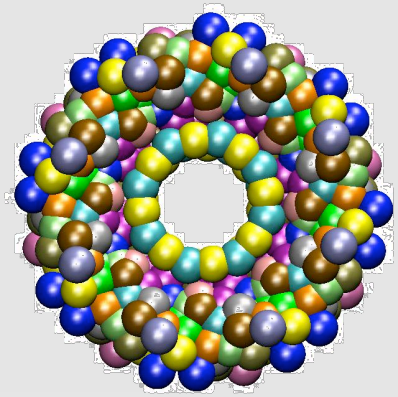
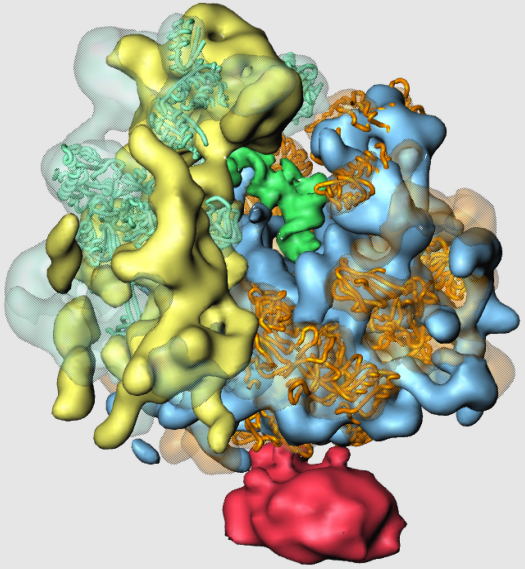
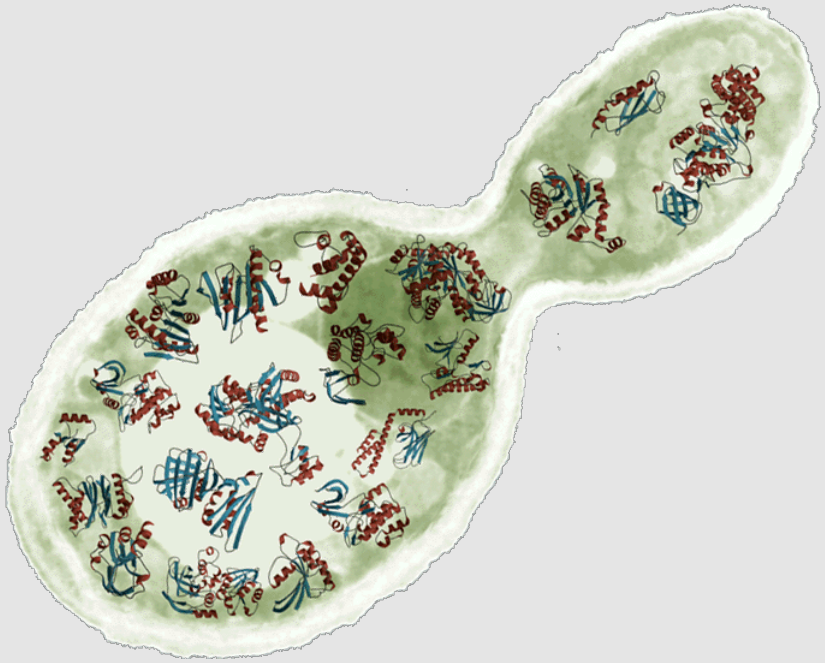
Aligning Sequences and Structures for Comparative Modeling

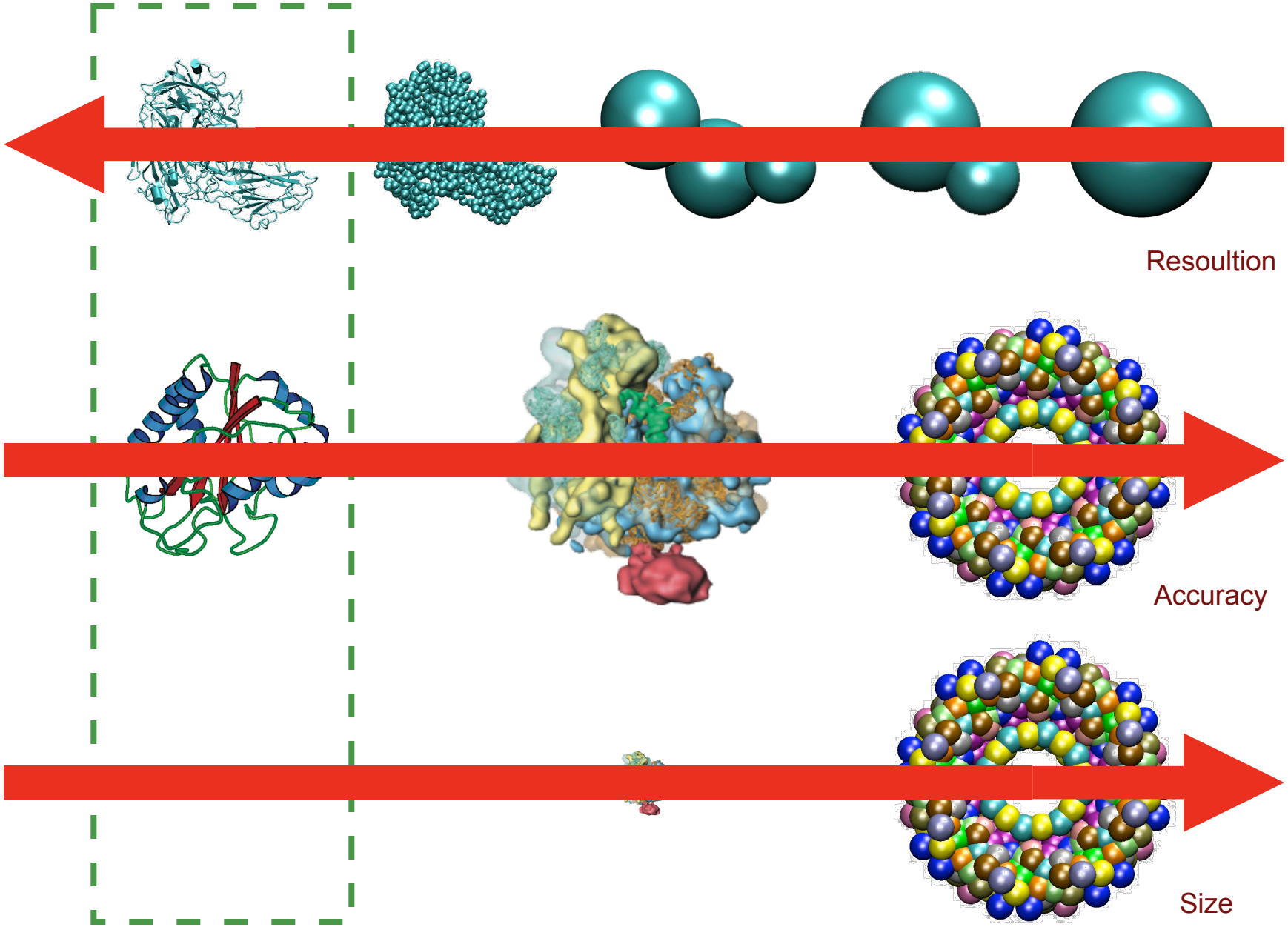






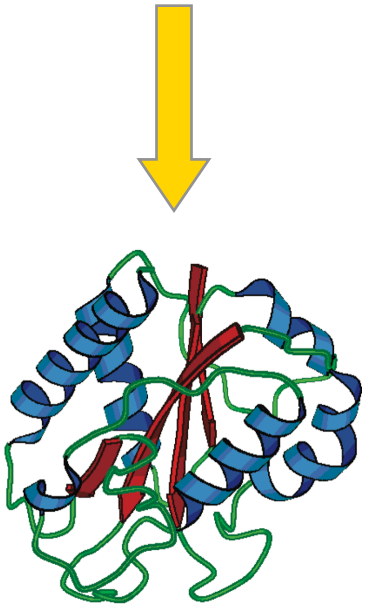
Marc A. Marti-Renom
http://salilab.org/~marcius

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


Resoultion
Accuracy
Size

Principles of protein structure

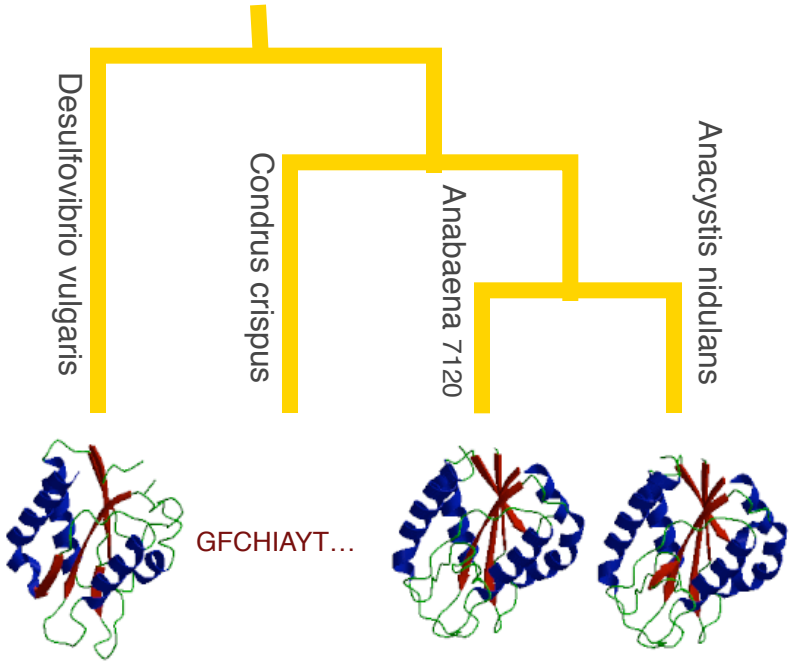
GFCHIKAYTRLIMVG…



Folding

(physics)

Ab initio prediction



Evolution

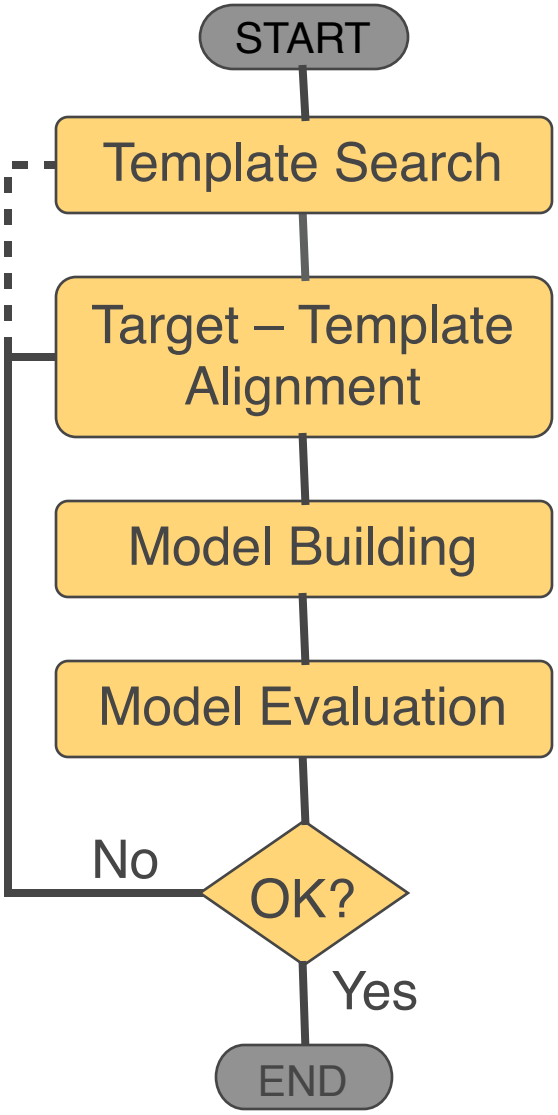
(rules)

Threading

Comparative Modeling

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

Steps in Comparative Protein Structure Modeling



TARGET

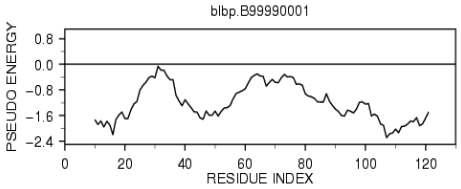
ASILPKRLFGNCEQTSDEG
LKIERTPLVPHISAQNVCLKI
DDVPERLIPERASFQWMN
DK

TEMPLATE

DBAli
SALIGN

ASILPKRLFGNCEQTSDEGLKIERTPLVPHISAQNVCLKIDDVPERLIPE
MSVIPKRLYGNCEQTSEEAIRIEDSPIV---TADLVCLKIDEIPERLVGE

MOULDER
SALIGN



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

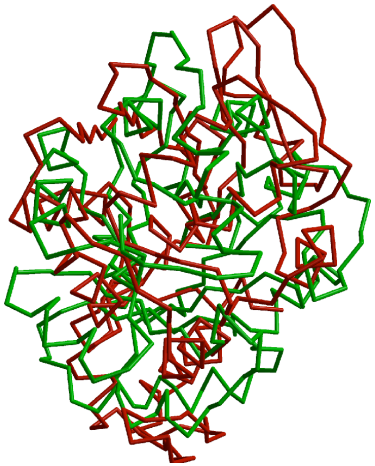
Typical errors in comparative models

MODEL

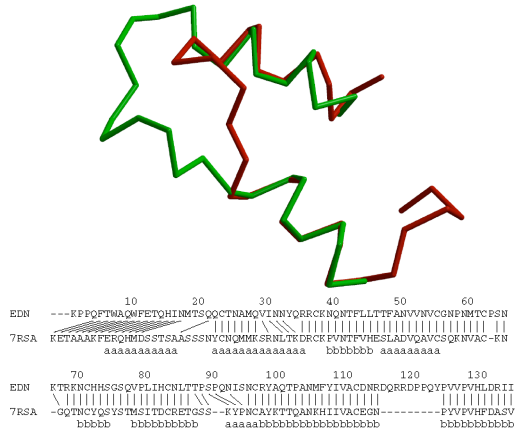
X-RAY

TEMPLATE

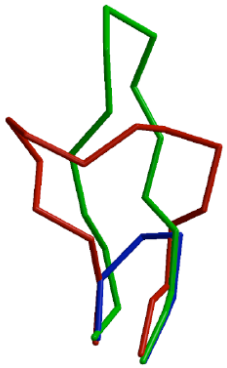
Incorrect template



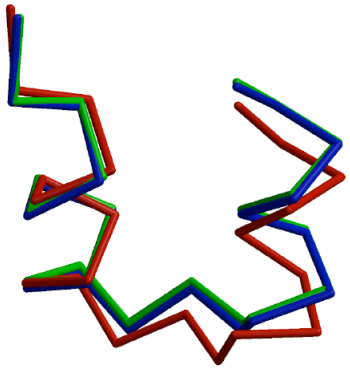
Misalignment



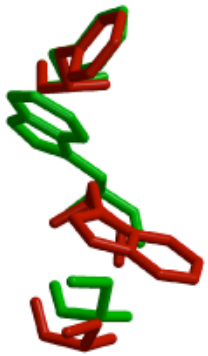
Region without a template



Distortion/shifts in aligned regions

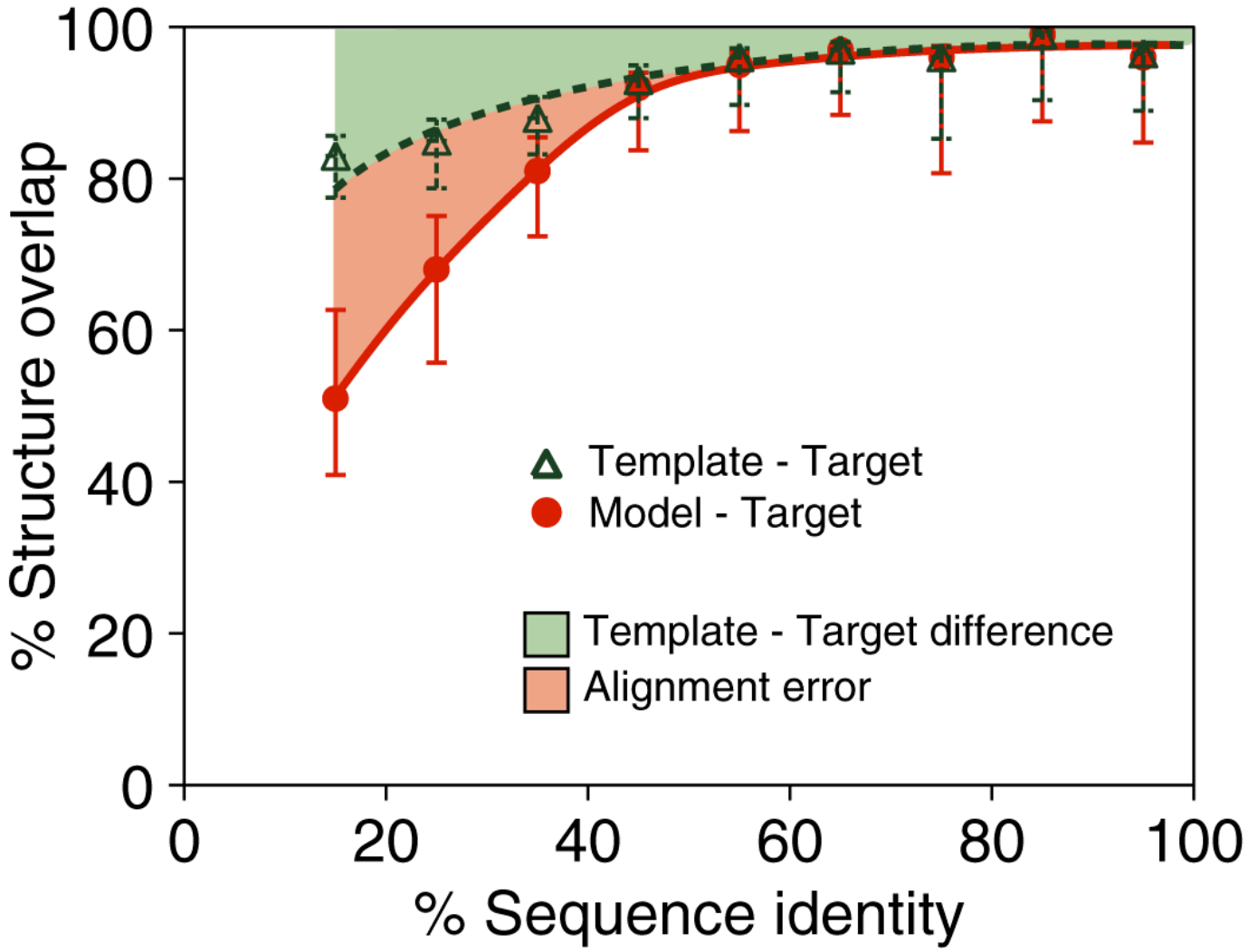


Sidechain packing



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

Alignment errors are frequent and large



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

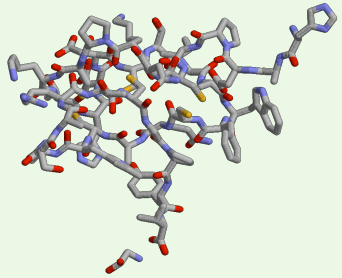
SALIGN & DBAli

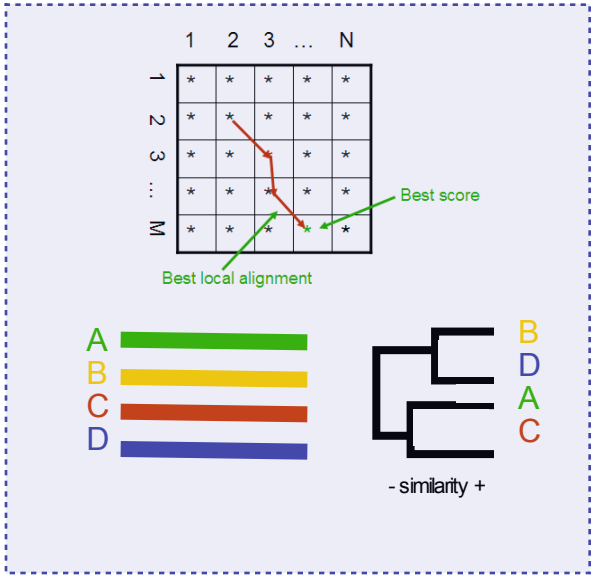
aligning structures

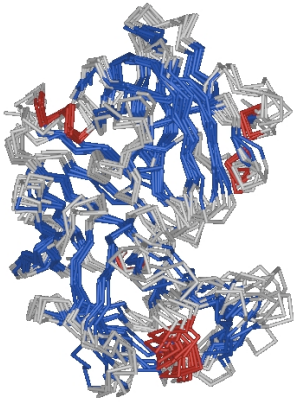
M.S. Madhusudhan, M.A. Marti-Renom, N. Eswar and A. Sali.
SALIGN: aligning structures with MODELLER.
in preparation

M.A. Marti-Renom and A. Sali.
DBAli: a comprehensive database of protein structure alignments.
in preparation

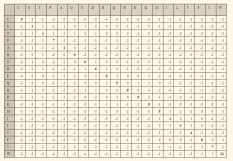
Structural alignment by properties conservation (SALIGN-MODELLER)

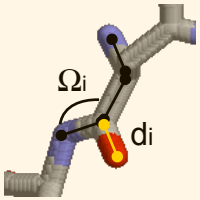


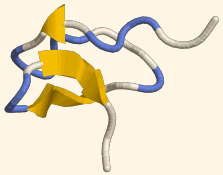


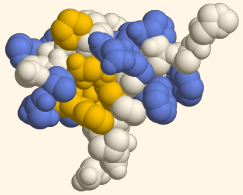


- ✓ **Uses all available structural information**
- ✓ **Provides the optimal alignment**
- • **Computationally expensive**









$$\text{RMSD} = \sqrt{\sum \left(x_i - \bar{x} \right)^2}$$

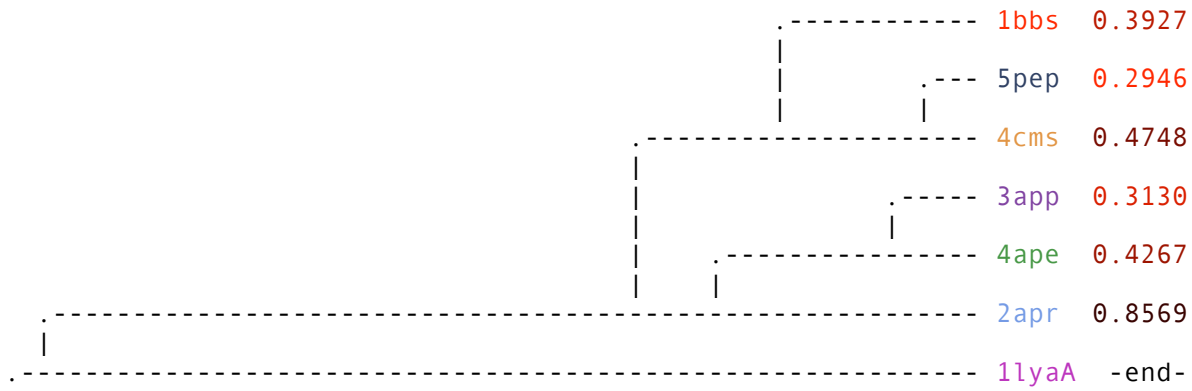
$R_{i,j}$ $D_{i(3),j(3)}$ $S_{i,j}$ $B_{i,j}$ $I_{i,j}$

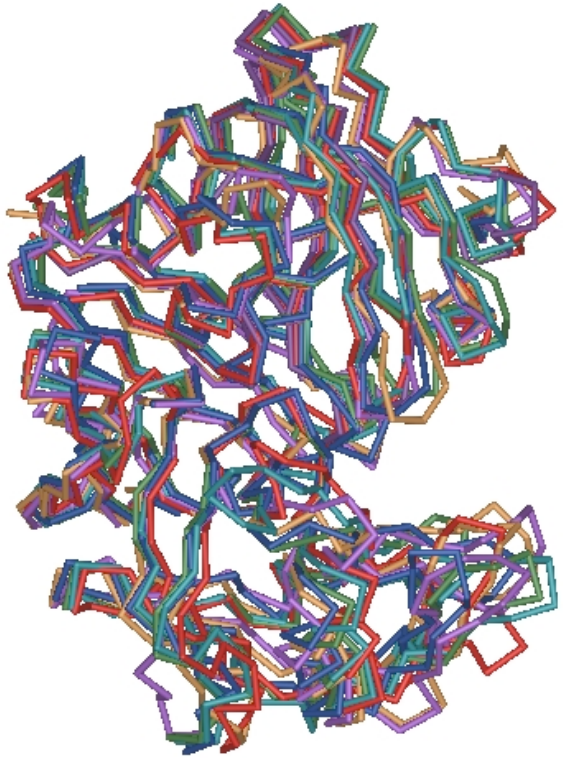
$$Score_{i,j} = w_1 * R_{i,j} + w_2 * D_{i(a),j(a)} + w_3 * S_{i,j} + w_4 * B_{i,j} + w_5 * I_{i,j} + w_6 * X_{i,j}$$

Madhusudhan et al. *in preparation*

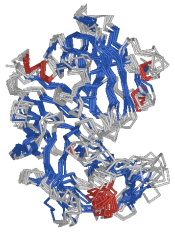
Multiple structure ‘tree’ alignment

	1bbs	1lyaA	5pep	4cms	3app	4ape	2apr
1bbs	-----	0.831	0.373	0.413	0.511	0.495	0.485
1lyaA		-----	0.847	0.839	0.885	0.875	0.874
5pep			-----	0.295	0.462	0.455	0.431
4cms				-----	0.486	0.482	0.447
3app					-----	0.313	0.424
4ape						-----	0.429
2apr							-----



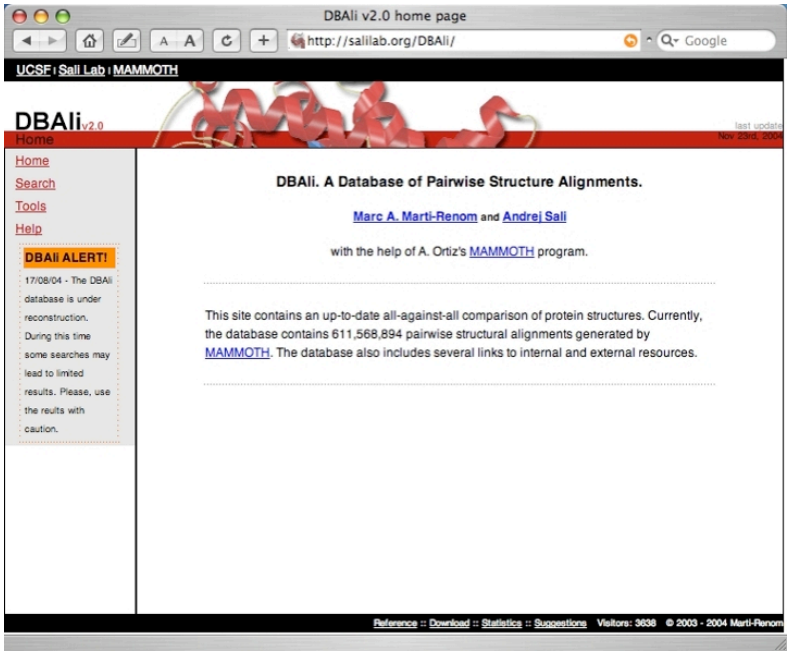


	1bbs	1lyaA	5pep	4cms	3app	4ape	2apr
1bbs	0	95	319	315	305	302	308
1lyaA	0	0	92	93	89	93	91
5pep	0	0	0	318	303	296	312
4cms	0	0	0	0	303	301	309
3app	0	0	0	0	0	319	310
4ape	0	0	0	0	0	0	313
2apr	0	0	0	0	0	0	0



DBAli$_{v2.0}$ database

http://salilab.org/DBAli/



- ✓ **Fully-automatic**
- ✓ **Data is kept up-to-date with PDB releases**
- ✓ **Tools for "on the fly" classification of families.**
- ✓ **Easy to navigate**
- ✓ **Provides tools for structural analysis**

- **Does not provide (yet) a stable classification**

Uses MAMMOTH for similarity detection

- ✓ **VERY FAST!!!**
- ✓ **Good scoring system with significance**

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

DBAli statistics as of Saturday 27th of November 2004

Last updated:
November 26th, 2004 (19:29h)

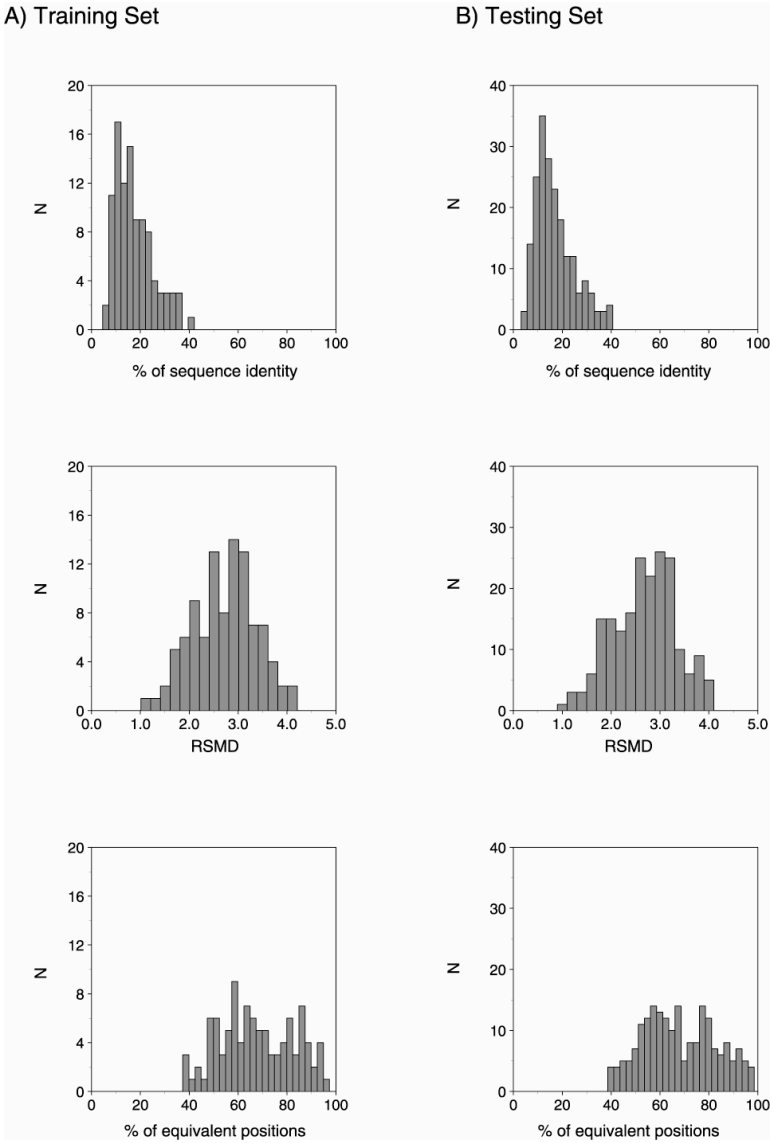
Number of chains in database:
58,545

Number of structure-structure comparisons:
612,899,530

SALIGN

aligning profiles

M.A. Marti-Renom, M.S. Madhusudhan, A. Sali.
Alignment of Protein Sequences by Their Profiles.
Protein Sciences **13**, 1071-1087, 2004.



Seq.-Seq.

ALIGN: DP pairwise method

BLAST2SEQ: Local heuristic method

Seq.-Str.

SEA: Local structure prediction method

Prof.-Seq.

SAM: HMM method

PSI-BLAST: Local search method that uses multiple sequence information for one of the sequences.

LOBSTER: HHM + Phylogeny Method

Prof.-Prof.

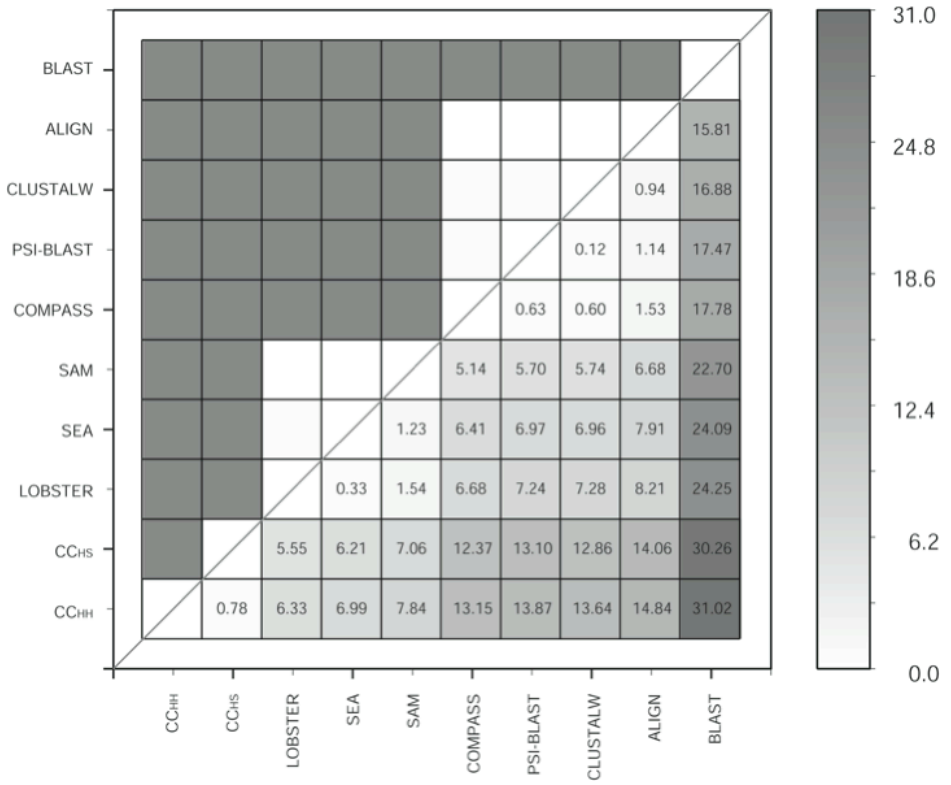
CLUSTALW: DP multiple sequence method.

COMPASS: DP profile-profile method

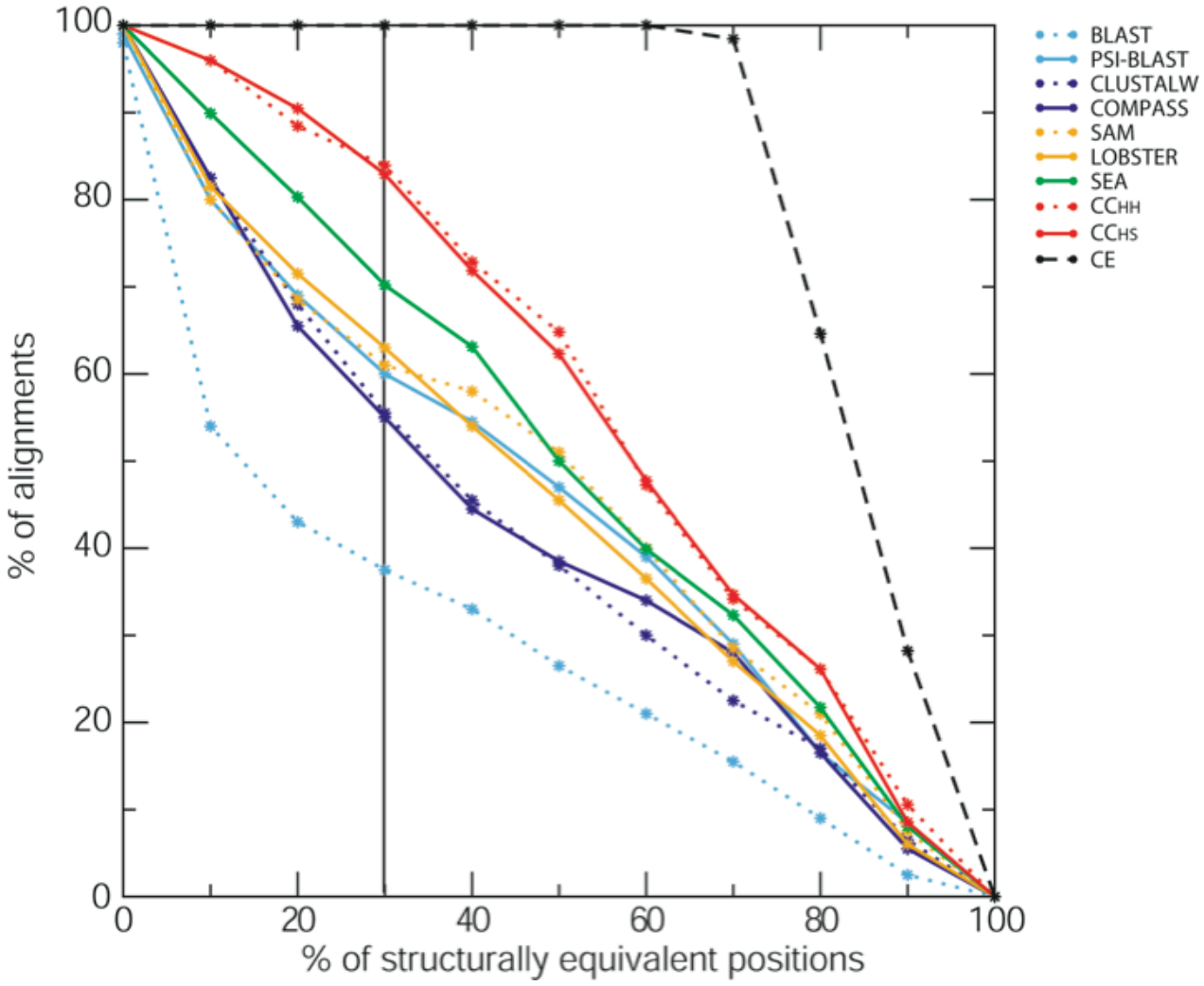
SALIGN: DP pairwise method that uses multiple sequence information for both sequences.

SALIGN accuracy

Method	CE overlap	Shift score
CE	100 ± 0	1.00 ± 0.00
BLAST	26 ± 29	0.32 ± 0.33
PSI-BLAST	43 ± 31	0.48 ± 0.35
SAM	48 ± 26	0.50 ± 0.34
LOBSTER	50 ± 27	0.51 ± 0.32
SEA	49 ± 27	0.53 ± 0.29
ALIGN	42 ± 25	0.44 ± 0.28
CLUSTALW	43 ± 27	0.44 ± 0.31
COMPASS	43 ± 32	0.49 ± 0.35
CC$_{HH}$	**56 ± 23**	**0.61 ± 0.24**
CC$_{HS}$	**56 ± 24**	**0.62 ± 0.24**
TOP	62 ± 20	0.67 ± 0.20



SALIGN success



Alignment accuracy (CE overlap)

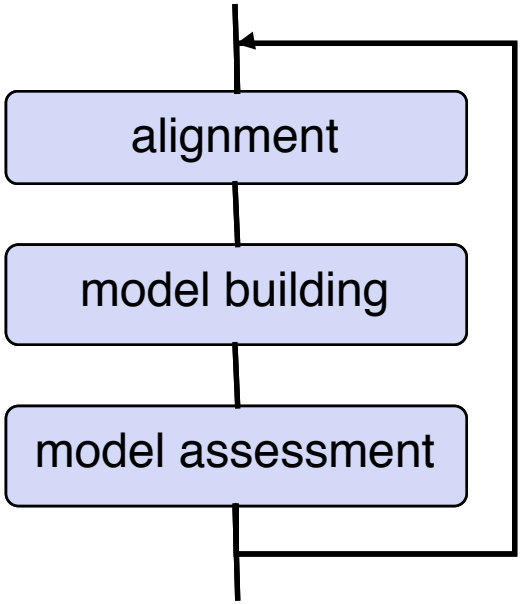
200 pairwise DBAli alignments

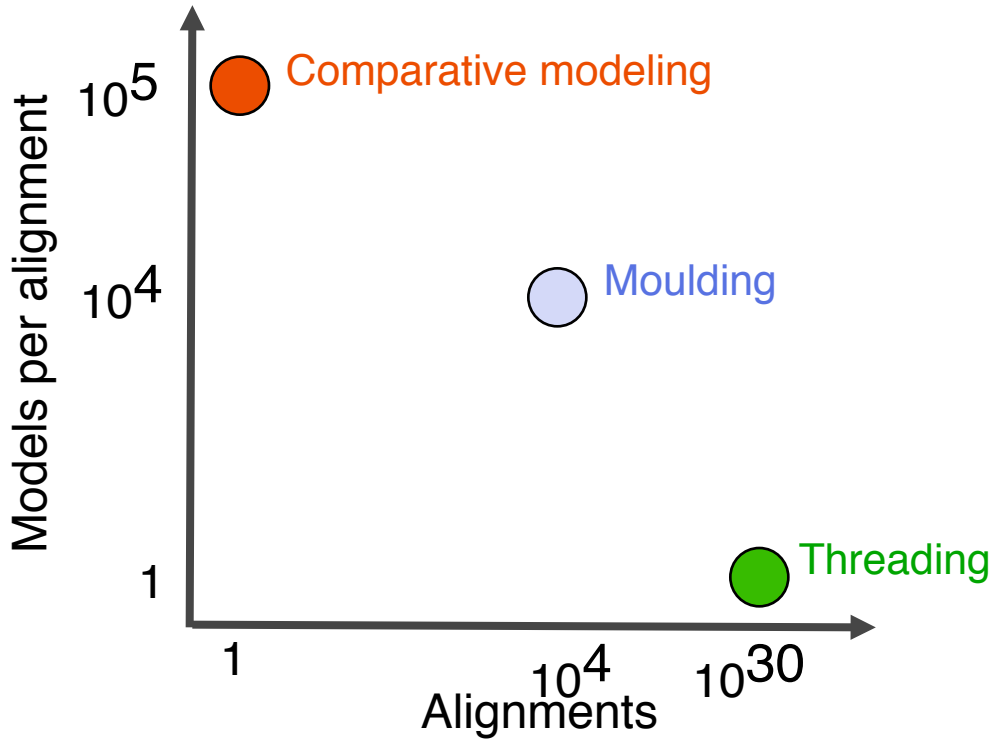
PSI-BLAST (sequence-profile alignment)	43%
SEA (local structure alignment)	49%
SALIGN (profile-profile alignment)	56%

MOULDER

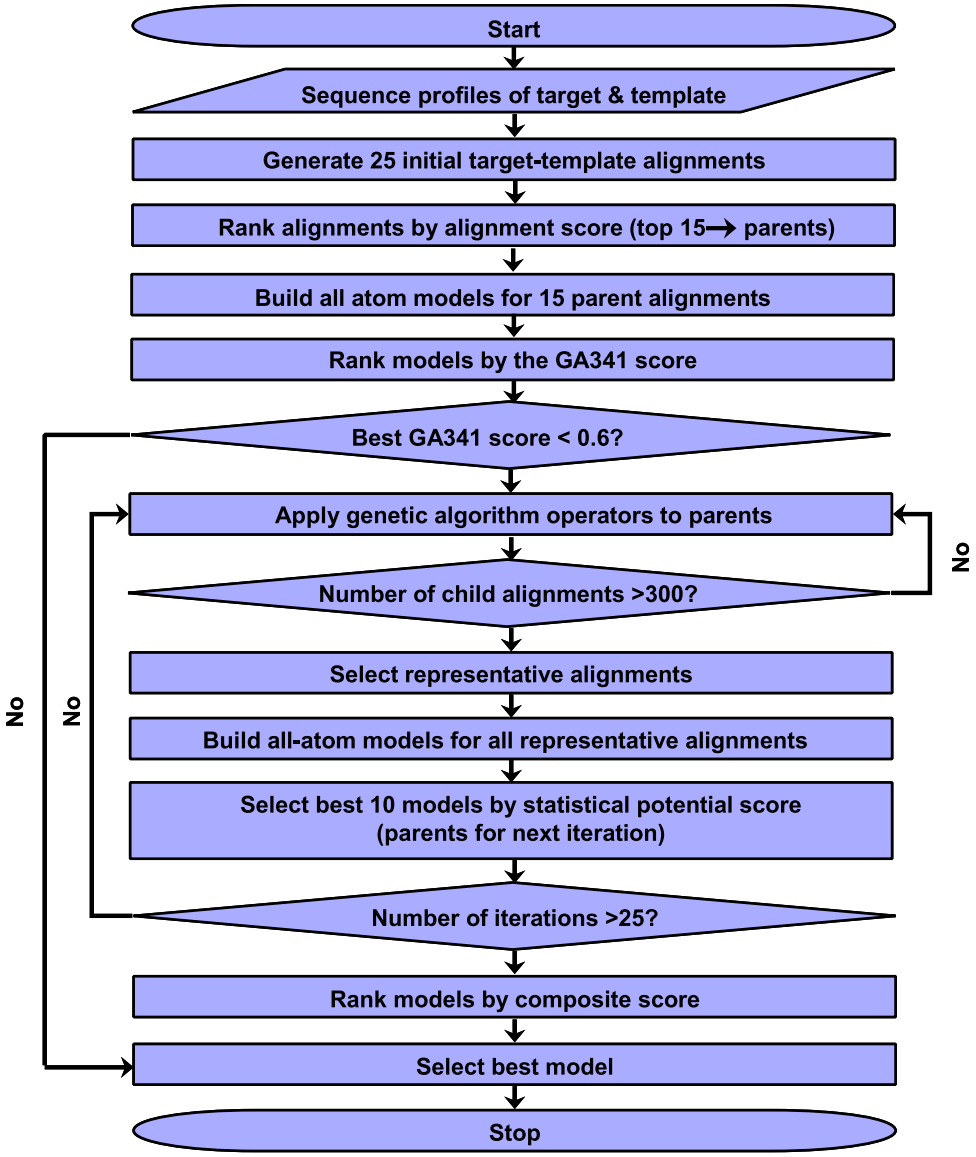
B. John, A. Sali.
Comparative Protein Structure Modeling by Iterative Alignment, Model Building, and Model Assessment.
Nucleic Acids Research **31**, 3982-3992, 2003.

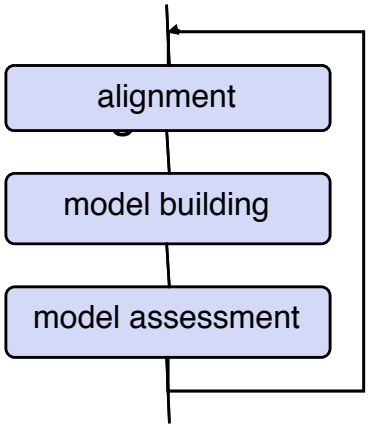
Moulding: iterative alignment, model building, model assessment





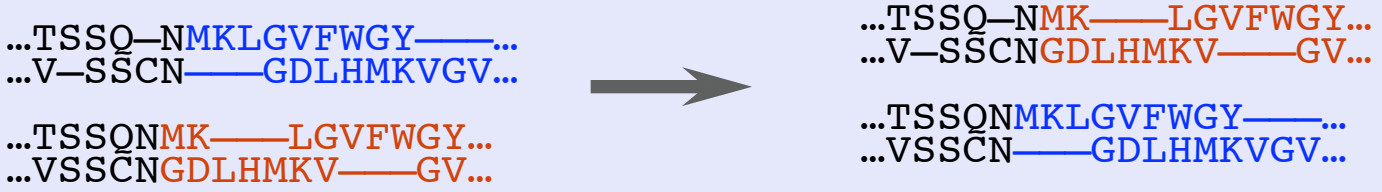
Moulding by a Genetic Algorithm approach





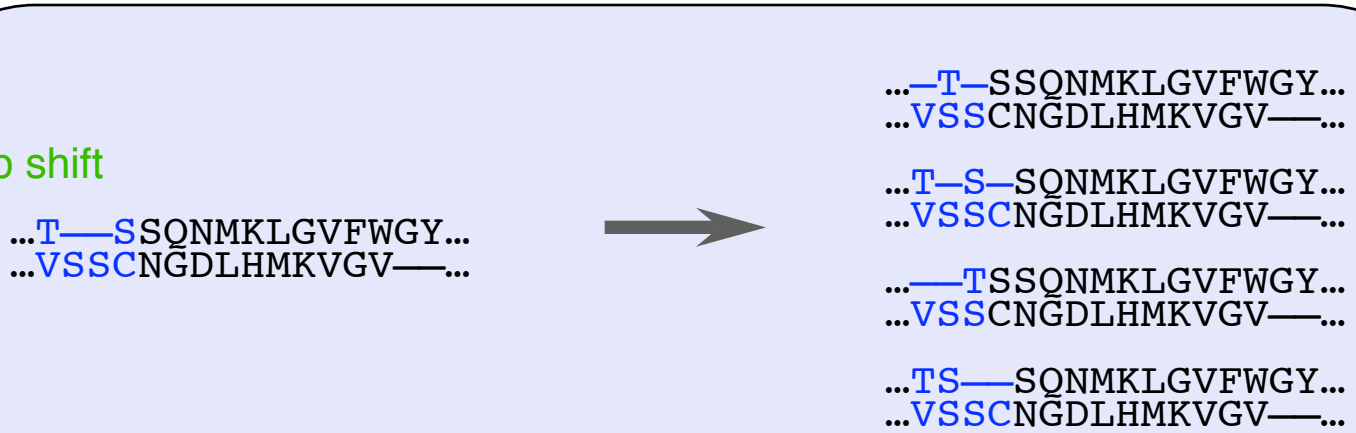
Genetic algorithm operators

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).

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

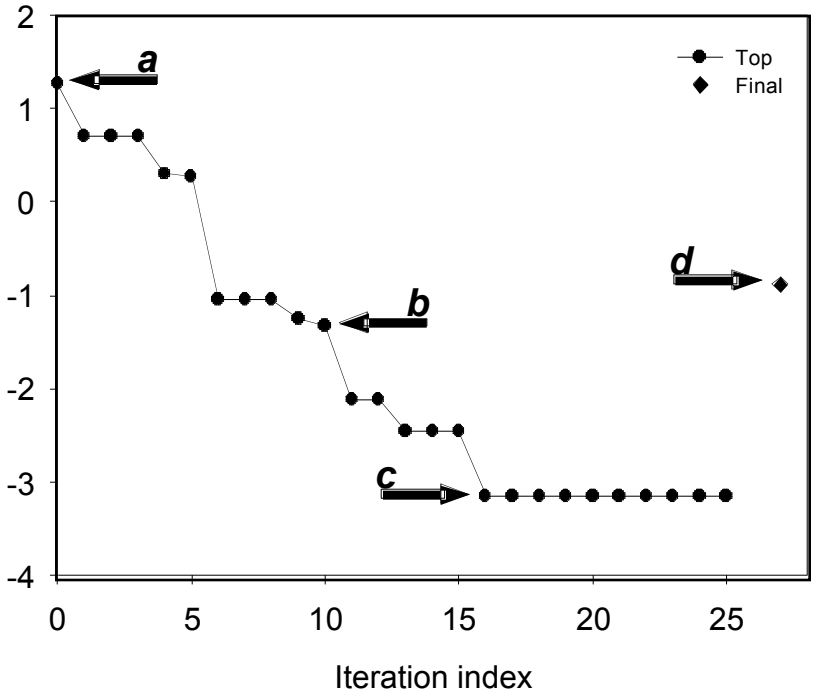
$Z(\text{score}) = (\text{score} - \mu)/\sigma$

μ … average score of all models

σ … standard deviation of the scores

Application to a difficult modeling case

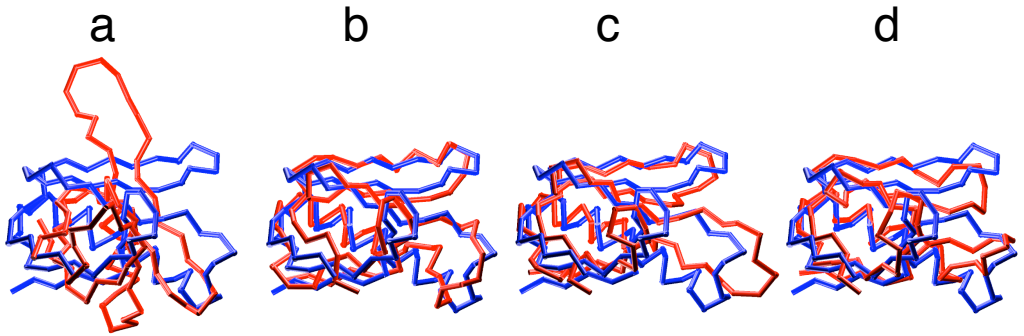
1BOV-1LTS



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%

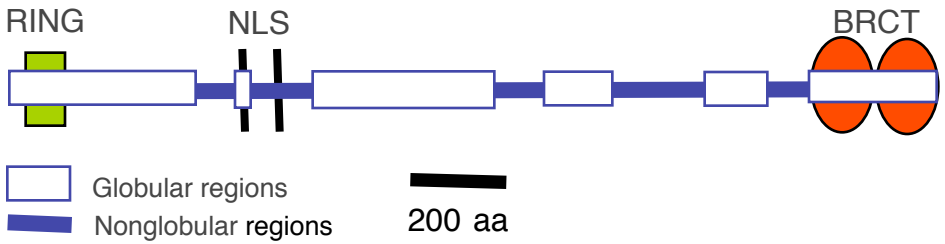
Structural analysis of missense mutations in human BRCA1 BRCT domains

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

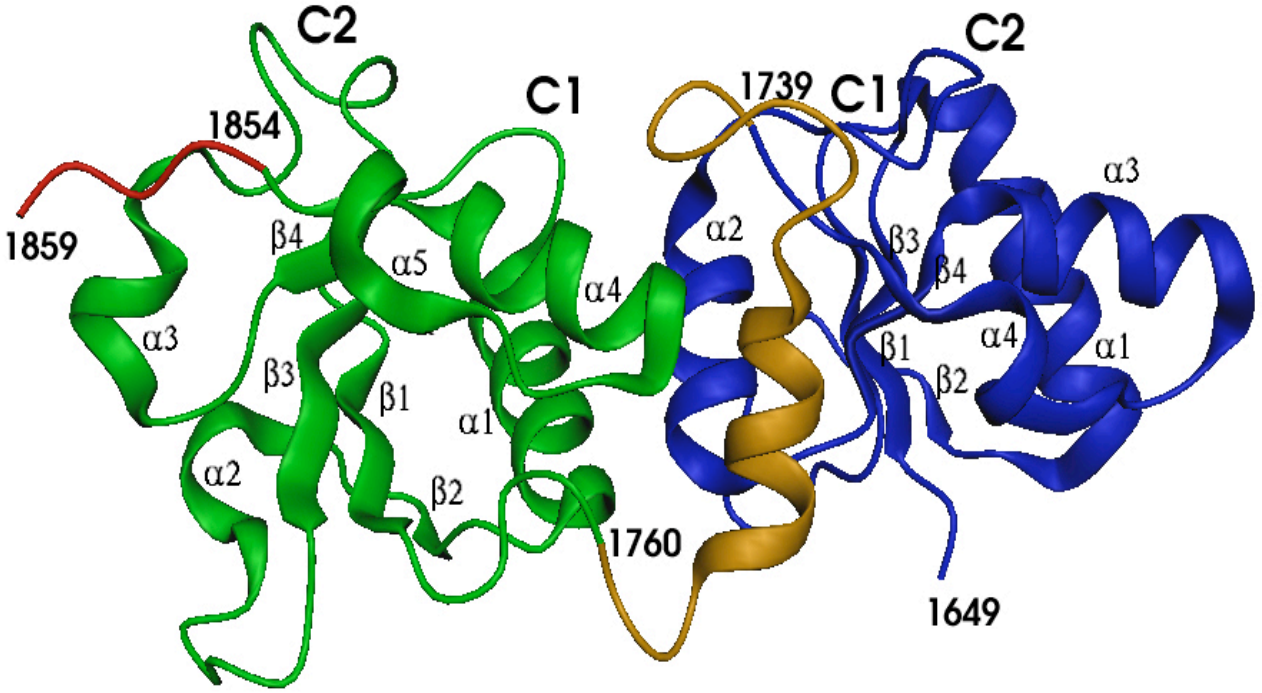
Cancer Research (June 2004). 64:3790-97

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.

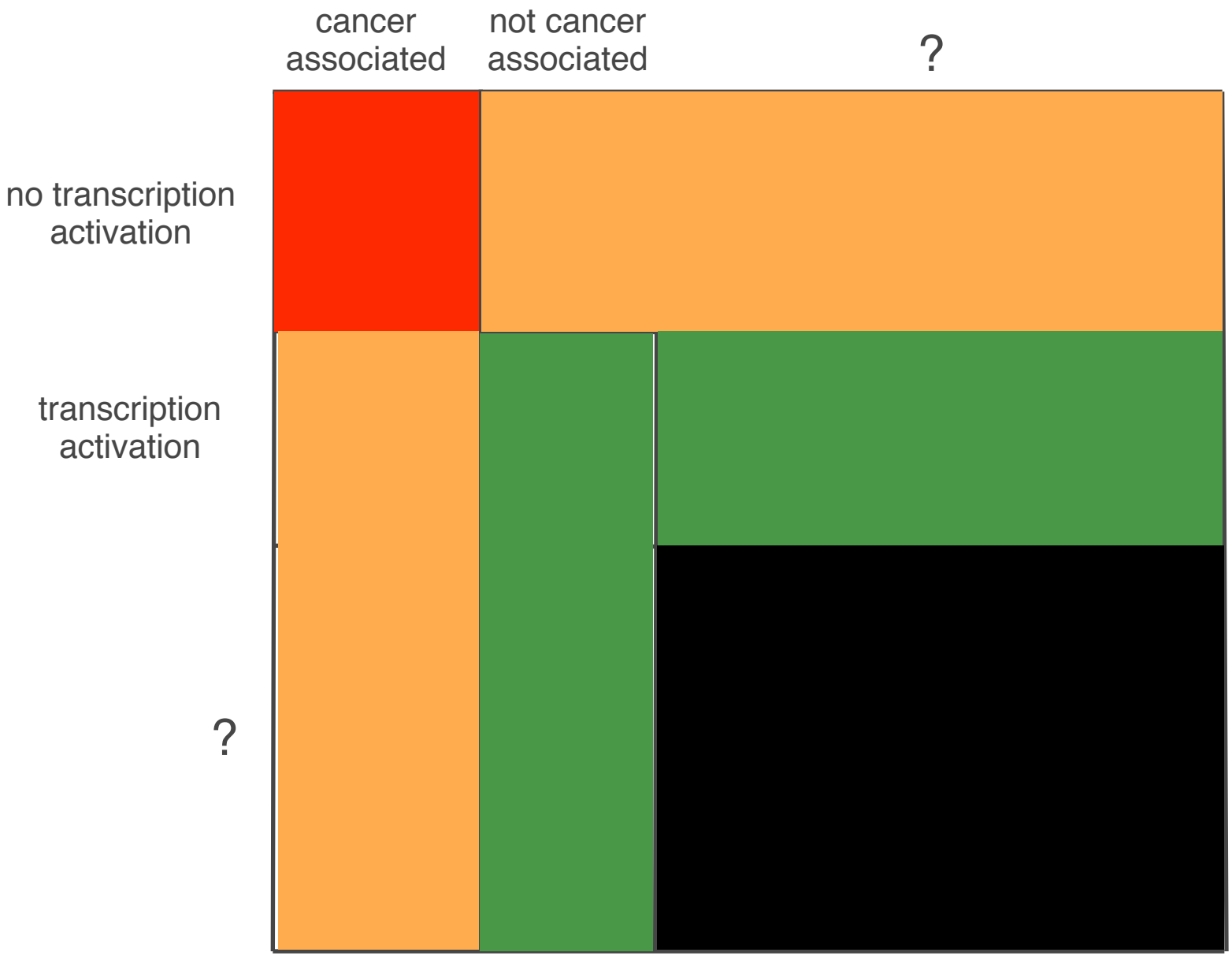
Authorized Signature:

Brian E. Ward, Ph.D.
Laboratory Director

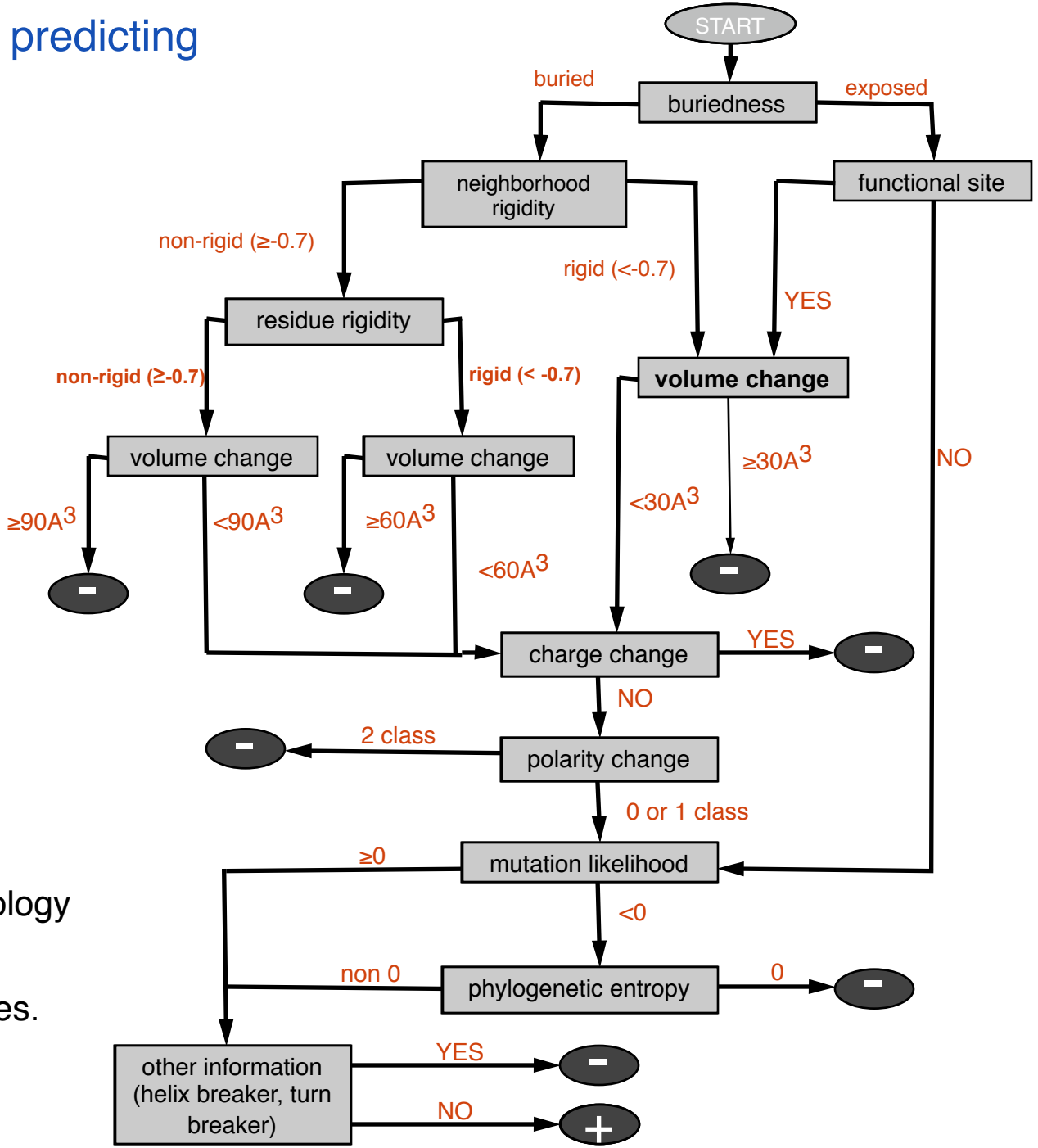
Thomas S. Frank, M.D.
Medical Director

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

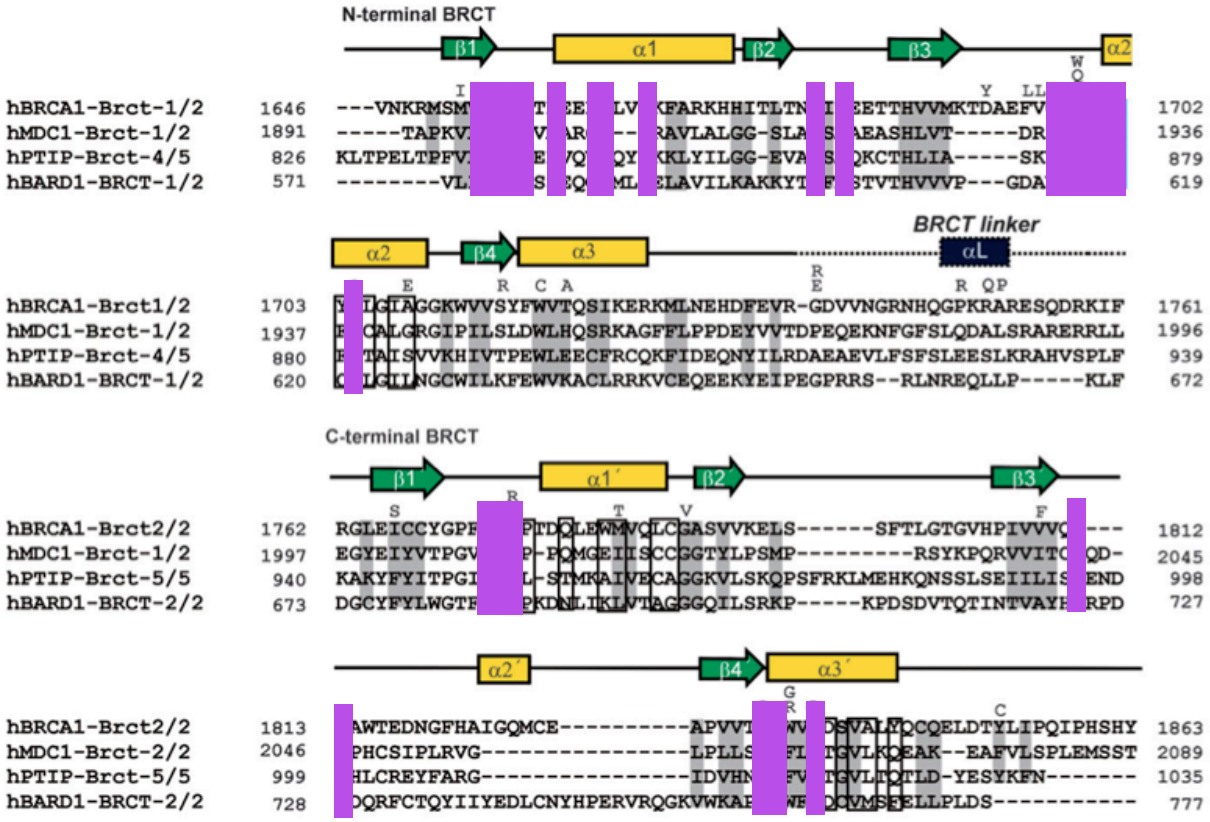


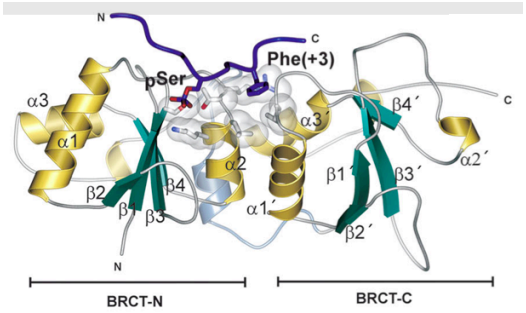
http://salilab.org/snpweb

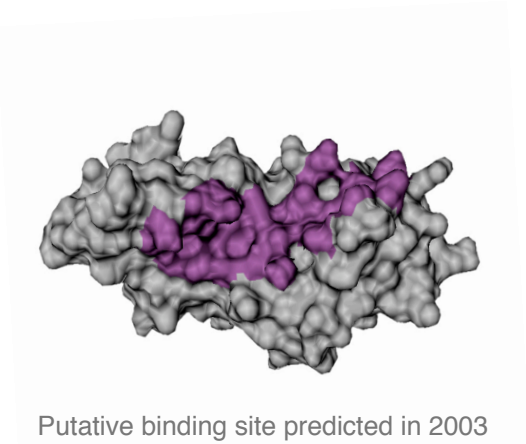
Mirkovic et al., Cancer Biology (2004) 64:3790-97

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

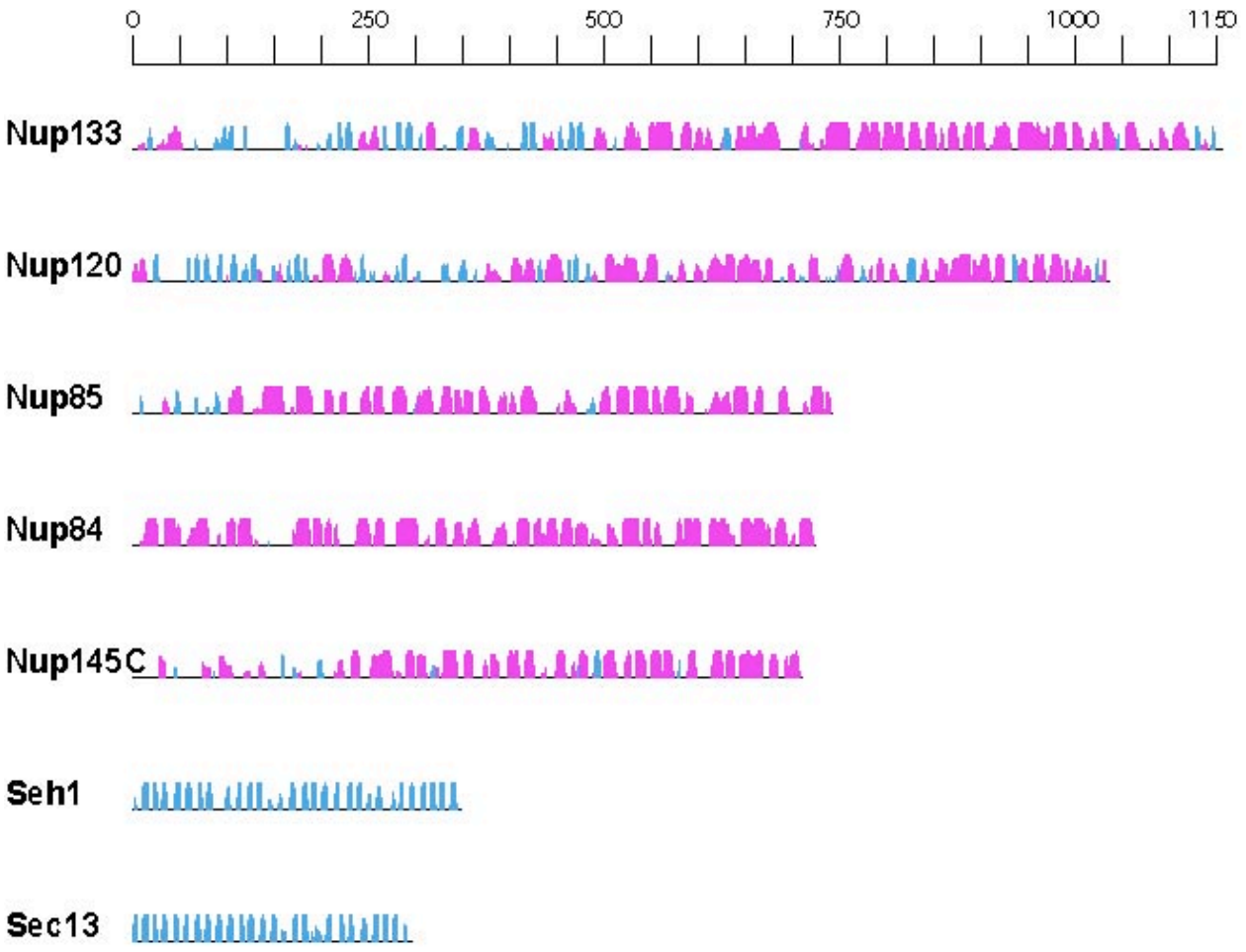
Mirkovic *et al.* 2004 Cancer Research. **June 2004 64**:3790

Common Evolutionary Origin of Coated Vesicles and Nuclear Pore Complexes

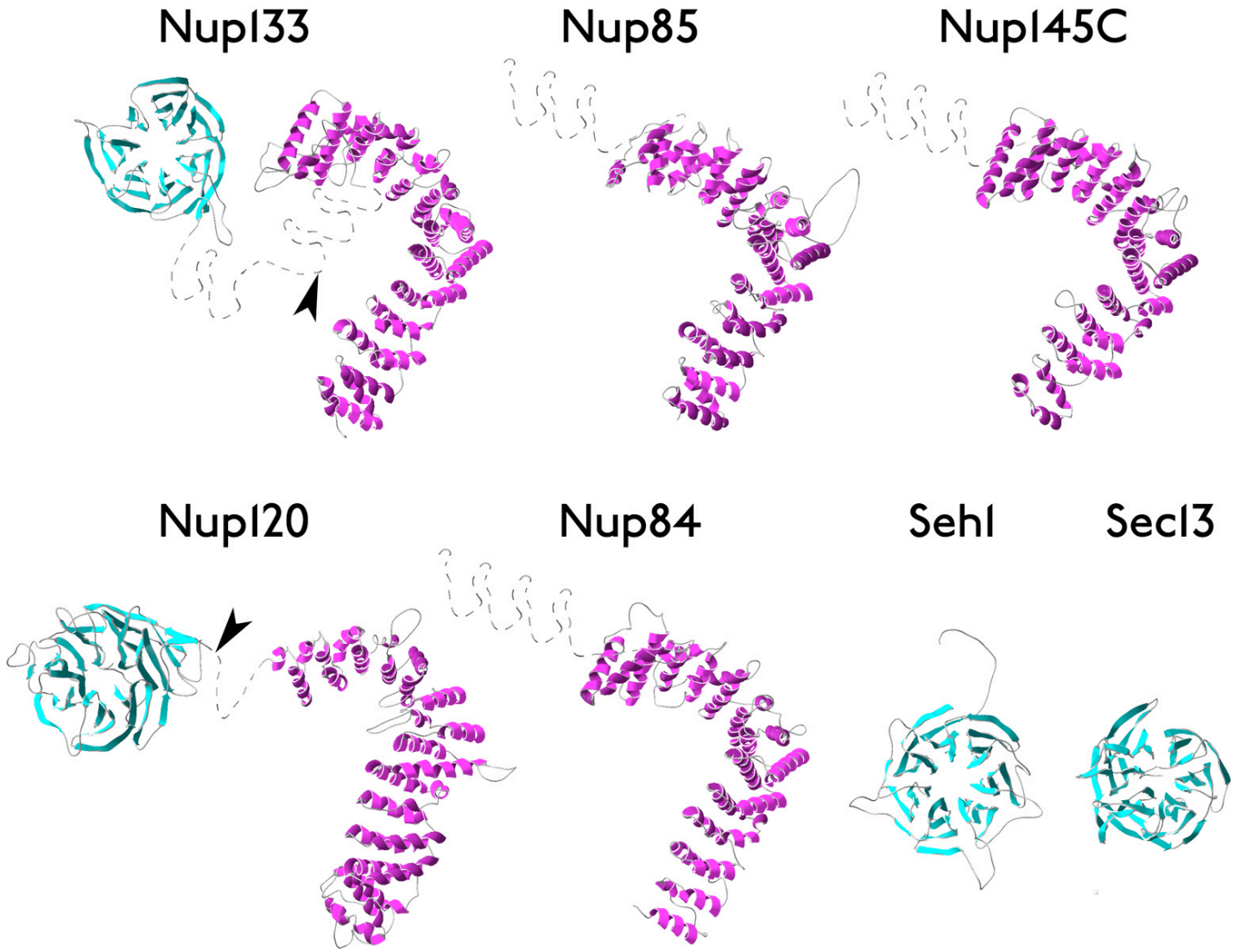
mGenThreader + SALIGN + MOULDER

D. Devos, S. Dokudovskaya, F. Alber, R. Williams, B.T. Chait, A. Sali, M.P. Rout.
Components of Coated Vesicles and Nuclear Pore Complexes Share a Common Molecular Architecture.
PLOS Biology **2(12):**e380, 2004

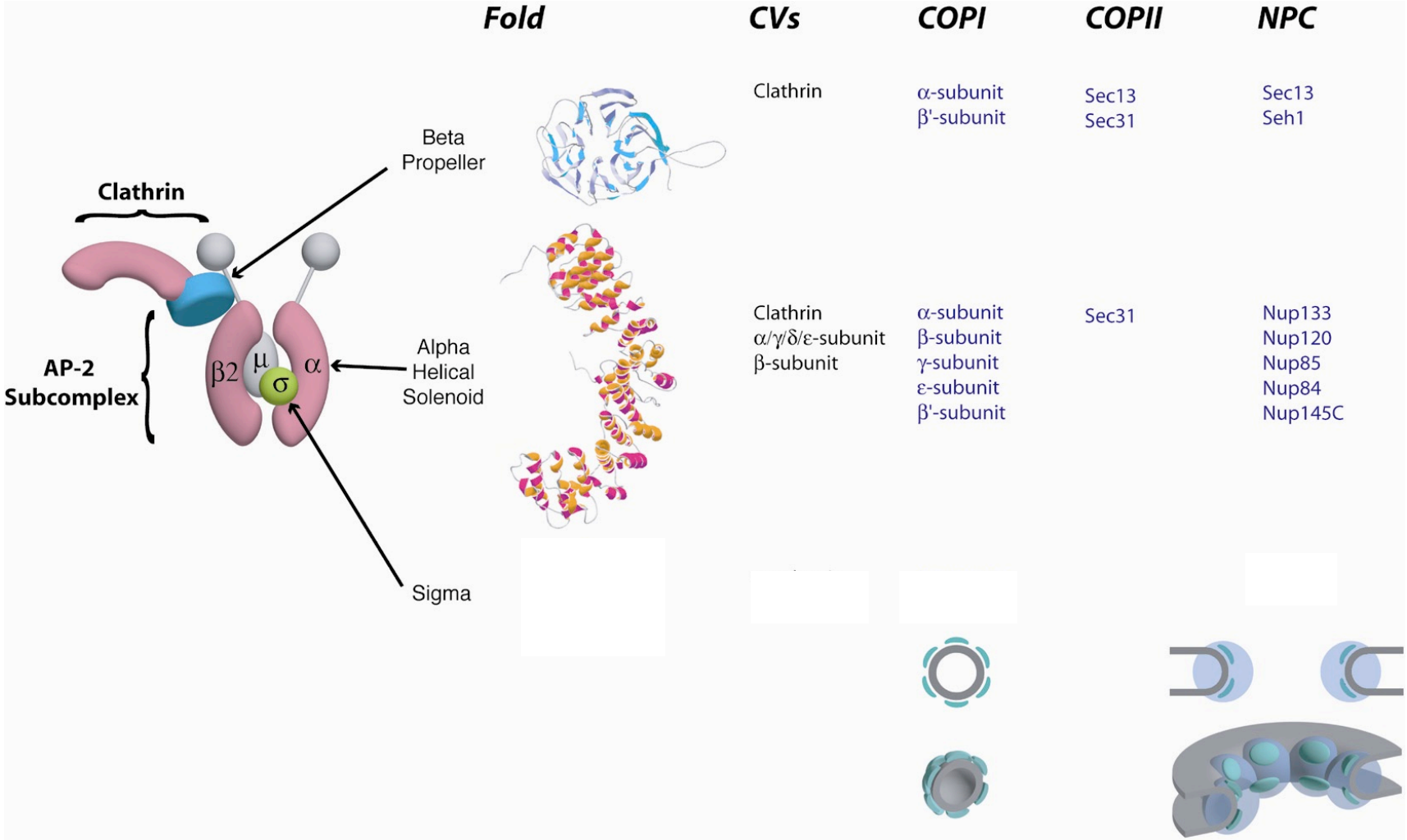
yNup84 complex proteins



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



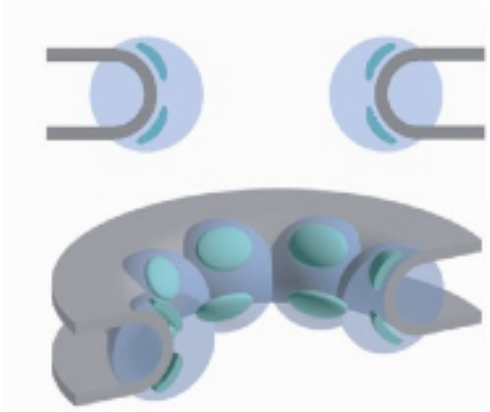
NPC and Coated Vesicles Share the β-**Propeller and** α-**Solenoid** Folds and Associate with Membranes



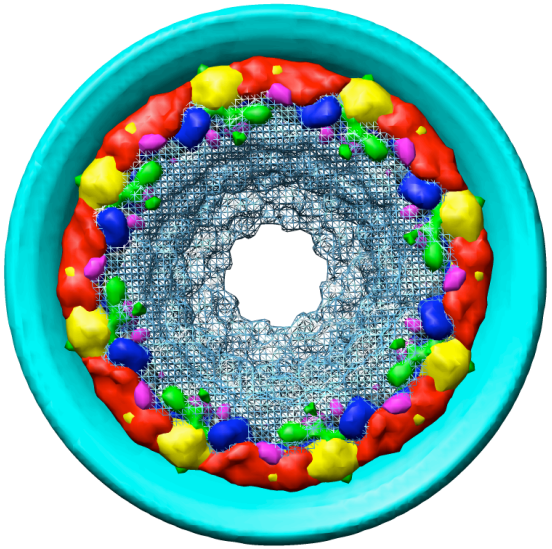
NPC and Coated Vesicles Both Associate with Membranes

Coated Vesicle

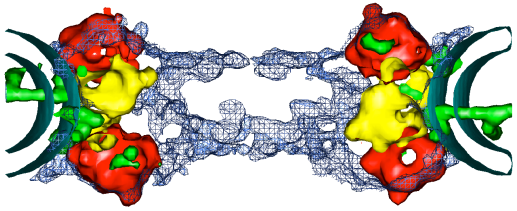
NPC model



top view

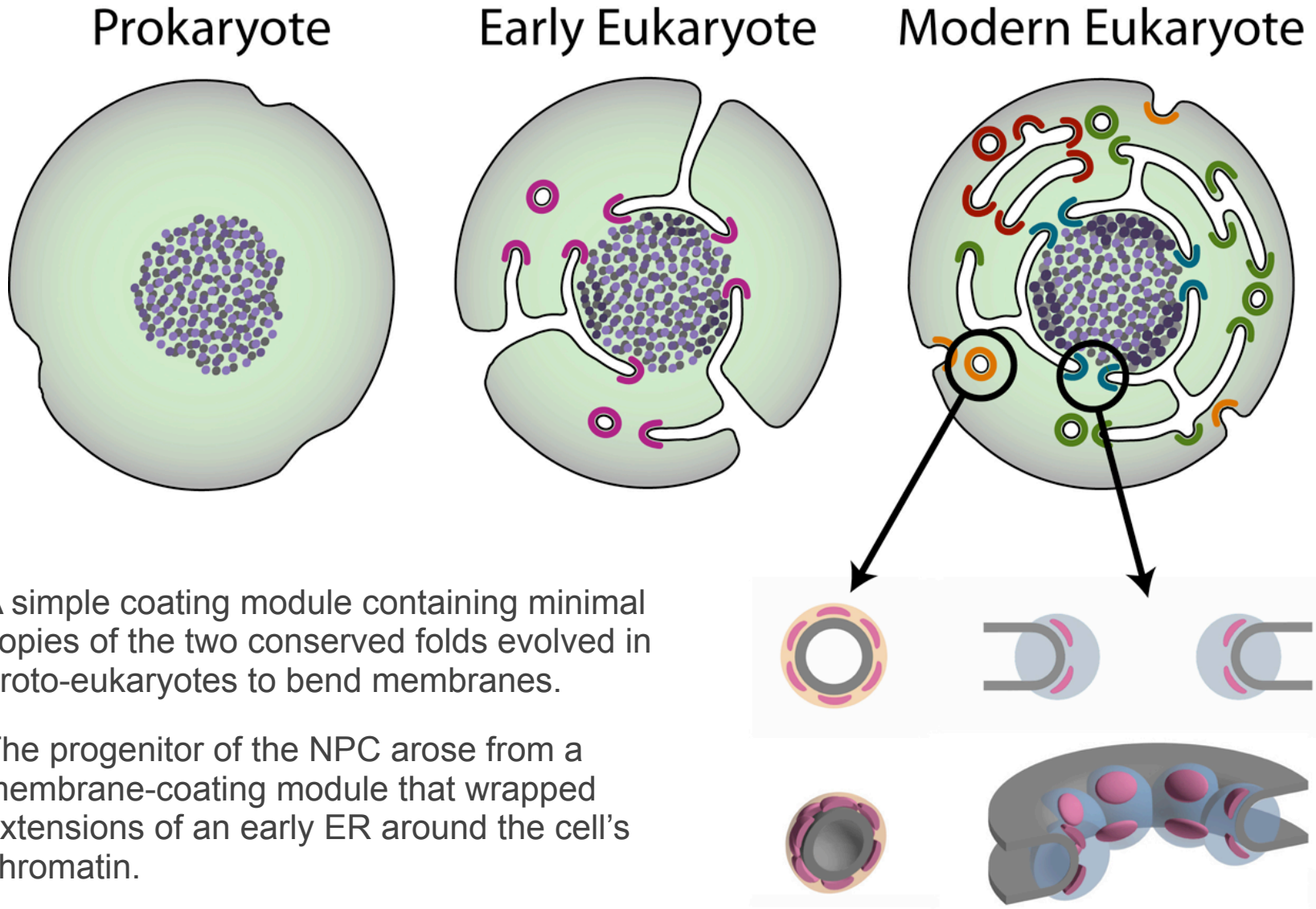


side view



Nup 84 complex

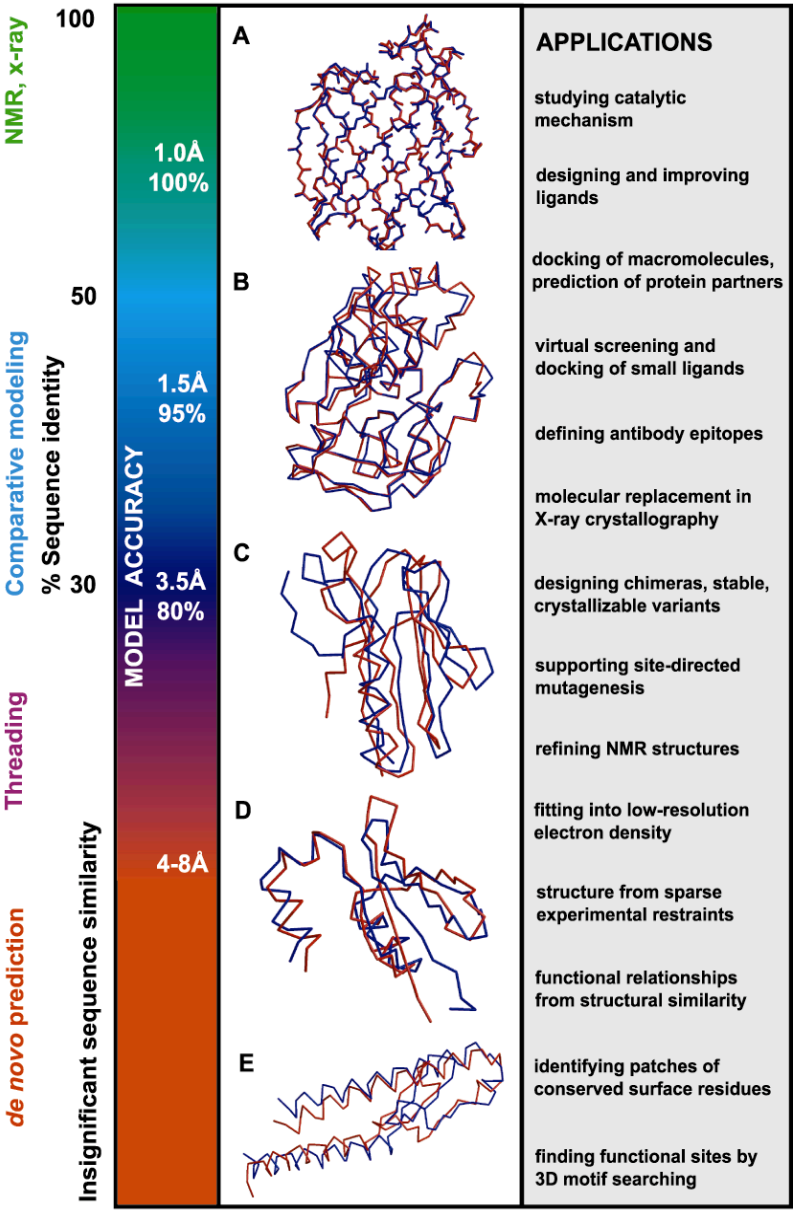
A Common Evolutionary Origin for Nuclear Pore Complexes and Coated Vesicles?
The proto-coatomer hypothesis



A simple coating module containing minimal copies of the two conserved folds evolved in proto-eukaryotes to bend membranes.

The progenitor of the NPC arose from a membrane-coating module that wrapped extensions of an early ER around the cell's chromatin.

Take home slide...



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

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 (Cornell)

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