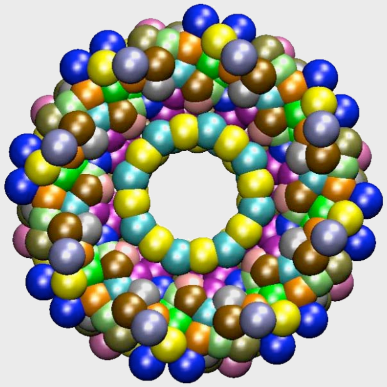
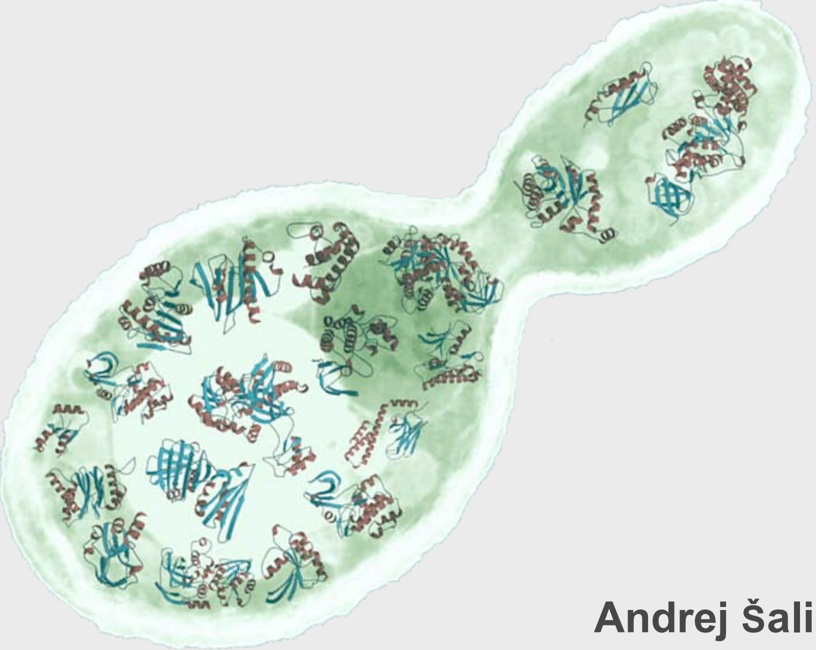
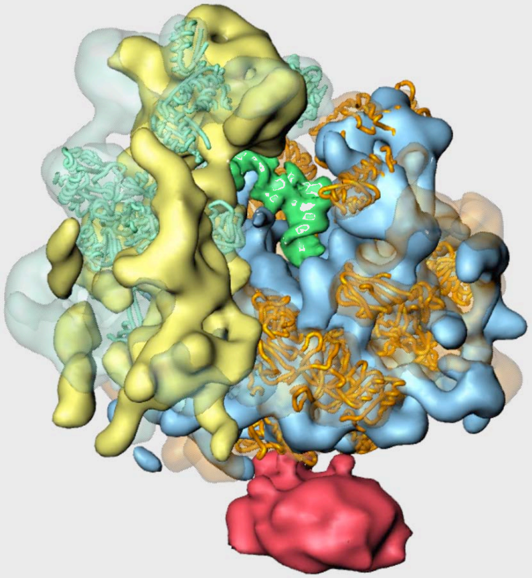
Modeling and Determining the Structures of Proteins and Macromolecular Assemblies







Andrej Šali

http://salilab.org/

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

UCSF

Structure characterization of macromolecular assemblies

1. Approach: integrated hierarchical system for structural biology.

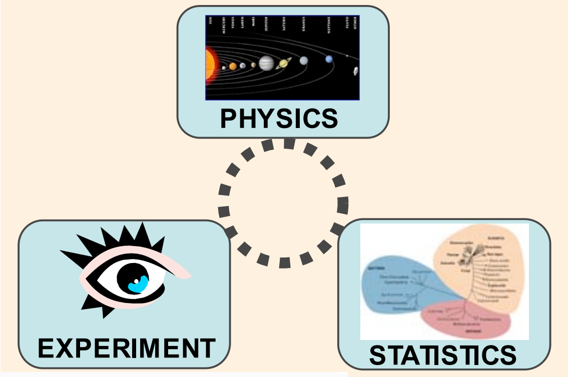
2. Medium resolution: by EM & comparative modeling.

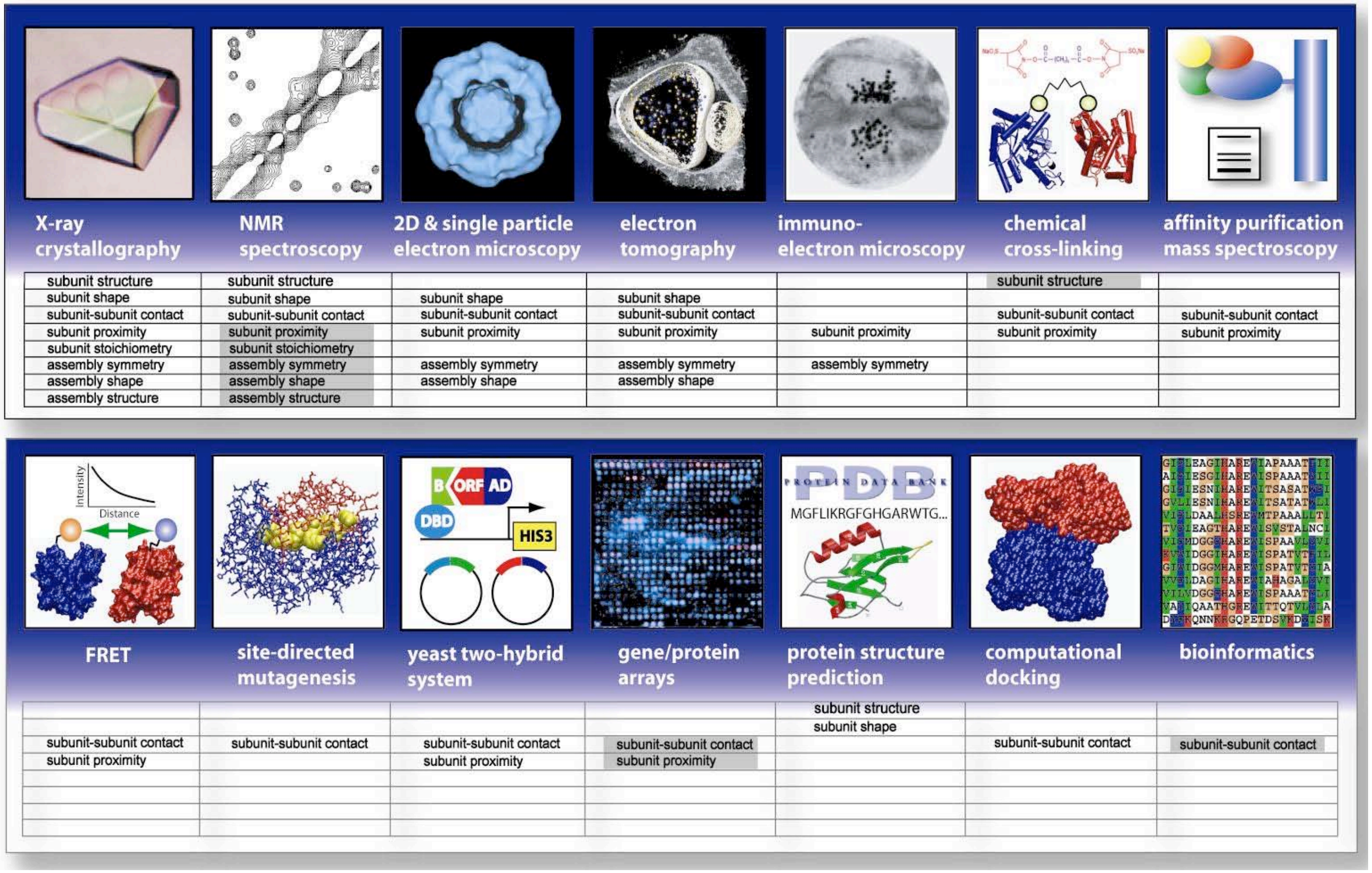
3. Low resolution: from "biochemical" information.

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.

Maximize efficiency, accuracy, resolution, and completeness
of the structural coverage of protein assemblies.





X-ray crystallography	NMR spectroscopy	2D & single particle electron microscopy	electron tomography	immuno-electron microscopy	chemical cross-linking	affinity purification mass spectroscopy
subunit structure	subunit structure				subunit structure	
subunit shape	subunit shape	subunit shape	subunit shape			
subunit-subunit contact	subunit-subunit contact	subunit-subunit contact	subunit-subunit contact		subunit-subunit contact	subunit-subunit contact
subunit proximity	subunit proximity	subunit proximity	subunit proximity	subunit proximity	subunit proximity	subunit proximity
subunit stoichiometry	subunit stoichiometry					
assembly symmetry	assembly symmetry	assembly symmetry	assembly symmetry	assembly symmetry		
assembly shape	assembly shape	assembly shape	assembly shape			
assembly structure	assembly structure					

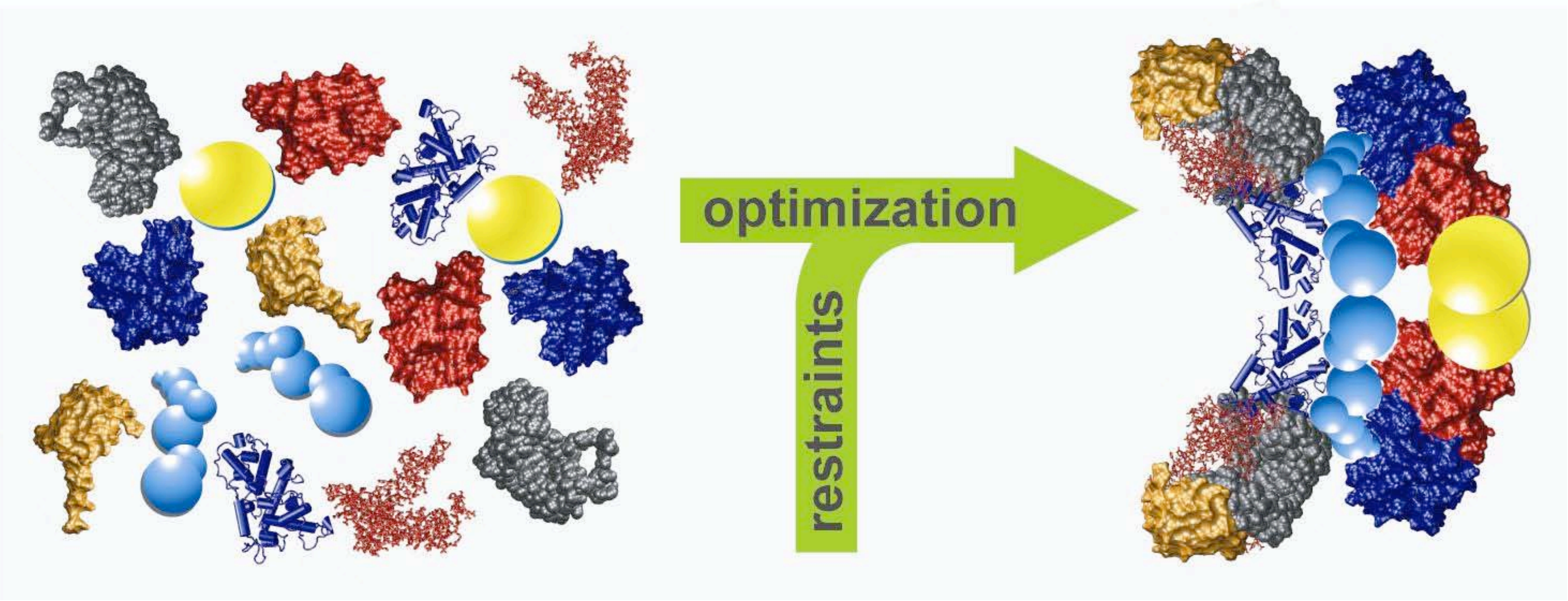
FRET	site-directed mutagenesis	yeast two-hybrid system	gene/protein arrays	protein structure prediction	computational docking	bioinformatics
				subunit structure		
				subunit shape		
subunit-subunit contact	subunit-subunit contact	subunit-subunit contact	subunit-subunit contact		subunit-subunit contact	subunit-subunit contact
subunit proximity		subunit proximity	subunit proximity			

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

Characterizing Macromolecular Assemblies by Satisfaction of Spatial Restraints

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

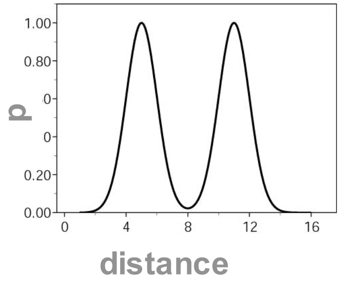
There is nothing but points and restraints on them.



5/17/05

Scoring Function

There is nothing but points and restraints on them.



$$P\,(R\,/\,I) \;=\; \prod_i p_i\,(r_i\,/\,I_i)$$

R … all degrees of freedom
I … all information
r_i … i^{th} restrained feature (*eg*, distance, angle, proximity, surface, density)
I_i … information about i^{th} restrained feature

http://salilab.org/modeller/

Sali, Blundell. *J. Mol. Biol.* 234, 779, 1993.
Alber, Kim, Sali. *Structure* 13, 435, 2005.

Challenges at the frontiers of structural biology

Andrej Šali and John Kuriyan

TIBS Millenium Issue, M20-M24, 1999.

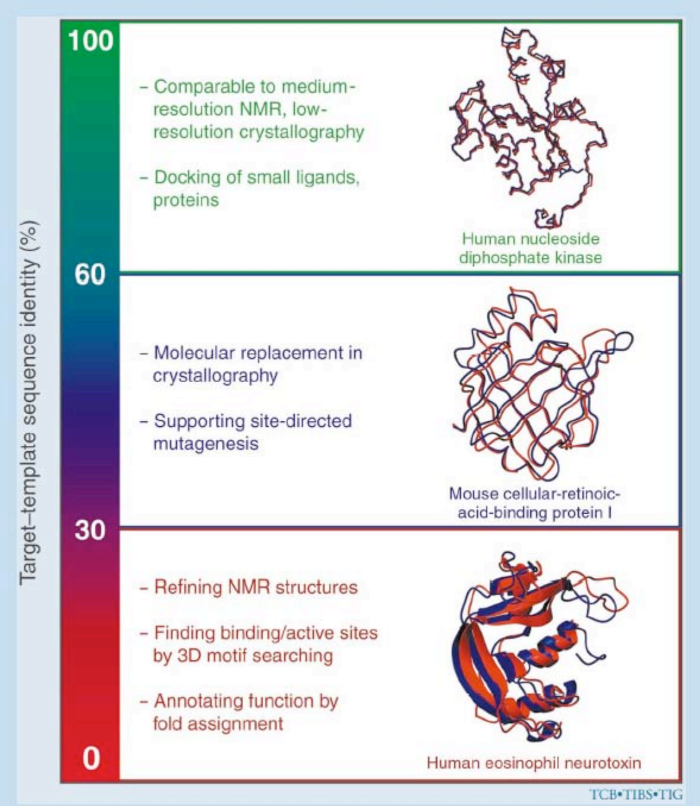

Target–template sequence identity (%)
100
60
30
0
– Comparable to medium-resolution NMR, low-resolution crystallography
– Docking of small ligands, proteins
Human nucleoside diphosphate kinase
– Molecular replacement in crystallography
– Supporting site-directed mutagenesis
Mouse cellular-retinoic-acid-binding protein I
– Refining NMR structures
– Finding binding/active sites by 3D motif searching
– Annotating function by fold assignment
Human eosinophil neurotoxin
TCB•TIBS•TIG

FIGURE 1. Schematic diagram showing the range of accuracy obtained by comparative modelling[23]. The potential uses of comparative models depend on their accuracy. This in turn depends significantly on the sequence identity between the sequence modelled and the known structure on which the model was based. Sample models (red) are compared with the actual structures (blue).

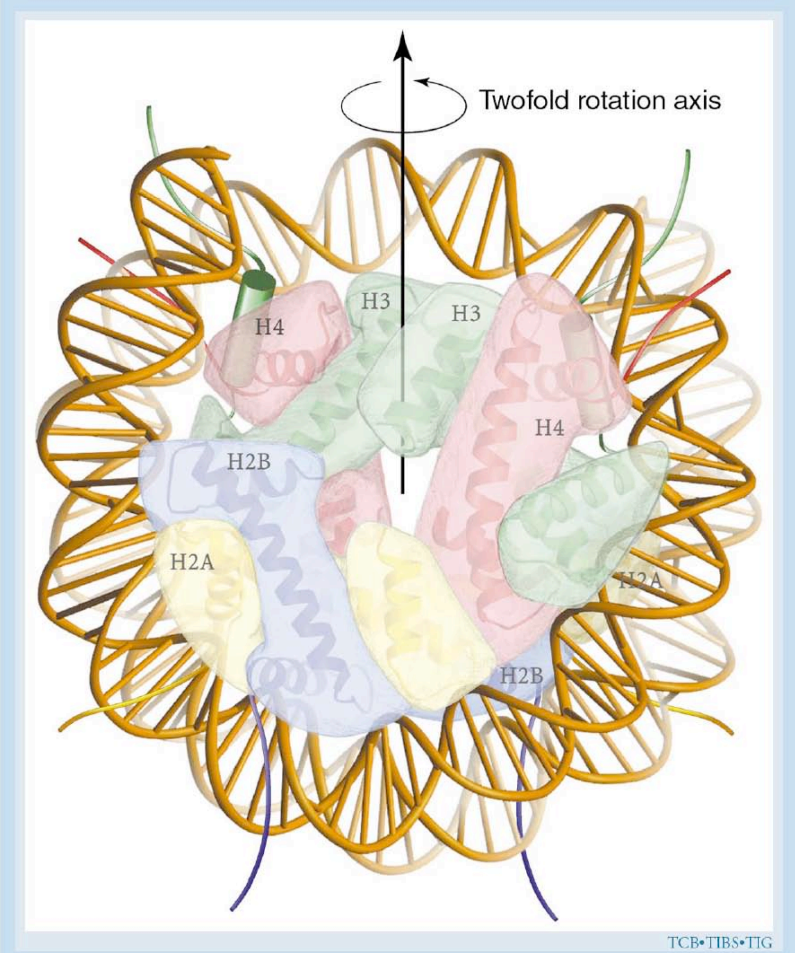
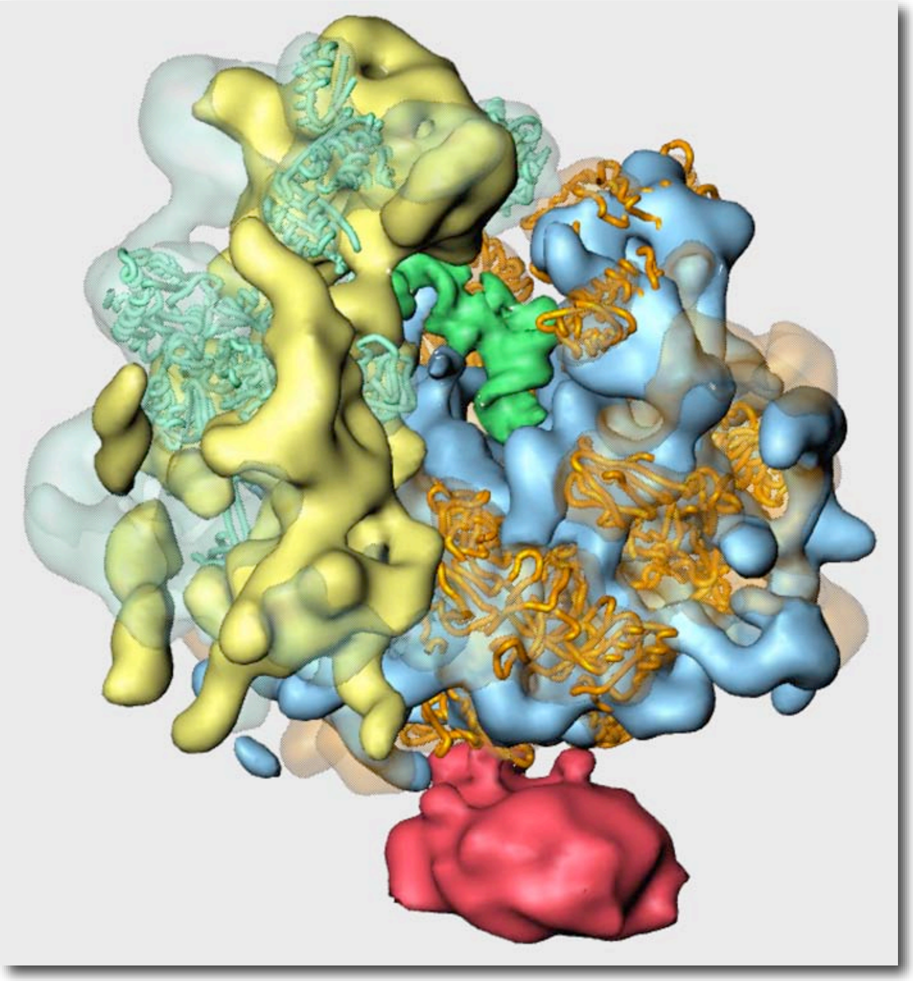


FIGURE 3. The structure of the nucleosome core, as determined by Richmond and colleagues[41]. The histone proteins form a spiral-shaped octameric assembly around which DNA is coiled. The histone octamer consists of two copies each of four different histone proteins – H2A, H2B, H3 and H4. These proteins contain tails that are shown protruding from the nucleosome. The tails are likely to be important in stabilizing the arrangement of nucleosomes in higher-order structures. Copyright 1999, Lore Leighton, used with permission.

S. cerevisiae ribosome



Fitting of comparative models into 15Å cryoEM 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%.

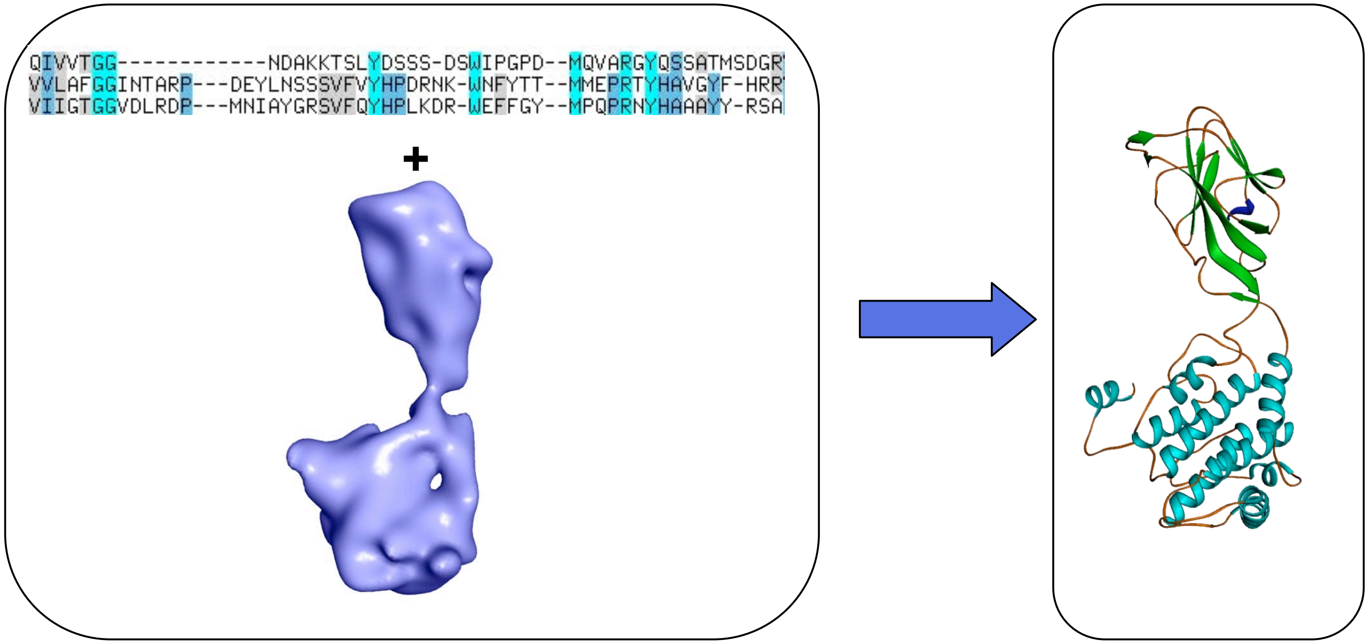
Architecture of the protein-conducting channel associated with the translating 80S Ribosome

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

Comparative modeling and fitting into EM density

Maya Topf, Frank Alber, Matt Baker, **Wah Chiu**

Improve comparative modeling by fitting models into the target EM density map;
Improve fitting into an EM density map by simultaneous model building.



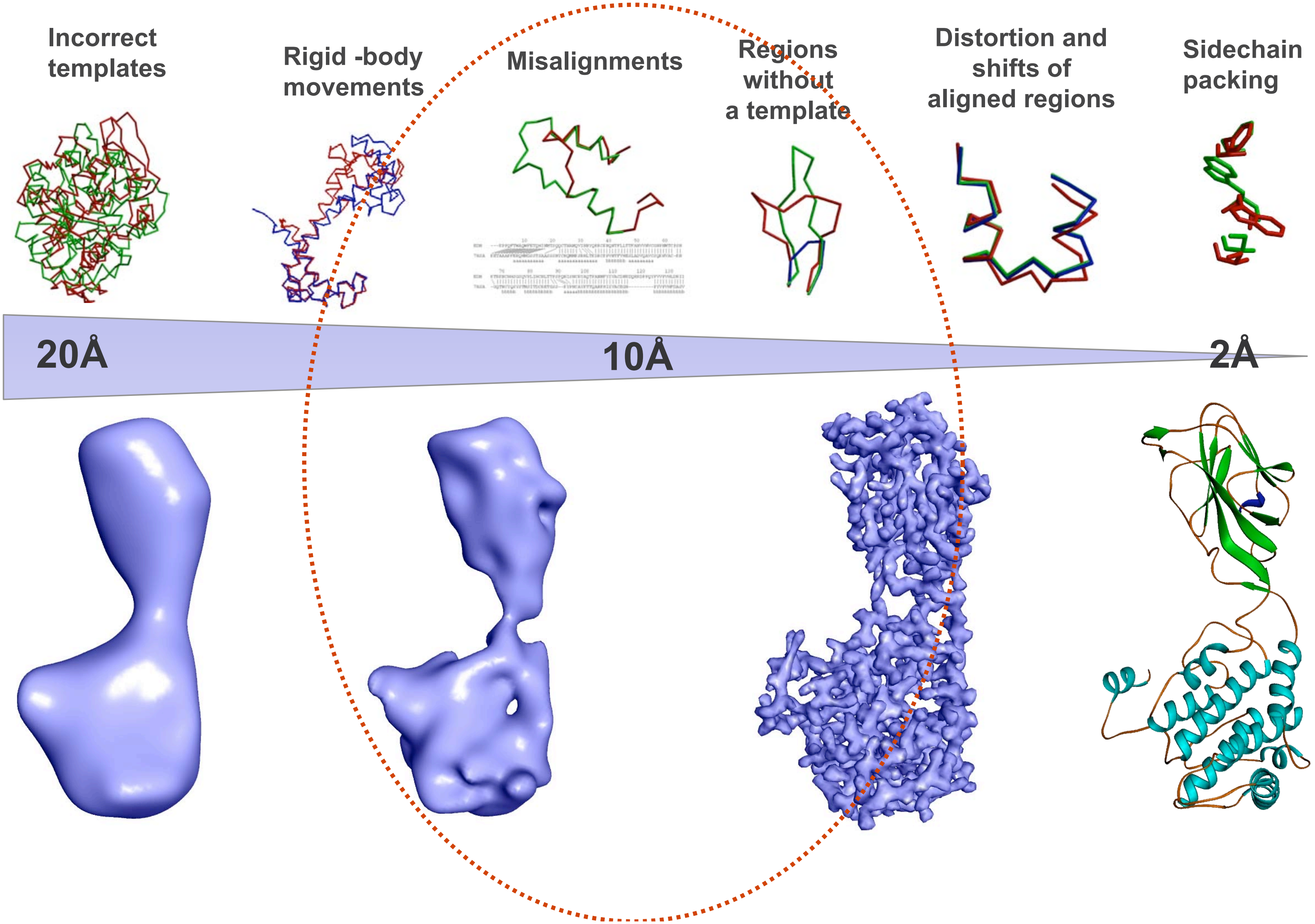
```
QIVVTGG------------NDAKKTSLYDSSS-DSWIPGPD--MQVARGYQSSATMSDGR
VVLAFGGINTARP---DEYLNSSSVFVYHPDRNK-WNFYTT--MMEPRTYHAVGYF-HRR
VIIGTGGVDLRDP---MNIAYGRSVFQYHPLKDR-WEFFGY--MPQPRNYHAAAYY-RSA
```

Motivation:

- Number of known structures in PDB: ~30,000
- Number of known sequences modeled by CM: ~850,000
  (Pieper et al., NAR 2004).

5/17/05

Errors in comparative models *vs.* resolution



Incorrect templates

Rigid -body movements

Misalignments

Regions without a template

Distortion and shifts of aligned regions

Sidechain packing

20Å

10Å

2Å

5/17/05

Fitting a model into an EM map (Mod-EM)

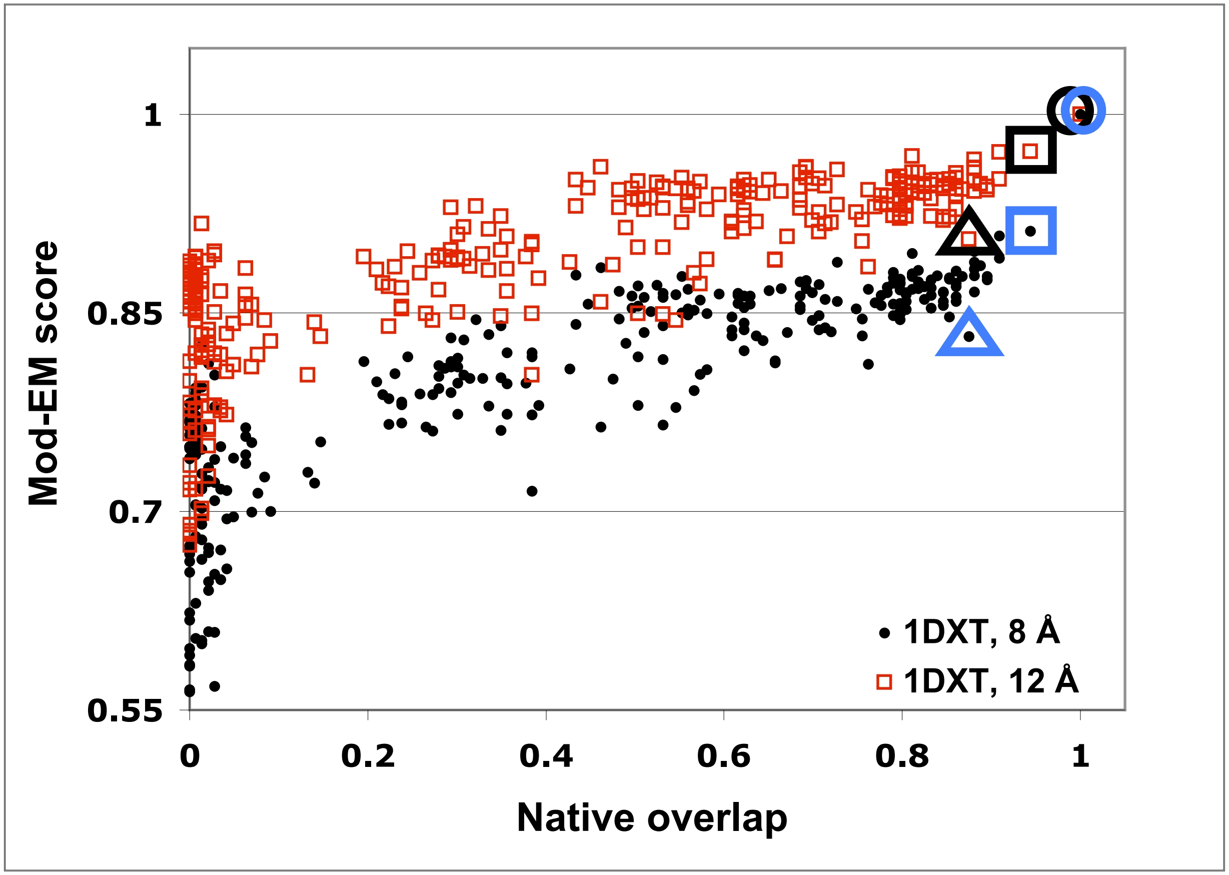
Developed a rigid body fitting procedure in MODELLER, MOD-EM, that optimizes a correlation coefficient between the map and a given model using a combination of grid search and Monte Carlo procedures.

Prepared a benchmark of 300 comparative models of varying accuracy covering the whole range of sequence-structure alignment accuracy for each of 20 test structures.

Tested how well is the best model selected by the quality of its fit into a given density map, as a function of resolution and noise.

Topf, Baker, John, Chiu, Sali. J. Str. Biol. **149**, 191-203, 2005.

Correlation between model accuracy and quality of a fit into density



R^2=0.6-0.7

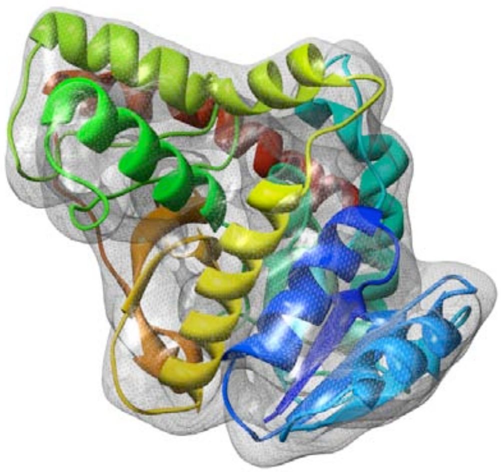
Native (1dxt, circle): 1

Best model (square): 2

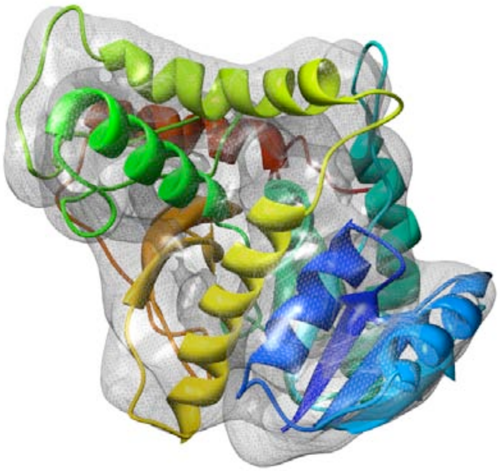
Template (1hbg, triangle): 132(8Å), 139(12Å)

10Å map
2cmd - 6ldh
310 aa

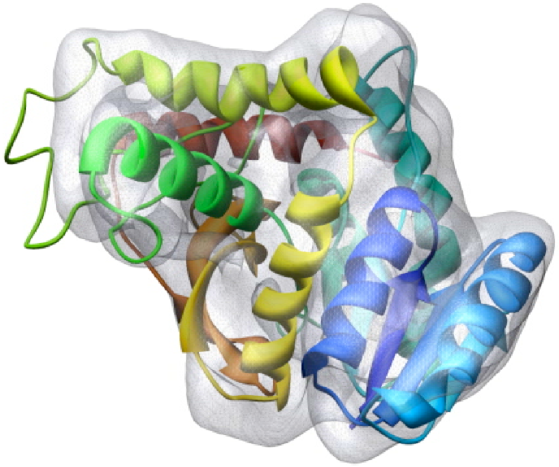
Native structure



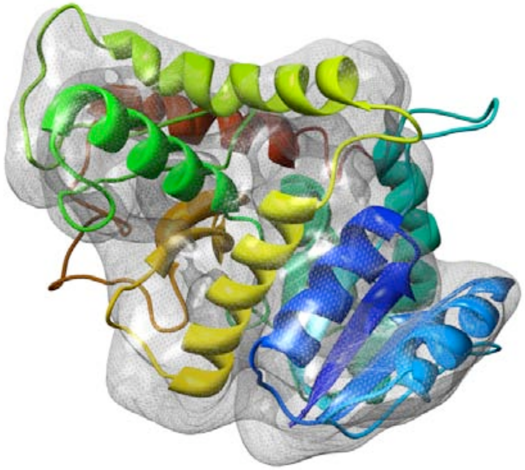
Most accurate model,
Best-fitting model (rank 1)



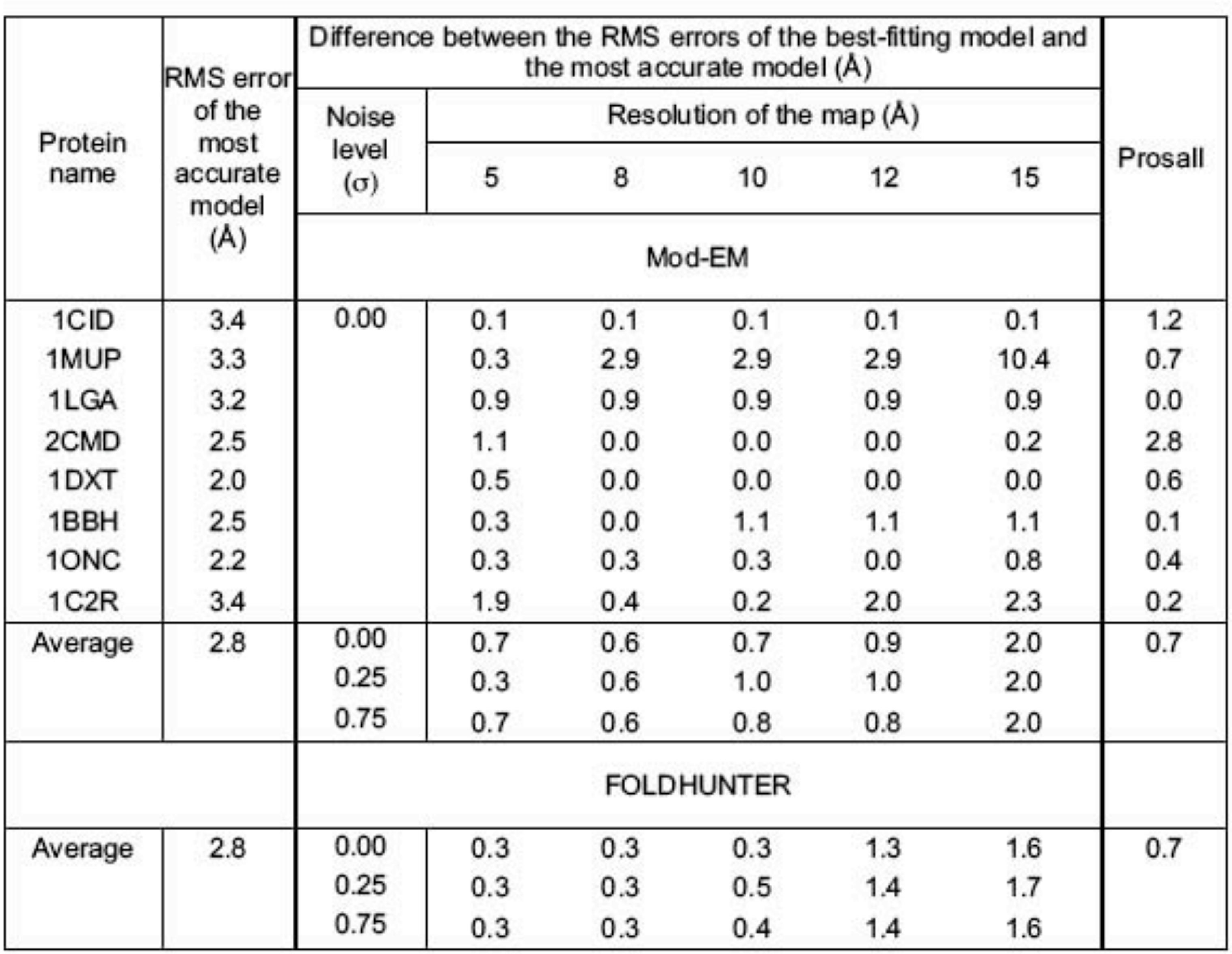
Template
(rank 5)



Best ProsaII model
(rank 256)



Quality of the best-fitting model

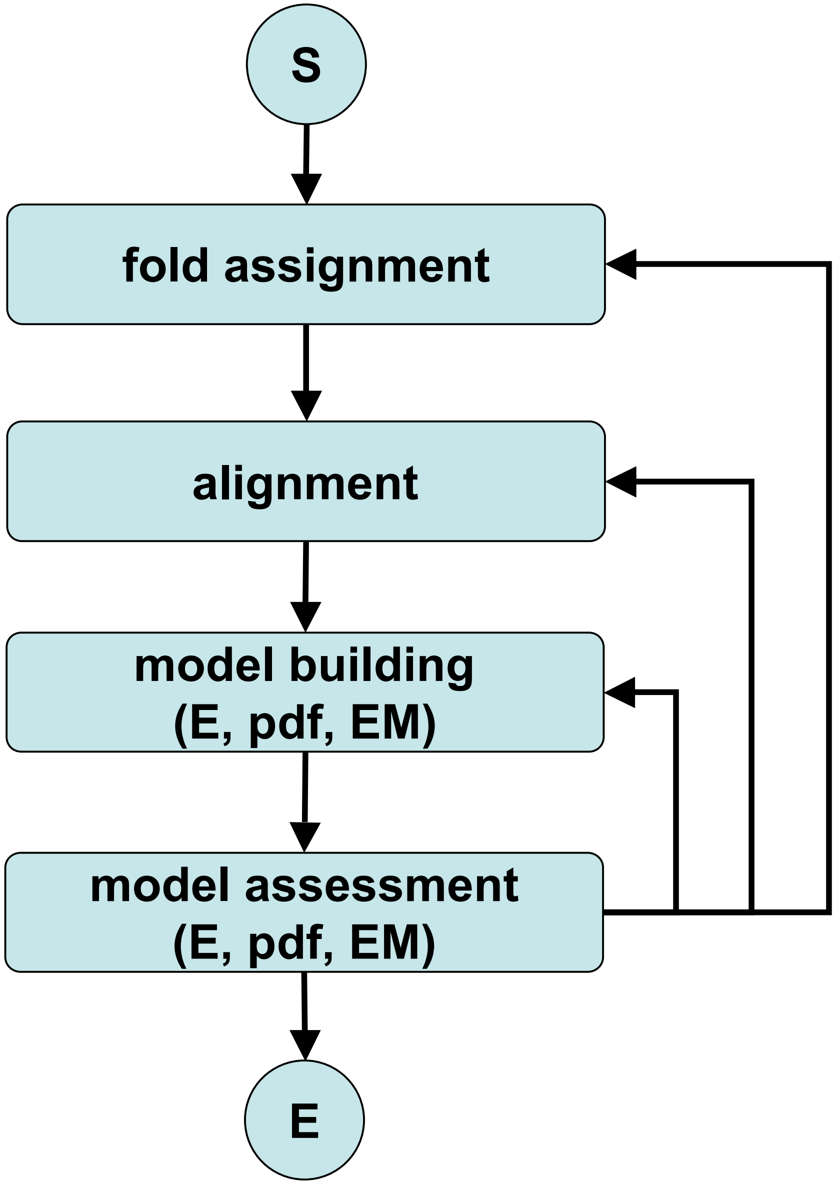


Protein name	RMS error of the most accurate model (Å)	Difference between the RMS errors of the best-fitting model and the most accurate model (Å)						Prosall
		Noise level (σ)	Resolution of the map (Å)					
			5	8	10	12	15	
		Mod-EM						
1CID	3.4	0.00	0.1	0.1	0.1	0.1	0.1	1.2
1MUP	3.3		0.3	2.9	2.9	2.9	10.4	0.7
1LGA	3.2		0.9	0.9	0.9	0.9	0.9	0.0
2CMD	2.5		1.1	0.0	0.0	0.0	0.2	2.8
1DXT	2.0		0.5	0.0	0.0	0.0	0.0	0.6
1BBH	2.5		0.3	0.0	1.1	1.1	1.1	0.1
1ONC	2.2		0.3	0.3	0.3	0.0	0.8	0.4
1C2R	3.4		1.9	0.4	0.2	2.0	2.3	0.2
Average	2.8	0.00	0.7	0.6	0.7	0.9	2.0	0.7
		0.25	0.3	0.6	1.0	1.0	2.0	
		0.75	0.7	0.6	0.8	0.8	2.0	
		FOLDHUNTER						
Average	2.8	0.00	0.3	0.3	0.3	1.3	1.6	0.7
		0.25	0.3	0.3	0.5	1.4	1.7	
		0.75	0.3	0.3	0.4	1.4	1.6	

Conclusions (CM & EM)

- EM density maps at 5-15 Å resolution contain information that can be exploited in comparative modeling, both for improving sequence-structure alignment and for model building.
- Fitting comparative models instead of template structures into EM maps can make a large difference in the accuracy of the final hybrid atomic models.
- Scope: ~60 times more sequences can be modeled than have been determined by crystallography or NMR spectroscopy, and most of them are modeled on less than 30% sequence identity to the closest known structure.

Combined comparative modeling and fitting



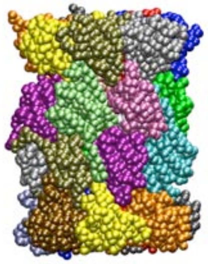
5/17/05

Very Low-Resolution Modeling of Large Assemblies

Many times the structures of some subunits are not available.

In such cases, we can only model the **configuration** of the subunits in the complex.

atoms



residues

proteins

The Yeast Nuclear Pore Complex

1. Structure
2. Evolution
3. Mechanism of assembly
4. Mechanism of action

Frank Alber, Damien Devos
UCSF

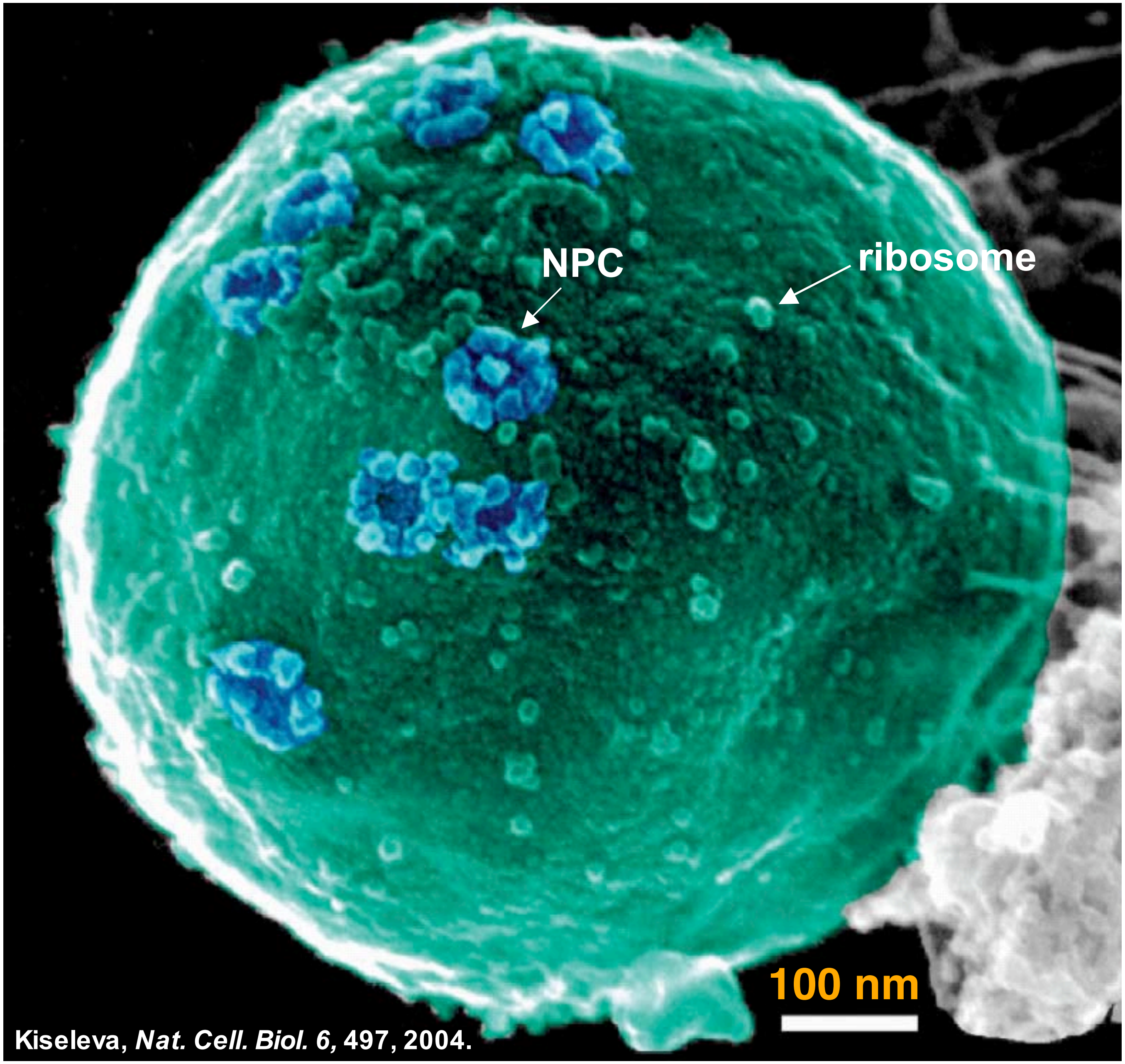
Jasmine Zhou
University of Southern California

Mike Rout
Tari Suprapto, Julia Kipper, Liesbeth Veenhoff, Svetlana Dokudovskaya

Brian Chait
Wenzhu Zhang
The Rockefeller University, New York

5/17/05

Nuclear Pore Complex (NPC)



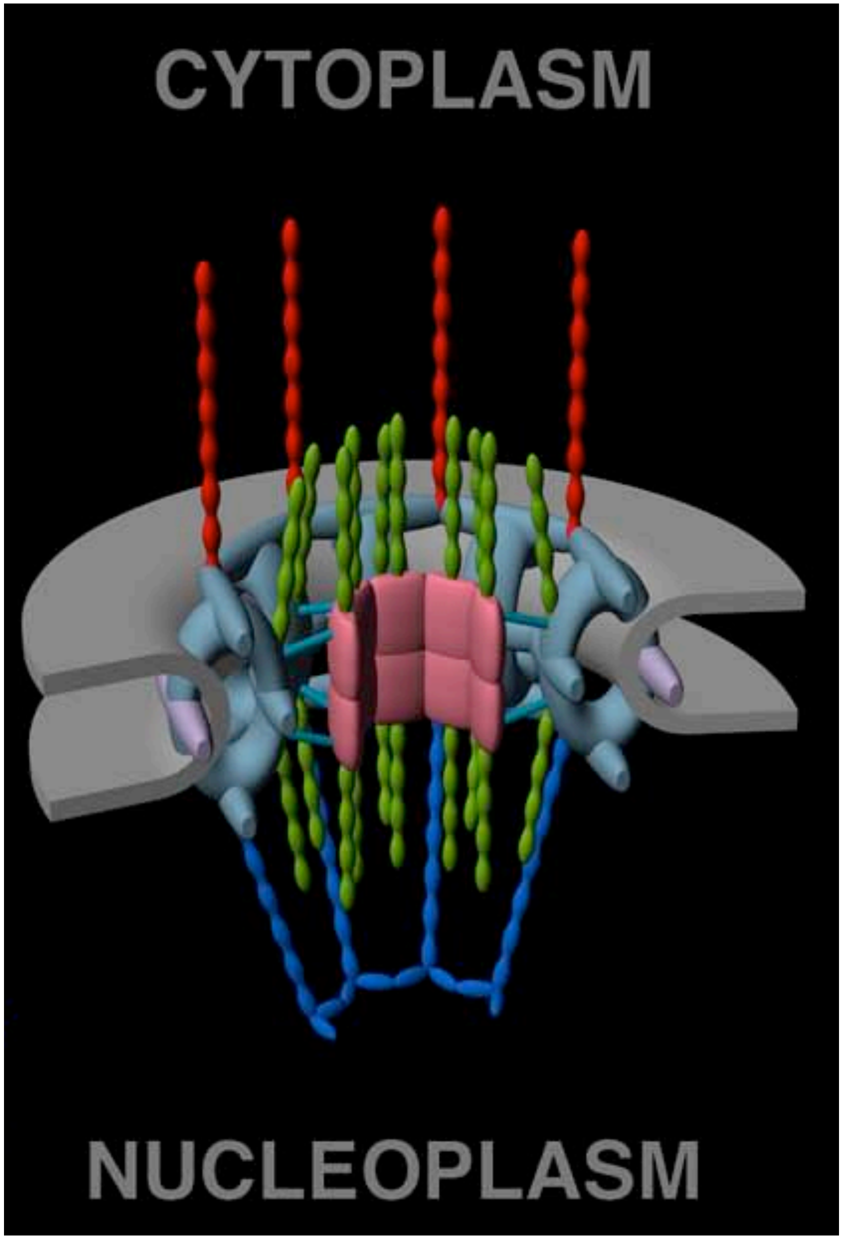
Kiseleva, *Nat. Cell. Biol. 6*, 497, 2004.

Consists of broadly conserved nucleoporins (nups).

50 MDa complex: ~480 proteins of 30 different types.

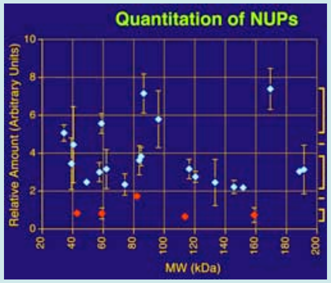
Mediates all known nuclear transport, *via* cognate transport factors.

NPC

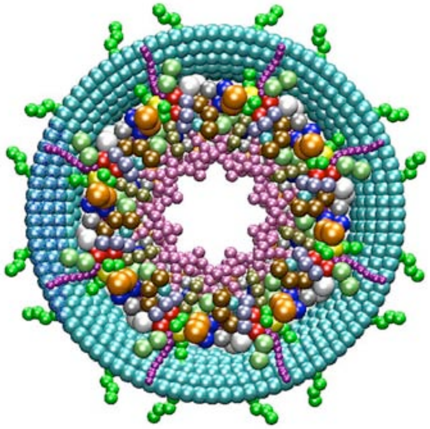


5/17/05

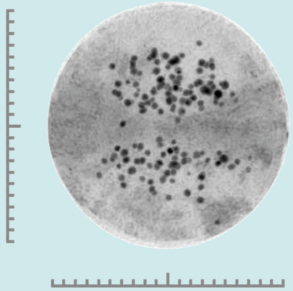
Use All Spatial Information

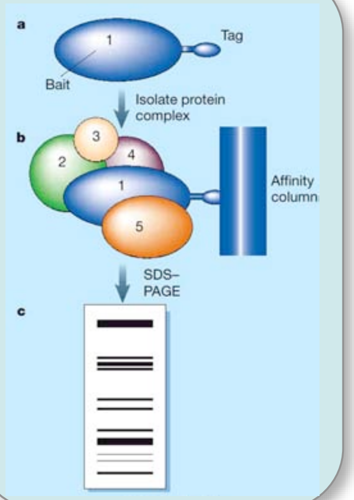


NUP Stoichiometry



NUP Localization

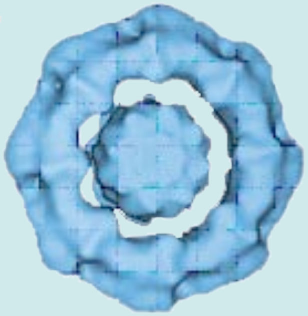


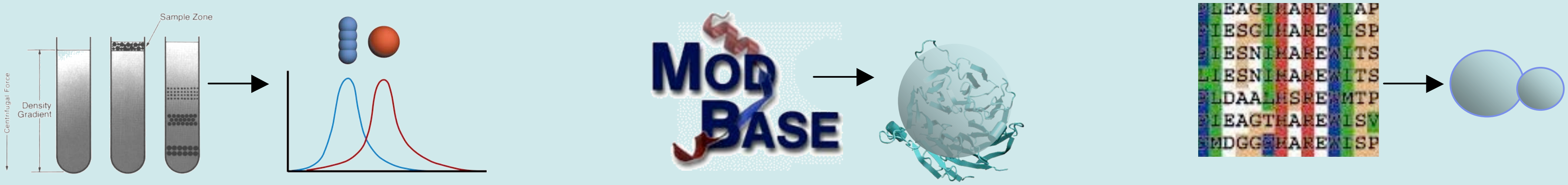


NUP- NUP Interactions

NUP Shape

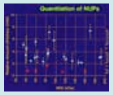
Symmetry Global shape



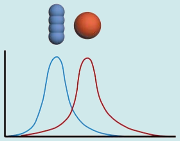


5/17/05

All Spatial Restraints on the NPC

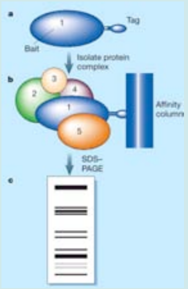


Stochiometry:
30 proteins, 456 copies in total



Protein (and subcomplex) shape from Stokes radii:
1,680 intra protein distance restraints and 5,776 lower bound distance restraints

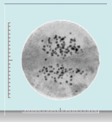
Excluded volume of proteins:
~456^2/2 distance lower bounds



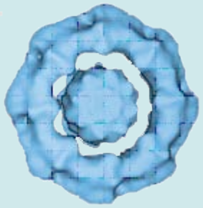
Protein-protein proximity: (immuno-purification)
5,472 upper distance bounds

Subcomplex connectivity: (immuno-purification)
3,344 binary restraints

Binary protein-protein contacts: from “overlay” experiments
208 binary restraints



Radial and axial localization of proteins: (IEM)
916 absolute positional restraints and 1,813 upper and lower distance restraints



Symmetry considerations: (cryo-EM)
~100,000 symmetry distance and ~100 symmetry dihedral angle restraints and 5,596 angle restraints

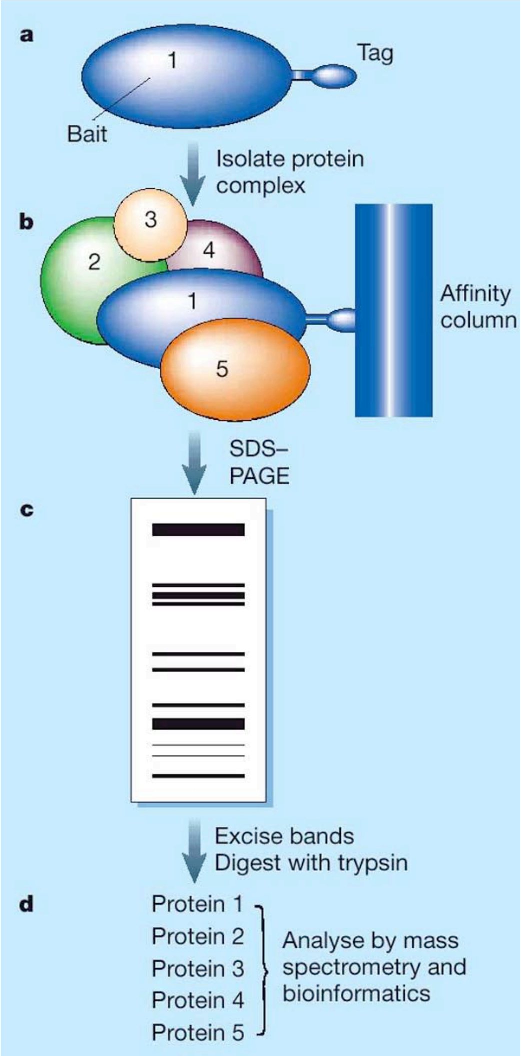
Modeling in the context of the nuclear envelope: NE shape and dimension (EM)
876 membrane particles

Membrane spanning protein regions:
48 surface restraints, 112 volume restraints

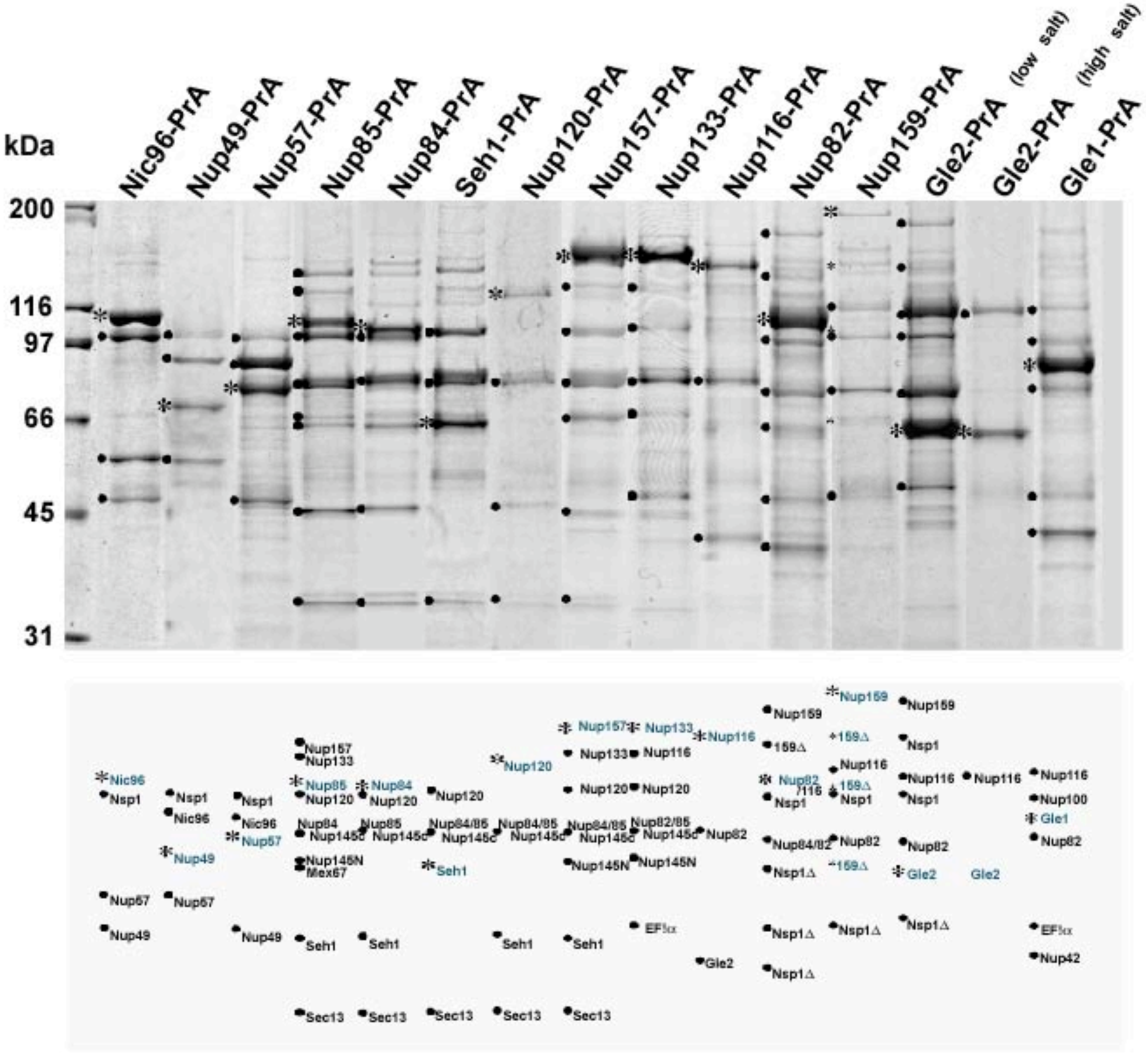
Luminal Pom152 ring: (EM)
16 binary restraints

Tagging, Immunopurification and Analysis of Nucleoporin Subcomplexes

nucleoporin	protein A tag



- **several *hundred* pullouts**
- **~1,300 protein bands identified by MS**

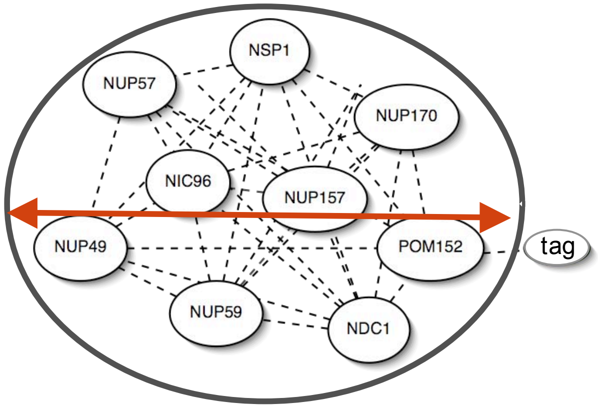


5/17/05

Structural Information from Pullouts

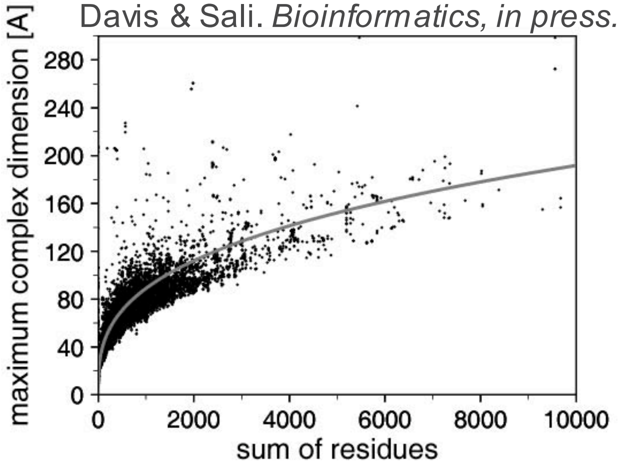
Subcomplex Proximity restraint

upper distance bound between all subunit beads in a pullout



derived from assemblies in PIBASE*

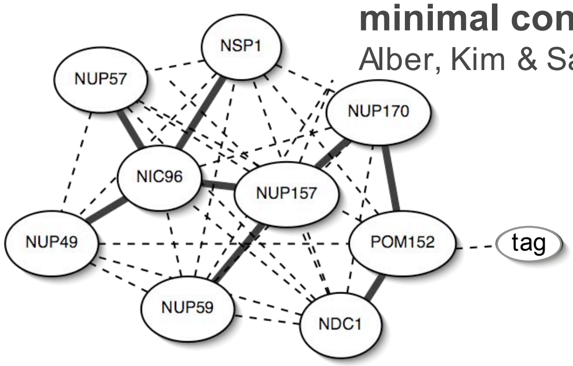
Davis & Sali. *Bioinformatics, in press.*



Subcomplex Connectivity restraint**

minimal connectivity between all subunits in a pullout

Alber, Kim & Sali. *Structure* 13, 435, 2005



5/17/05

Optimization

- Start with a random configuration of protein centers.
- Minimize violations of input restraints by conjugate gradients and molecular dynamics with simulated annealing.
- Obtain an “ensemble” of many independently calculated models (~300,000).

Membrane spanning proteins:

Pom152 Pom34
Ndc1

FG repeat proteins:

Nup159
Nsp1
Nup1
Nup100
Nup116
Nup145N
Nup60
Nup59
Nup57
Nup53
Nup49
Nup42

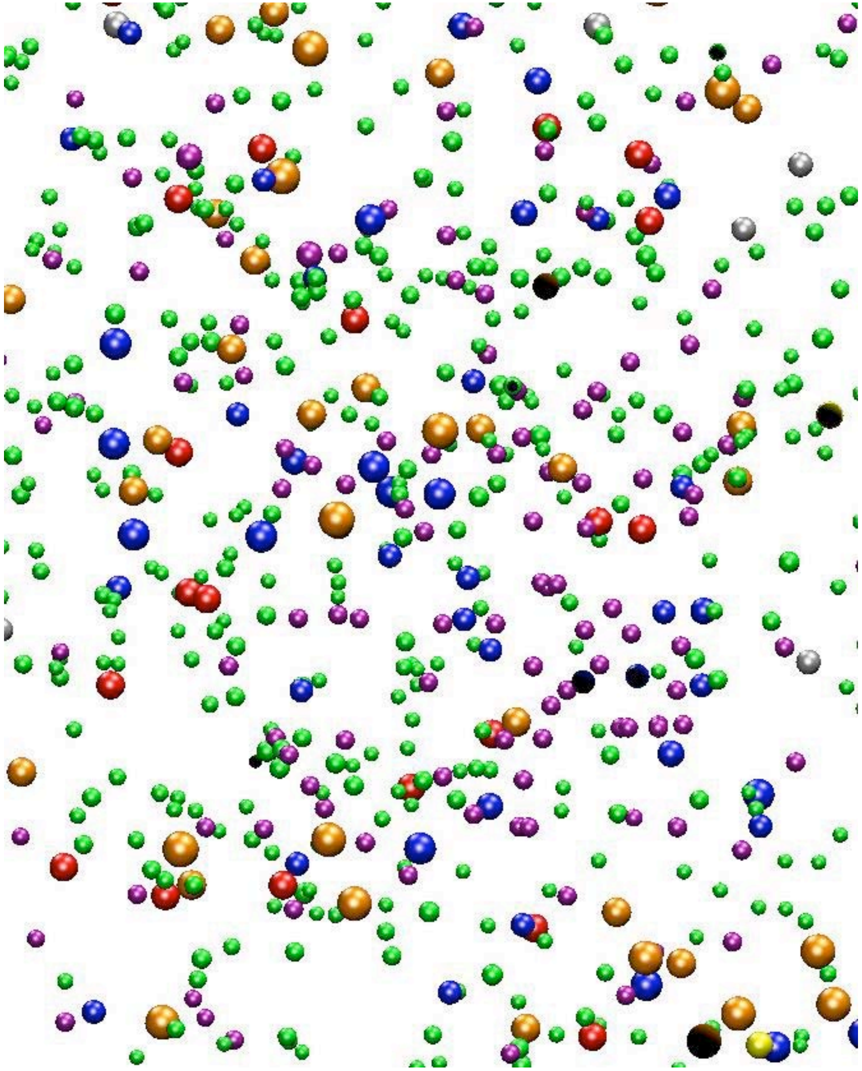
Nup84 complex:

Nup84
Nup85
Nup120
Nup133
Seh1
Sec13
Nup145C

Large Core proteins:

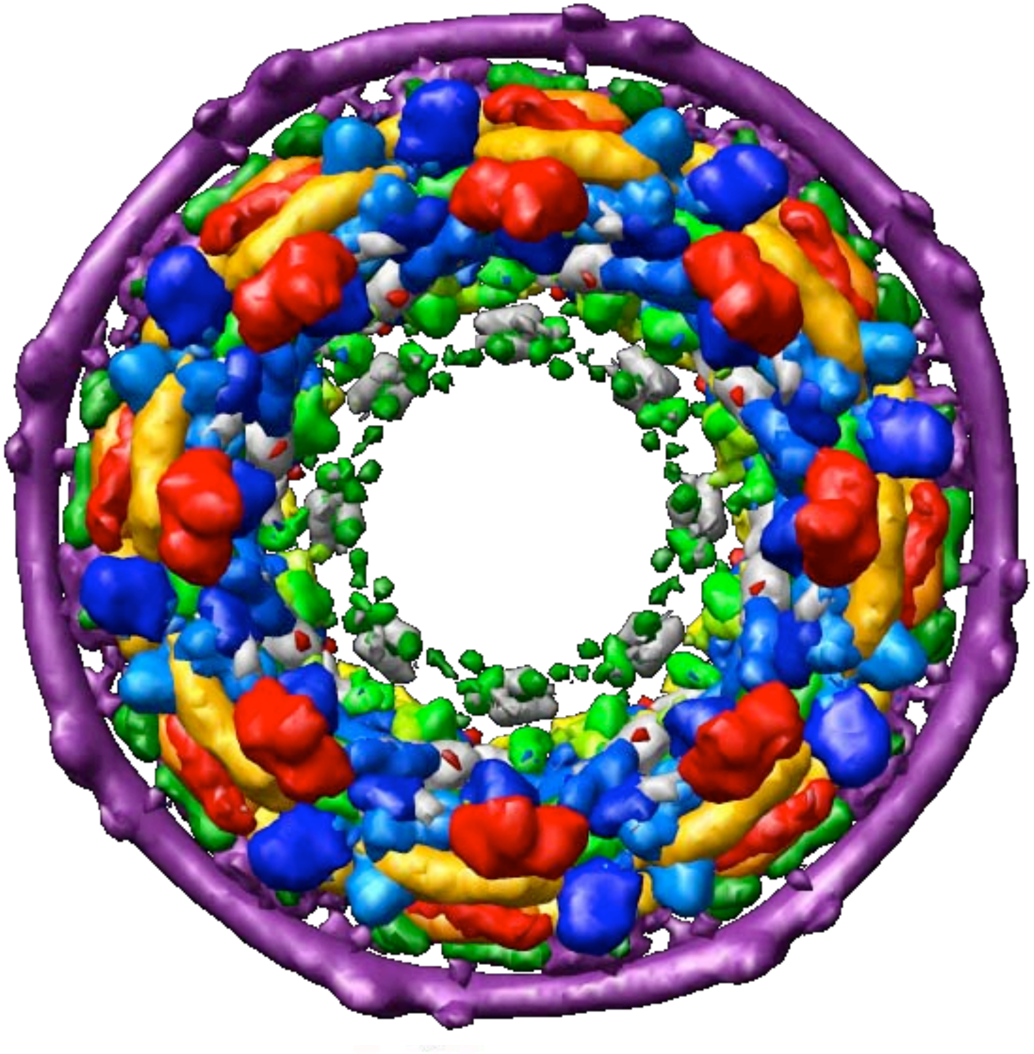
Nup192
Nup188
Nup170
Nup157

Nup82
Nic96



Protein Localization Probability

Calculated from the structural superposition of the ensemble of models that satisfy all input restraints

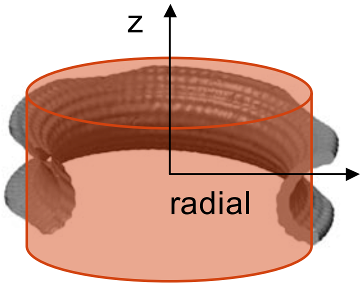


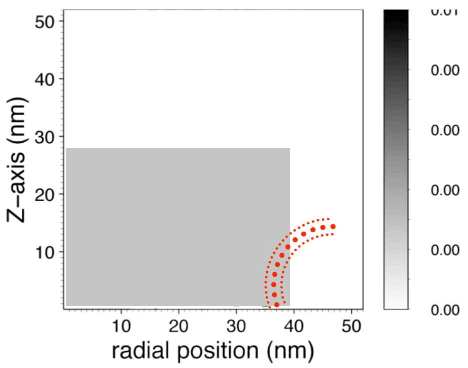
Protein Localization Probability

There is enough information to localize most nups

Nup188:

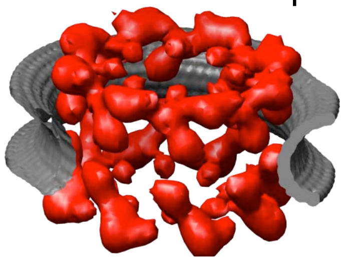
Immuno-EM

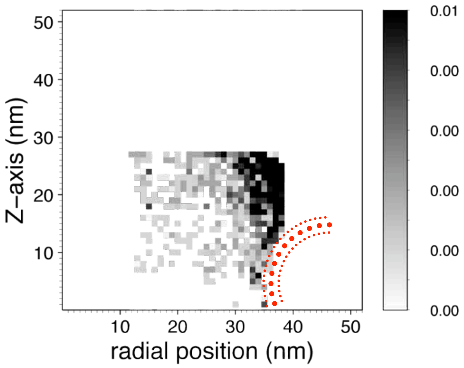




H: 10.01

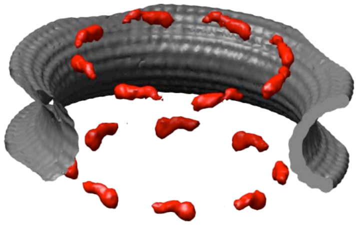
Immuno-EM
Stochiometry
Excluded volume
Symmetry
Nuclear Envelope

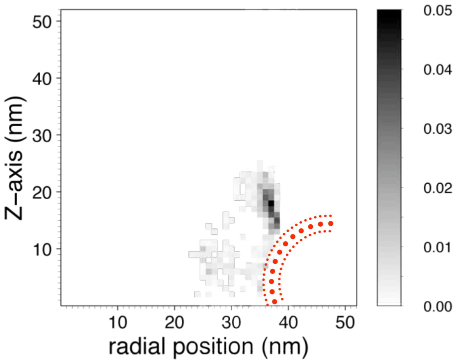




7.8

Immuno-EM
Stochiometry
Excluded volume
Symmetry
Pullouts





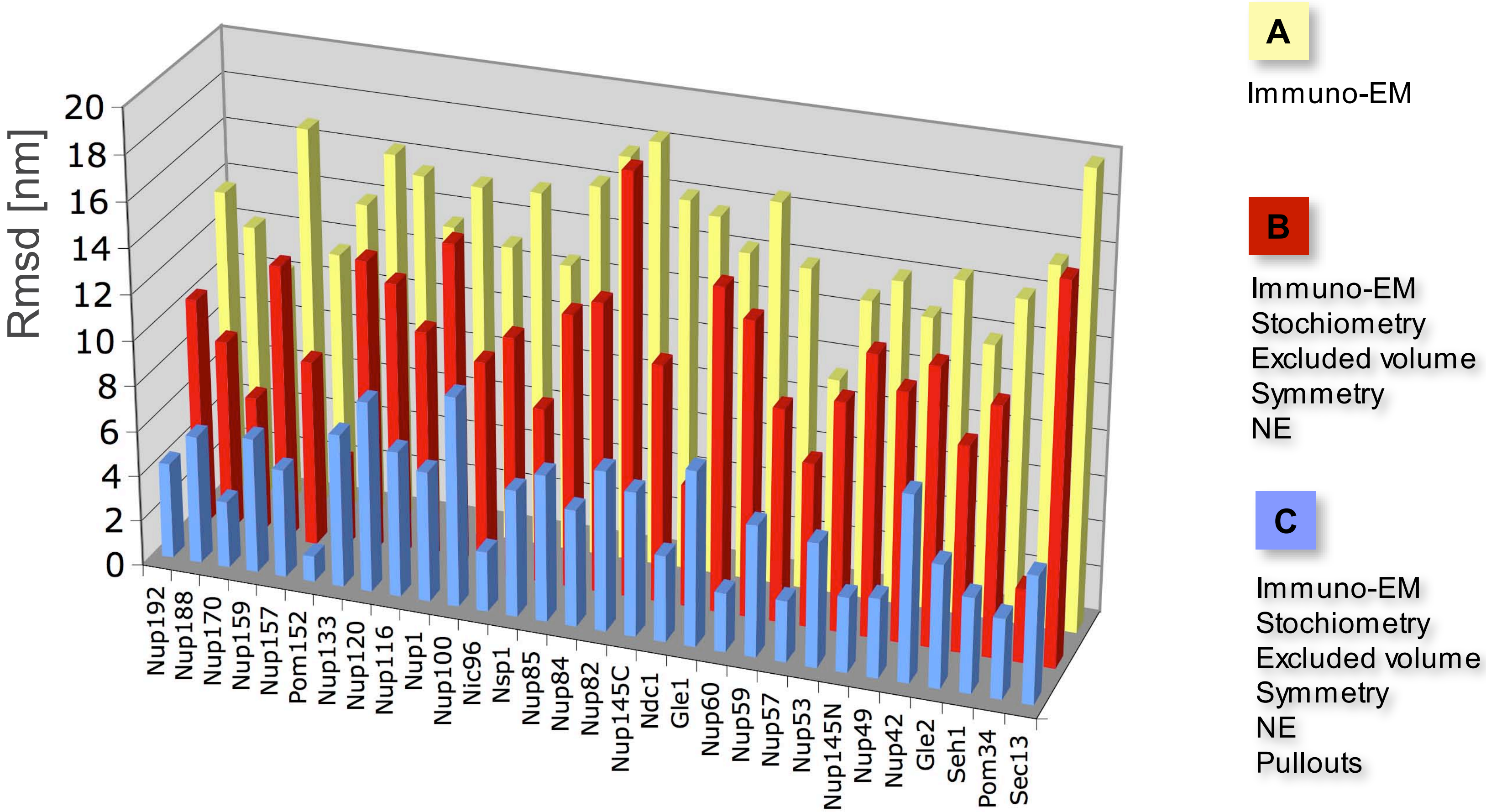
4.5

$H = -\Sigma_i \, p_i \log_2 p_i$

5/17/05

Average Mean Displacement of each Protein

There is enough information to localize most nups



Assessing the Well Scoring Models

1. How similar are the models to each other?

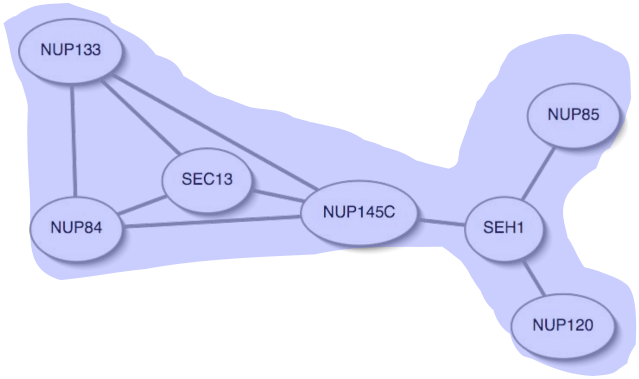
2. Do the models make sense given other data?

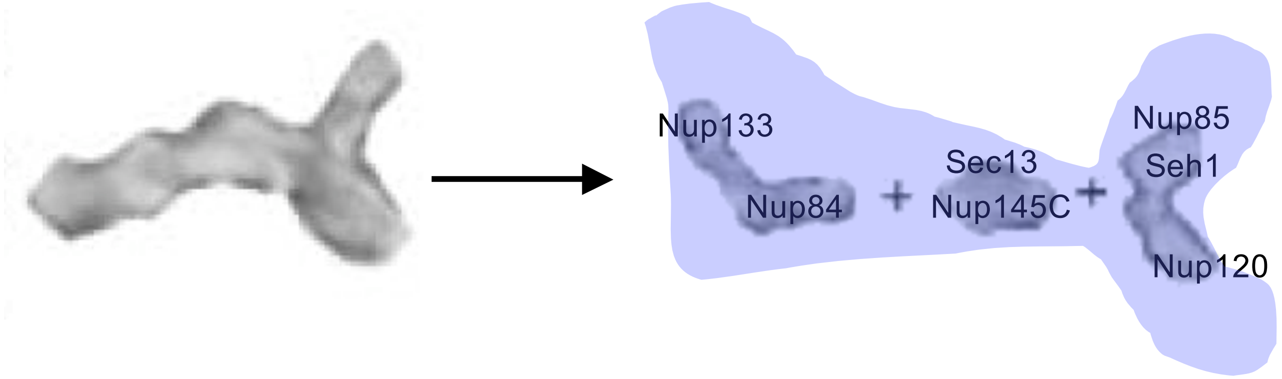
3. Using simple models as benchmarks.

   Alber, Kim, Sali. *Structure* 13, 435, 2005.

Nup84 Complex Topology

Consistent with experimental data (not included in the calculations)

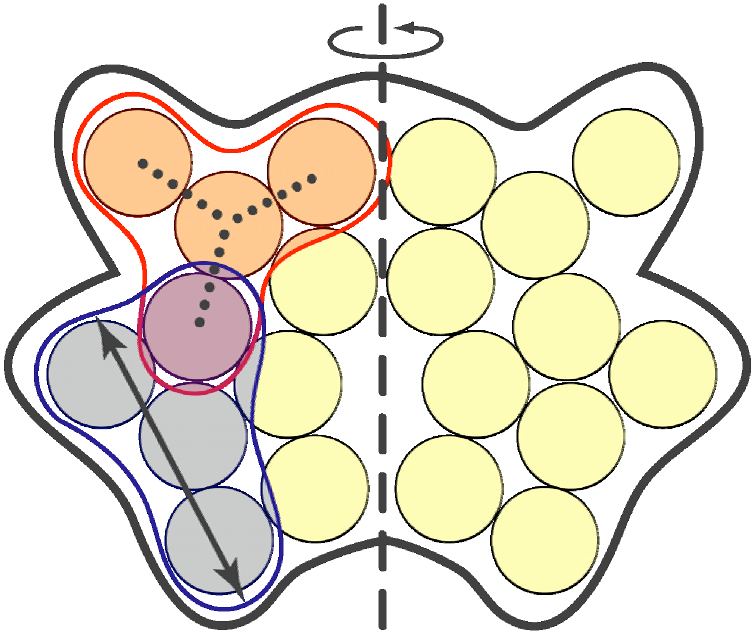




M. Lutzmann, R. Kunze, A. Buerer, U. Aebi & E. Hurt, *EMBO J.* 21, 387, 2002.

Structural characterization of assemblies from overall shape and subcomplex compositions

F. Alber, M. Kim, A. Sali. *Structure* 13, 435, 2005.



(i) the subunit excluded volume,
(ii) the assembly shape,
(iii) the subunit proximity in the subcomplex (the proximity restraint),
(iv) the subunit connectivity in the subcomplex (the connectivity restraint),
(v) the symmetry.

Test case

Data set: 27 pullouts

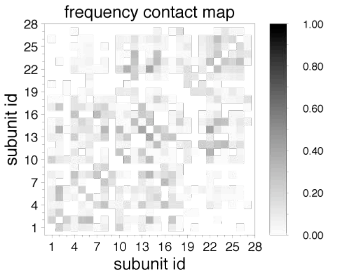
representative model

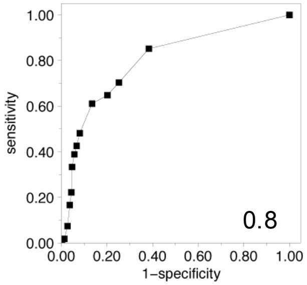
Frequency contact maps

ROC-curves

True positive rate: TPR
False positive rate: FPR
DRMS: smallest (average)

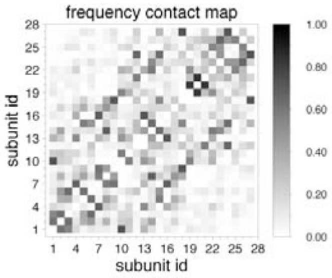
Subunit excluded volume
Subcomplex proxmity

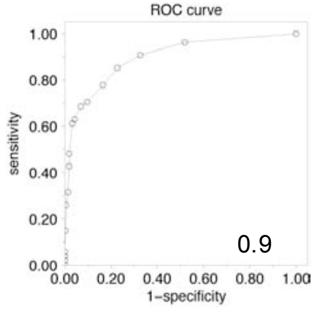




TRP: 22.2 %
FPR: 52.2%
DRMS: 1.6 (1.9)

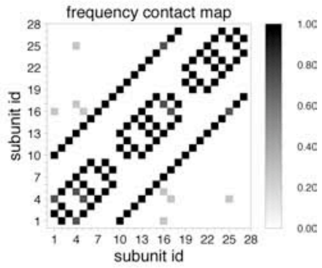
Subunit excluded volume
Subcomplex proximity
Assembly shape

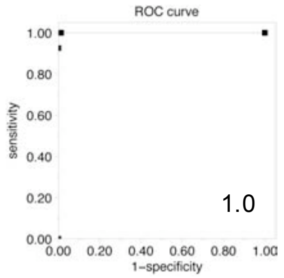




TPR: 48.0 %
FPR: 18.8%
DRMS: 0.6 (1.2)

Subunit excluded volume
Subcomplex proximity
Assembly shape
Subcomplex connectivity





TPR: 48.0 %
FPR: 18.8%
DR 0.0 (0.1)

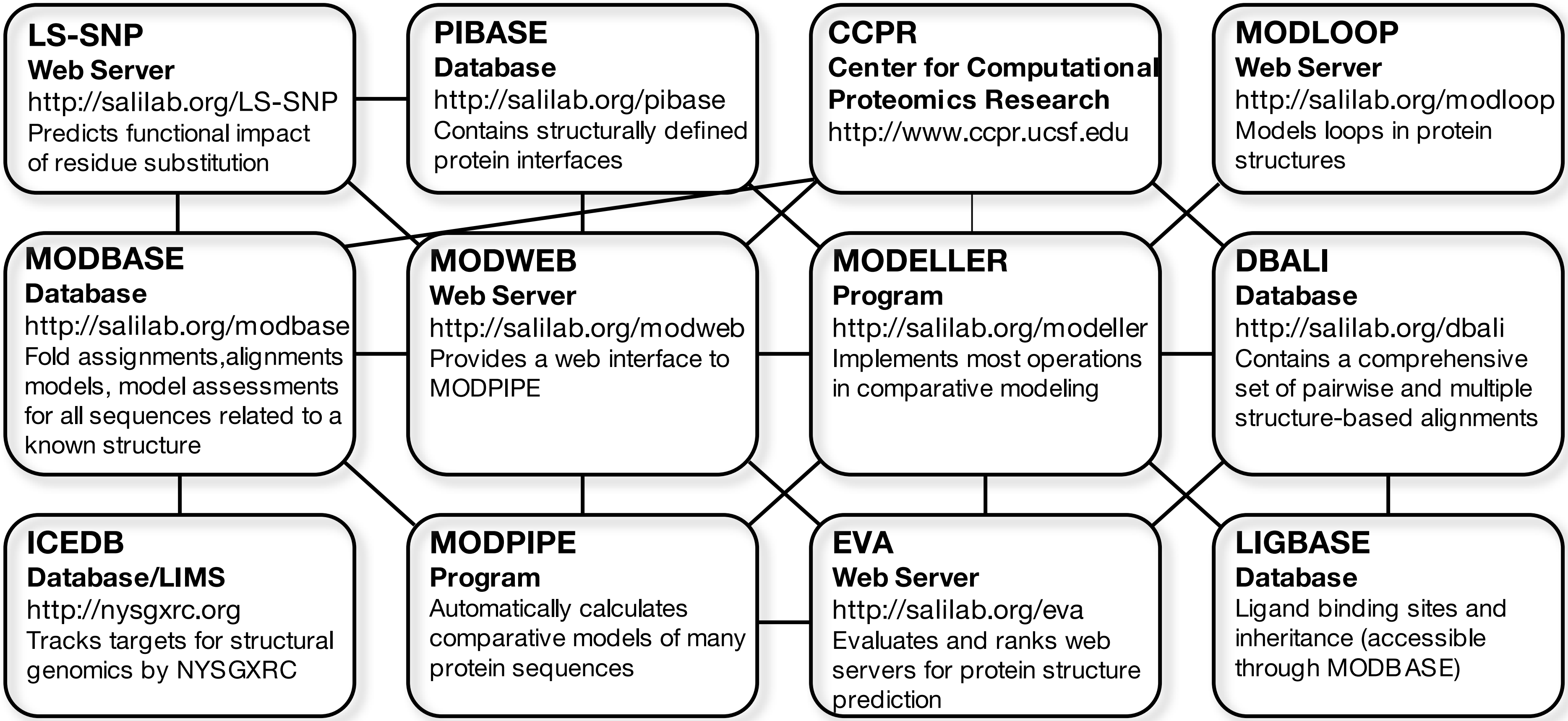
Alber, Kim, Sali, *Structure*, 2005

Towards a higher resolution structure of NPC

Characterize structures of the individual subunits, then fit them into the current low-resolution model.

A suite of programs, servers and databases for comparative protein structure modeling

http://salilab.org



LS-SNP
Web Server
http://salilab.org/LS-SNP
Predicts functional impact of residue substitution

PIBASE
Database
http://salilab.org/pibase
Contains structurally defined protein interfaces

CCPR
Center for Computational Proteomics Research
http://www.ccpr.ucsf.edu

MODLOOP
Web Server
http://salilab.org/modloop
Models loops in protein structures

MODBASE
Database
http://salilab.org/modbase
Fold assignments,alignments models, model assessments for all sequences related to a known structure

MODWEB
Web Server
http://salilab.org/modweb
Provides a web interface to MODPIPE

MODELLER
Program
http://salilab.org/modeller
Implements most operations in comparative modeling

DBALI
Database
http://salilab.org/dbali
Contains a comprehensive set of pairwise and multiple structure-based alignments

ICEDB
Database/LIMS
http://nysgxrc.org
Tracks targets for structural genomics by NYSGXRC

MODPIPE
Program
Automatically calculates comparative models of many protein sequences

EVA
Web Server
http://salilab.org/eva
Evaluates and ranks web servers for protein structure prediction

LIGBASE
Database
Ligand binding sites and inheritance (accessible through MODBASE)

External Resources
PDB, Uniprot, GENBANK, NR, PIR, INTERPRO, Kinase Resource
UCSC Genome Browser, Pfam, SCOP, CATH

5/17/05

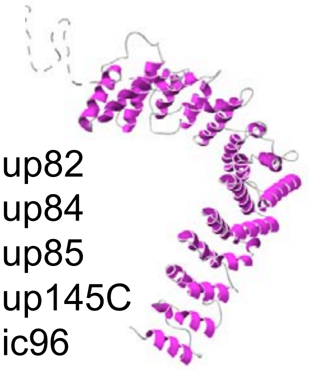
Fold Prediction

Devos, Dokudavskaya, Alber, Williams, Chait, Sali, Rout. *PLoS Biology 12, 1,* 2004

1) Simplicity of fold organization: 5 fold types describe 95 % of all residues in the NPC.

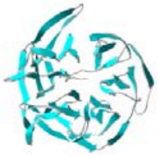
2) NPC has evolved through extensive gene duplication.

α-solenoid



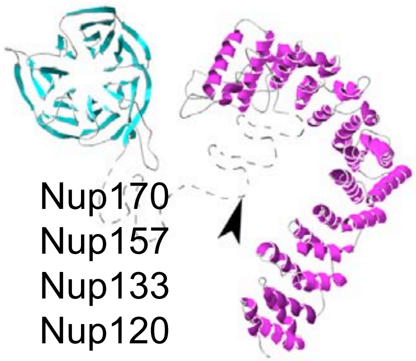
Nup82
Nup84
Nup85
Nup145C
Nic96

β-propeller



Seh1
Sec13

Clathrin-like



Nup170
Nup157
Nup133
Nup120

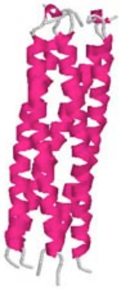
unstructured-FG repeat regions

Nup100
Nup1
Nup116
Nup159

Nsp1
Nup57
Nup60
Nup53

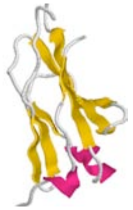
Nup145N
Nup53

Coiled-coiled



Nsp1
Nup1
Nup60
Nup159
Nup57
Nup53

IgG-fold

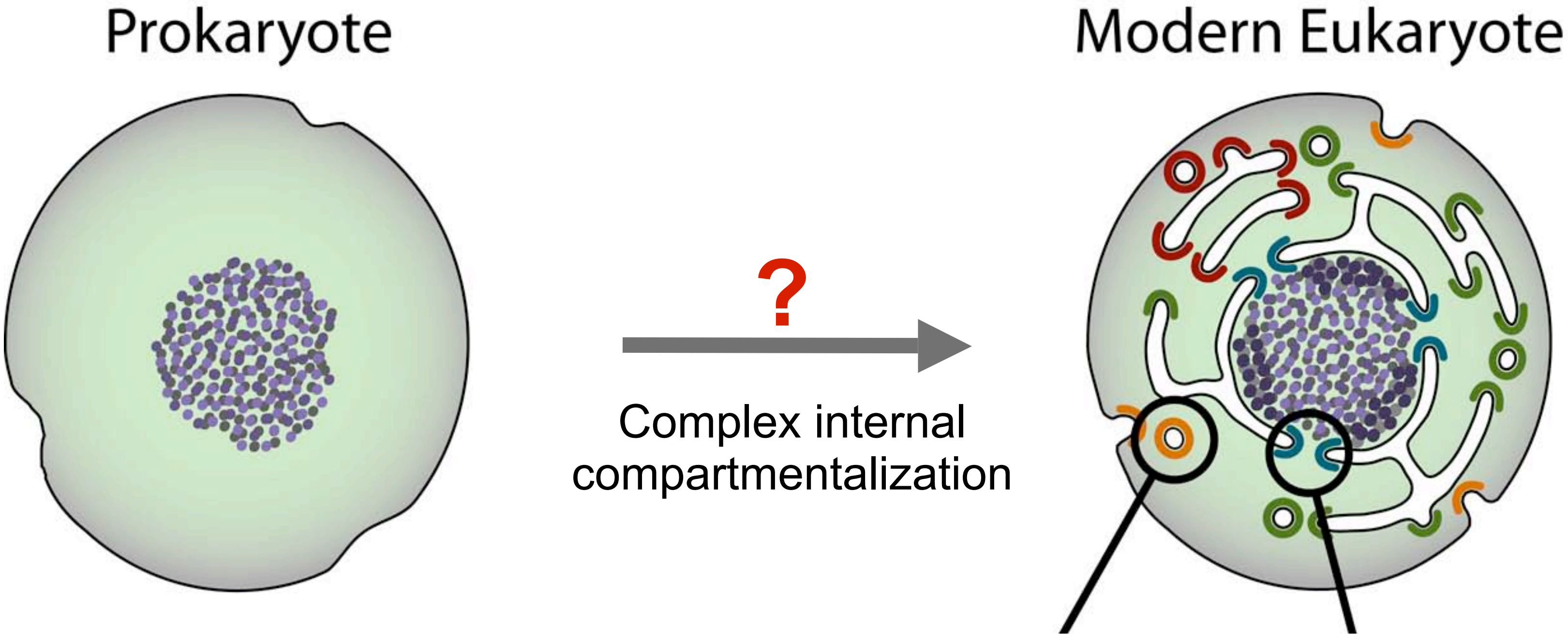


Pom152

Trans-membrane helices

Pom152
Ndc1
Pom34

Eukaryotic evolution



How could such a complicated system evolve in organisms with no analogous transport system?

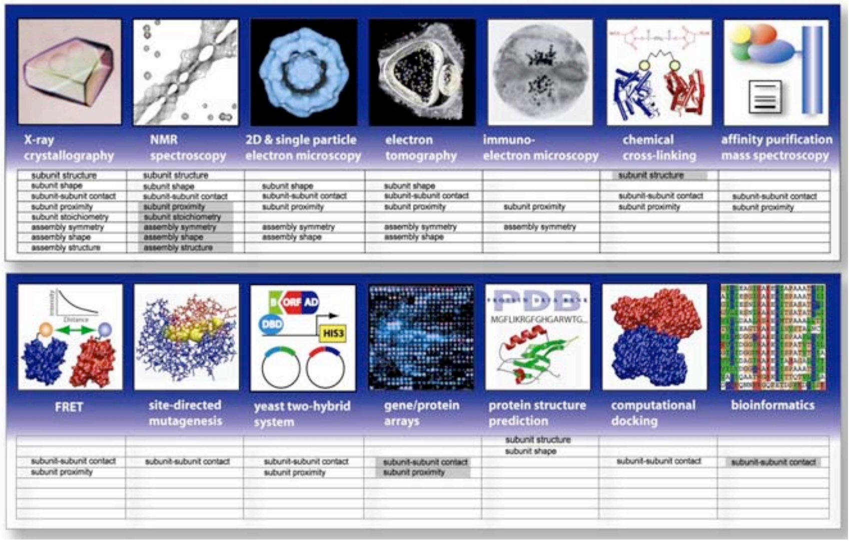
Summary: NPC Structure

- There are models (configurations) that satisfy all input restraints.
- These models are similar to each other in terms of protein-protein contacts.
- The model is in harmony with some other data.
- Simple models indicate feasibility.
- The model inspired hopefully testable hypotheses about evolution of the NPC and coated vesicles (as well as the mechanism of pore formation).
- The model will hopefully provide a starting point for a higher resolution characterization of the assembly (*eg*, EM, tomography, x-ray, cross-linking).

In Conclusion

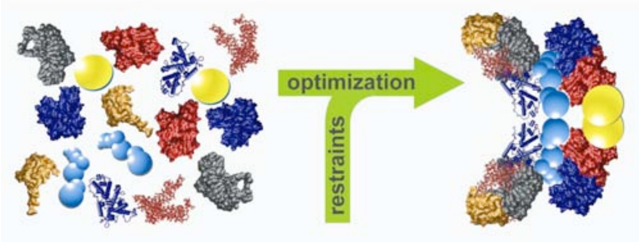
The goal is a comprehensive description of the multitude of interactions between molecular entities, which in turn is a prerequisite for the discovery of general structural principles that underlie all cellular processes.

This goal will be achieved by a ***tight*** integration of experimental and computational approaches, spanning all relevant size and time scales.



X-ray crystallography	NMR spectroscopy	2D & single particle electron microscopy	electron tomography	immuno-electron microscopy	chemical cross-linking	affinity purification mass spectroscopy
subunit structure	subunit structure				subunit structure	
subunit shape	subunit shape	subunit shape	subunit shape			
subunit-subunit contact	subunit-subunit contact	subunit-subunit contact	subunit-subunit contact		subunit-subunit contact	subunit-subunit contact
subunit proximity	subunit proximity	subunit proximity	subunit proximity	subunit proximity	subunit proximity	subunit proximity
subunit stoichiometry	subunit stoichiometry					
assembly symmetry	assembly symmetry	assembly symmetry	assembly symmetry	assembly symmetry		
assembly shape	assembly shape	assembly shape	assembly shape			
assembly structure	assembly structure					

FRET	site-directed mutagenesis	yeast two-hybrid system	gene/protein arrays	protein structure prediction	computational docking	bioinformatics
				subunit structure		
				subunit shape		
subunit-subunit contact	subunit-subunit contact	subunit-subunit contact	subunit-subunit contact		subunit-subunit contact	subunit-subunit contact
subunit proximity		subunit proximity	subunit proximity			



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