit	α-subunit β-subunit γ-subunit ε-subunit β'-subunit	Sec31	Nup133 Nup120 Nup85 Nup84 Nup145C



Evolution?



Pore Formation

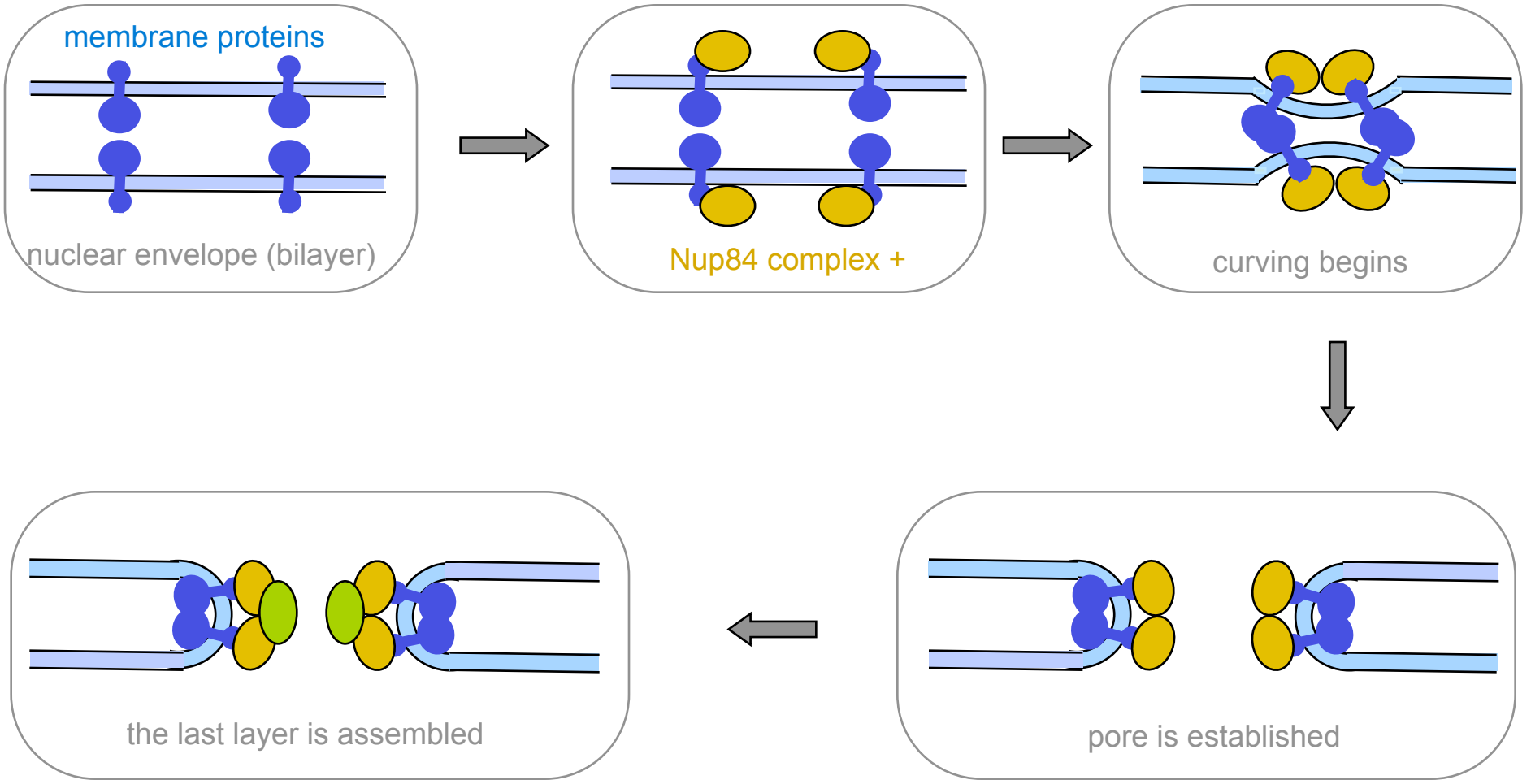
Need to maintain the integrity of the nuclear envelope.

1. From analogy with clathrins: Likely membrane-bending activity of the Nup84 complex;

2. From the NPC model: Nup84 complex interacts with the membrane proteins and/or the membrane.

3. From the expression profile clustering and the model: the order of assembly of NPC.

Pore Formation Hypothesis



Concluding remarks

- At present, useful 3D models can be obtained for domains in ~ 50% of the proteins (20% of domains).
- Completeness in structural coverage (structural genomics).
- Assembly of domains into higher order complexes.

Acknowledgments

http://salilab.org

Protein Structure Modeling
Andrej Sali
Bino John
Narayanan Eswar
Ursula Pieper
Roberto Sánchez (MSSM)
András Fiser (AECOM)
Francisco Melo (CU, Chile)
Azat Badretdinov (Accelrys)
M. S. Madhusudhan
Ash Stuart
Nebojša Mirkovic
Valentin Ilyin (NE)
Eric Feyfant (GI)
Min-Yi Shen
Ben Webb
Rachel Karchin
Mark Peterson

Assemblies
Frank Alber
Damien Devos
Maya Topf
Dmitry Korkin
Narayanan Eswar
Fred Davis
M.S. Madhusudhan
Mike Kim

1D to 3D for biologists
David Huassler (UCSC)
Jim Kent (UCSC)
Daryl Thomas (UCSC)
Mark (UCSC)
Rolf Apweiler (EBI)

Yeast NPC
Tari Suprapto (RU)
Julia Kipper (RU)
Wenzhu Zhang (RU)
Liesbeth Veenhoff (RU)
Sveta Dokudovskaya (RU)
J. Zhou (USC)
Mike Rout (RU)
Brian Chait (RU)

Ribosomes
J. Frank

Chimera
P. Babbitt
T. Ferrin

Brain Lipid Binding Protein
Liang Zhu (RU)
Nat Heintz (RU)

Structural Genomics
Stephen Burley (SGX)
John Kuriyan (UCB)
NY-SGXRC

BRCA1
A. Monteiro (Cornel)

Mast Cell Proteases
Rick Stevens (BWH)

Fly p53
Shengkan Jin (RU)
Arnie Levine (RU)

NIH
NSF
Sinsheimer Foundation
A. P. Sloan Foundation
Burroughs-Wellcome Fund
Merck Genome Res. Inst.
Mathers Foundation
I.T. Hirschl Foundation
The Sandler Family Foundation
Human Frontiers Science Program
SUN
IBM
Intel
Structural Genomix