### Aligning Sequences for Comparative Modeling





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

## Principles of protein structure



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/

#### Typical errors in comparative models

MODEL X-RAY TEMPLATE

Region without a template



Incorrect template



Misalignment



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.

# Minimizing errors in sequence-structure alignment

- Multiple sequence profiles.
- Iterative alignment model building model assessment.

# SALIGN

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

# Reference set



#### http://salilab.org/DBAli

CE alignments from Phil Bourne and Ilya Shindyalov Shindyalov IN, Bourne PE (1998) *Protein Engineering* **11**(9) 739-747.

# SALIGN protocols

**Profile generation** 

- PSI-Blast (PBP)
- Henikoff & Henikoff (HH)
- Henikoff & Henikoff + Similarity (HS)
- Henikoff & Henikoff substitution matrix (MAT)

#### Profile comparison

- Correlation coefficient (CC)
- Euclidean distance (ED)
- Dot product (DP)
- Jensen-Shannon distance (JS)
- Average value (Ave)

# **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
ССнн	56 ± 23	0.61 ± 0.24
ССнз	56 ± 24	0.62 ± 0.24
ТОР	62 ± 20	0.67 ± 0.20





## Alignment accuracy (CE overlap)

200 pairwise DBAli alignments



# 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

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



### Genetic algorithm operators



### Composite model assessment score

Weighted linear combination of several scores:

- Pair (P<sub>p</sub>) and surface (P<sub>s</sub>) statistical potentials;
- Structural compactness (S<sub>C</sub>);
- Harmonic average distance score (H<sub>a</sub>);
- Alignment score (A<sub>S</sub>).

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

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

- $\mu$  ... average score of all models
- $\sigma \ldots$  standard deviation of the scores

# Application to a difficult modeling case 1BOV-1LTS



Sequence identity 4.4%

Initial model Ca RMSD 10.1Å

Final model Ca 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 [%]	0		Initial prediction		Final prediction		Best prediction	
	Coverage [% aa]	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%

examples...

# 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



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

CONFIDENTIAL



#### BRACAnalysis <sup>™</sup> Comprehensive BRCA1-BRCA2 Gene Sequence Analysis Result



#### 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



These testresults 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 sering that includes appropriate counseling. The accompanying Technical Specifications summary describes the analysis, method, performance characteristics, nomencicity, early interpretive optimic of this test. This test may be considered investigational by some states. This test was developed and its performance characteristics determined by Mynad Genetic Laboratores. It has not been reviewed by the U.S. Food and Drug Administration. The FDA has determined that such Genarace or approval is not necessary.

#### Missense mutations in BRCT domains by function



05/10/2004



#### Putative binding site on BRCA1



Williams *et al.* 2004 Nature Structure Biology. **June 2004 11**:519 Mirkovic *et al.* 2004 Cancer Research. **June 2004 64**:3790





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

What is the physiological ligand of Brain Lipid-Binding Protein? Predicting features of a model that are not present in the template



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.

## S. cerevisiae ribosome



Fitting of comparative models into 15Å cryoelectron 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.

#### Utility of protein structure models, despite errors



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

#### 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

## Acknowledgments

http://salilab.org

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)

> Chimera P. Babbitt T. Ferrin

> > NIH

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

#### Structural Genomix

#### 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

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

> BRCA1 A. Monteiro (Cornel)

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

#### Structural Genomics

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

> Mast Cell Proteases Rick Stevens (BWH)