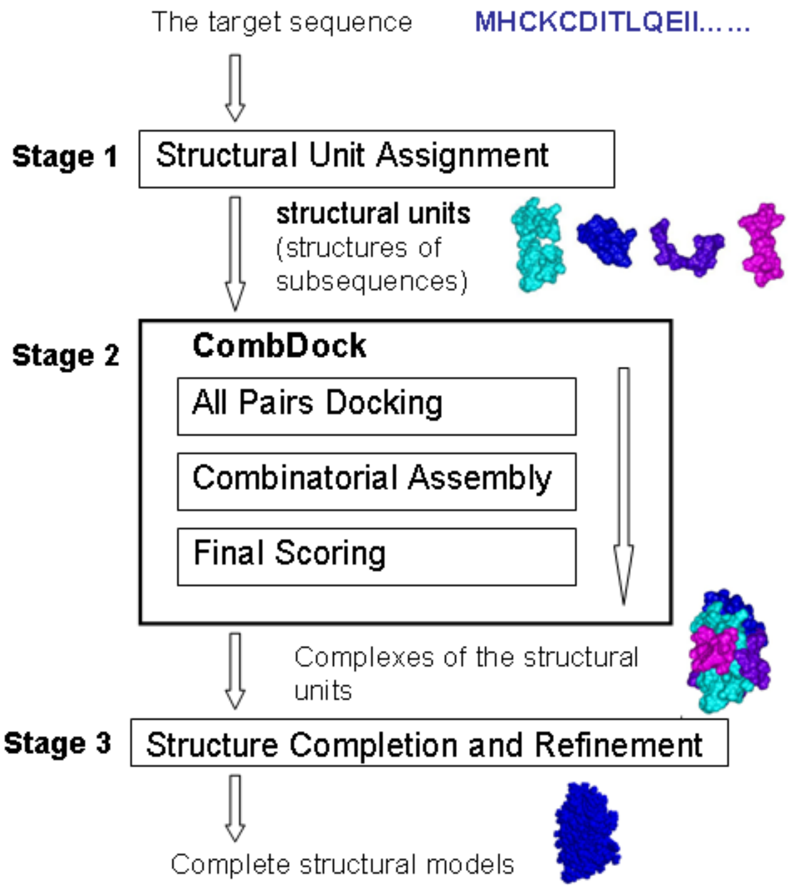
V20 Predicting Structures of Protein Complexes from Connectivities

`CombDock:` automated approach for predicting 3D structure of heterogenous multimolecular assemblies.

Input: structures of *N* individual proteins

Problem appears more difficult than the pairwise docking problem.

Idea: exploit additional geometric constraints that are part of the combinatorial problem.



Haim Wolfson
Tel Aviv University
http://www.cs.tau.ac.il/~wolfson/

Inbar et al., J. Mol. Biol. 349, 435 (2005)

Review: pairwise docking: Katchalski-Kazir algorithm

Discretize proteins A and B on a grid.

Every node is assigned a value

$$f_{A_{l,m,n}} = \begin{cases} 1 & : \quad \text{surface of molecule} \\ \rho & : \quad \text{core of molecule} \\ 0 & : \quad \text{open space} \end{cases}$$

and

$$f_{B_{l,m,n}} = \begin{cases} 1 & : \quad \text{inside molecule} \\ 0 & : \quad \text{open space} \end{cases}$$

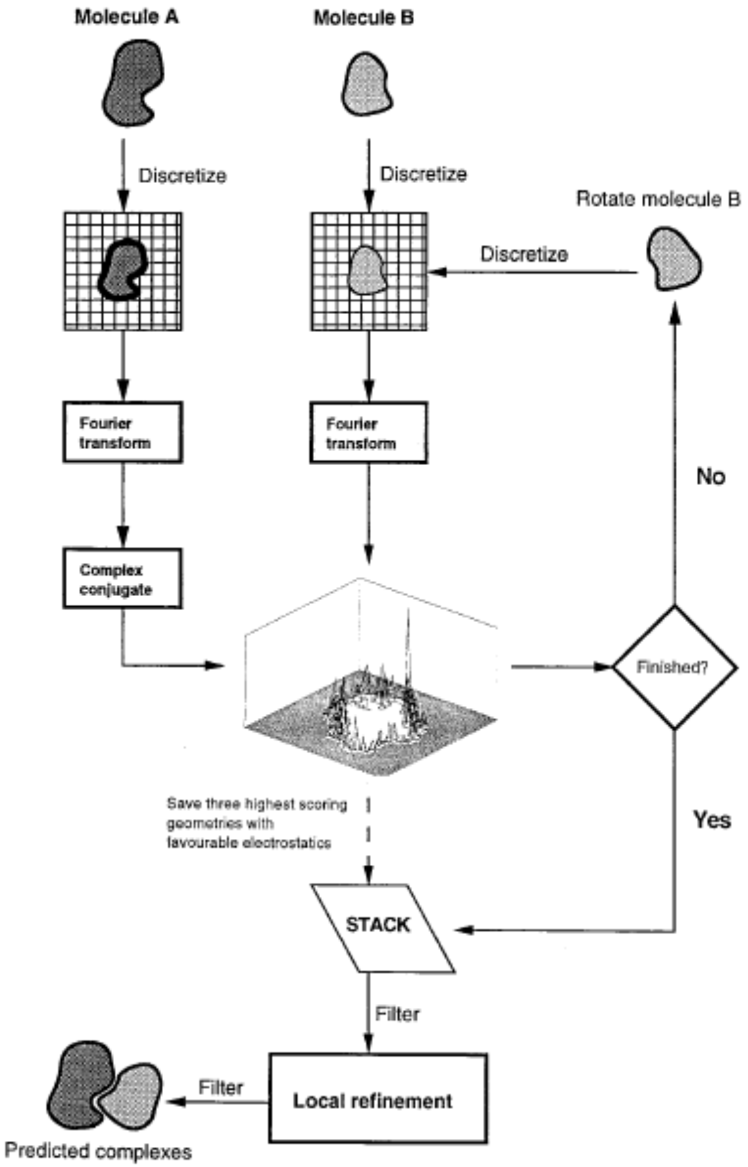
The correlation function of f_A and f_B is:

$$f_{C_{\alpha,\beta,\gamma}} = \sum_{l=1}^{N} \sum_{m=1}^{N} \sum_{n=1}^{N} f_{A_{l,m,n}} \times f_{B_{l+\alpha,m+\beta,n+\gamma}}$$

Use FFT to compute correlation efficiently.

Output: solutions with best surface complementarity.

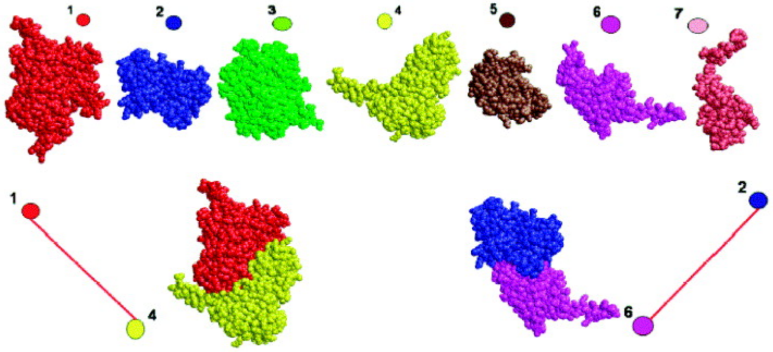
Gabb et al. J. Mol. Biol. (1997)



(1) All pairs docking module

Aim: predict putative pairwise interactions

Based on the *N* individual protein structures perform pairwise docking for each of the *N* (*N* - 1) / 2 pairs of proteins



Since the correct scoring of pairwise-docking is difficult, the correct solution may be among the first few hundred solutions.

→ keep *K* best solutions for each pair of proteins.

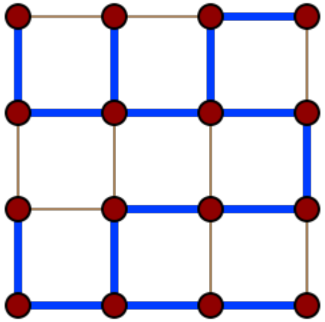
Here, *K* was varied from dozens to hundreds.

Inbar et al., J. Mol. Biol. 349, 435 (2005)

(2) Combinatorial assembly module

Input: N subunits and $N (N - 1) / 2$ sets of K scored transformations. These are the candidate interactions.

Reduction to a spanning tree



Spanning tree = a graph that connects all vertices and has no circles

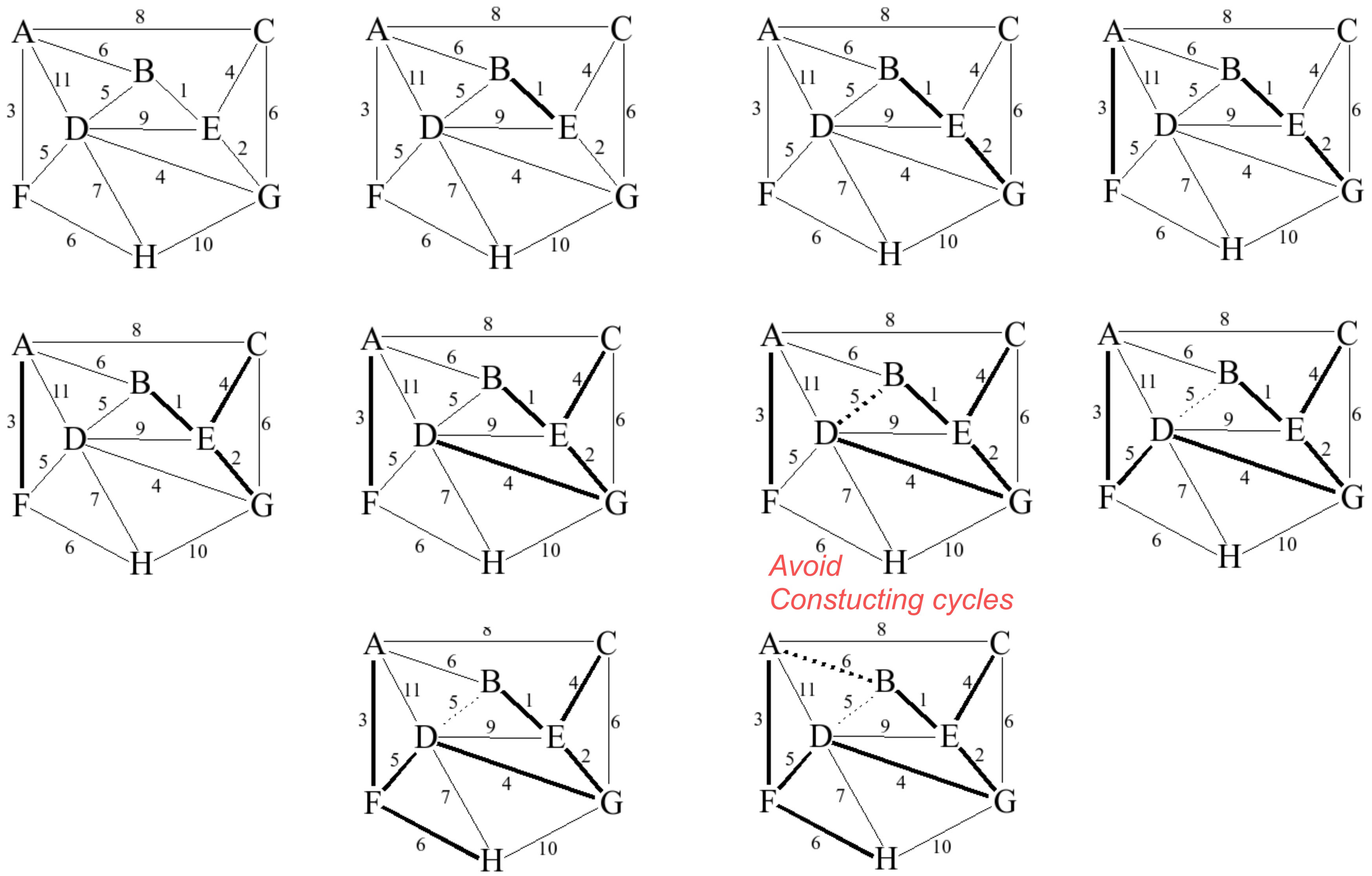
Build weighted graph representing the input:

- each protein structure = vertex
- each transformation (docking orientation) = edge connecting the corresponding vertices
- edge weight = docking score of the transformation

→ Since the input contains K transformations for each pair of subunits, we get a complete graph with ***K* parallel edges** between each pair of vertices.

Inbar et al., J. Mol. Biol. 349, 435 (2005)
www.wikipedia.org

Review: Spanning tree – algorithm of Kruskal



Avoid
Constucting cycles

(2) Combinatorial assembly module

For 2 subunits, each candidate binary docking complex is represented by an **edge** and the 2 vertices.

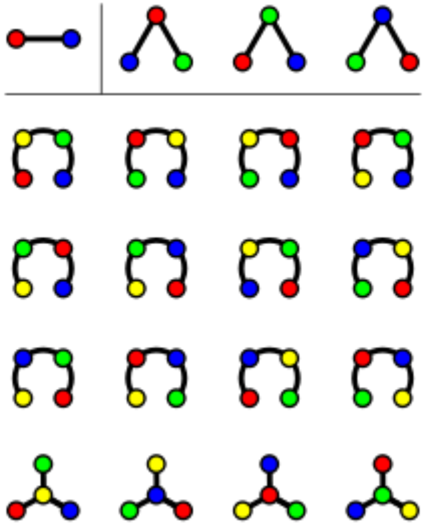
For the full complex, a candidate complex is represented by a **spanning tree**. Each spanning tree of the input graph represents a particular **3D structure** for the complex of all input structures.

→ Problem of finding 3D structures of complexes is equivalent to finding spanning trees.

The number of spanning trees in a complete graph with *N* nodes and **no parallel edges** is N^{N-2} (Cayley's formula).

Here, the input graph has *K* parallel edges between each pair of vertices. → the number of spanning trees is $N^{N-2}\ K^{N-1}$.

→ Exhaustive searches are infeasible!



Cayley's formula (the number of different trees on *n* vertices is n^{n-2}, graphically demonstrated for graphs with 2, 3 and 4 nodes.

Inbar et al., J. Mol. Biol. 349, 435 (2005)
www.wikipedia.org

(2) Combinatorial assembly module:algorithm

`CombDock` algorithm uses 2 basic principles:
(1) hierarchical construction of the spanning tree
(2) greedy selection of subtrees

→ 2 subtrees of smaller size (that were previously generated) are connected with an input edge to generate trees with *i* vertices

In this way, the common parts of different trees are generated only once.

When connecting subtrees, check whether there are severe **penetrations** between pairs of subunits that are represented by different subtrees.

Inbar et al., J. Mol. Biol. 349, 435 (2005)

(2) Combinatorial assembly module:algorithm

Stage 1: algorithm start with trees of size 1.
Each tree contains a single vertex that represents a subunit.

Stage *i*: the tree complexes that consist of exactly *i* vertices (subunits) are generated by connecting 2 trees generated at a lower stage with an input edge transformation.

Tree complexes that fulfil the penetration constraint are kept for the next stages.

Because it is impractical to search all valid spanning trees, the algorithm performs a greedy selection of subtrees.

For each subset of vertices, the algorithm keeps only the *D* best-scoring valid trees that connect them.

The **tree score** is the sum of its edge weights.

Inbar et al., J. Mol. Biol. 349, 435 (2005)

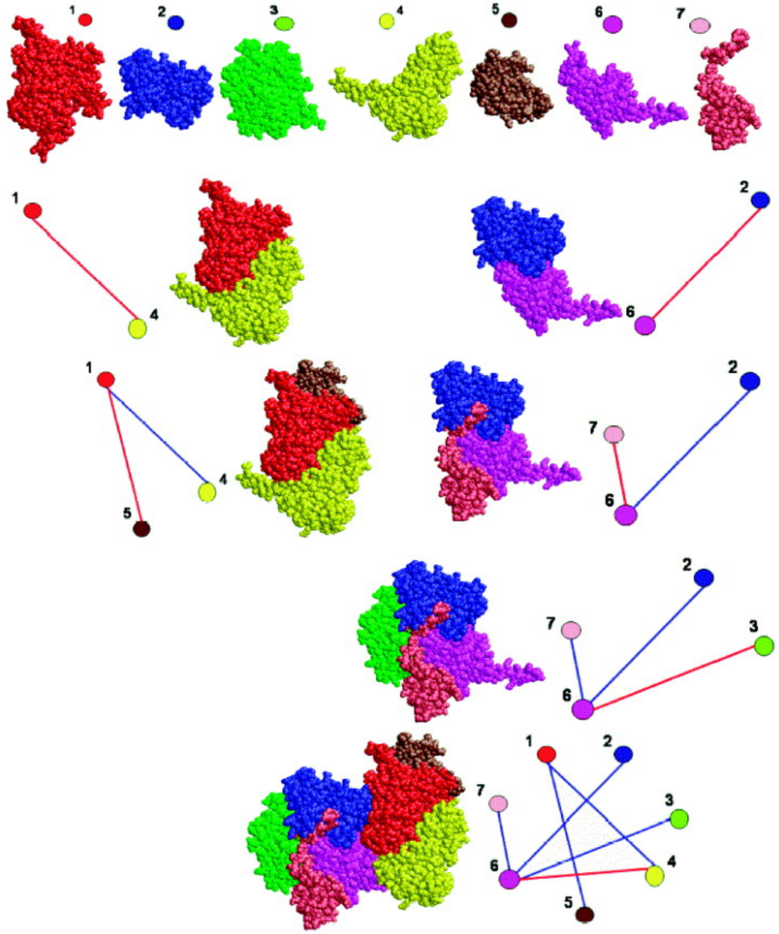
Example: arp2/3 complex

The arp2/3 complex consists of 7 subunits (top).

Shown are only the complexes of the different stages that were relevant to the construction of the third-best scoring solution with RMSD 1.2 Å (bottom).

Red edge: transformation of the current stage,

Blue edges: transformations of previous stages.



Inbar et al., J. Mol. Biol. 349, 435 (2005)

Final scoring

A **geometric score** evaluates the shape complementarity between the subunits:

- check distances between surface points on adjacent subunits.

- close surface points increase score,

- penetrating surface points decrease score.

Physico-chemical component of the final score counts all surface points that belong to non-polar atoms = this gives an estimate of the hydrophobic effect.

Inbar et al., J. Mol. Biol. 349, 435 (2005)

Clustering of solutions

Clustering of solutions:

(1) compute **contact maps** between subunits: array of $N (N - 1)$ bins.

If two subunits are in contact within the complex,
set the corresponding bit to 1, and to 0 otherwise.

(2) superimpose complexes that have the same contact map
and compute RMSD between C^{α} atoms.

If this distance is less than a threshold, consider complexes
as members of a **cluster**.

For each cluster, keep only the complex with the highest score.

Inbar et al., J. Mol. Biol. 349, 435 (2005)

Performance for known complexes

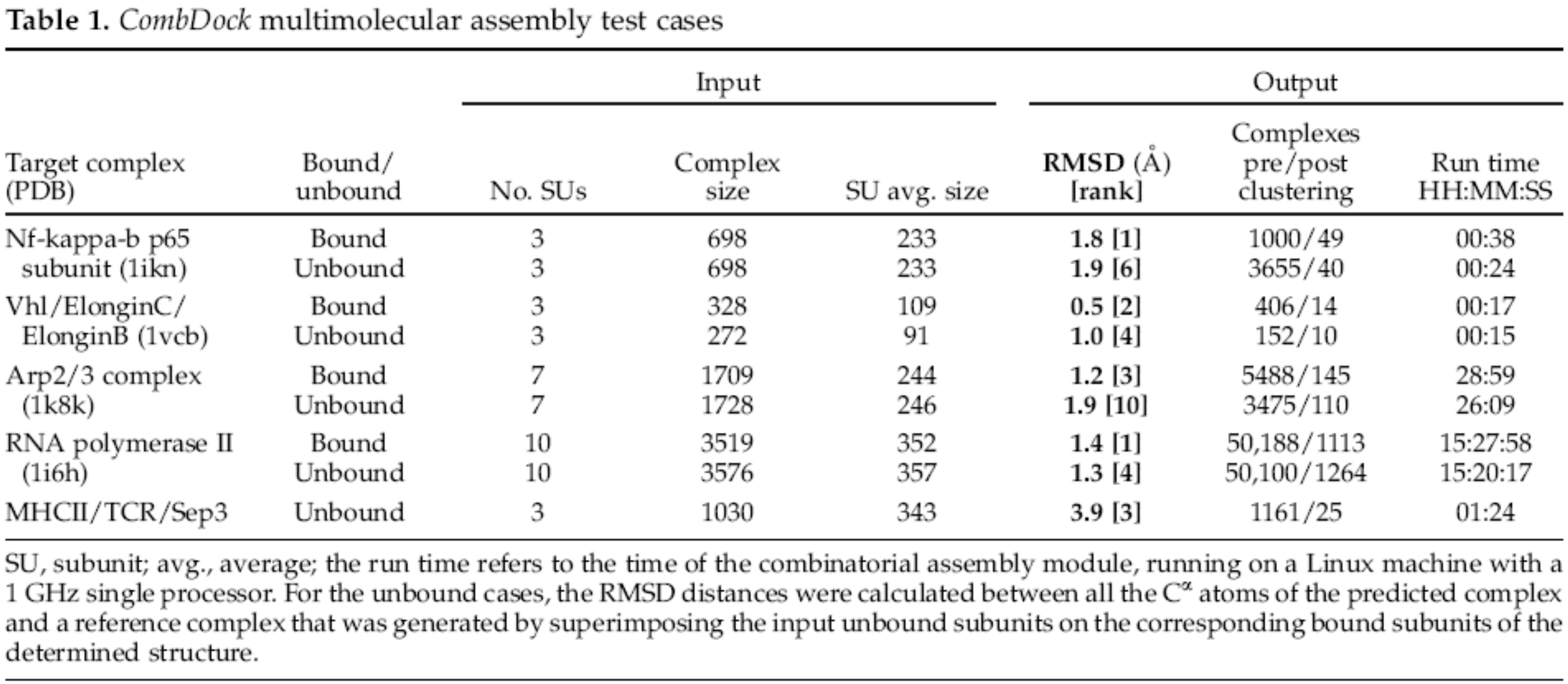


Table 1. *CombDock* multimolecular assembly test cases

		Input			Output		
Target complex (PDB)	Bound/ unbound	No. SUs	Complex size	SU avg. size	**RMSD** (Å) **[rank]**	Complexes pre/post clustering	Run time HH:MM:SS
Nf-kappa-b p65 subunit (1ikn)	Bound	3	698	233	**1.8 [1]**	1000/49	00:38
	Unbound	3	698	233	**1.9 [6]**	3655/40	00:24
Vhl/ElonginC/ ElonginB (1vcb)	Bound	3	328	109	**0.5 [2]**	406/14	00:17
	Unbound	3	272	91	**1.0 [4]**	152/10	00:15
Arp2/3 complex (1k8k)	Bound	7	1709	244	**1.2 [3]**	5488/145	28:59
	Unbound	7	1728	246	**1.9 [10]**	3475/110	26:09
RNA polymerase II (1i6h)	Bound	10	3519	352	**1.4 [1]**	50,188/1113	15:27:58
	Unbound	10	3576	357	**1.3 [4]**	50,100/1264	15:20:17
MHCII/TCR/Sep3	Unbound	3	1030	343	**3.9 [3]**	1161/25	01:24

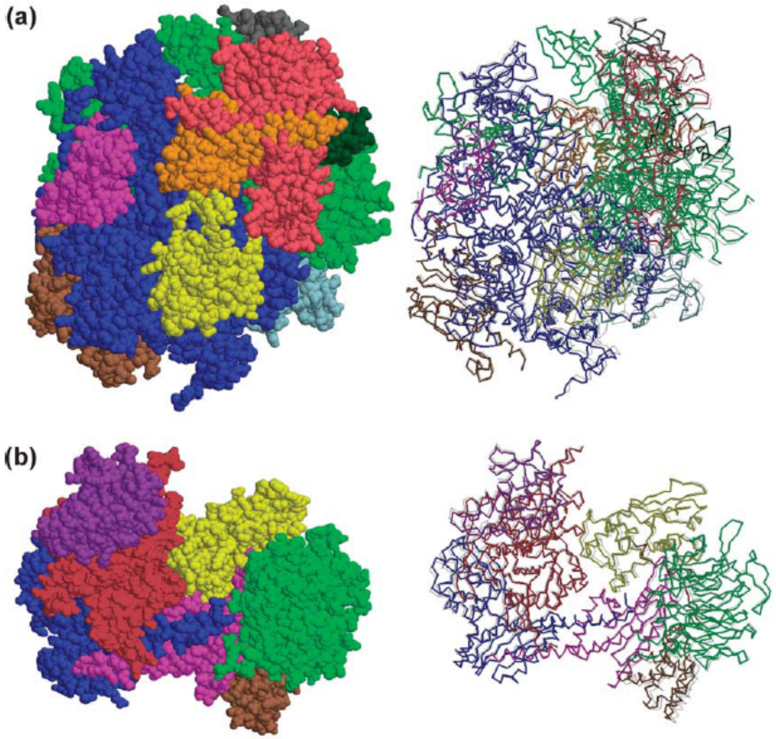
SU, subunit; avg., average; the run time refers to the time of the combinatorial assembly module, running on a Linux machine with a 1 GHz single processor. For the unbound cases, the RMSD distances were calculated between all the C^{α} atoms of the predicted complex and a reference complex that was generated by superimposing the input unbound subunits on the corresponding bound subunits of the determined structure.

Inbar et al., J. Mol. Biol. 349, 435 (2005)

Examples of large complexes

CombDock solution

solution superposed on the crystal structure (gray thiner lines)



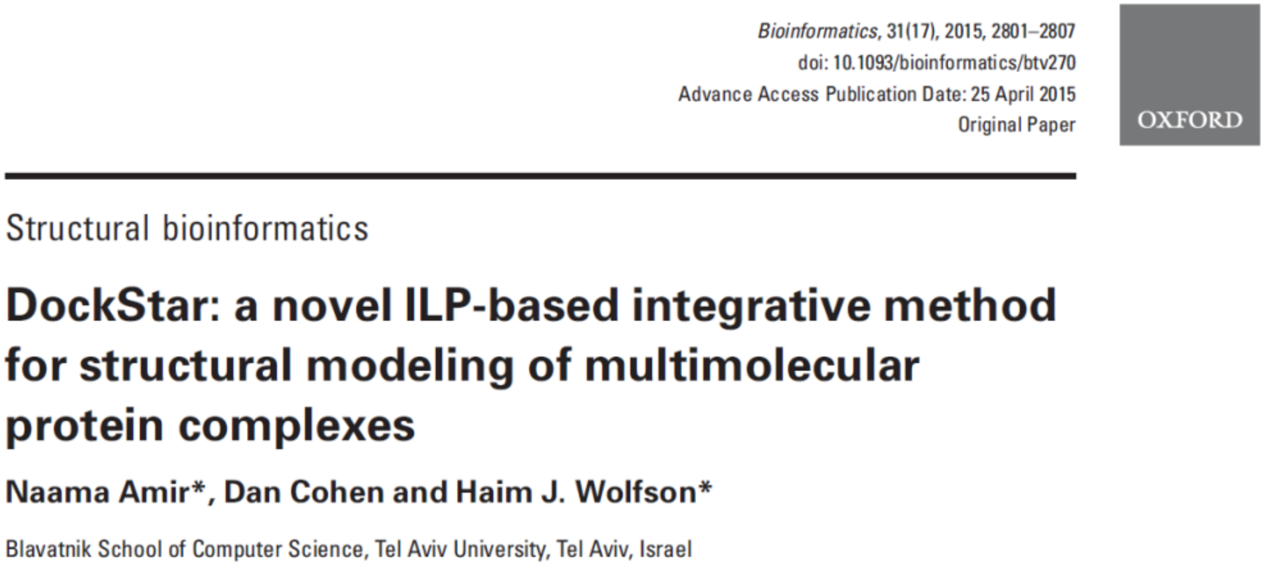
(a) the bestranked complex of the 10 subunits of **RNA polymerase II**, RMSD 1.4 Å.

(b) the third-best scoring assembly of the 7 subunits of the **arp2/3 complex**, RMSD 1.2 Å.

CombDock is not as succesful for docking „unbound“ subunit structures that structurally differ from „bound“ conformations.

Inbar et al., J. Mol. Biol. 349, 435 (2005)

DockStar: overcome limitations of CombDock



Bioinformatics, 31(17), 2015, 2801–2807
doi: 10.1093/bioinformatics/btv270
Advance Access Publication Date: 25 April 2015
Original Paper

OXFORD

Structural bioinformatics

DockStar: a novel ILP-based integrative method for structural modeling of multimolecular protein complexes

Naama Amir*, Dan Cohen and Haim J. Wolfson*

Blavatnik School of Computer Science, Tel Aviv University, Tel Aviv, Israel

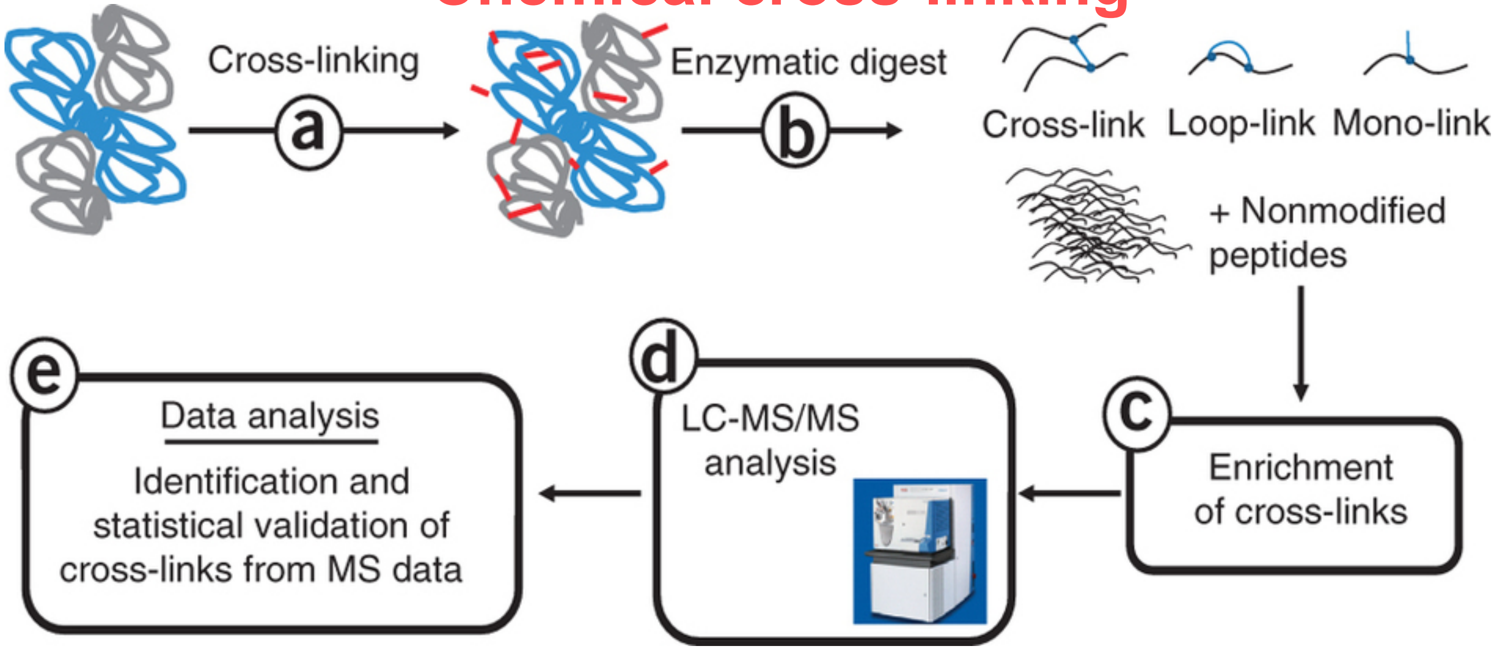
2 subtasks for generation of macromolecular complex structures:

(a) Detection of the protein-protein interaction graph between the individual subunits; use additional data from **chemical cross-linking** for this,

(b) Detection of a globally consistent pose of the subunits, so that there are no steric clashes between them and the binding energy of the whole complex is optimized.

Amir et al., Bioinformatics 31, 2801 (2015)

Chemical cross-linking



(**a**) cross-linking reaction using a chemical cross-linking reagent. These molecules have a certain length, have two reactive groups at both ends of the molecule and may covalently bind either to cysteine or lysine residues of a single protein or of two proteins)

(**b**) enzymatic digestion of the proteins to peptides,

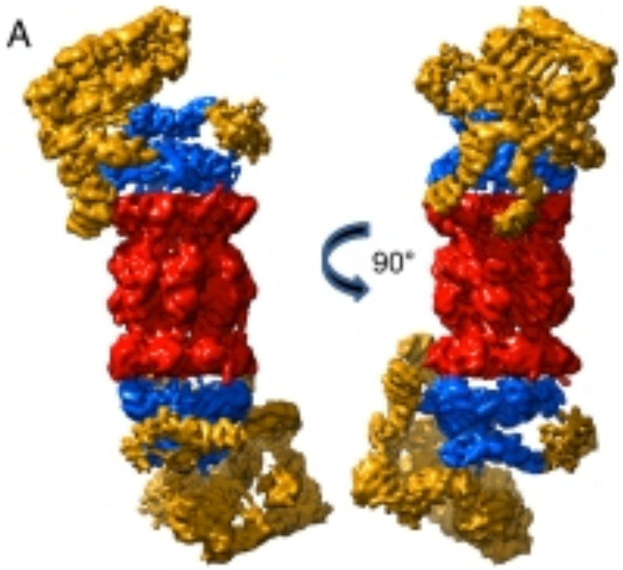
(**c**) enrichment of cross-linked peptides,

(**d**) analysis of cross-linked peptides by LC-MS/MS,

(**e**) data analysis.

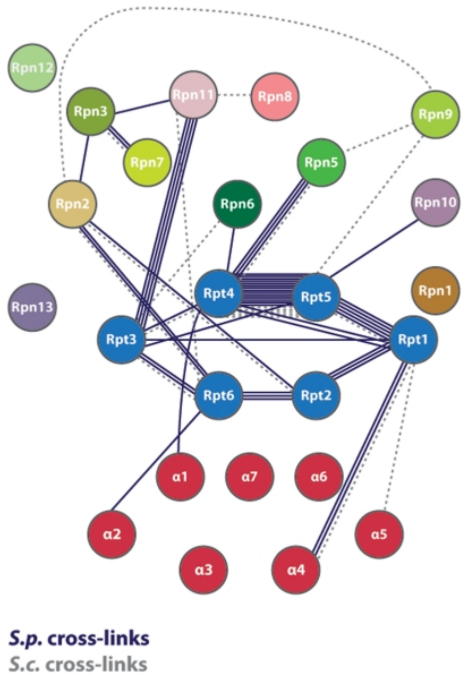
Leitner et al. Nature Protocols
9, 120–137 (2014)

Iterative refinement of the 3D structure of S26 proteasome

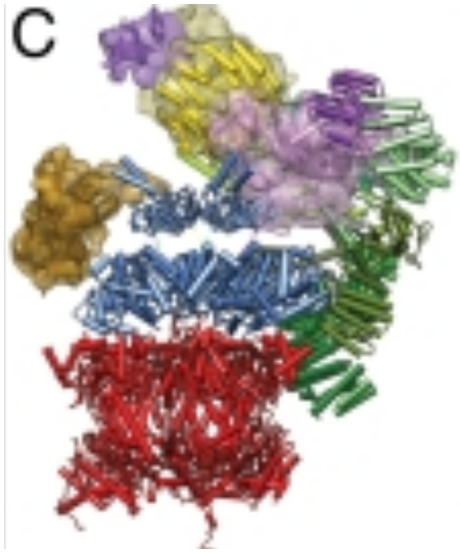


Low resolution

EM structure



Chemical cross-links for the *S. pombe* and *S. cerevisiae* 26S proteasomes. 55 (21) pairs of cross-linked lysines from the *S. pombe* (*S. cerevisiae*) 26S proteasome subunits. Multiple edges between a pair of subunits indicate multiple cross-linked lysine pairs.



Atomistic structure generated

Lasker et al., PNAS (2012) 109: 1380

StarDock

Low resolution MS data is especially suitable to assist in subtask (a).

-Native MS of intact protein complexes and their subcomplexes (→TAP-MS) can determine the **stoichiometry** of the complex subunits and deduce the **interaction graph** of the multimolecular complex.

- Chemical cross-linking combined with MS provides **distance constraints** between surface residues both on the same and on neighboring subunits.

This provides information both for the detection of the interaction graph as well as constraints on the relative spatial poses of neighboring subunits.

Such constraints have been successfully e.g. exploited in the modeling of the

-26S proteasome,

-the proteasome lid,

-the TRiC/CCT chaperonin,

-the RNA polymerase II–TFIIF complex and more.

Amir et al., Bioinformatics 31, 2801 (2015)

2.1 Generation of transformation sets

Generate for each subunit a set of candidate rigid transformations.

One subunit is chosen as an **anchor subunit**. Preferably, the anchor subunit should have the largest number of neighbors in the multimolecular assembly interaction graph. All other subunits which are known to interact with the anchor are then docked to it.

This requires a **star shaped spanning tree** topology of the interaction graph.

Pairwise docking step is carried out by `PatchDock`, which optimizes shape complementarity, while satisfying maximal distance constraints between residues of neighboring subunits from cross-linking.

The top 1000 `PatchDock` transformations are refined, rescored and re-ranked by the `FiberDock` tool

-> pairwise scores

Amir et al., Bioinformatics 31, 2801 (2015)

2.3 Selection of best global solution

For each of the *n* subunits, let

- P_i $(0 \leq i < n)$ be **subunit** *i*,
- $T(P_i)$ be the set of **candidate transformations** received from the previous stage for subunit P_i.
- $T_{i,r}$ be **transformation** *r* of subunit P_i .
- $S(T_{i,r}, T_{j,s})$ be the **pairwise interaction score** of subunits P_i and P_j transformed by $T_{i,r}$ and $T_{j,s}$, respectively.

The **globally optimal solution** Sol includes one transformation per subunit and maximizes score(Sol) defined as:

$$\text{score}(\text{Sol}) = \sum_{T_{i,r}, T_{j,s} \in \text{Sol} \cap i \neq j} S(T_{i,r}, T_{j,s})$$

Amir et al., Bioinformatics 31, 2801 (2015)

2.3 Selection of best global solution

This optimization task can be formulated as the following graph theoretic problem:

Let G = (V,E) be an undirected *n*-partite graph with vertex set $V = V_0 \cup \ldots \cup V_{n-1}$, so that for each transformation $T_{i,r} \in T(P_i)$ there is a vertex $u_{i,r} \in V_i$.

(Each V_i contains all transformations *r* of subunit P_i as its vertices $u_{i,r}$.)

Each pair of vertices is joined by an edge:

$$E = \{(u_{i,r}, v_{j,s}) | u_{i,r} \in V_i; v_{j,s} \in V_j; i \neq j\}$$

with the weight $w(u_{i,r}, v_{j,s}) = S(T_{i,r}, T_{j,s}) \qquad \forall (u_{i,r}, v_{j,s}) \in E$

The optimal solution is achieved by choosing one vertex per V_i that maximizes the edge-weight of the induced sub-graph.

Amir et al., Bioinformatics 31, 2801 (2015)

ILP formulation

This graph theoretic task can be formulated as an ILP (Nemhauser and Wolsey, 1988). Define a variable $X_{i,r}$ for each vertex $u_{i,r} \in V$ and a variable $Y_{i,r,j,s}$ for each edge $e(u_{i,r,}, v_{j,s}) \in E$ as follows

$$X_{i,r} = \begin{cases} 1 & \text{if } u_{i,r} \text{ is chosen} \\ 0 & \text{otherwise} \end{cases}$$

$$Y_{i,r,j,s} = \begin{cases} 1 & \text{if both } u_{i,r} \text{ and } v_{j,s} \text{ are chosen} \\ 0 & \text{otherwise} \end{cases}$$

The ILP **objective function** is

$$\text{Maximize} \qquad \text{score(Sol)} = \sum_{(u_{i,r}, v_{j,s}) \in E} w(u_{i,r}, v_{j,s}) Y_{i,r,j,s}$$

Subject to the constraints:

$$\sum_{u_{i,r} \in V_i} X_{i,r} = 1 \qquad \forall i, 0 \le i < n$$

$$\sum_{u_{i,r} \in V_i} Y_{i,r,j,s} = X_{j,s} \qquad \forall j, s, i, \quad j \neq i$$

Amir et al., Bioinformatics 31, 2801 (2015)

The objective function is exactly the edge-weight of the chosen sub-graph. The first constraint ensures that exactly one transformation is chosen for each subunit. The second constraint ensures that an edge is chosen if and only if both vertices that it connects are chosen as well.

The ILP step was solved by the CPLEX 12.5 package

ILP formulation – alternative solutions

The ILP method outputs one single highest scoring global solution.

To retrieve additional high scoring solutions, the ILP step is applied iteratively to find a solution that maximizes the objective function and was not chosen before.

For this, a **linear constraint** is used (see paper by Amir et al.).

Sofar we considered complexes having a **star shaped spanning tree**, where an anchor subunit, which interacts with all the other subunits, can be chosen. However, this is a special case.

Arbitrary complexes are divided into overlapping sub-complexes, each with a star shaped spanning tree, which are solved separately as above.

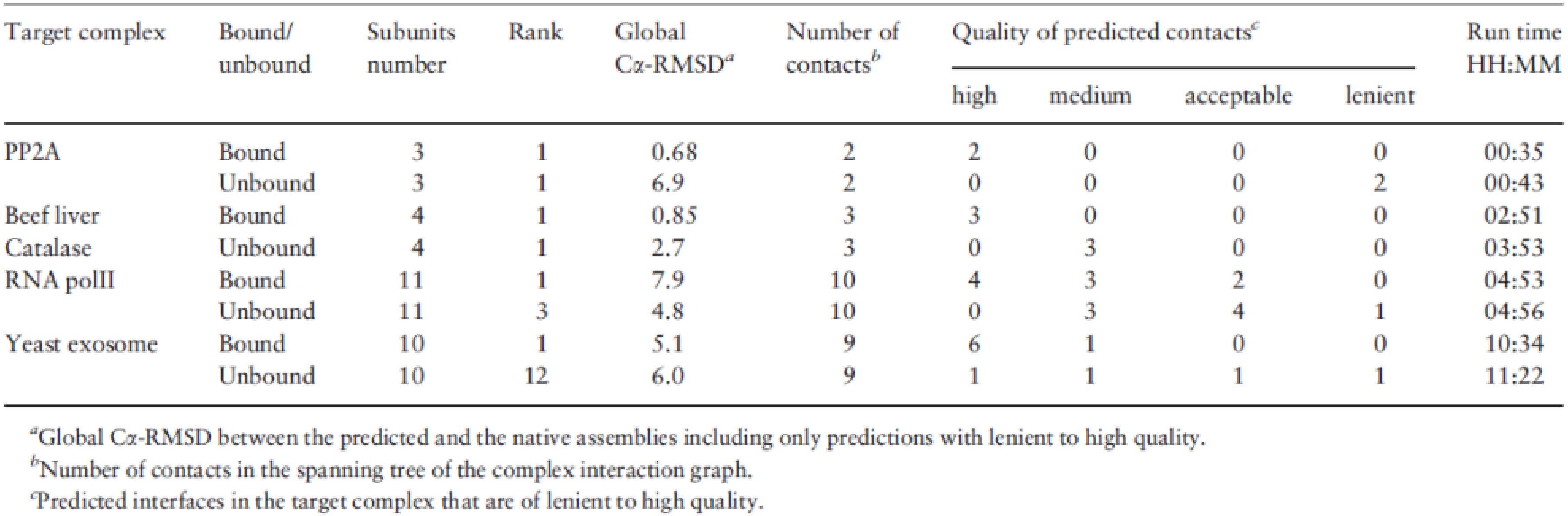
Then, top solutions of subcomplexes that share a subunit are merged, while defining the shared subunit as the new ‘anchor’. All the transformations in the merged (new) subcomplex are recalculated vis-a-vis the reference frame of the new ‘anchor’. These new transformation sets are used as input for steps 2–4 of the algorithm in order to solve the larger sub-complex.

In several such iterations one can cover all the subunits of the assembly.

Amir et al., Bioinformatics 31, 2801 (2015)

DockStar applications

Table 1. Summary of the DockStar's results



Target complex	Bound/ unbound	Subunits number	Rank	Global Cα-RMSD[a]	Number of contacts[b]	Quality of predicted contacts[c]				Run time HH:MM
						high	medium	acceptable	lenient	
PP2A	Bound	3	1	0.68	2	2	0	0	0	00:35
	Unbound	3	1	6.9	2	0	0	0	2	00:43
Beef liver	Bound	4	1	0.85	3	3	0	0	0	02:51
Catalase	Unbound	4	1	2.7	3	0	3	0	0	03:53
RNA polII	Bound	11	1	7.9	10	4	3	2	0	04:53
	Unbound	11	3	4.8	10	0	3	4	1	04:56
Yeast exosome	Bound	10	1	5.1	9	6	1	0	0	10:34
	Unbound	10	12	6.0	9	1	1	1	1	11:22

[a] Global Cα-RMSD between the predicted and the native assemblies including only predictions with lenient to high quality.

[b] Number of contacts in the spanning tree of the complex interaction graph.

[c] Predicted interfaces in the target complex that are of lenient to high quality.

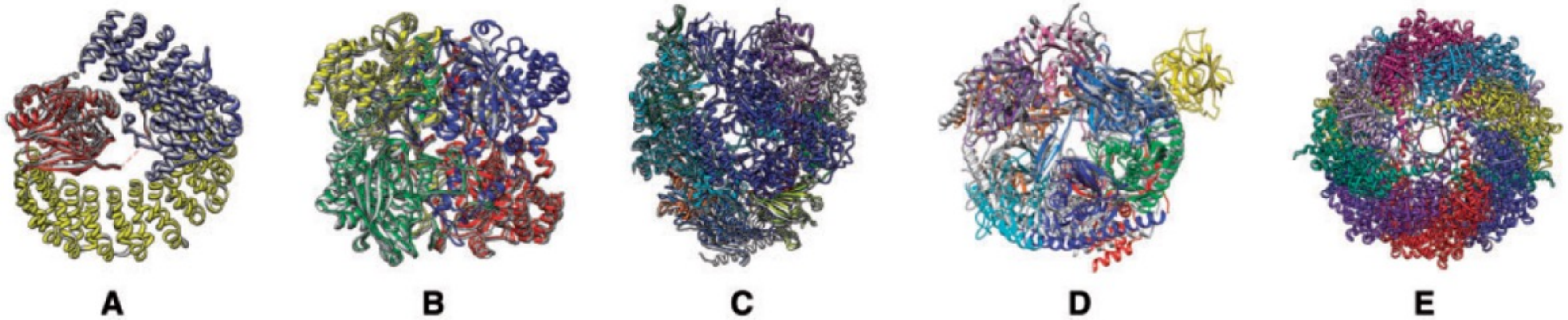
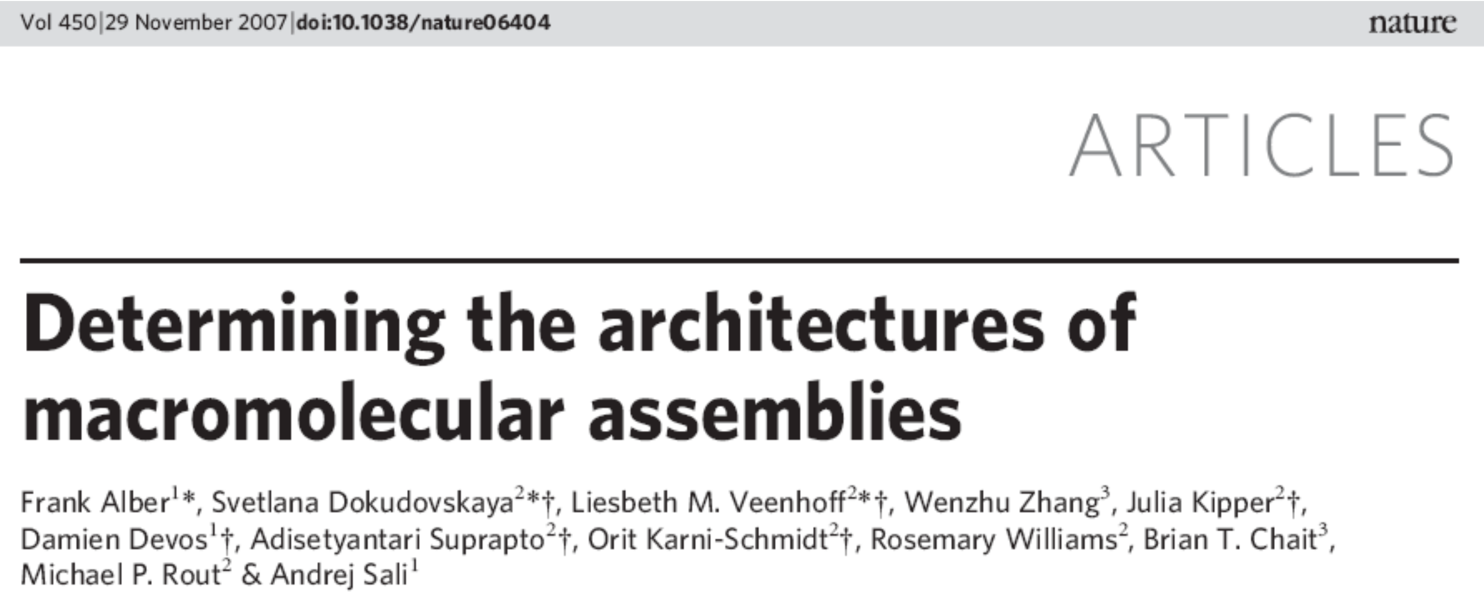


Fig. 1. The predicted models of the bound cases (coloured by chains) superimposed on the correct complex structures taken from the PDB (grey). (**A**) PP2A (A(yellow), B(blue), C(red)), (**B**) The Beef Liver Catalase [A(yellow), B(blue), C(red), D(green)], (**C**) RNA polymerase II [Rbp1(blue), Rbp2(cyan), Rbp3(light blue), Rbp5(purple), Rbp6(green), Rbp7(pink), Rbp8(yellow), Rbp9(dark green), Rbp10(orange), Rbp11(brown), Rbp12(red)], (**D**) The Yeast Exosome [Rrp45(blue), Rrp41(cyan), Rrp43(light blue), Rrp46(green), Rrp42(purple), Mtr3(pink), Rrp40(red), Rrp4(orange), Csl4(yellow), Dis3(dark green)]. (**E**) The predicted order of chains in the model of the TRiC/CCT Chaperonin: Z(red) Q(blue) H(yellow) E(light blue) B(pink) D(grey) A(green) G(purple)

Amir et al., Bioinformatics 31, 2801 (2015)

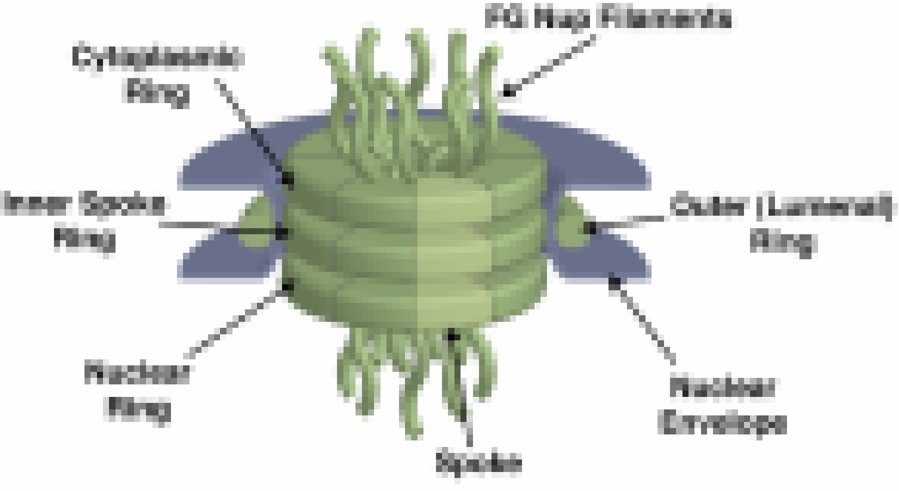


Vol 450|29 November 2007|doi:10.1038/nature06404 nature

ARTICLES

Determining the architectures of macromolecular assemblies

Frank Alber[1]*, Svetlana Dokudovskaya[2]*†, Liesbeth M. Veenhoff[2]*†, Wenzhu Zhang[3], Julia Kipper[2]†, Damien Devos[1]†, Adisetyantari Suprapto[2]†, Orit Karni-Schmidt[2]†, Rosemary Williams[2], Brian T. Chait[3], Michael P. Rout[2] & Andrej Sali[1]

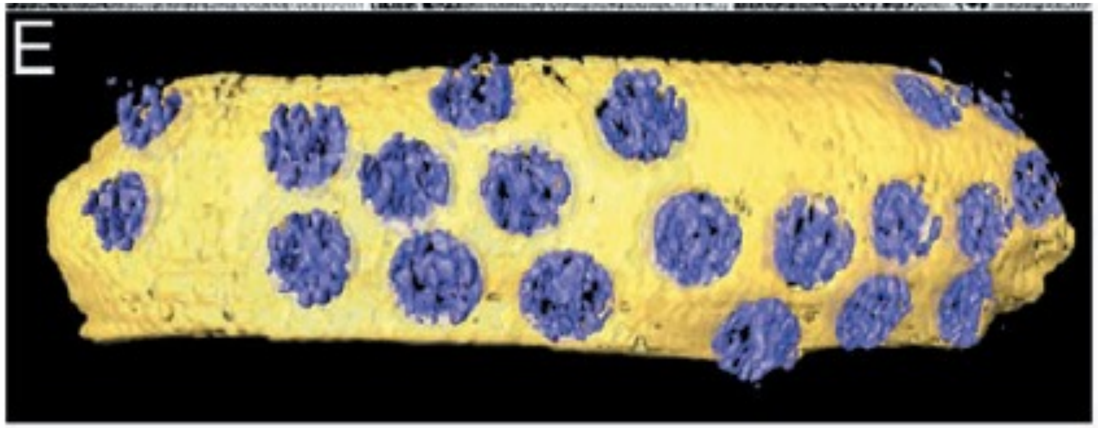


Alber et al., Nature 450, 683 (2007)

The nuclear pore complex (NPC)

NPCs are large assemblies of ca. 30 different proteins, the nucleoporins (ca. 120 megadaltons in metazoa).

Each NPC contains at least **456** individual protein molecules.



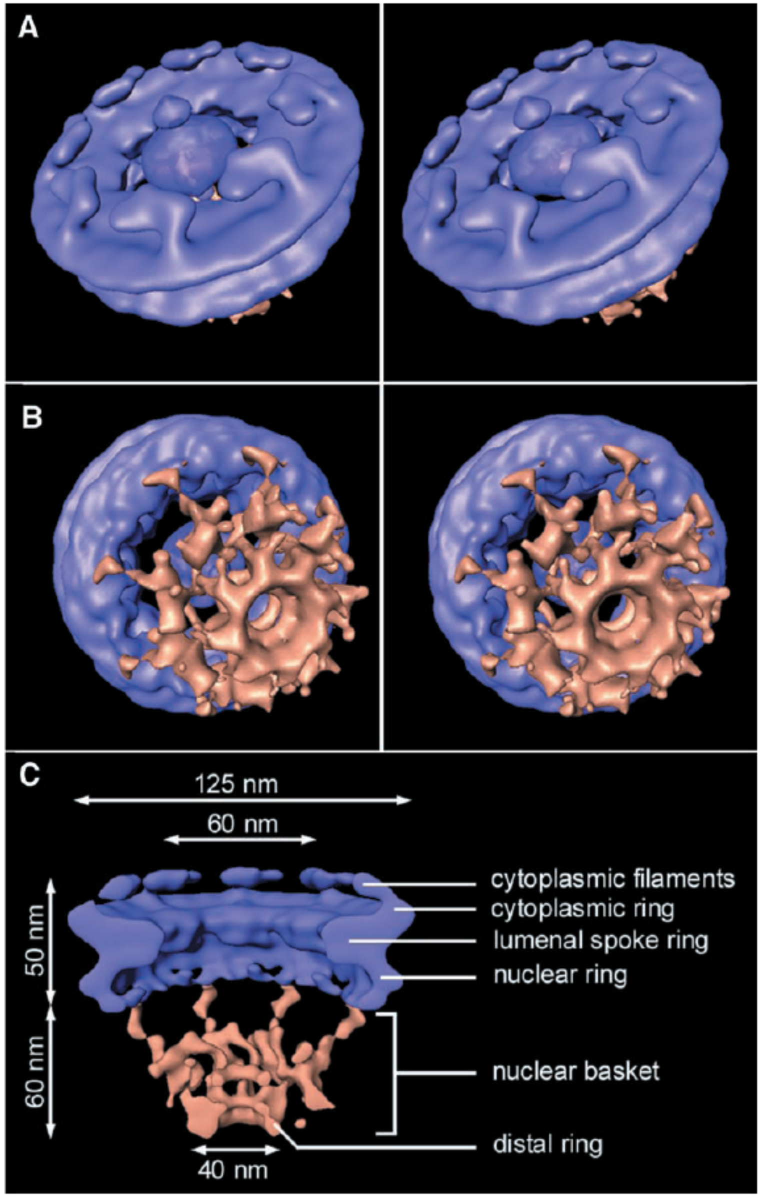
Surface rendered representation of a segment of nuclear envelope (NPCs in blue, membranes in yellow). The dimensions of the rendered volume are 1680 nm 984 nm 558 nm. The number of NPCs was ca. 45/µm2.

Nucleocytoplasmic transport of macromolecular cargoes between the nucleus and the cytoplasm depends on their recognition by **transport factors** (exportins and importins). These interact with the NPC to carry cargoes across the nuclear envelope.

NPCs show a broad degree of compositional and structural **conservation** among all eukaryotes studied.

Beck et al. Science 306, 1387 (2004)

Structure of the Dictyostelium NPC



(A) **Cytoplasmic face** of the NPC in stereo view. The cytoplasmic filaments are arranged around the central channel.

(B) **Nuclear face** of the NPC in stereo view. The distal ring of the basket is connected to the nuclear ring by the nuclear filaments.

(C) Cutaway view of the NPC.

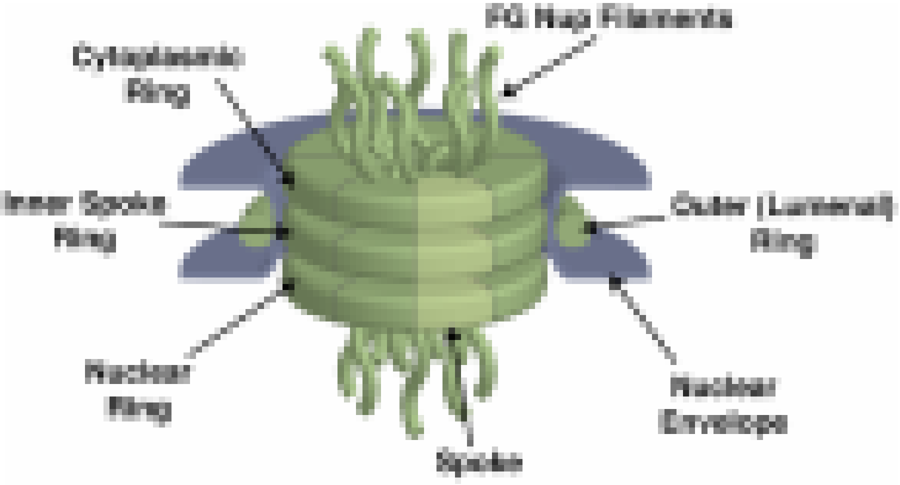
Beck et al. Science 306, 1387 (2004)

Nuclear Pore Complex

CryoEM shows the NPC as a **doughnut-shaped structure** with an eight-fold rotational axis perpendicular to the NE plane.
This symmetry indicates that the NPC is composed of 8 identical building blocks, termed **spokes** arranged radially around a central **channel** that serves as the conduit for macromolecular transport.

Each NPC spans the nuclear envelope through a pore formed by the **fusion** of the inner and outer nuclear envelope membranes.

Numerous **filamentous structures** project from the NPC into the cytoplasm and nucleoplasm.

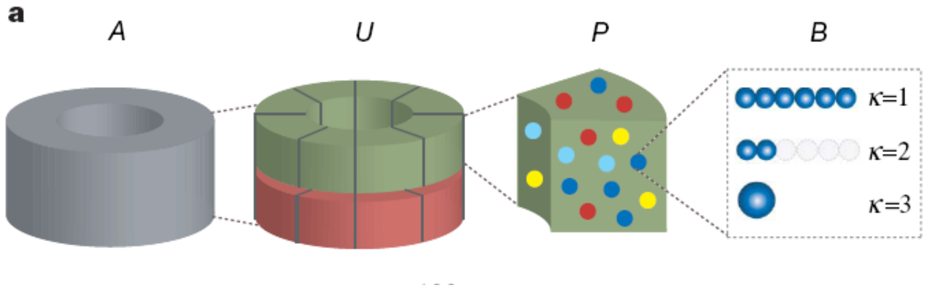


The NPC consists of **30 different proteins**.

4-level hierarchical representation of the NPC

ImmunoEM experiments localized each nup to the nucleoplasmic, cytoplasmic, or both sides of the equatorial plane → formally represent the NPC composition and protein stoichiometry with a 4-level hierarchy, consisting of

- the whole NPC (assembly, A),
- the 16 half spokes (unit, U),
- the nup (protein, P),
- and bead (particle, B) levels.

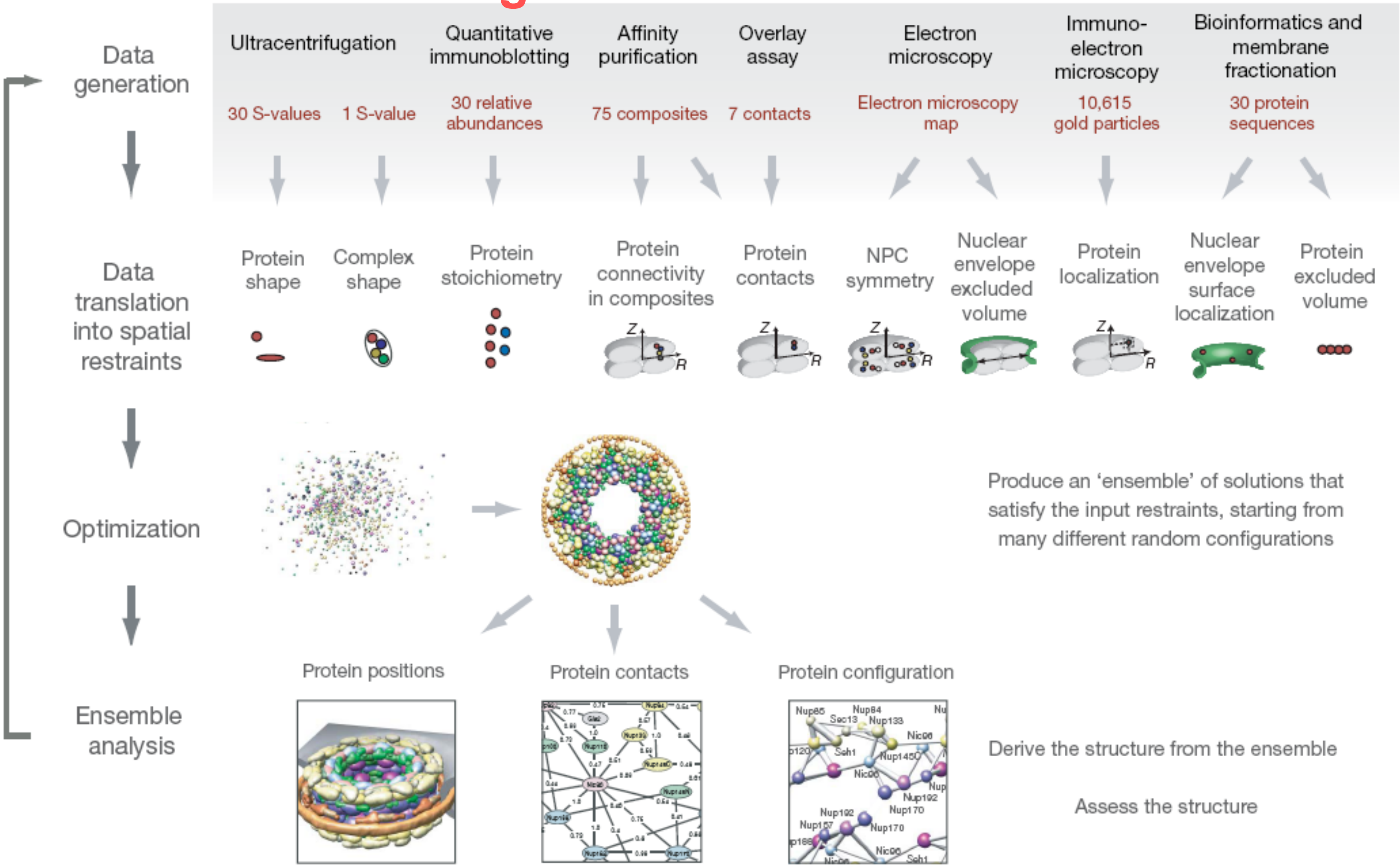


Each of the 8 half-spoke units U at the cytosolic side is composed of 27 different types of nups, of which 2 are present in 2 copies each, totaling 29 protein instances.

Similarly, each of the 8 half-spoke units U at the nucleoplasmic side contains 28 protein instances of 25 different types.

8 x (29 + 28) = 456 proteins in total

Overview of integrative structure determination



Data generation

Ultracentrifugation: 30 S-values, 1 S-value

Quantitative immunoblotting: 30 relative abundances

Affinity purification: 75 composites

Overlay assay: 7 contacts

Electron microscopy: Electron microscopy map

Immuno-electron microscopy: 10,615 gold particles

Bioinformatics and membrane fractionation: 30 protein sequences

Data translation into spatial restraints

Protein shape, Complex shape, Protein stoichiometry, Protein connectivity in composites, Protein contacts, NPC symmetry, Nuclear envelope excluded volume, Protein localization, Nuclear envelope surface localization, Protein excluded volume

Optimization

Produce an ‘ensemble’ of solutions that satisfy the input restraints, starting from many different random configurations

Ensemble analysis

Protein positions, Protein contacts, Protein configuration

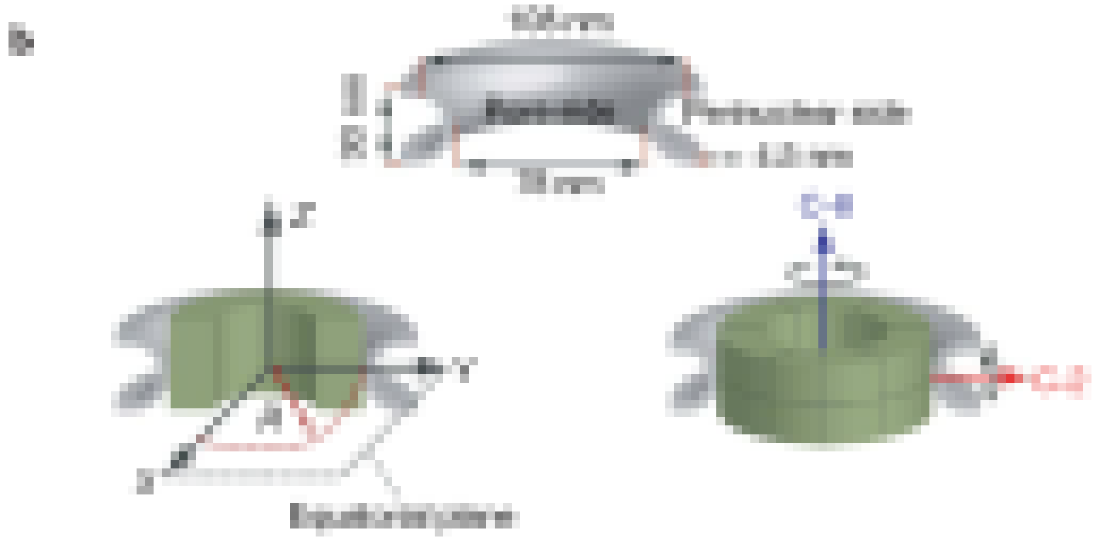
Derive the structure from the ensemble

Assess the structure

Figure 1 | Determining the architecture of the NPC by integrating spatial restraints from proteomic data. First, structural data (red) are generated by various experiments (black). Second, the data are translated into spatial restraints. Third, an ensemble of structural solutions that satisfy the data are obtained by minimizing the violations of the spatial restraints, starting from many different random configurations. Fourth, the ensemble is clustered into distinct sets of structures on the basis of their similarities, and analysed in terms of protein positions, contacts and configuration.

Dimensions and symmetry

Top : the dimensions of the nuclear envelope, as taken from cryo-EM images.



Bottom-left: the coordinate system used has the origin at the centre of the nuclear envelope pore. The nuclear envelope is indicated in grey.

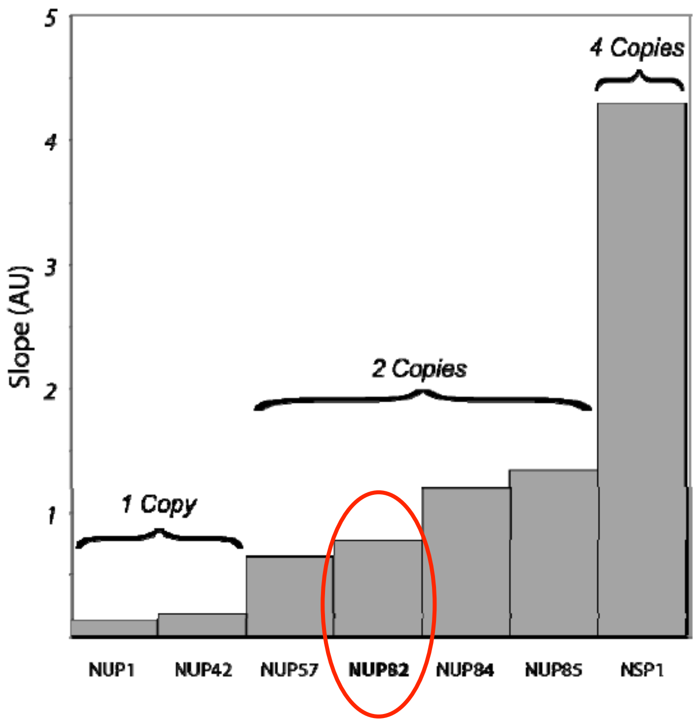
Bottom-right: the eight-fold (C-8) and two-fold (C-2) symmetry axes of the NPC, as revealed primarily by cryo-EM.

Stochiometry of each component in the NPC

Aim: E.g. identification of Nup82 copy number.

Aliquots of nuclear envelope preparations from PrA tagged strains equivalent to 3.6, 6, 10 and 15 µg were processed for **immunoblot analysis**.

Result: Nup82 falls into the same range as the values of the 2 copy per spoke Nups.



Strains with known copy numbers – Nup42, Nup1 (1 copy per spoke), Nup57, Nup84, Nup85 (2 copies per spoke) and Nsp1 (4 copies per spoke) were used as a **control**.

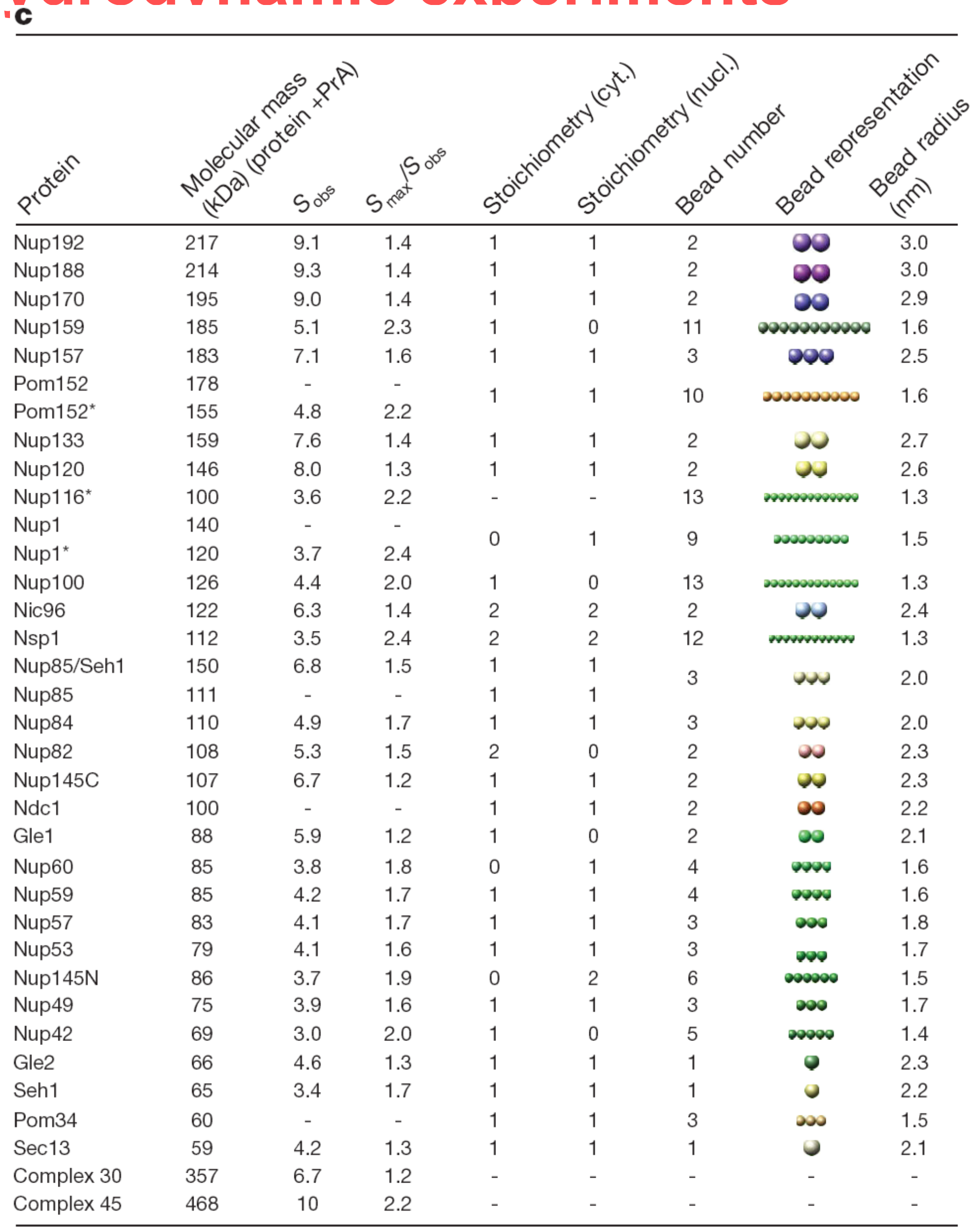
Protein shape from hydrodynamic experiments

Shown in colums 5 and 6 are the **stoichiometry** of a protein in the cytoplasmic (cyt.) and nucleoplasmic (nucl.) half-spoke, as measured by quantitative immunoblotting.

Protein shapes:

S_{max} values were calculated based on the molecular mass (kDa) of each protein;

S_{max}/S_{obs} < 1.4 indicates a globular protein;
1.6–1.9, moderately elongated;
> 2, highly elongated.

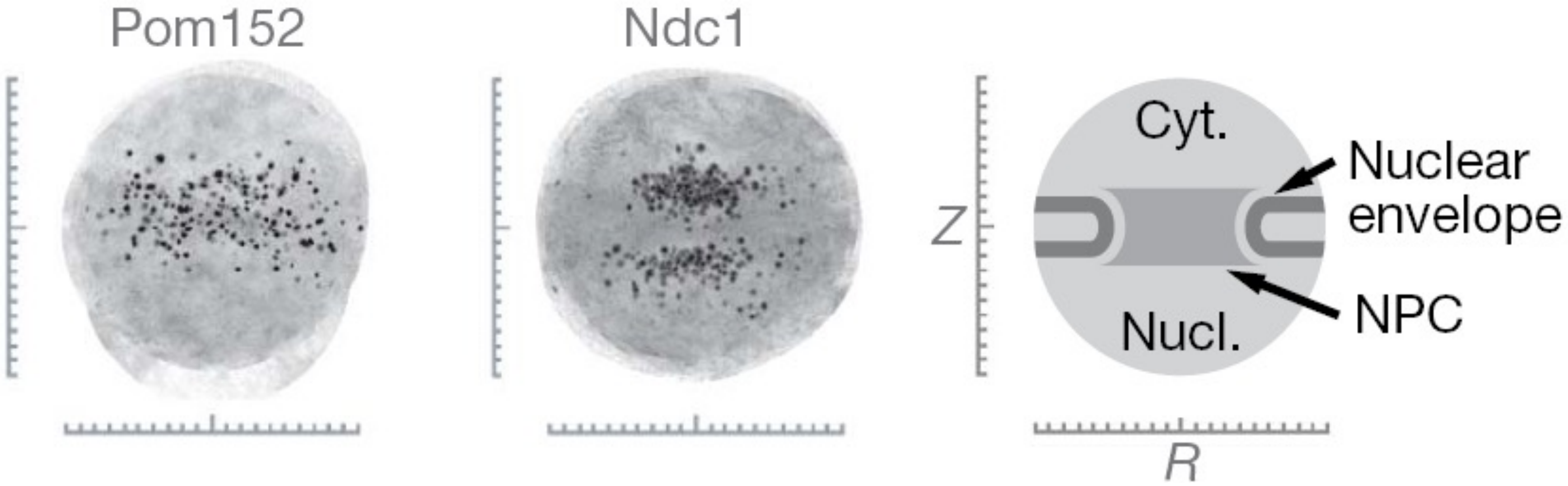


c

Protein	Molecular mass (kDa) (protein +PrA)	S_{obs}	S_{max}/S_{obs}	Stoichiometry (cyt.)	Stoichiometry (nucl.)	Bead number	Bead representation	Bead radius (nm)
Nup192	217	9.1	1.4	1	1	2	●●	3.0
Nup188	214	9.3	1.4	1	1	2	●●	3.0
Nup170	195	9.0	1.4	1	1	2	●●	2.9
Nup159	185	5.1	2.3	1	0	11	●●●●●●●●●●●	1.6
Nup157	183	7.1	1.6	1	1	3	●●●	2.5
Pom152	178	-	-	1	1	10	●●●●●●●●●●	1.6
Pom152*	155	4.8	2.2					
Nup133	159	7.6	1.4	1	1	2	●●	2.7
Nup120	146	8.0	1.3	1	1	2	●●	2.6
Nup116*	100	3.6	2.2	-	-	13	●●●●●●●●●●●●●	1.3
Nup1	140	-	-	0	1	9	●●●●●●●●●	1.5
Nup1*	120	3.7	2.4					
Nup100	126	4.4	2.0	1	0	13	●●●●●●●●●●●●●	1.3
Nic96	122	6.3	1.4	2	2	2	●●	2.4
Nsp1	112	3.5	2.4	2	2	12	●●●●●●●●●●●●	1.3
Nup85/Seh1	150	6.8	1.5	1	1	3	●●●	2.0
Nup85	111	-	-	1	1			
Nup84	110	4.9	1.7	1	1	3	●●●	2.0
Nup82	108	5.3	1.5	2	0	2	●●	2.3
Nup145C	107	6.7	1.2	1	1	2	●●	2.3
Ndc1	100	-	-	1	1	2	●●	2.2
Gle1	88	5.9	1.2	1	0	2	●●	2.1
Nup60	85	3.8	1.8	0	1	4	●●●●	1.6
Nup59	85	4.2	1.7	1	1	4	●●●●	1.6
Nup57	83	4.1	1.7	1	1	3	●●●	1.8
Nup53	79	4.1	1.6	1	1	3	●●●	1.7
Nup145N	86	3.7	1.9	0	2	6	●●●●●●	1.5
Nup49	75	3.9	1.6	1	1	3	●●●	1.7
Nup42	69	3.0	2.0	1	0	5	●●●●●	1.4
Gle2	66	4.6	1.3	1	1	1	●	2.3
Seh1	65	3.4	1.7	1	1	1	●	2.2
Pom34	60	-	-	1	1	3	●●●	1.5
Sec13	59	4.2	1.3	1	1	1	●	2.1
Complex 30	357	6.7	1.2	-	-	-	-	-
Complex 45	468	10	2.2	-	-	-	-	-

Localization of each component in the NPC

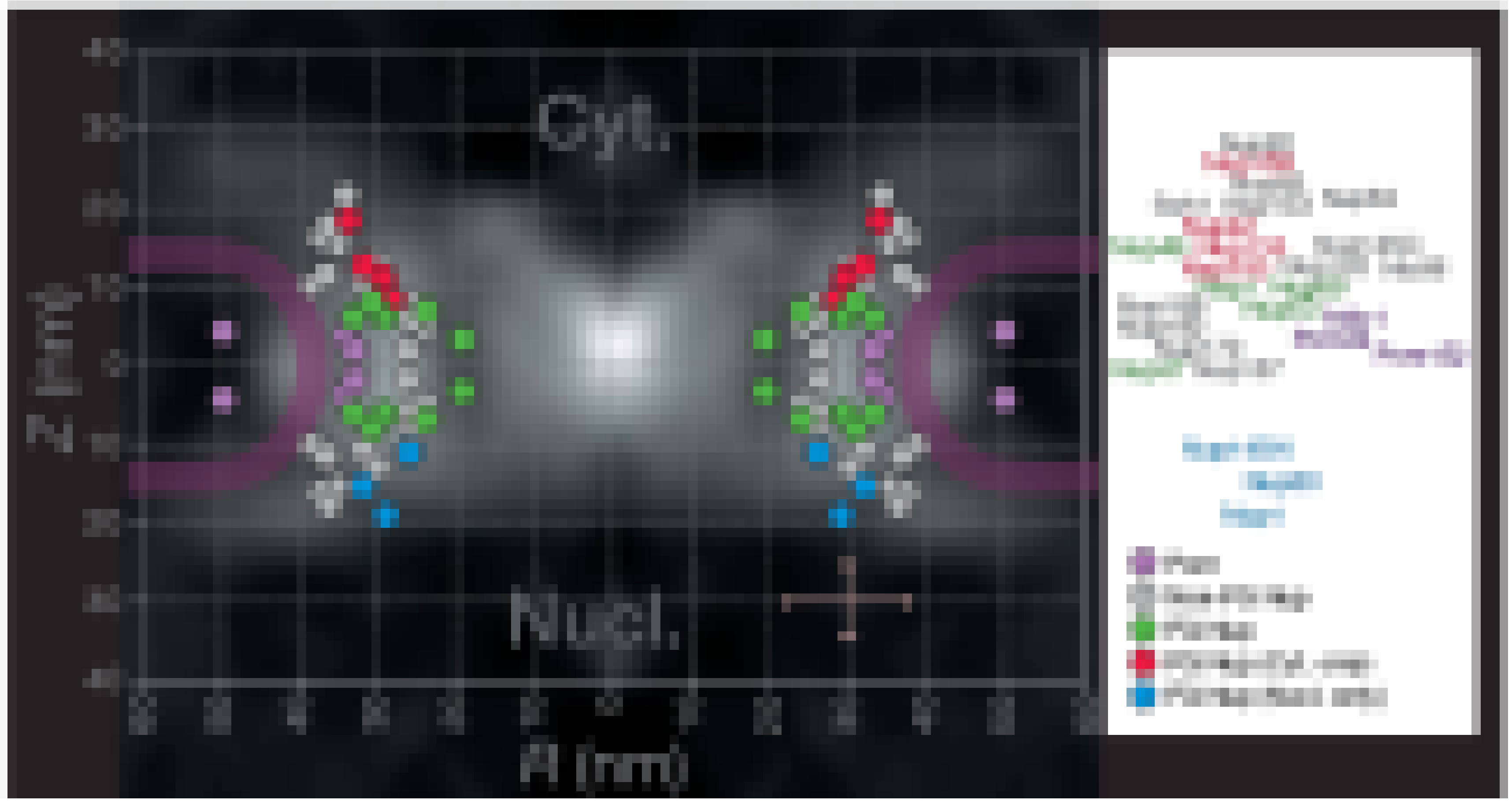
The coarse localization of most nucleoporins within the NPC was obtained by **immuno-EM**, relying on a gold-labelled antibody that specifically interacted with the localized protein through its carboxy-terminal PrA tag.



Immuno-EM montages for Pom152–PrA nuclei and Ndc1–PrA nuclear envelopes.

Right: the position of every gold particle in each montage was measured from both the central Z-axis of the NPC (R) and from the equatorial plane of the nuclear envelope (Z).

Localization of each component in the NPC



Estimated position of the C terminus of each protein in the NPC relative to the central Z-axis of the NPC (R) and the equatorial plane (Z) superimposed on the protein density map of a cross-section of the yeast NPC obtained by cryo-EM.

How do the NPC components fit together?

The coarse shape, approximate position and stoichiometry of each nucleoporin are not enough to build an accurate picture of the NPC.
Like the pieces in a jigsaw puzzle, we also need information about **physical interactions** between nucleoporins.

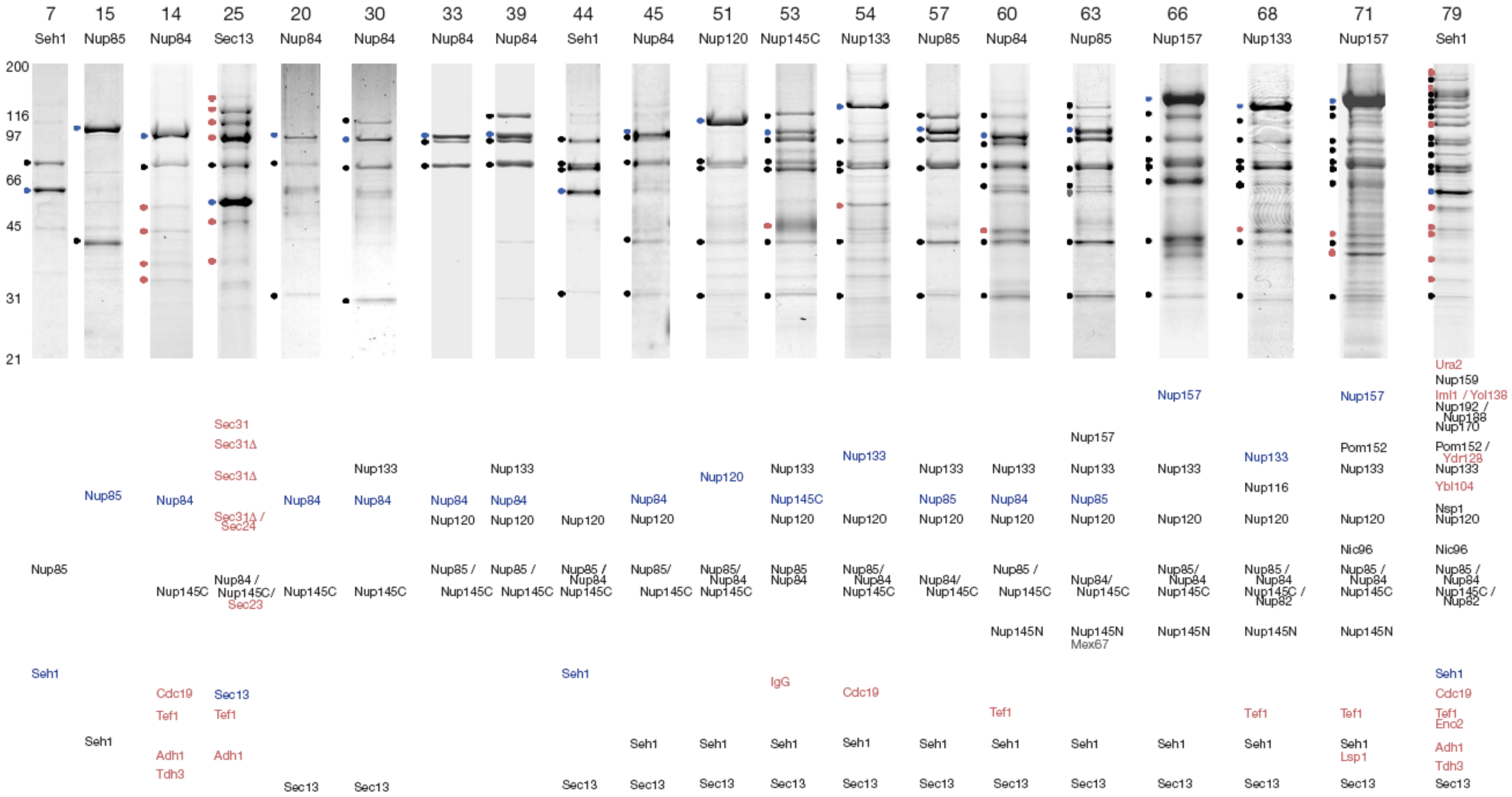
→ Obtain information about interactions from overlay assays and **affinity purification** experiments, as well as from the composition of the Pom rings (consisting of Pom34 and Pom152).

Overlay assay: identifies a pair of proteins that interact with each other.
Affinity purification: identifies one or more proteins that interact directly or indirectly with the bait protein.

An affinity purification produces a distinctive set of co-isolating proteins, which we term a **composite**.
A composite may represent a single complex of physically interacting proteins or a mixture of such complexes overlapping at least at the tagged protein.

Protein interactions of the Nup84 complex

aabove lanes: name of PrA-tagged protein and identification number for composite



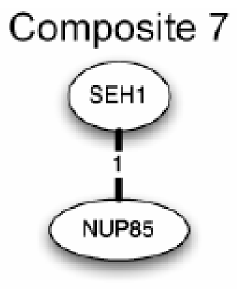
Below each lane: identity of co-purifying proteins; PrA-tagged proteins are **blue**, co-purifying nucleoporins black, NPC-associated proteins **grey**, and other proteins (e.g. contaminants) **red**. Affinity-purified PrA-tagged proteins and interacting proteins were resolved by SDS–PAGE and visualized with Coomassie blue. Molecular mass standards (kDa) are indicated to the left of the panel. The bands marked by filled circles at the left of the gel lanes were identified by mass spectrometry.

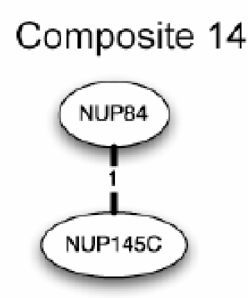
Localization of each component in the NPC

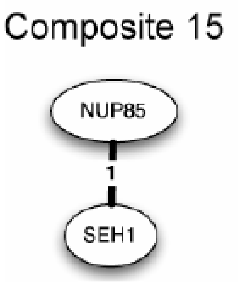
A good example of the compositional overlap is the Nup84 complex.

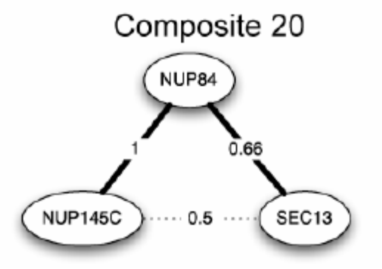
The smallest building blocks of this complex are **heterodimers** (composites 7, 14, 15).

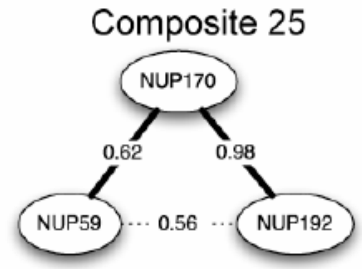
Under different isolation conditions, these dimers can be purified with an increasing number of additional proteins, such as **trimers** (25, 20), a **tetramer** (33), a **pentamer** (39), **hexamers** (44, 45, 51), and the full **septameric** Nup84 complex (53, 54, 57).

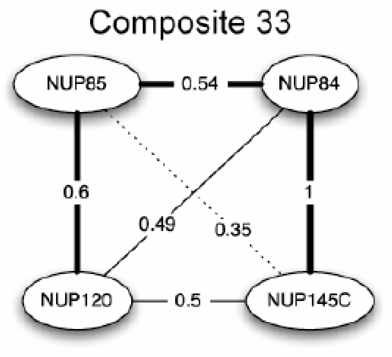






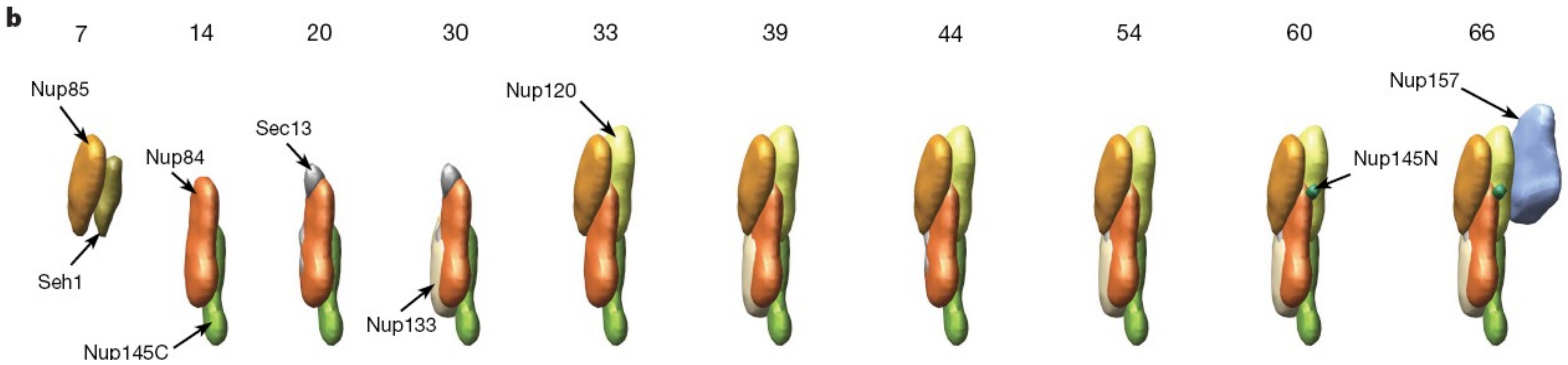


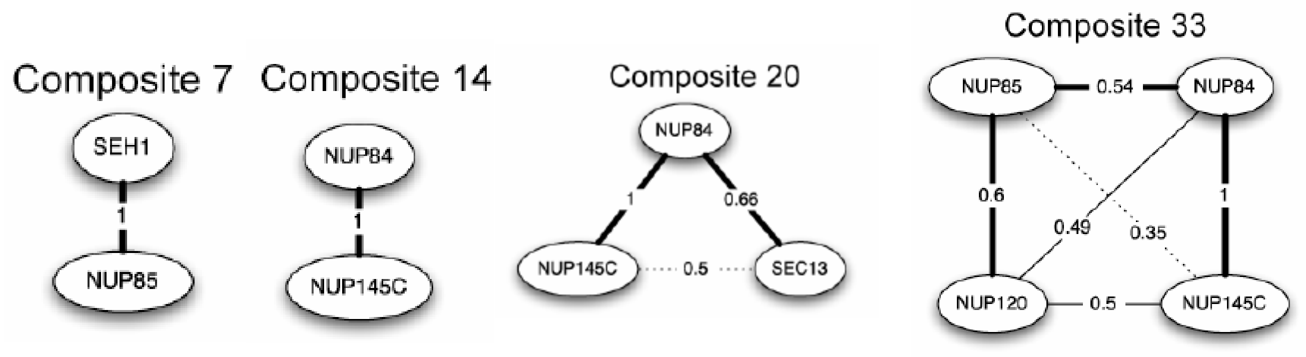




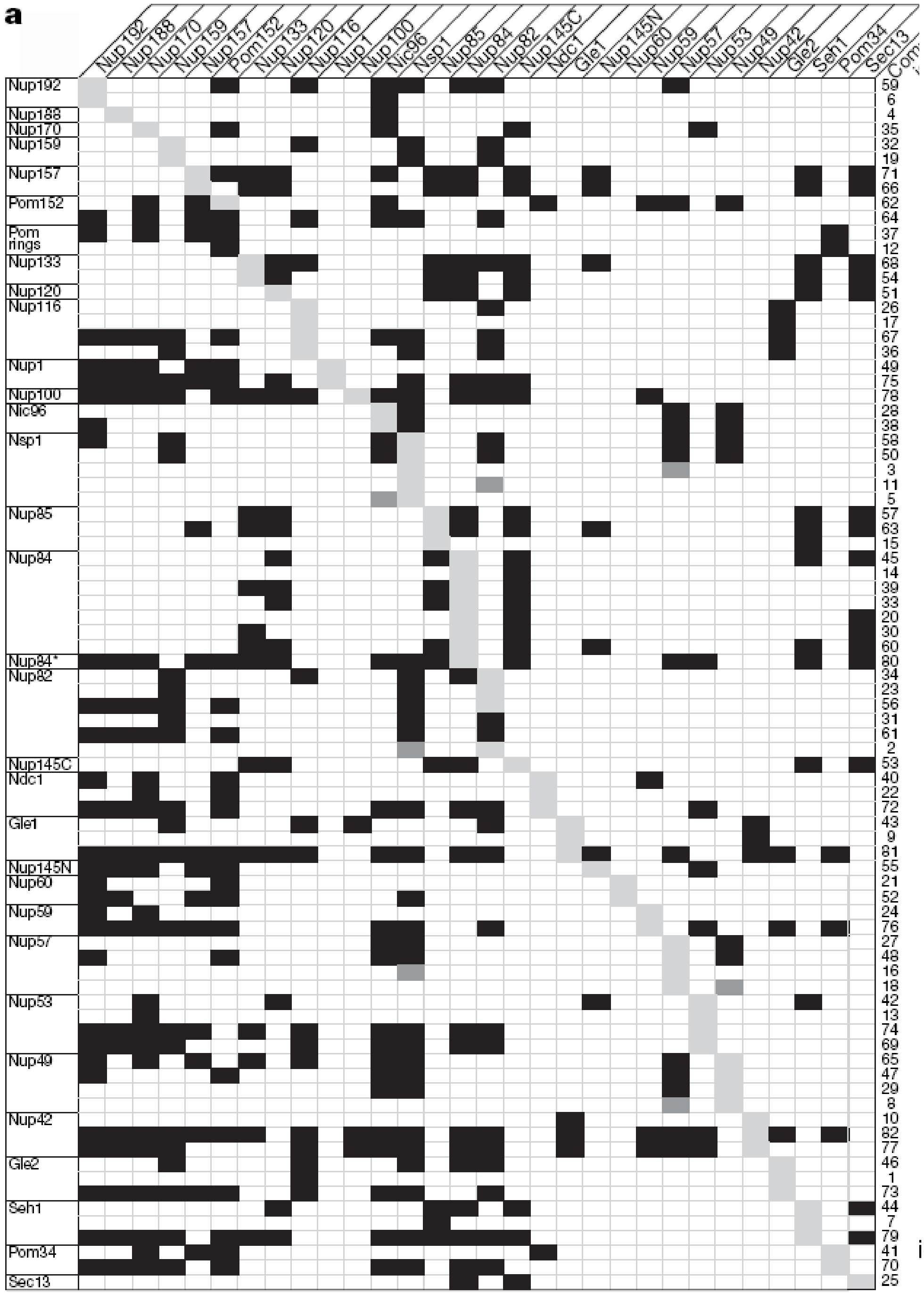
Protein interactions of the Nup84 complex

The mutual arrangement of the Nup84-complex-associated proteins as visualized by their localization volumes in the final NPC structure.





Bioinformatics III



Protein proximity by affinity purification + Composites determined by affinity purification.

Vertical: affinity-purified nucleoporin–PrA

Horizontal: corresponding nucleoporins in each composite.

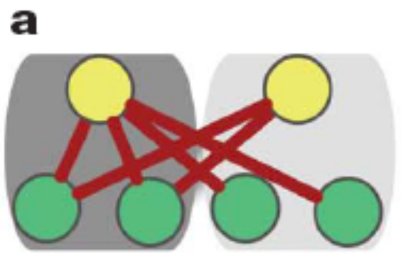
Right: composite identifiers

Black box: indicates presence of a nucleoporin in a composite,

Grey box: tagged nucleoporin.

Dark grey box: a direct interaction determined by overlay assay.

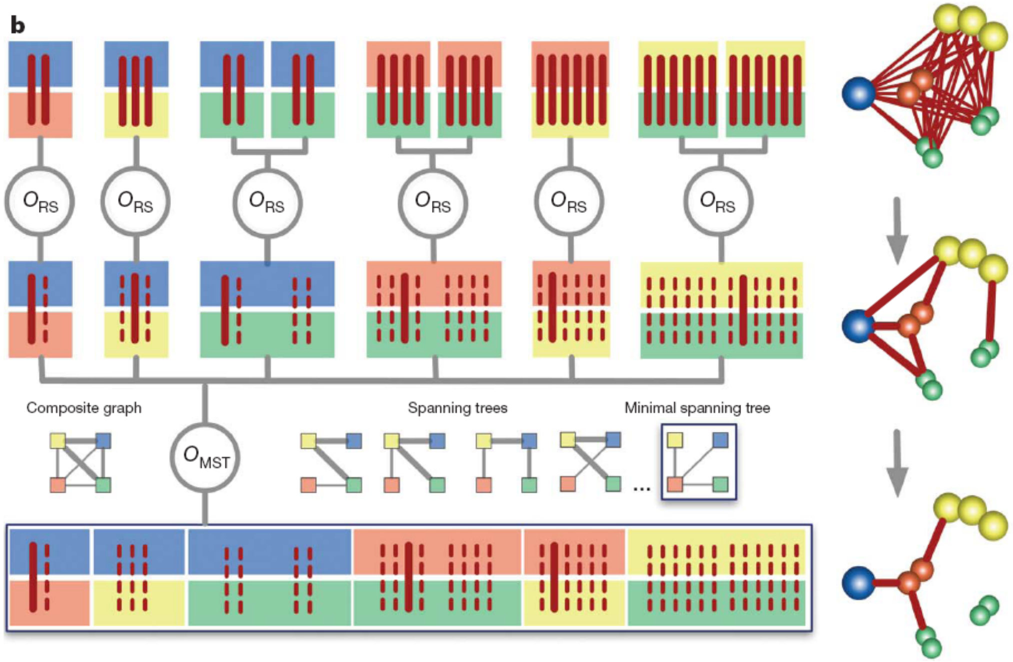
Ambiguity in data interpretation and conditional restraints



Shown is the **ambiguity** for a protein interaction between proteins of green and yellow types.

The ambiguity results from the presence of **multiple copies** of the same protein in the same or neighbouring symmetry unit.

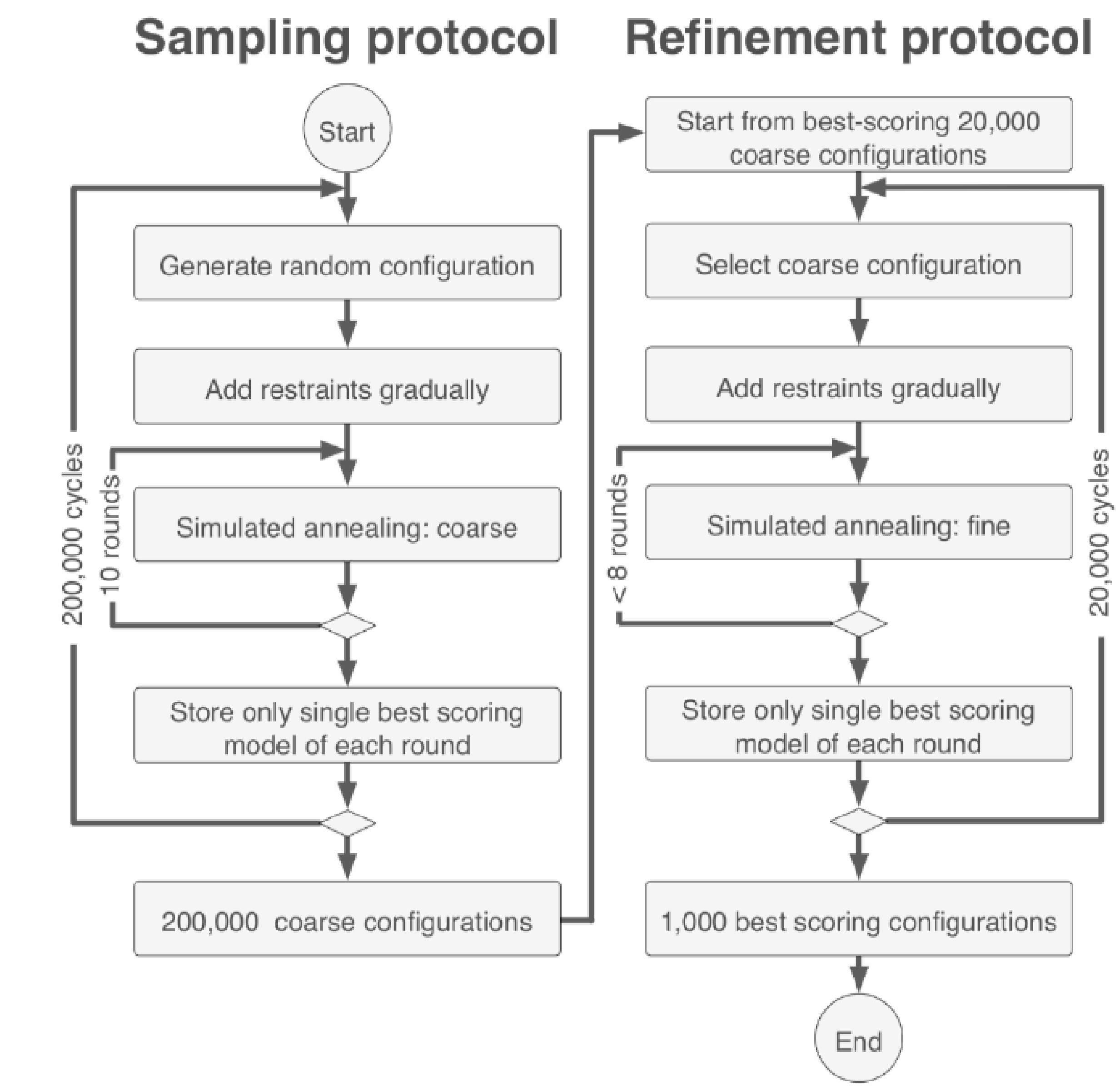
In our NPC calculations, both neighbouring half-spokes on the cytoplasmic and nucleoplasmic sides are considered, for a total of four neighbouring half-spokes.



b, The conditional restraint is illustrated by an example of a **composite of four protein types** (yellow, blue, red, green), derived from an assembly containing a single copy of the yellow, blue, and red protein and two copies of the green protein; proteins are represented by a single bead (blue protein), a pair of beads (green and red proteins), and a string of three beads (yellow protein) (right panel).

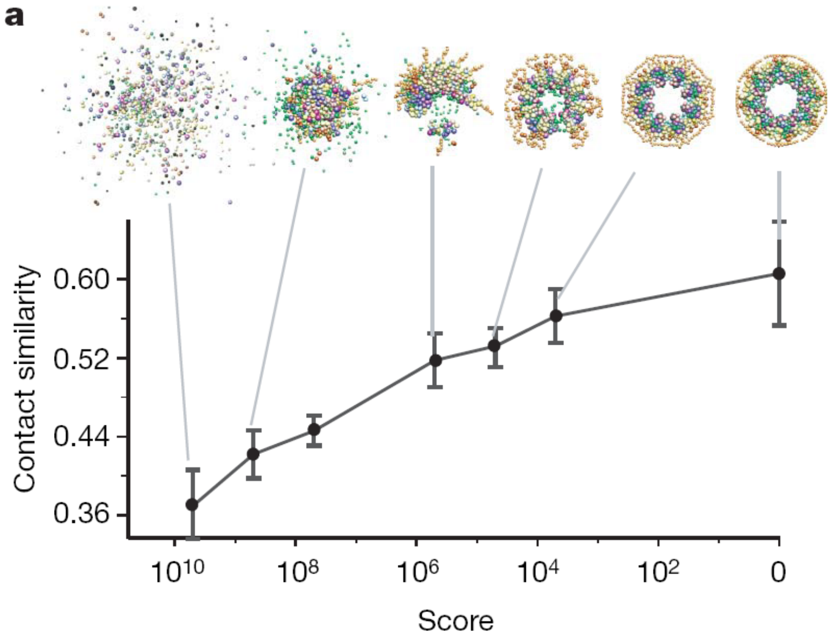
This composite implies that at least 3 of the following 6 possible types of interaction must occur: blue–red, blue–yellow, blue–green, red–green, red–yellow and yellow–green. In addition, (1) the 3 selected interactions must form a '**spanning tree**' of the 'composite graph'; (2) each type of interaction can involve either copy of the green protein; and (3) each protein can interact through any of its beads. These considerations can be encoded through a tree-like evaluation of the conditional restraint. At the top level, all optional bead–bead interactions between all protein copies are clustered by protein types. Each alternative bead interaction is restrained by a harmonic upper bound on the distance between the beads; these are 'optional restraints', because only a subset is selected for contribution to the final value of the conditional restraint. Next, a 'rank-and-select' operator (O_{RS}) selects only the **least violated optional restraint** from each interaction type, resulting in six restraints (thick red line) at the middle level of the tree. Finally, the minimal spanning tree operator (O_{MST}) finds the combination of 3 restraints that are most consistent with the composite data (thick red line); here the edge weights in the minimal spanning tree correspond to the restraint values given the current assembly structure. The column on the right shows a structural interpretation of the composite with proteins represented by their coloured beads and alternative interactions indicated by edges between them. The composite graph (left) is a fully connected graph that consists of nodes for all identified protein types and edges for all pairwise interactions between protein types; in the context of the conditional restraint, the edge weights correspond to the restraint values. 5 of the 16 possible spanning trees are also shown. This restraint evaluation process is executed at each optimization step based on the current configuration, thus resulting in possibly different subsets of selected optional restraints at each step.

Simulation protocol


Sampling protocol
Start
Generate random configuration
Add restraints gradually
Simulated annealing: coarse
Store only single best scoring model of each round
200,000 coarse configurations
200,000 cycles
10 rounds
Refinement protocol
Start from best-scoring 20,000 coarse configurations
Select coarse configuration
Add restraints gradually
Simulated annealing: fine
Store only single best scoring model of each round
1,000 best scoring configurations
End
< 8 rounds
20,000 cycles

Calculation of the NPC bead structure by satisfaction of spatial restraints

The contact similarity quantifies how similar 2 configurations are in terms of the number and types of their protein contacts; a contact between two proteins occurs if the distance between their closest beads is less than 1.4 times the sum of the bead radii.

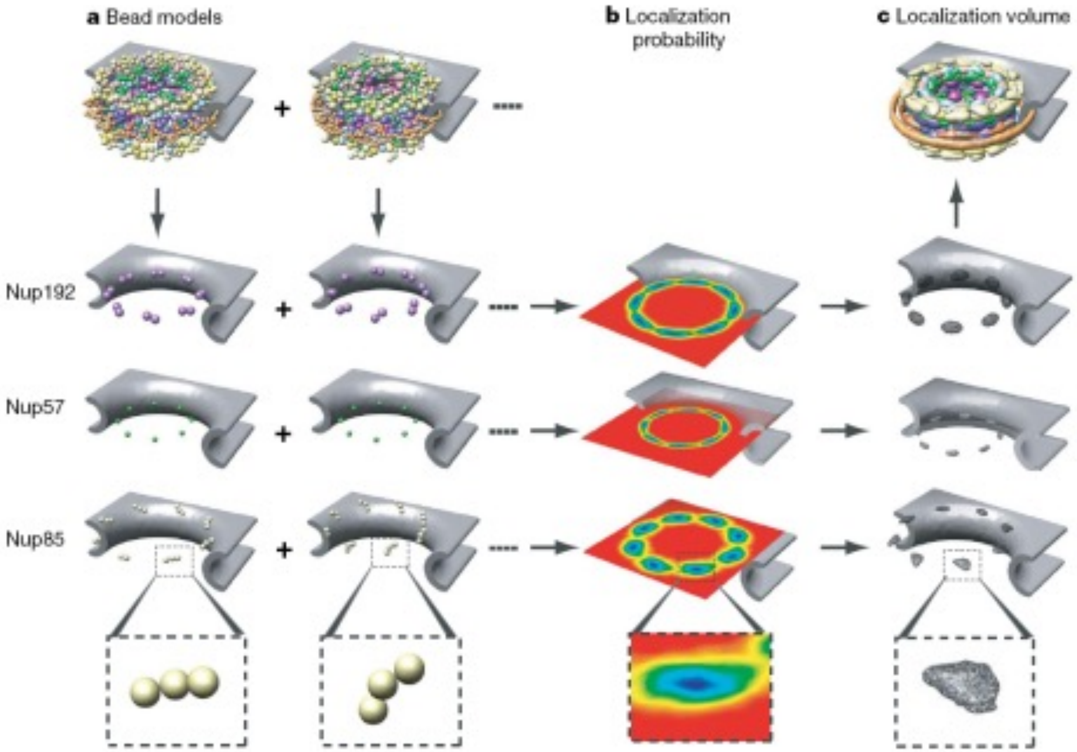


Representative configurations at various stages of the optimization process from left (very large scores) to right (with a score of 0); a score of 0 indicates that all input restraints have been satisfied.

Representation of the optimization process as it progresses from an initial random configuration to an optimal structure.

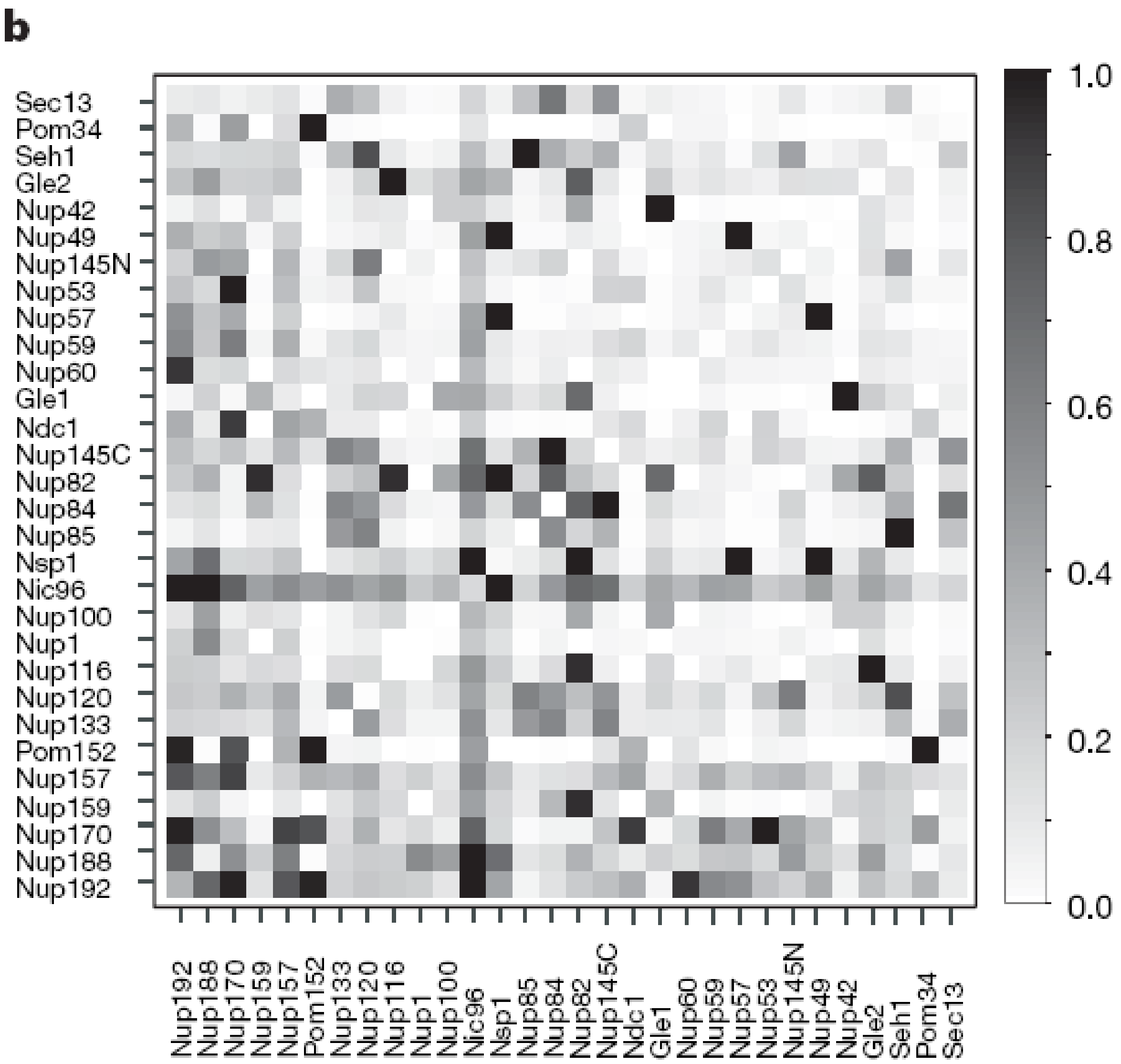
As the score approaches zero, the contact similarity increases, showing that there is only a single cluster of closely related configurations that satisfy the input data.

Final bead models



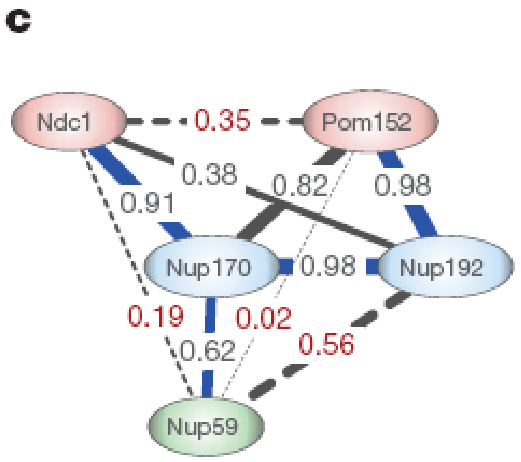
a, Top: 2 representative bead models of the NPC (excluding the FG-repeat regions) from the ensemble of 1,000 superposed structures satisfying all restraints. The 8 positions of 3 sample proteins (Nup192, Nup57 and Nup85) on the cytoplasmic side are shown, with a detailed view of the bead representation of 1 copy of Nup85 at the bottom.

Contact frequencies for all pairs of proteins



Contact frequency : fraction of structures in the ensemble that contains at least one protein contact between any protein instances of the 2 types.

Contact frequencies between proteins in composite 40



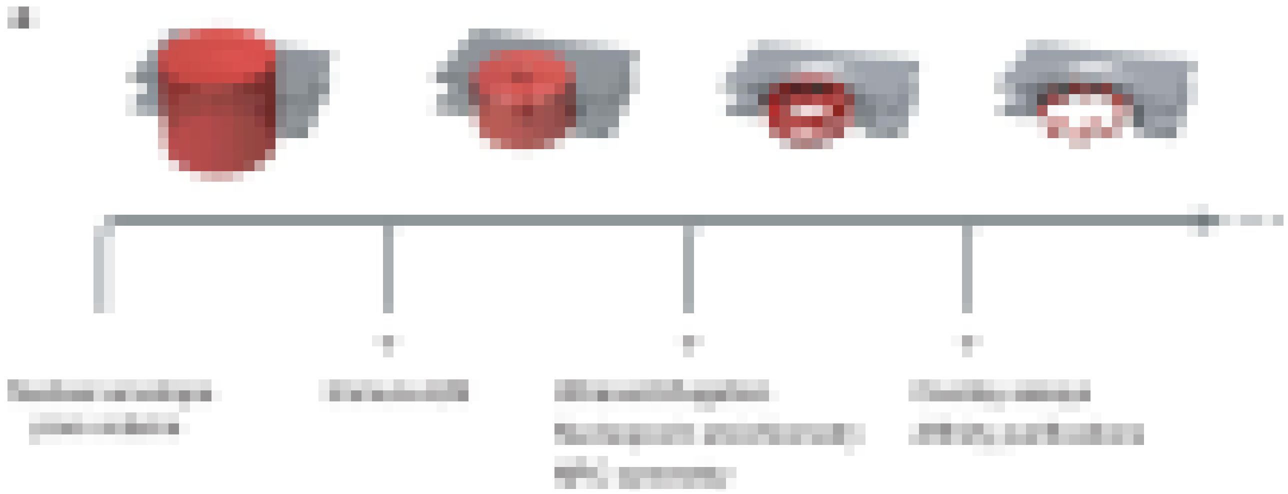
Proteins are nodes connected by edges with the observed contact frequency as the edge weight (indicated by its thickness).

Edges that are part of the maximal spanning tree are shown by thick blue lines.

The maximal spanning tree is the spanning tree that maximizes the sum of the edge weights.

Dotted red lines: edges with a statistically significant reduction in contact frequency from their initial values implied by the composite data alone (P-value $<10^{-3}$).

Bead model, ensemble, localization probability

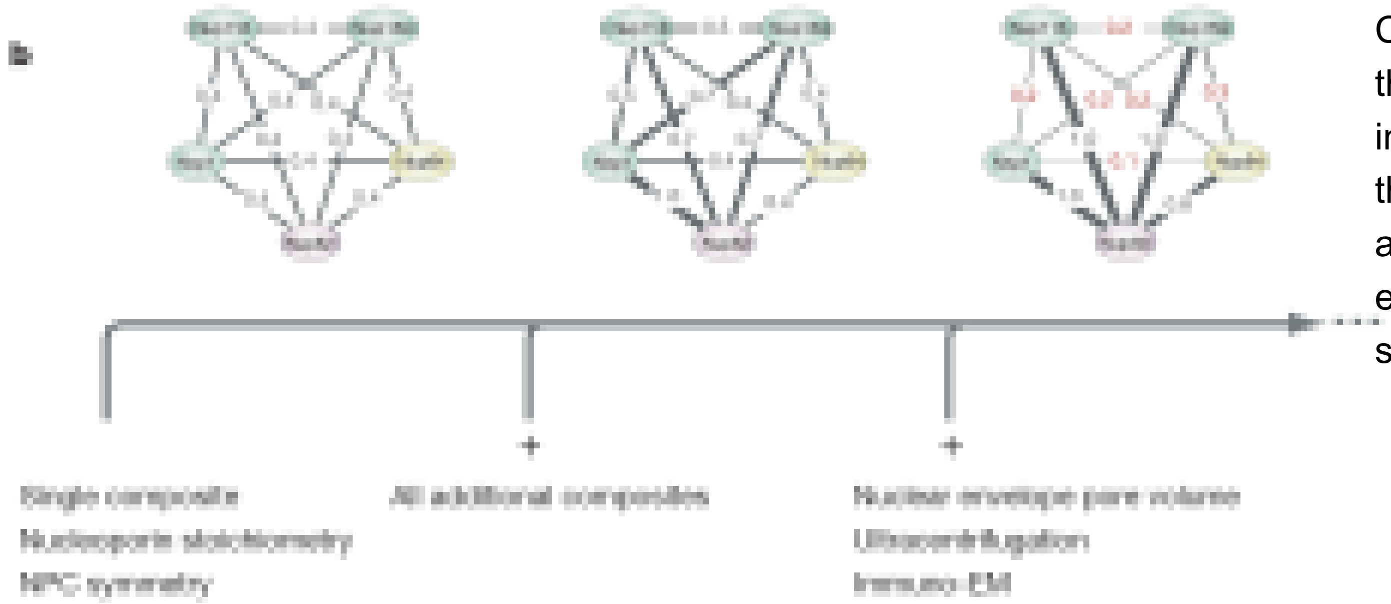


The structure is increasingly specified by the addition of different types of synergistic experimental information.

As an example, each panel illustrates the localization of 16 copies of Nup192 in the ensemble of NPC structures, generated using the data sets indicated below.

The localization probability is contoured at 65% of ist maximal value (red). The smaller the volume, the better localized are the proteins.

Protein contacts


b
+
+
Single composite
Nucleoporin stoichiometry
NPC symmetry
All additional composites
Nuclear envelope pore volume
Ultracentrifugation
Immuno-EM

Contact frequencies reflect the likelihood that a protein interaction is formed given the data considered and are calculated from the ensemble of optimized structures.

Prediction of protein interactions from contact frequencies improves as more data are used. This figure shows as an example the contact frequencies between proteins found in composite 34. Contact frequencies are shown as edge weights and indicated by the thickness of the lines connecting the proteins.

Left: only a single composite is used (together with stoichiometry and symmetry information), all interactions are equally likely.

Middle: the highest likelihood of interaction between a particular protein pair from all composites is used, the uncertainty about the interactions is reduced.

Right: all data are used, the contact frequencies are either very high (>0.65) or very low (<0.25), thus allowing a strong prediction of protein interactions. As before, numbers in red indicate final contact frequencies that significantly decreased (at a P-value $<10^{-3}$) from their initial values.

Evaluation by experimental data not used sofar

Finally, the structure was tested by comparing it to experimental data that were not included in the structure calculation.

1 **omission of a randomly chosen subset** of 10% of the protein interaction data still results in structures with contact frequencies essentially identical to those derived from the complete data set → the structure is **robust**.

2 the **shape** of our NPC structure strongly resembles the published EM maps of the NPC, even though these data were not used here.

3 the **diameter** of the transport channel in our structure is ca. 38 nm (excluding the FG-repeat regions), in good agreement with the experimentally reported maximal diameter of transported particles.

4 **Nup133**, which has been experimentally shown to interact with highly curved membranes via its ALPS-like motif, is **adjacent** to the nuclear envelope in our structure.

5 Our configuration for **Nup84** complex is completely consistent with previous results.

Evaluation by experimental data not used sofar

Frank Alber:

„Together these assessments indicate that our data are sufficient to determine the configuration of the proteins comprising the NPC.

Indeed, it is hard to conceive of any combination of errors that could have biased our structure towards a single solution that resembles known NPC features in so many ways."

Conclusions

Integrative approach to solve the structure of an extremely large supracomplex using diverse biophysical and proteomic data. Advantages:

1 it benefits from the synergy among the input data. **Data integration** is in fact necessary for structure determination. None of the individual data sets contains sufficient spatial information on its own.

2 the integrative approach can potentially survey all the structures that are consistent with the data. Alternatively, if no structure is consistent with the data, then some experiments or their interpretations are incorrect.

3 this approach can make the process of structure determination more efficient, by indicating which measurements would be most informative.

4 the approach can, in principle, incorporate essentially any structural information about a given assembly. Thus, it is straightforward to adapt it for calculating higher resolution structures by including additional spatial restraints from higher resolution data sets, such as atomic structures of proteins, chemical crosslinking, footprinting, small angle X-ray scattering (SAXS) and cryo-EM.

Additional slides: Shape and size of each component

No atomic structures for most nucleoporins,
→ estimate their **shapes** based on their **sedimentation coefficients** determined by ultracentrifugation of the purified proteins.

Good: the sedimentation behaviour of most FG nucleoporins agrees with their predicted filamentous, native disordered structure.

Good: Pom152, an integral membrane component, with multiple domains modelled as b-cadherin-like folds, appeared to be a highly elongated structure, consistent.

Most of the other nucleoporins appear to have a relatively compact 3D structure in agreement with their predicted fold assignments.

Overview of integrative structure determination

Our approach to structure determination can be seen as an iterative series of 4 steps:

- data generation by experiment,
- translation of the data into spatial restraints,
- calculation of an ensemble of structures by satisfaction of these restraints, and
- an analysis of the ensemble to produce the final structure.

The structure calculation part of this process is expressed as an optimization problem, a solution of which requires three main components:

(1) a **representation** of the assembly in terms of its constituent parts;

(2) a **scoring function**, consisting of individual spatial restraints that encode all the data; and

(3) an **optimization** of the scoring function, which aims to yield structures that satisfy the restraints.

Protein representation

Every protein P is represented as a set of **beads** B, each with associated attributes (e.g., radius, mass).
The number of beads and their attributes determine the **resolution** (granularity) of the protein representation.

The most detailed data about the shape of most nups come from hydrodynamic experiments → approximate the coarse shape and volume of each protein with a linear chain of equally-sized **beads** that best reproduce the observed **sedimentation coefficients** and are also consistent with our 3D fold assignments.

Protein conformations in the NPC may differ from their conformations in solution. Therefore each protein is represented as a flexible chain, to allow for maximally extended to maximally compact conformations.

The bead chain describes a protein at the highest resolution in our representation (the “root” representation $\kappa = 1$).

Figure legend: Optimization in two stages

First, a coarse sampling protocol (left column) generates 200,000 coarse configurations, starting each time from a different random configuration.

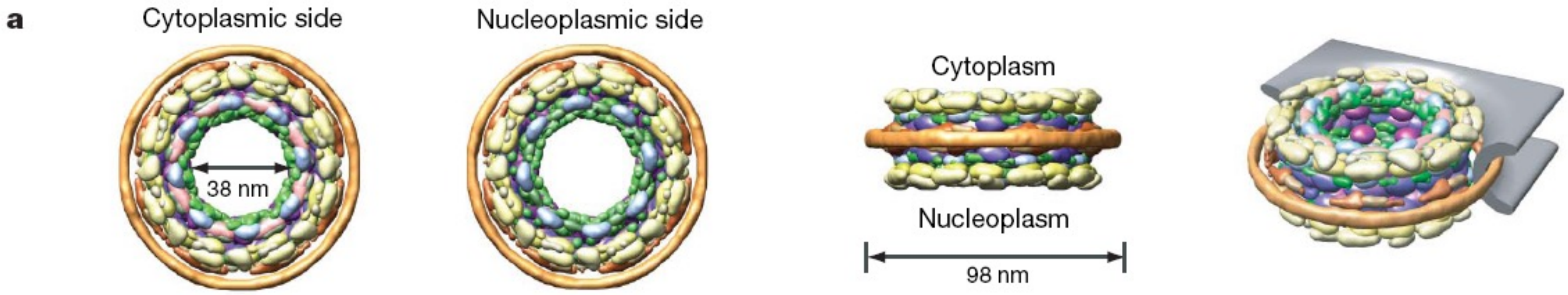
This protocol relies on a variable target function method that consists of gradually increasing the number of restraints that are included in the scoring function, finally culminating in the full scoring function F.

At each stage of the variable target function method, a combination of the conjugate gradient (CG) minimization and a molecular dynamics (MD) simulation with simulated annealing is applied.

In total, a single optimization of an initial random configuration consists of an iteration of approximately 10.000 small shifts of protein particles (guided by either CG or MD).

Second, a refinement protocol (right column) further refines the best 10% configurations from the sampling stage.

localization volumes



Ensemble interpretation in terms of protein positions, contacts and configuration.

a, Localization volumes of all 456 proteins in the NPC (excluding the FG-repeat regions) in 4 different views.