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]

To understand the workings of a living cell, we need to know the architectures of its macromolecular assemblies. Here we show how proteomic data can be used to determine such structures. The process involves the collection of sufficient and diverse high-quality data, translation of these data into spatial restraints, and an optimization that uses the restraints to generate an ensemble of structures consistent with the data. Analysis of the ensemble produces a detailed architectural map of the assembly. We developed our approach on a challenging model system, the nuclear pore complex (NPC). The NPC acts as a dynamic barrier, controlling access to and from the nucleus, and in yeast is a 50 MDa assembly of 456 proteins. The resulting structure, presented in an accompanying paper, reveals the configuration of the proteins in the NPC, providing insights into its evolution and architectural principles. The present approach should be applicable to many other macromolecular assemblies.

A mechanistic understanding of the cell requires the structural characterization of the thousands of its constituent biological assemblies[1]. So far, conventional approaches have provided a valuable but limited window into the structures of these assemblies. For example, X-ray crystallography and nuclear magnetic resonance (NMR) spectroscopy can resolve the atomic details of individual proteins and small complexes, whereas electron microscopy produces morphological maps but can lack the ability to identify and detail specific components in the map of the whole assembly. As a result, we do not yet have atomic-resolution structures, or even low-resolution representations, for the vast majority of complexes in the cell. How, then, are we to resolve the molecular architectures of these assemblies?

In an attempt to address this problem, we have taken the yeast (*Saccharomyces cerevisiae*) nuclear pore complex (NPC) as a case in point. The NPC is among the largest macromolecular assemblies in the cell, mediating the exchange of molecules that pass between the nuclear and cytoplasmic compartments. Yeast NPCs are ~50 MDa structures built of multiple copies of some 30 different proteins (nucleoporins), totalling at least 456 protein molecules[2]. Each NPC is a plastic structure embedded in the nuclear envelope and is composed of eight morphologically similar 'spokes' surrounding a central tube[3–6]. Filling this tube and projecting into both the cytoplasmic and nuclear sides are flexible filamentous domains from proteins termed FG (phenylalanine-glycine) repeat nucleoporins; these domains form the docking sites for transport factors that carry macromolecular cargoes through the NPC.

The NPC represents a significant challenge for conventional structure determination approaches owing to its large size and the high degree of flexibility of the complex and its components. Thus, although electron microscopy has provided valuable insights into the overall shape of the NPC, its molecular architecture (that is, the spatial configuration of its component proteins) has yet to be revealed, and atomic structures have only been solved for domains covering ~5% of its component protein sequences[7]. The NPC therefore encapsulates many of the obstacles that will be encountered in the detailed structural examination of other macromolecular assemblies.

We describe here a set of proteomics experiments and a computational platform for converting the resulting data into the structures of macromolecular assemblies. Central to this approach is the realization that many kinds of biophysical and proteomic data contain valuable structural information about assemblies.

Overview of integrative structure determination

Our approach to structure determination can be seen as an iterative series of four 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 (Fig. 1). 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.

Formally, our approach is similar to the determination of protein structures by NMR spectroscopy, in which the folding of the polypeptide chain is determined by satisfying distance restraints between pairs of atoms[8]. As with NMR spectroscopy, a structure is computationally determined from experimental data. Here, atoms

[1]Department of Bioengineering and Therapeutic Sciences, Department of Pharmaceutical Chemistry, and California Institute for Quantitative Biosciences, Byers Hall, Suite 503B, 1700 4th Street, University of California at San Francisco, San Francisco, California 94158-2330, USA. [2]Laboratory of Cellular and Structural Biology, and [3]Laboratory of Mass Spectrometry and Gaseous Ion Chemistry, The Rockefeller University, 1230 York Avenue, New York, New York 10065, USA. †Present addresses: Laboratory of Nucleocytoplasmic Transport, Institut Jacques Monod, 2 place Jussieu, Tour 43, Paris 75251, France (S.D.); Department of Biochemistry, University of Groningen, Nijenborgh 4, 9747 AG Groningen, The Netherlands (L.M.V.); German Aerospace Center (PT-DLR), Heinrich-Konen-Strasse 1, D-53227 Bonn, Germany (J.K.); Structural Bioinformatics, EMBL, Meyerhofstrasse 1, D-69117 Heidelberg, Germany (D.D.); Office of Technology Transfer, The Rockefeller University, 1230 York Avenue, New York, New York 10065, USA (A.S.); Herbert Irving Comprehensive Cancer Center, Columbia University, 1130 St Nicholas Avenue, New York, New York 10032, USA (O.K.-S.).
*These authors contributed equally to this work.

are replaced by proteins, and their positions and relative proximities are restrained on the basis of data from a variety of proteomics and other experiments, including affinity purification, ultracentrifugation, electron microscopy and immuno-electron microscopy (immuno-EM).

Data generation. The most important aspect of our approach is its potential to use simultaneously almost any conceivable type of information to determine assembly structures. For example, sedimentation analysis of the isolated proteins can be used to infer their shapes; immuno-EM can give an approximate localization of each protein in the assembly; and affinity purification of tagged proteins and protein complexes can yield information about the arrangement and interactions of proteins within the assembly. These data can be of a kind not normally used for structure determination (for example, complexes identified by affinity purification), can refer to different levels in the structural hierarchy (for example, a protein domain, a whole protein, or a protein complex), and can be ambiguous in terms of their structural interpretation (for example, the uncertainty as to which copy of the protein is involved in an interaction, when multiple copies exist).

The use of such data for structure determination presented us with four major challenges. First, large amounts of suitable data must be collected to give sufficient spatial information to define structures; fortunately, the proteomic revolution has provided methodologies that allow us to garner enough information. Second, much of the data can be of relatively low precision; thus, to avoid over-interpretation, appropriate tolerances must be used in its structural interpretation. Third, the possibility of false-positive data must be minimized and taken into consideration. Fourth, ambiguity of the data in terms of its structural interpretation must be treated when multiple copies of the same protein are present in an assembly and the experiment does not determine which specific instance of a protein is detected. All of these challenges can be addressed by an integrative approach that incorporates information varying greatly in terms of its accuracy and precision; limitations of any particular type of data can be overcome by the use of large and diverse data sets derived from synergistic experimental methods[1,9].

Data translation into spatial restraints. The data can be used to restrain many different features of the assembly, such as the positions of proteins, protein contacts, proximity between proteins, and the shape and symmetry of the whole assembly. A 'restraint' specifies values of the restrained feature that are consistent with the experimental information about it; a perfectly satisfied restraint is indicated here by 0, whereas values larger than 0 correspond to a violated restraint. Thus, a restraint encodes our uncertainty in the restrained feature. In essence, restraints can be thought of as generating a 'force' on each component in the assembly, to mould them into a configuration that satisfies the data used to define the restraints.

Optimization. All the restraints are summed to obtain a scoring function, which determines the degree of consistency between the restrained spatial features in a structure and the experimental information; a perfect structure is indicated by 0, reflecting the summed values of perfectly satisfied restraints, whereas values larger than 0 correspond to a structure that increasingly violates restraints. The scoring function is then optimized to calculate a structure that minimizes violations of the restraints. It is necessary to generate many such structures to provide a good sampling of structures that are consistent with the data (that is, the 'ensemble').

Ensemble analysis. All of the structures that satisfy the input restraints are clustered into distinct sets, on the basis of their similarities. There are three possible outcomes of such clustering. First, if only a single cluster of structures satisfies all the input information,

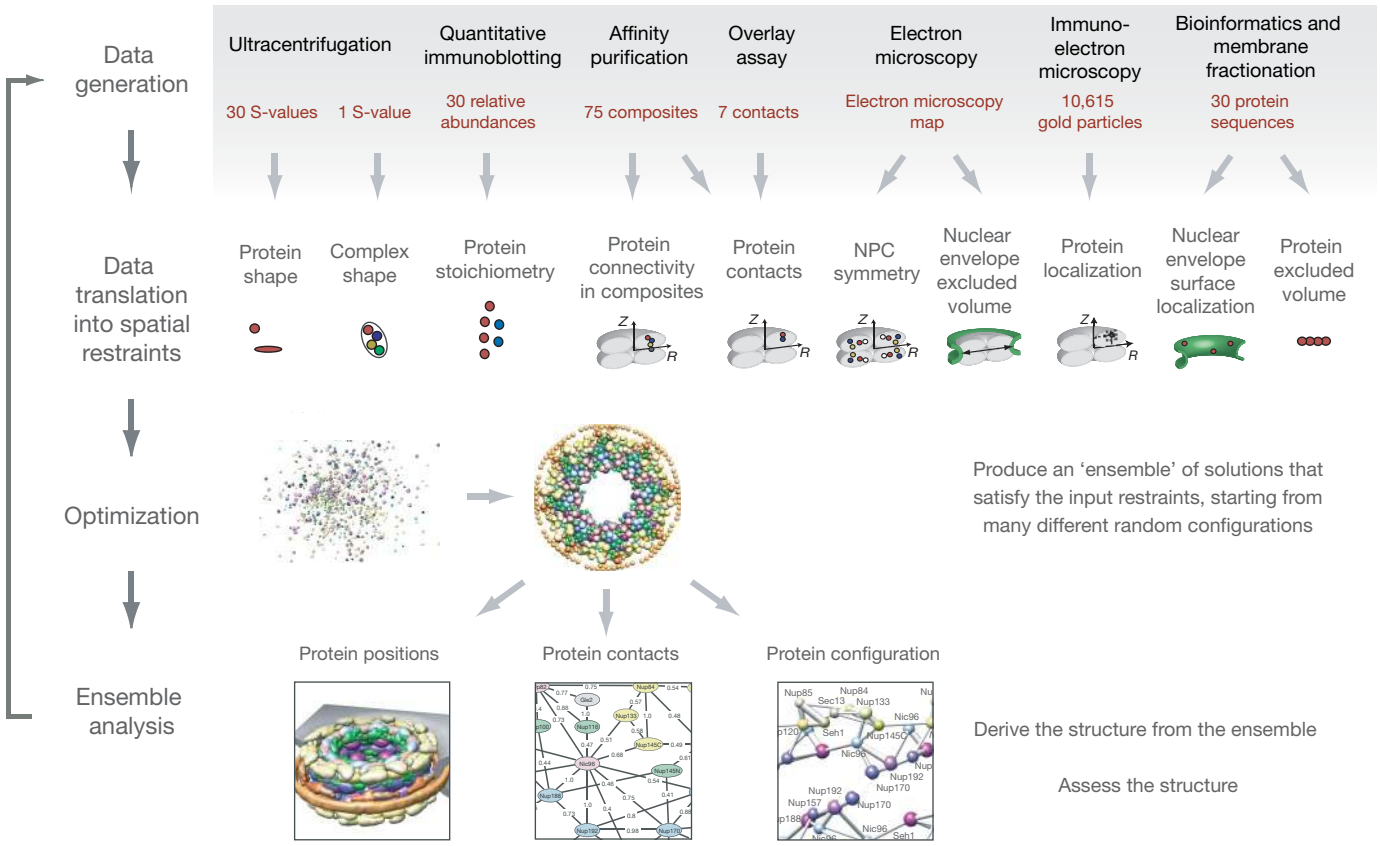


Figure 1 | Determining the architecture of the NPC by integrating spatial restraints from proteomic data. First, structural data (red) are generated by various experiments (black). Second, the data are translated into spatial restraints. Third, an ensemble of structural solutions that satisfy the data 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.

there is probably sufficient data for determining the unique native state. Second, if different clusters are consistent with the input information, either the data are insufficient to define the single native state or there are multiple native structures. If the number of clusters is small, the structural differences between them may suggest additional experiments so as to narrow down the possible solutions. Third, if no structures satisfy all input information, either the data or their interpretation in terms of the restraints is incorrect. Given the first two outcomes, the ensemble can be analysed to determine different aspects of the native state, such as protein positions, contacts and configuration. The variability of the ensemble provides an estimate of the precision of the structure determination.

We illustrate our approach by determining the configuration of the protein components in the NPC from the yeast *S. cerevisiae* (Fig. 1).

Data generation

As no single experimental technique has been sufficient to solve the molecular architecture of the NPC, we used a variety of techniques, each of which gave different and synergistic information about the structure; the techniques were chosen to generate the needed structural information with a defined level of accuracy.

An NPC component list. To determine any structure, we must first define its parts (Fig. 2). In the case of the NPC, we have already determined that some 30 nucleoporins constitute the assembly[2]. Although the exact composition is still uncertain because some proteins interact relatively transiently with the NPC, potential omission of a small fraction of such transient components is unlikely to interfere with structure determination.

The stoichiometry of each component in the NPC. The stoichiometry of each nucleoporin in each half-spoke has been previously established[2]. However, having found the stoichiometry of Nup82 to be ambiguous, we re-examined it with new strains and found that Nup82 is present in two copies per spoke (Fig. 3 and Supplementary Fig. 7).

The shape and size of each component. Next, we must represent the structures of the constituent nucleoporins. Because atomic structures have not yet been solved for most nucleoporins, we estimated their shapes based primarily on their sedimentation coefficients determined by ultracentrifugation of the purified proteins (Fig. 3 and Supplementary Information). The sedimentation behaviour of most FG nucleoporins agrees with their predicted filamentous, native disordered structure[10,11]. Pom152, an integral membrane component, appeared to be a highly elongated structure, consistent with its multiple domains modelled as β-cadherin-like folds[7]. Most of the other nucleoporins appear to have a relatively compact tertiary structure that is again in agreement with their predicted fold assignments[7,12]. The seven-protein Nup84 complex[13] could be separated into two smaller complexes on sedimentation: an elongated tetramer (composite 30, see below) and an elongated hexamer (composite 45, see below), consistent with their elongated appearance when visualized by electron microscopy[14].

The size, shape and symmetry of the NPC. It is also helpful to have some information on the overall shape and symmetry of the NPC. The position of the nuclear envelope membrane relative to the NPC and the NPC's symmetry are based on our electron microscopy and cryo-electron microscopy (cryo-EM) data[5]. These studies have revealed an eight-fold rotational symmetry of the yeast NPC and an approximate two-fold rotational symmetry between the nucleoplasmic and cytosolic halves of the NPC, defining the 'half-spoke' as a 16-fold pseudo-symmetry unit of the NPC (Fig. 2). We have also previously shown that heparin treatment of isolated NPCs produced a ring-like substructure ('Pom rings'), which is associated with the pore membrane and perinuclear space in the intact NPC[15]. We isolated and examined these rings (Supplementary Information), and found that they had a maximum diameter of ~106 nm, consistent with the measured maximum NPC diameter of ~97 nm[5].

The localization of each component in the NPC. We have previously obtained the coarse localization of most nucleoporins within the NPC by immuno-EM, relying on a gold-labelled antibody that specifically interacted with the localized protein through its carboxy-terminal PrA tag (Fig. 4a)[2]. We have now generated a more accurate and complete immunolocalization map of the NPC, in which its constituent proteins, except Sec13, have been localized using a larger data set and improved analysis (Fig. 4b and Supplementary Information).

Inherent limitations in the immuno-EM method allow it to provide only a broad range of allowed axial and radial values for each nucleoporin. Nevertheless, these ranges are smaller than the dimensions of the half-spoke and so are still informative. Notably, most nucleoporins are found on both the nuclear and cytoplasmic sides of the NPC and are tightly packed within a region adjacent to the nuclear membrane (Fig. 4). Most of the FG nucleoporins are found on both sides of the NPC, with a small number found exclusively on the cytoplasmic or nuclear side; for simplicity, we consider Nup116 and Nup100 to be cytoplasmically disposed and Nup145N to be nucleoplasmically disposed, although ~20% of the signal of each is found on the opposite side. Most of the non-FG nucleoporins are also found on both sides. The membrane proteins are found close to the nuclear envelope membrane, and Pom152–PrA is localized to the lumen of the nuclear envelope. Our immuno-EM map agrees almost entirely with independent localizations performed by other groups. For example, Nup159 and Nup82 have previously been shown to be restricted to the peripheral cytoplasmic face[16]; Nup1 was found on the peripheral nuclear face[17]; and Nup157, Nup170, Nup53 and Nup59 were shown to localize proximally to both sides of the

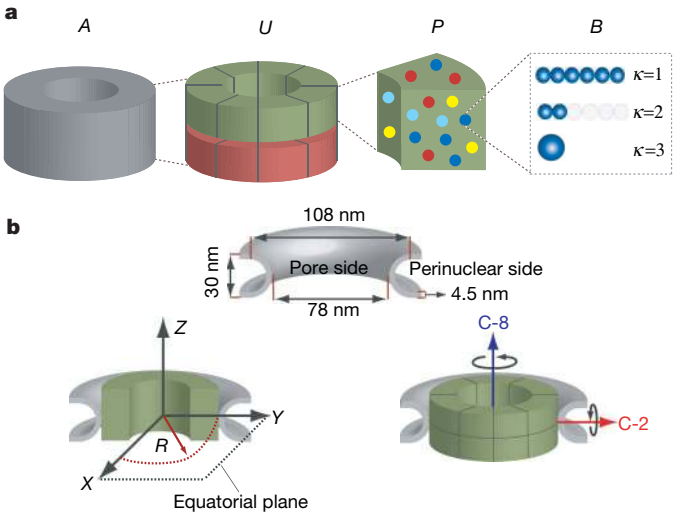
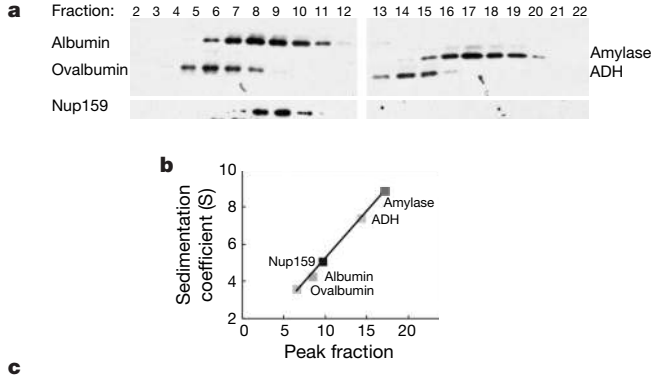


Figure 2 | Structural representation of the NPC. a, Hierarchical representation of the NPC that facilitates the expression of the experimental data in terms of spatial restraints. Formally, we define the whole NPC assembly *A* as a set of symmetry units *U* of two different types with eight instances each, referred to as half-spokes. Half-spokes of the first type (green) reside at the cytoplasmic side and half-spokes of the second type (red) reside at the nucleoplasmic side of the nuclear envelope. Two adjacent half-spokes, one of each type, form a spoke. Each of the 16 NPC half-spokes consists of a set of proteins *P* that are described by their type and index. Each protein is represented by a flexible string of beads *B* in the root representation $\kappa = 1$. Additional representations $\kappa > 1$ can be derived from the root representation (for example, by omitting some beads as in $\kappa = 2$ or by combining beads as in $\kappa = 3$). For the NPC, each protein is described with up to nine different representations. **b**, Top panel: the dimensions of the nuclear envelope, as taken from cryo-EM images (ref. 5). Bottom-left panel: the coordinate system we use has the origin at the centre of the nuclear envelope pore. The nuclear envelope is indicated in grey. Bottom-right panel: the eight-fold (C-8) and two-fold (C-2) symmetry axes of the NPC, as revealed primarily by cryo-EM[5]. We apply the two-fold symmetry only to proteins that appear with identical stoichiometry in both the nucleoplasmic and cytoplasmic half-spokes.

NPC[18] (other independent localizations are listed in Supplementary Table 10).

How 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: rather like the pieces in a jigsaw puzzle, we also need information on the interactions between nucleoporins. We obtained this information from a large number of overlay assays and affinity purification experiments, as well as from the composition of the Pom rings (consisting of Pom34 and Pom152). An overlay assay identifies a pair of proteins that interact with each other, whereas an affinity purification identifies one or more proteins that interact directly or indirectly with the bait protein (Figs 5 and 6 and Supplementary Information). 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. We only used overlay and affinity purification data with a signal-to-noise ratio above a demanding threshold (Supplementary Information).

We designed several affinity purification methods to obtain a large and diverse set of composites (Supplementary Information). PrA was used as a high-affinity C-terminal purification tag on each nucleoporin. Different cell fractions from the tagged strains served as starting materials, although most fractions were produced by whole-cell cryolysis, which proved to be rapid and convenient, yielding high amounts of each complex with minimal losses and proteolytic damage. We generated ~20 variants of extraction buffers with diverse properties to release different kinds of complexes from the fractions. Complexes were isolated via the tagged nucleoporins using antibody conjugated to either Sepharose or magnetic beads, although we preferred magnetic beads as it permitted rapid, high-yield isolations, and eliminated an upper size limit on the purified complexes (Supplementary Information). We also performed affinity purifications from diploid compared with haploid strains to detect a potential second, untagged copy of a given nucleoporin in the complex—a strong indication of a homotypic interaction for that nucleoporin; Pom152–PrA and Nup82–PrA were the only two nucleoporins giving composites containing a second untagged copy. Although



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	-	-	-	-	-

Figure 3 | Protein shape and stoichiometry information. **a**, Protein shape from hydrodynamic experiments. Purified native PrA-tagged nucleoporins were sedimented on sucrose gradients, together with a set of biotin-labelled marker proteins. Fractions were collected and analysed by immunoblotting of the biotin and PrA tags. An immunoblot of fractions from a typical sedimentation analysis is shown, indicating the position of the tagged protein (Nup159–PrA) together with the markers ovalbumin (3.6 S), bovine serum albumin (4.3 S), alcohol dehydrogenase (ADH, 7.4 S) and β-amylase (8.9 S). **b**, Peak positions for the sedimenting proteins were determined and linear regression was used to calibrate the sedimentation coefficients of the PrA-tagged nucleoporin. **c**, Bead representations $\kappa = 1$ of the NPC proteins and their stoichiometries per half-spoke. The stoichiometry of a protein in the cytoplasmic (cyt.) and nucleoplasmic (nucl.) half-spoke, as measured by quantitative immunoblotting[2], is shown. 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[45]. An asterisk indicates that C-terminal fragments were measured. Also shown is a visualization of the protein as a flexible bead chain (shown here in its most extended configuration), which is based on sedimentation analysis, identification of domains by sequence comparison and secondary structure prediction.

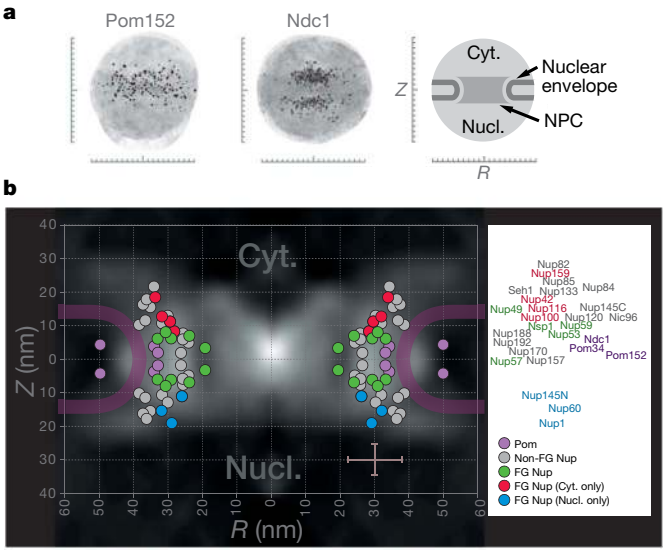


Figure 4 | Localization of proteins by immuno-EM. **a**, Immuno-EM montages for Pom152–PrA nuclei and Ndc1–PrA nuclear envelopes. Scale bars are graduated in 10-nm intervals using the coordinate system defined in Fig. 2b. The major features in each montage are shown schematically at the right, showing how 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*). **b**, 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[5]. The average allowed ranges along the *R* and *Z* coordinates (±8 nm and ±4.5 nm, respectively) are indicated by the brown bars in the bottom right corner.

we originally designed our approaches for the purification of NPC complexes, they have proved to be useful for the isolation of many types of complexes from different cells[7,12,19–24].

Identification of proteins was performed by mass spectrometry[25,26]. Generally, the most vicinal associates of the tagged protein should be approaching stochiometric amounts in the purified complexes; conversely, distally associating proteins may be less abundant. By concentrating on only Coomassie-stainable SDS–polyacrylamide gel electrophoresis (PAGE) bands, we ensured that we identified only the more abundant proteins in any given affinity purification and avoided trace residuals (Fig. 5a). Polypeptides below ~20 kDa were excluded from this analysis for technical reasons[27]; however, due to their small volume, their exclusion is not likely to significantly affect structure determination.

Affinity purifications of tagged versions of all yeast nucleoporins, as well as the NPC-associated messenger RNA transport factors Gle1 and Gle2 (refs 28, 29), yielded 73 distinct composites; together with overlay assays and Pom ring data, we have defined a total of 82 composites (Fig. 6a and Supplementary Information). The composites varied in complexity from dimers to those containing 20 proteins (composite 82) and, importantly, shared significant overlap in composition (Fig. 6b). Therefore, we expect considerable synergy among the composites when used to map the architecture of the whole assembly.

A good example of the compositional overlap is the Nup84 complex (Fig. 5a, b)[13,14,30]. The smallest building blocks of this complex are heterodimers (Fig. 5, 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). This full complex interacts with Nup157 (63, 66) and Nup145N (60). Finally, the entire Nup84 complex co-precipitates together with the Nup170 complex and an Nsp1-containing complex (79). Our data also agree with composites generated by other groups. For example, the Nup84 composites[13,14,30], a

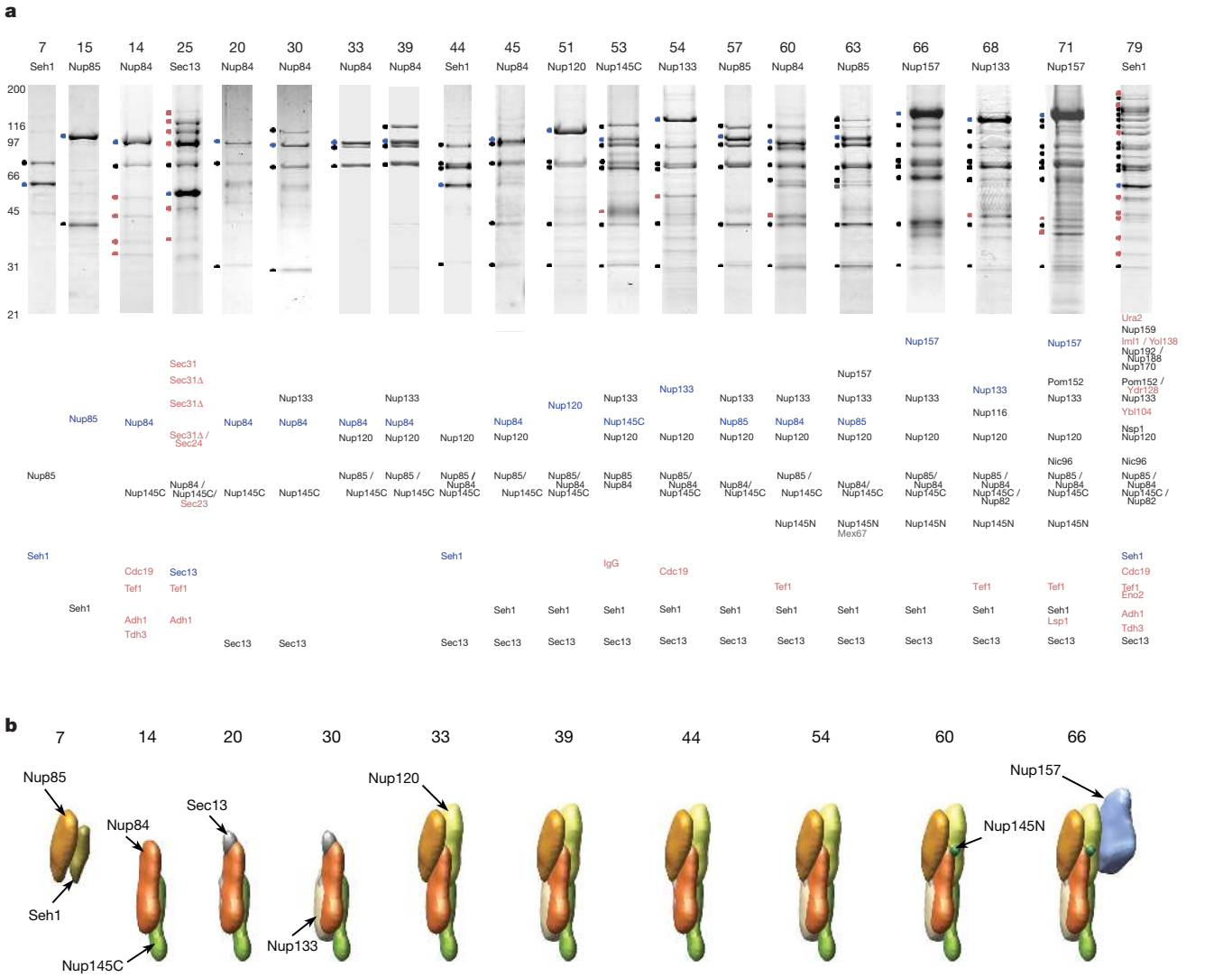


Figure 5 | Protein interactions of the Nup84 complex. a, A sample of affinity purifications containing Nup84 complex proteins. Affinity-purified PrA-tagged proteins and interacting proteins were resolved by SDS–PAGE and visualized with Coomassie blue. The name of the PrA-tagged protein together with a corresponding identification number for the composite is indicated above each lane (Supplementary Information). 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 (either of the example shown here or of a parallel version; Supplementary Information). The identity of the co-purifying proteins is indicated in order below each lane; PrA-tagged proteins are indicated in blue, co-purifying nucleoporins in black, NPC-associated proteins in grey, and other proteins (including contaminants) in red. Each individual gel image was differentially scaled along its length so that its molecular mass standards aligned to a single reference set of molecular mass standards, and contrast-adjusted to improve visibility. **b**, The mutual arrangement of the Nup84-complex-associated proteins as visualized by their localization volumes. The localization volumes, obtained from the final NPC structure (Fig. 9), allow a visual interpretation of the relative proximities of the proteins.

Nup116 composite[31], a Nup170 composite[18], a Nup42–Gle1 dimer[29], a Nic96 composite[32] and others (Supplementary Table 9) have been previously described, and are completely consistent with the composites identified here.

Data translation into spatial restraints

The next step is to translate the experimental data about the NPC structure into spatial restraints (Fig. 1). These restraints were numerous, overlapping and varied in type, and thus were expected to be sufficient for defining the architecture of the NPC.

Restraints and the scoring function. Structure determination is enabled by expressing information as a scoring function, the global optimum of which corresponds to the structure of the native assembly[33]. One such function is a joint probability density function (PDF) of protein positions, given the available information I about the system, $p(C/I)$, where $C = (c_1, c_2, \ldots, c_n)$ is the list of the cartesian coordinates (c_i) of the n component proteins in the assembly (that is, the configuration of the proteins). This joint PDF gives the probability density that a component i of the native configuration is positioned very close to c_i, given the information I we wish to consider in the calculation. In general, I may include any structural information from experiments, physical theories, or statistical preferences. The complete joint PDF is generally unknown, but can be approximated as a product of PDFs p_f that describe individual assembly features (for example, distances or relative orientations of proteins):

$$p(C/I) = \prod_f p_f(C/I_f)$$

The scoring function $F(C)$ is then defined as the logarithm of the joint PDF:

$$F(C) = -\ln \prod_f p_f(C/I_f) = \sum_f r_f(C)$$

For convenience, we refer to the logarithm of a feature PDF as a restraint r_f and the scoring function is therefore the sum of the individual restraints.

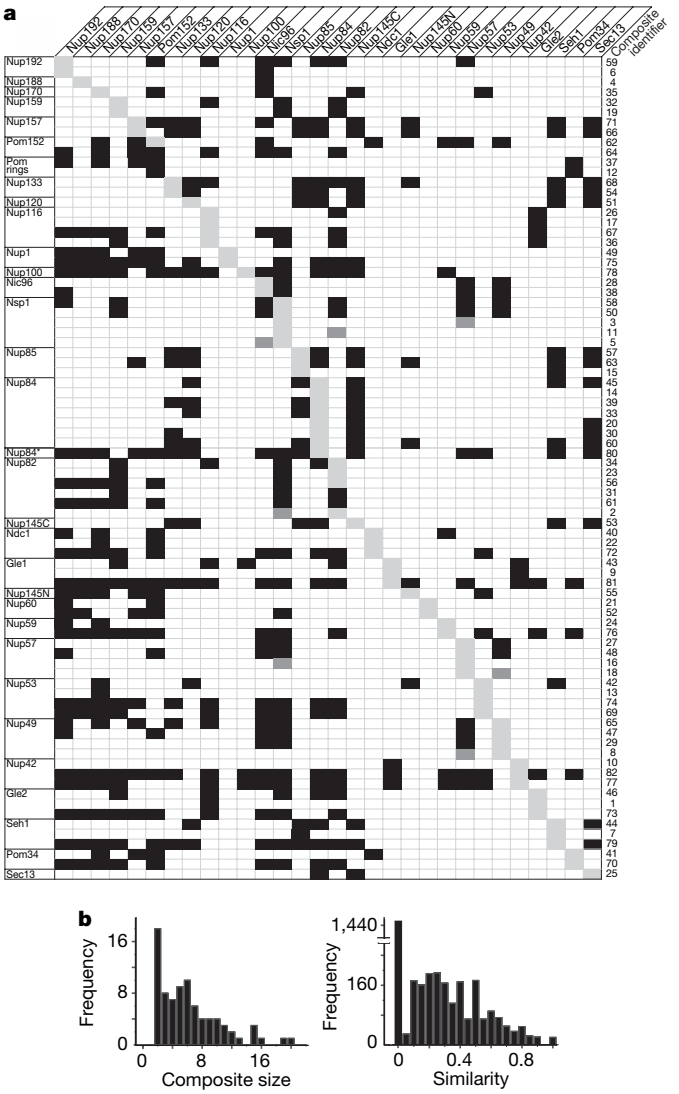


Figure 6 | Protein proximity by affinity purification. a, Composites determined by affinity purification. The affinity-purified nucleoporin–PrA is indicated on the vertical axis, and the corresponding nucleoporins in each composite are shown on the horizontal axis. Composite identifiers are indicated to the right. Presence of a nucleoporin in a composite is indicated by a black box, and the tagged nucleoporin is indicated by a light grey box. In composite 64 (Pom152) and in composites 31 and 61 (Nup82), a second untagged copy of a corresponding protein is present, indicated by a black box. A direct interaction determined by overlay assay is indicated by a dark grey box. The asterisk for Nup84 indicates that the data were obtained with GFP-tagged Nup84. **b**, Distributions of composite size (left) and composite similarity (right). The similarity between two composites is defined by $2a/(2a + b + c)$, where a is the number of proteins that occur in both composites, b is the number of proteins present only in the first composite, and c is the number of proteins present only in the second composite.

Setting up the representation of the NPC. To define restraints on the components of an assembly, we must first specify the symmetry unit of the assembly (that is, the half-spoke in the case of the NPC) (Fig. 2a) and the stoichiometry of its components (Fig. 3). In addition, we must define the representations of the components. Each nucleoporin was represented by a flexible chain consisting of a small number of connected beads (Figs 2a and 3). The number and radii of the beads were chosen to reproduce the protein masses and the sedimentation coefficients[34]. The flexibility of the representation and the low granularity of the NPC structure are sufficient to accommodate uncertainties in the measured S-values and their interpretation. For the FG nucleoporins, no restraints other than the chain connectivity and excluded volume were imposed on the beads representing the FG-repeat regions.

Given the symmetry unit and the protein representations, we can formally represent the NPC with a four-level hierarchy corresponding to the whole NPC, the half-spokes, proteins and beads representing each protein (Fig. 2a). In addition, the nuclear envelope was represented as a rigid surface of many small beads, providing a mould in which the NPC forms (Fig. 2b).

Symmetry of the NPC. The eight-fold and approximate two-fold rotational symmetries of the NPC (Fig. 2b) were imposed by requiring essentially identical configurations of the proteins in common within each half-spoke; the corresponding restraint is formally the root-mean-square of the differences between equivalent intra-half-spoke distances. Although any individual NPC assembly may be perturbed from this perfect symmetry at any given point in time, restraints on the symmetry are nevertheless justified by the relatively low-resolution structure reported here, our intent to characterize the average structure, and exclusion of the FG-repeat regions from the symmetry restraints.

Protein positions from immuno-EM. To reflect the uncertainty in the immuno-EM data, we do not restrain a protein to a specific position. Instead, the C-terminal bead of each protein, corresponding to the tag position, was restrained by imposing lower and upper harmonic bounds on its Z and R coordinates (Fig. 2b), corresponding to the ranges allowed by the immuno-EM data. On average, the allowed area spans 16 and 9 nm along the R and Z coordinate, respectively (Fig. 4 Supplementary Tables 2 and 7, and Supplementary Fig. 8). With such large allowed ranges, the immuno-EM data provide little more information to the structure calculation than which side of the nuclear envelope each nucleoporin is on, and whether it is close to or distal from the NPC equatorial plane and the NPC axis.

Protein positions using the nuclear envelope as a mould. The transmembrane-spanning helices of the three membrane proteins Pom152, Ndc1 and Pom34 were predicted by the program TMHH[35]. The corresponding beads were then restrained to the surface of the nuclear envelope by harmonic positional restraints. In addition, the terminal regions of each protein were restrained either to the pore or perinuclear sides of the nuclear envelope, on the basis of the immuno-EM data and the number of predicted transmembrane helices[2].

Protein proximities from overlay assays and affinity purifications. The overlay assays and affinity purifications carry information about protein proximities, and so are encoded by the same type of spatial restraint. These data provide the richest set of restraints for our NPC structure.

To interpret each composite in terms of a spatial restraint, we must consider three ambiguities. First, there is an ambiguity as to what contacts are present in a composite when it contains more than two proteins. A composite implies only that a copy of each protein in the composite must directly interact with at least one copy of another protein in the composite; any structure that satisfies this condition is consistent with the observed composite. In other words, a composite of n proteins implies at least $n-1$ such interactions between proteins of all types in the composite. Thus, each allowed combination of protein interactions corresponds to a 'spanning tree' of a 'composite graph' (as explained in Fig. 7b). Second, when there are multiple copies of the same protein in the assembly, there is an ambiguity as to which copy is involved in a given type of interaction (Fig. 7a). A measured interaction implies only that at least one copy of the protein is involved in that interaction. Third, when multiple beads are used to represent a protein, there is an ambiguity as to which bead is involved in the interaction (Fig. 2a). A measured interaction implies only that at least one bead of the protein is involved in that interaction. As a result of these three ambiguities, we need to encode a composite by a 'conditional restraint', ensuring that all allowed combinations of alternative assignments of interacting bead pairs are considered (Fig. 7b). Finding the assignment of interactions to specific beads that satisfies the data becomes part of the optimization process (see below). Other minor restraints were also derived from

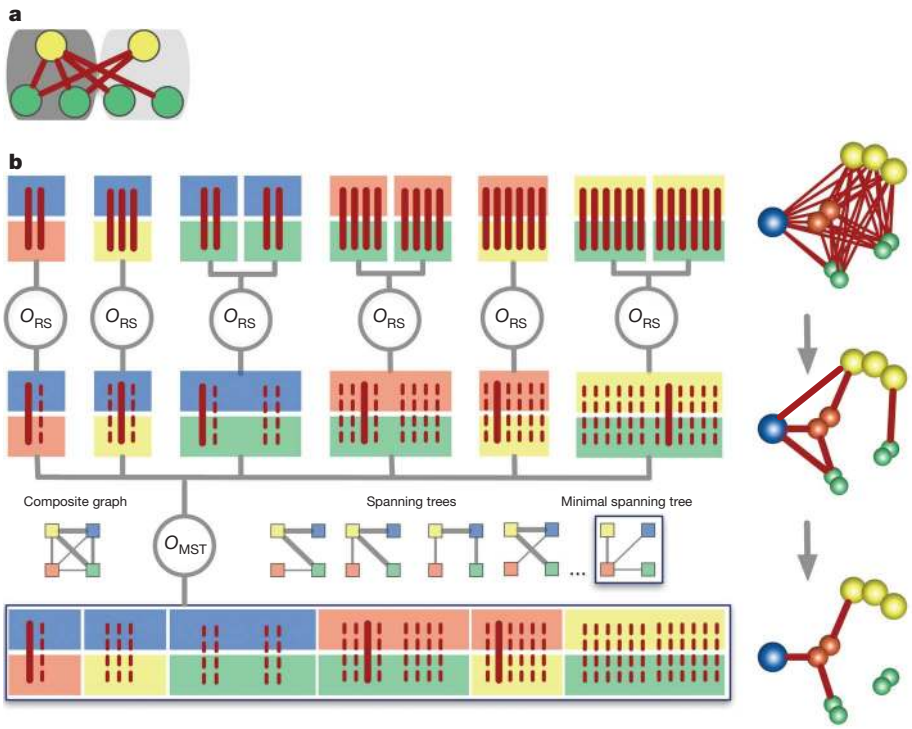


Figure 7 | Ambiguity in data interpretation and conditional restraints. **a**, The ambiguity for a protein interaction between proteins of green and yellow types is illustrated. The ambiguity results from the presence of multiple copies of the same protein in the same or neighbouring symmetry unit. In our NPC calculations, both neighbouring half-spokes on the cytoplasmic and nucleoplasmic sides are considered, for a total of four neighbouring half-spokes (not shown). **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 three of the following six possible types of interaction must occur: blue–red, blue–yellow, blue–green, red–green, red–yellow and yellow–green. In addition, (1) the three selected interactions must form a 'spanning tree' of the 'composite graph' (defined below); (2) each type of interaction can involve either copy of the green protein (in general, all alternatives must be considered as illustrated in **a**); 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 three restraints that are most consistent with the composite data (thick red line); here the edge weights in the minimal spanning tree (defined below) 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 (shown on the 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. Five of the sixteen possible spanning trees are also shown. A spanning tree is a graph with the smallest possible number of edges that connect all nodes. The minimal spanning tree is the spanning tree with the minimal sum of edge weights. 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.

the overlay assay and affinity purification data (Supplementary Information).

Optimization

With the scoring function in hand, the positions of the proteins are determined by optimization of the scoring function (Supplementary Information), resulting in structures that are consistent with the data (Fig. 1). The optimization starts with a random configuration of the constituent proteins' beads, and then iteratively moves them so as to minimize violations of the restraints (Fig. 8). In essence, the restraints cooperate to slowly 'pull together' the proteins into a good-scoring configuration. We use standard methods of conjugate gradients and molecular dynamics with simulated annealing (Supplementary Information). These methods allow the evolving structure some 'breathing room' to explore the scoring function landscape, minimizing the likelihood of getting caught in local scoring function minima (Fig. 8a). To comprehensively sample structures consistent with the data, independent optimizations of randomly generated initial configurations were performed until an ensemble of 1,000 structures satisfying the input restraints was obtained (approximately 200,000 trials were required, running for approximately 30 days on 200 CPUs) (Fig. 8b).

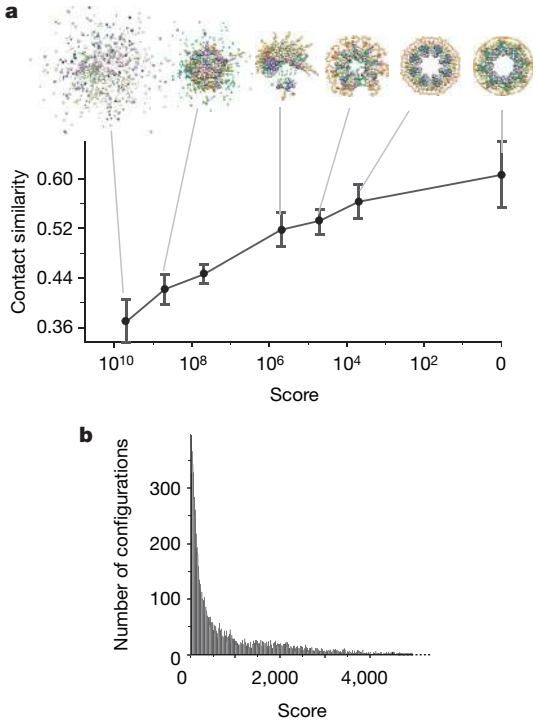


Figure 8 | Calculation of the NPC bead structure by satisfaction of spatial restraints. a, Representation of the optimization process as it progresses from an initial random configuration to an optimal structure. The graph shows the relationship between the score (a measure of the consistency between the configuration and the input data) and the average contact similarity. The contact similarity quantifies how similar two 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 (Supplementary Information). The average contact similarity at a given score is determined from the contact similarities between the lowest scoring configuration and a sample of 100 configurations with the given score. Error bars indicate standard deviation. Representative configurations at various stages of the optimization process from left (very large scores) to right (with a score of 0) are shown above the graph; a score of 0 indicates that all input restraints have been satisfied. 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. **b**, Distribution of configuration scores. The presence of configurations with the score close to 0 demonstrates that our sampling procedure finds configurations consistent with the input data. These configurations satisfy all the input restraints within the experimental error.

Ensemble interpretation

We analysed the ensemble of 1,000 structures that satisfy the input data (Fig. 8b) in terms of protein positions, contacts and configuration (Figs 9 and 10).

Protein positions. These 1,000 structures were first superposed (Fig. 9a) (Supplementary Information). Next, the superposed structures were converted into the probability of any volume element being occupied by a given protein (that is, the 'localization probability') (Fig. 9b). The spread around the maximum localization probability of each protein describes how precisely its position was defined by the input data. The positions that have a single narrow maximum in their probability distribution in the ensemble are determined most precisely. When multiple maxima are present in the distribution at the precision of interest, the input restraints are insufficient to define the single native state of that protein (or there are multiple native states).

The actual localization probabilities yielded single pronounced maxima for almost all proteins, demonstrating that the input restraints define one predominant structure. The average standard deviation for the distance between neighbouring protein centroids is 5 nm; the precision of the larger, centrally positioned proteins seems to be higher than that of the anchor domains of some FG nucleoporins. This level of precision defines a region smaller than the diameters of many nucleoporins. Thus, our map is sufficient to determine the relative positions of proteins in the NPC; we do not interpret features smaller than this precision. On the basis of the localization probabilities (Fig. 9b), we also define the volume most likely occupied by each protein, termed the 'localization volume' (Figs 9c and 10a). The localization volumes of the proteins overlap only to a small degree, such that only 10% of the NPC volume is assigned to two or more proteins, again underscoring how well the position of each nucleoporin is resolved. On the basis of our current data, we are not able to distinguish between the two possible mirror-symmetric structures; here, we present one of them.

Protein contacts. The proximities of any two proteins in the structure can be measured by their relative 'contact frequency', which is defined by how often the two proteins contact each other in the ensemble (Fig. 10b). Contacts are highly conserved among the ensemble structures, despite some variability; 32 protein pairs have a contact frequency higher than 65%. Of all the 435 contact frequencies, 7% are high (65–100%) and 73% are low (0–25%); this again demonstrates that the structure is well defined, as an ensemble of varied structures would yield mainly medium contact frequencies. Notably, few high-contact frequencies are seen between proteins of the same type, indicating that the NPC is held together primarily by heterotypic interactions.

We can improve our determination of contacts by considering not only the contact frequencies but also the composite data (Fig. 10c). More specifically, we define two proteins to be 'adjacent' if their relative contact frequency is larger than 65% or if they appear in the maximal spanning tree of any composite graph whose edge weights correspond to contact frequencies (as explained in Fig. 10c). If two proteins are adjacent, they are more likely to interact with each other in the native NPC structure than when they are not adjacent[36]. In total, 51 types of adjacencies were found (Fig. 10d). A particularly large number of adjacencies are observed for Nic96 and Nup82, which both appear in two copies per symmetry unit, as well as for the core proteins Nup192 and Nup188. Whereas the latter two proteins bridge the bulk of the NPC to the membrane proteins and also provide anchor sites for FG nucleoporins, Nic96 bridges major ring structures of the NPC and also serves as an anchor site for FG nucleoporins[37]. Most FG nucleoporins are peripherally located and therefore show only a few adjacencies.

Protein configuration. We can now combine the protein positions and adjacencies into a configuration of the NPC proteins (Fig. 10e, f). This representation allows us to deconvolute the composites into their constituent complexes (for example, see Figs 5b and 10g).

Synergy among restraints. How our data act synergistically is best demonstrated by the progressive increase in the certainty about the protein positions, as a result of an incremental addition of information (Fig. 11a). Hence, the variability among the 1,000 structures is significantly smaller than the uncertainties in any of the original data. For example, the allowed ranges for protein localization by immuno-EM are reduced from ±4.5 and ±8 nm along the *Z*-axis and the radial coordinate, respectively, to ±2 and ±3 nm in the ensemble, as a direct result of data integration. Similarly, data integration also improves the prediction of protein interactions (Fig. 11b).

Assessment of precision and accuracy

The accuracy of a model is defined as the difference between the model and the native structure. Therefore, it is currently impossible to know with certainty the accuracy of the determined NPC structure. Nevertheless, five lines of evidence indicate that the accuracy of our structure is similar to its precision, and thus representative of the true configuration of the NPC.

Self-consistency of the experimental data. Inconsistencies in the experimental data or its interpretation can be identified when the optimization generates only frustrated structures that do not satisfy the input restraints. This is not the case for our NPC calculations; we find only a single cluster of NPC structures that satisfy all the input restraints. To show that it is not trivial to find structures satisfying all restraints, we repeated the calculations with a comparable, but partly incorrect set of restraints (Supplementary Information). Specifically, all untagged proteins were randomly swapped between composites, leaving the number of composites, the number of proteins in each composite, and all other restraints unchanged. An optimization using this modified restraint set failed to produce any structures that satisfied all restraints.

Variability in the ensemble. We have confirmed that the ensemble of 1,000 structures is sufficiently large for the precision of the NPC architecture to be determined reliably: the reproducibility of contact frequencies calculated from random subsets of the ensemble was plotted as a function of the subset size (Supplementary Information). The similarity between two sets of contact frequencies converges for random subsets of ~100 structures.

The ability of a restraint set to define a native state. We have previously described an approach to test whether or not a given restraint set is sufficient to reconstruct a known native state[36]. In this approach, a native structure is assumed, the restraints to be tested are simulated from this structure, the structure is then reconstructed based only on these restraints, and finally the reconstruction is compared to the original assumed structure. Using this approach, we have simulated composite restraints based on our NPC structure, reproducing the number of composites and the distribution of their size in the original data set; all other restraints were kept the same as in the real application. The accuracy of the reconstructed model was comparable to the precision of the current NPC model.

Patterns unlikely to occur by chance. The distribution of nucleoporins in our structure is expected to reflect their functionality and evolution, and so should be decidedly nonrandom. Indeed, as discussed at length in the accompanying paper[37], there is a striking co-segregation of proteins by fold type to particular locations in the

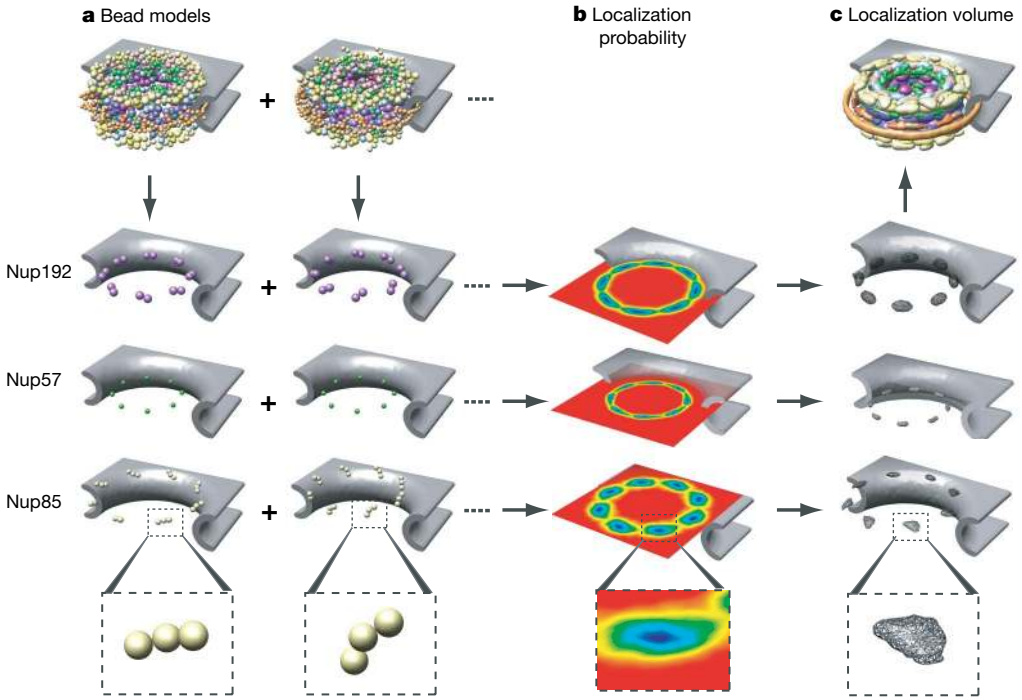


Figure 9 | Bead model, ensemble, localization probability and localization volume. **a**, Top: two representative bead models of the NPC (excluding the FG-repeat regions) from the ensemble of 1,000 superposed structures satisfying all restraints (Fig. 8b). The eight positions of three sample proteins (Nup192, Nup57 and Nup85) on the cytoplasmic side are shown, with a detailed view of the bead representation of one copy of Nup85 at the bottom. **b**, The localization probability for each protein type is obtained by converting the ensemble into the probability of any volume element being occupied by the protein. Shown are contour maps of the cross-sections in the plane parallel to the equatorial plane that contains the maximum value of the protein localization probability. **c**, The localization volume of the sample proteins, derived from the localization probability. The volume elements are first sorted by their localization probability values. The localization volume then corresponds to the top-ranked elements, the volume of which sums to the protein volume, estimated from its molecular mass. The localization volume of a protein reveals its most probable localization. Because of the limited precision of the information used here, the localization volume of a protein should not be mistaken for its density map, such as that derived by cryo-EM.

NPC, although no fold information (except for the transmembrane domains) was used in the generation of the structure.

Experimental data not used in the calculation of the model. Finally, our structure can be most directly tested by comparing it to experimentally determined data that were not included in the structure calculation. First, our structure is robust, in the sense that 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. Second, the shape of our NPC structure[37] strongly resembles the published electron microscopy maps of the NPC[5,38–42], even though these data were not used here (Supplementary Fig. 22). Third, the diameter of the transport channel in our structure is ~38 nm (excluding the FG-repeat regions), in good agreement with the experimentally reported maximal diameter of transported particles[43]. Fourth, 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[44]. Moreover, three of the four additional scaffold nucleoporins that are predicted to contain the ALPS-like motif are also close to the nuclear envelope. Finally, perhaps the best example is that of the Nup84 complex. Our configuration for this complex (Fig. 5b)[37] is completely consistent with previous results[13,14,30]. Specifically, Nup85 and Seh1 form a dimer that together with Nup120 forms the trimeric 'head' of the complex, consistent with the top two arms of the 'Y'-shaped Nup84 complex (Fig. 5b)[14]. Similarly, Nup145C, Nup84, Sec13 and Nup133 form the 'tail' in

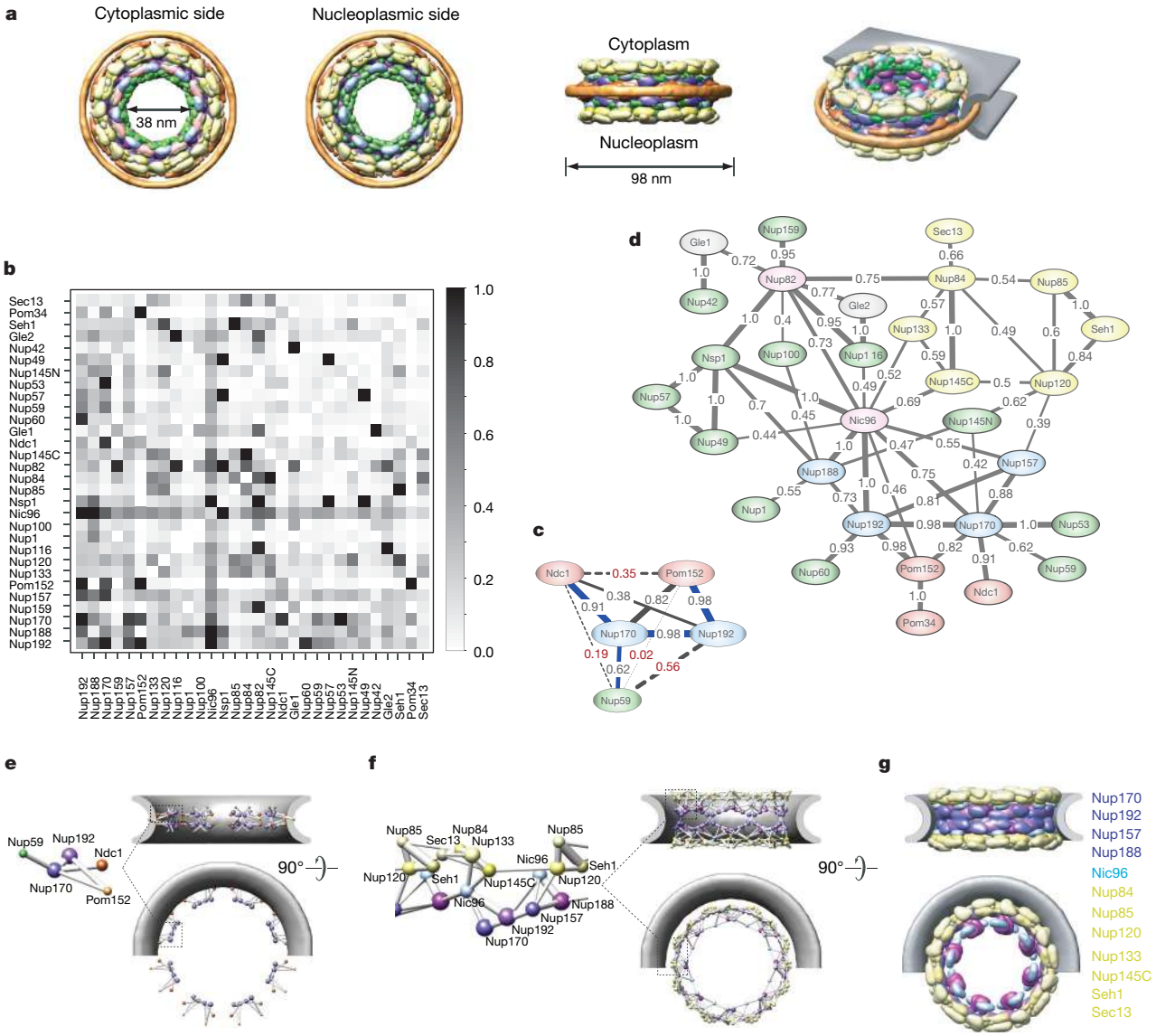


Figure 10 | Ensemble interpretation in terms of protein positions, contacts and configuration. **a**, Localization volumes of all 456 proteins in the NPC (excluding the FG-repeat regions) in four different views. The diameter of the transport channel and the NPC are also indicated. The proteins are colour-coded according to their assignment to the six NPC modules[37]. **b**, Contact frequencies for all pairs of proteins. The contact frequency of a pair of protein types is the fraction of structures in the ensemble that contains at least one protein contact between any protein instances of the two types. **c**, 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. All edges with a statistically significant reduction in contact frequency from their initial values implied by the composite data alone (*P*-value $< 10^{-3}$; Supplementary Information) are indicated by dotted lines with contact frequencies shown in red. **d**, Protein adjacencies for the whole NPC, with proteins as nodes and edges connecting proteins that are determined to be adjacent to each other. The edge weight is the observed contact frequency. **e**, Configuration of the proteins in composite 40. The location of a protein corresponds to the average position of the beads representing non-FG repeats of the protein. **f**, Configuration of Nic96 and the NPC scaffold proteins. **g**, Localization volume of Nic96 and the NPC scaffold proteins[37].

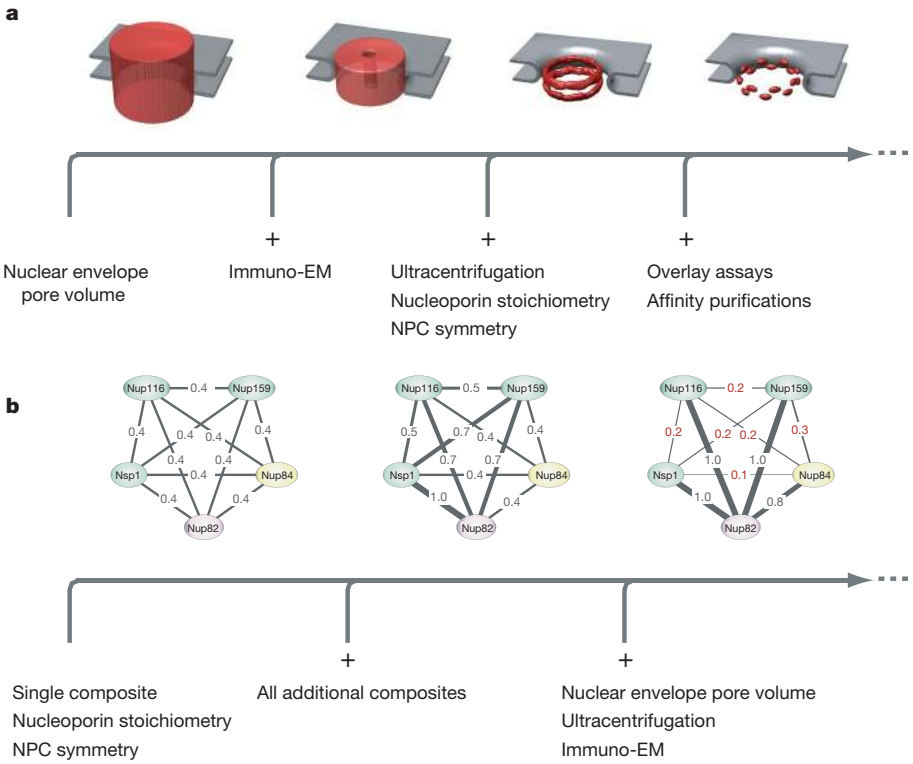


Figure 11 | The structure is increasingly specified by the addition of different types of synergistic experimental information. **a**, Protein positions. As an example, each panel illustrates the localization of 16 copies of Nup192 in the ensemble of NPC structures, generated using the data sets indicated below. The localization probability is contoured at 65% of its maximal value (red). The smaller the volume, the better localized are the proteins. The NPC structure is therefore essentially moulded into shape by the large amount of diverse experimental data. **b**, Protein contacts. Prediction of protein interactions from contact frequencies improves as more data are used. As an example, each panel illustrates 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: when only a single composite is used (together with stoichiometry and symmetry information), all interactions are equally likely (initial contact frequency, Supplementary Information). Middle: when the highest likelihood of interaction between a particular protein pair from all composites is used, the uncertainty about the interactions is reduced. Right: when 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. Contact frequencies reflect the likelihood that a protein interaction is formed given the data considered and are calculated from the ensemble of optimized structures. Numbers in red indicate final contact frequencies that significantly decreased (at a P-value $<10^{-3}$) from their initial values (Supplementary Information).

both our structure and the Y-shaped complex (Fig. 5b)[14]. Here, we resolve the relative positions of the proteins in this complex and show how the complex is integrated into the architecture of the entire NPC.

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

We have devised an integrative approach to solve the structure of the NPC using diverse biophysical and proteomic data. This approach has several advantages. First, it benefits from the synergy among the input data. Data integration is in fact necessary for structure determination, because none of the individual data sets contains sufficient spatial information on its own. Despite the little structural information in each individual restraint, the concurrent satisfaction of all restraints derived from independent experiments markedly reduces the degeneracy of the final structures. Second, 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. Third, this approach can make the process of structure determination more efficient, by indicating which measurements would be most informative. Fourth, 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 cross-linking, footprinting, small angle X-ray scattering (SAXS) and cryo-EM. It is conceivable that these additional data sets might allow us to determine pseudo-atomic structures of assemblies as complex as the NPC. Furthermore, by obtaining detailed structural information concerning different stages of a dynamic process, our approach may animate the NPC's assembly and transport mechanisms[6].

The molecular architecture of many macromolecular complexes could, in principle, be resolved using a similar integrative approach. With regards to the NPC, the resulting structure has already provided abundant insights into the function and evolution of the cell[37].

METHODS SUMMARY

See Supplementary Information for a detailed description of our Methods. The experimental data, the Integrative Modelling Platform software and the NPC structural model are available at http://ncdir.org/npc.

Received 30 August; accepted 22 October 2007.

1. Sali, A., Glaeser, R., Earnest, T. & Baumeister, W. From words to literature in structural proteomics. *Nature* **422**, 216–225 (2003).
2. Rout, M. P. *et al.* The yeast nuclear pore complex: composition, architecture, and transport mechanism. *J. Cell Biol.* **148**, 635–651 (2000).
3. Macara, I. G. Transport into and out of the nucleus. *Microbiol. Mol. Biol. Rev.* **65**, 570–594 (2001).
4. Weis, K. Nucleocytoplasmic transport: cargo trafficking across the border. *Curr. Opin. Cell Biol.* **14**, 328–335 (2002).

5. Yang, Q., Rout, M. P. & Akey, C. W. Three-dimensional architecture of the isolated yeast nuclear pore complex: functional and evolutionary implications. *Mol. Cell* **1,** 223–234 (1998).
6. Beck, M., Lucic, V., Förster, F., Baumeister, E. & Medalia, O. Snapshots of nuclear pore complexes in action captured by cryo-electron tomography. *Nature* **449,** 611–615 (2007).
7. Devos, D. *et al.* Simple fold composition and modular architecture of the nuclear pore complex. *Proc. Natl Acad. Sci. USA* **103,** 2172–2177 (2006).
8. Havel, T. F. & Wüthrich, K. A distance geometry program for determining the structures of small proteins and other macromolecules from nuclear magnetic resonance measurements of intramolecular 1H–1H proximities in solution. *Bull. Math. Biol.* **46,** 673–698 (1984).
9. Malhotra, A., Tan, R. K. & Harvey, S. C. Prediction of the three-dimensional structure of *Escherichia coli* 30S ribosomal subunit: a molecular mechanics approach. *Proc. Natl Acad. Sci. USA* **87,** 1950–1954 (1990).
10. Denning, D. P., Patel, S. S., Uversky, V., Fink, A. L. & Rexach, M. Disorder in the nuclear pore complex: the FG repeat regions of nucleoporins are natively unfolded. *Proc. Natl Acad. Sci. USA* **100,** 2450–2455 (2003).
11. Lim, R. Y. *et al.* Flexible phenylalanine-glycine nucleoporins as entropic barriers to nucleocytoplasmic transport. *Proc. Natl Acad. Sci. USA* **103,** 9512–9517 (2006).
12. Devos, D. *et al.* Components of coated vesicles and nuclear pore complexes share a common molecular architecture. *PLoS Biol.* **2,** e380 (2004).
13. Siniossoglou, S. *et al.* Structure and assembly of the Nup84p complex. *J. Cell Biol.* **149,** 41–54 (2000).
14. Lutzmann, M., Kunze, R., Buerer, A., Aebi, U. & Hurt, E. Modular self-assembly of a Y-shaped multiprotein complex from seven nucleoporins. *EMBO J.* **21,** 387–397 (2002).
15. Strambio-de-Castillia, C., Blobel, G. & Rout, M. P. Isolation and characterization of nuclear envelopes from the Yeast *Saccharomyces. J. Cell Biol.* **131,** 19–31 (1995).
16. Miller, A. L. *et al.* Cytoplasmic inositol hexakisphosphate production is sufficient for mediating the Gle1-mRNA export pathway. *J. Biol. Chem.* **279,** 51022–51032 (2004).
17. Solsbacher, J., Maurer, P., Vogel, F. & Schlenstedt, G. Nup2p, a yeast nucleoporin, functions in bidirectional transport of importin alpha. *Mol. Cell. Biol.* **20,** 8468–8479 (2000).
18. Marelli, M., Aitchison, J. D. & Wozniak, R. W. Specific binding of the karyopherin Kap121p to a subunit of the nuclear pore complex containing Nup53p, Nup59p, and Nup170p. *J. Cell Biol.* **143,** 1813–1830 (1998).
19. Archambault, V. *et al.* Genetic and biochemical evaluation of the importance of Cdc6 in regulating mitotic exit. *Mol. Biol. Cell* **14,** 4592–4604 (2003).
20. Archambault, V. *et al.* Targeted proteomic study of the cyclin-Cdk module. *Mol. Cell* **14,** 699–711 (2004).
21. Tackett, A. J. *et al.* I-DIRT, a general method for distinguishing between specific and nonspecific protein interactions. *J. Proteome Res.* **4,** 1752–1756 (2005).
22. Cristea, I. M., Williams, R., Chait, B. T. & Rout, M. P. Fluorescent proteins as proteomic probes. *Mol. Cell. Proteomics* **4,** 1933–1941 (2005).
23. Niepel, M., Strambio-de-Castillia, C., Fasolo, J., Chait, B. T. & Rout, M. P. The nuclear pore complex-associated protein, Mlp2p, binds to the yeast spindle pole body and promotes its efficient assembly. *J. Cell Biol.* **170,** 225–235 (2005).
24. Cristea, I. M. *et al.* Tracking and elucidating alphavirus-host protein interactions. *J. Biol. Chem.* **281,** 30269–30278 (2006).
25. Zhang, W. & Chait, B. T. ProFound: an expert system for protein identification using mass spectrometric peptide mapping information. *Anal. Chem.* **72,** 2482–2489 (2000).
26. Krutchinsky, A. N., Kalkum, M. & Chait, B. T. Automatic identification of proteins with a MALDI-quadrupole ion trap mass spectrometer. *Anal. Chem.* **73,** 5066–5077 (2001).
27. Stelter, P. *et al.* Molecular basis for the functional interaction of dynein light chain with the nuclear-pore complex. *Nature Cell Biol.* **9,** 788–796 (2007).
28. Murphy, R., Watkins, J. L. & Wente, S. R. GLE2, a *Saccharomyces cerevisiae* homologue of the *Schizosaccharomyces pombe* export factor RAE1, is required for nuclear pore complex structure and function. *Mol. Biol. Cell* **7,** 1921–1937 (1996).
29. Murphy, R. & Wente, S. R. An RNA-export mediator with an essential nuclear export signal. *Nature* **383,** 357–360 (1996).
30. Lutzmann, M. *et al.* Reconstitution of Nup157 and Nup145N into the Nup84 complex. *J. Biol. Chem.* **280,** 18442–18451 (2005).
31. Bailer, S. M. *et al.* Nup116p associates with the Nup82p-Nsp1p-Nup159p nucleoporin complex. *J. Biol. Chem.* **275,** 2354–23548 (2000).
32. Grandi, P., Doye, V. & Hurt, E. C. Purification of NSP1 reveals complex formation with 'GLFG' nucleoporins and a novel nuclear pore protein NIC96. *EMBO. J.* **12,** 3061–3071 (1993).
33. Shen, M. Y. & Sali, A. Statistical potential for assessment and prediction of protein structures. *Protein Sci.* **15,** 2507–2524 (2006).
34. Harding, S. E. Determination of macromolecular homogeneity, shape, and interactions using sedimentation velocity analytical ultracentrifugation. *Methods Mol. Biol.* **22,** 61–73 (1994).
35. Krogh, A., Larsson, B., von Heijne, G. & Sonnhammer, E. L. Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. *J. Mol. Biol.* **305,** 567–580 (2001).
36. Alber, F., Kim, M. F. & Sali, A. Structural characterization of assemblies from overall shape and subcomplex compositions. *Structure* **13,** 435–445 (2005).
37. Alber, F. *et al.* The molecular architecture of the nuclear pore complex. *Nature* doi:10.1038/nature06405 (this issue).
38. Akey, C. W. & Radermacher, M. Architecture of the *Xenopus* nuclear pore complex revealed by three-dimensional cryo-electron microscopy. *J. Cell Biol.* **122,** 1–19 (1993).
39. Stoffler, D. *et al.* Cryo-electron tomography provides novel insights into nuclear pore architecture: implications for nucleocytoplasmic transport. *J. Mol. Biol.* **328,** 119–130 (2003).
40. Kiseleva, E. *et al.* Yeast nuclear pore complexes have a cytoplasmic ring and internal filaments. *J. Struct. Biol.* **145,** 272–288 (2004).
41. Hinshaw, J. E., Carragher, B. O. & Milligan, R. A. Architecture and design of the nuclear pore complex. *Cell* **69,** 1133–1141 (1992).
42. Beck, M. *et al.* Nuclear pore complex structure and dynamics revealed by cryoelectron tomography. *Science* **306,** 1387–1390 (2004).
43. Pante, N. & Kann, M. Nuclear pore complex is able to transport macromolecules with diameters of about 39 nm. *Mol. Biol. Cell* **13,** 425–434 (2002).
44. Drin, G. *et al.* A general amphipathic α-helical motif for sensing membrane curvature. *Nature Struct. Mol. Biol.* **14,** 138–146 (2007).
45. Schurmann, G., Haspel, J., Grumet, M. & Erickson, H. P. Cell adhesion molecule L1 in folded (horseshoe) and extended conformations. *Mol. Biol. Cell* **12,** 1765–1773 (2001).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We thank H. Shio for performing the electron microscopic studies; J. Fanghänel, M. Niepel and C. Strambio-de-Castillia for help in developing the affinity purification techniques; M. Magnasco for discussions and advice; A. Kruchinsky for assistance with mass spectrometry; M. Topf, D. Korkin, F. Davis, M.-Y. Shen, F. Foerster, N. Eswar, M. Kim, D. Russel, B. Peterson and B. Webb for many discussions about structure characterization by satisfaction of spatial restraints; C. Johnson, S. G. Parker and C. Silva, T. Ferrin and T. Goddard for preparation of some figures; and S. Pulapura and X. J. Zhou for their help with the design of the conditional diameter restraint. We are grateful to J. Aitchison for discussion and insightful suggestions. We also thank all other members of the Chait, Rout and Sali laboratories for their assistance. We acknowledge support from an Irma T. Hirschl Career Scientist Award (M.P.R.), a Sinsheimer Scholar Award (M.P.R.), a grant from the Rita Allen Foundation (M.P.R.), a grant from the American Cancer Society (M.P.R.), the Sandler Family Supporting Foundation (A.S.), the Human Frontier Science Program (A.S., L.M.V.), NSF (A.S.), and grants from the National Institutes of Health (B.T.C., M.P.R., A.S.), as well as computer hardware gifts from R. Conway, M. Homer, Intel, Hewlett-Packard, IBM and Netapp (A.S.).

Author Information Reprints and permissions information is available at www.nature.com/reprints. Correspondence and requests for materials should be addressed to A.S. (sali@salilab.org), M.P.R. (rout@rockefeller.edu), or B.T.C. (chait@rockefeller.edu).