doi:10.1016/j.jmb.2009.03.001 J. Mol. Biol. (2009) **388**, 283–298

JMB

Available online at www.sciencedirect.com

ScienceDirect

ELSEVIER

An Efficient Strategy for the Determination of the Three-Dimensional Architecture of Ribonucleoprotein Complexes by the Combination of a Few Easily Accessible NMR and Biochemical Data: Intermolecular Recognition in a U4 Spliceosomal Complex

Ping Li[1]†, John Kirkpatrick[2]† and Teresa Carlomagno[1,2]*

[1]*Department of NMR-Based Structural Biology, Max Planck Institute for Biophysical Chemistry, Am Fassberg 11, 37077 Göttingen, Germany*

[2]*European Molecular Biology Laboratory, Meyerhofstrasse 1, 69117 Heidelberg, Germany*

Received 5 November 2008; received in revised form 2 March 2009; accepted 3 March 2009
Available online 10 March 2009

Edited by M. F. Summers

Ribonucleoprotein (RNP) complexes are involved in several cellular processes, including RNA processing, transcription and translation. RNP structures are often dynamic in nature, undergoing significant remodeling during the course of their function. Visualization of the three-dimensional arrangement of single components in the complex and characterization of the intermolecular interactions are essential for understanding the mechanisms of operation. Crystallization either is not always achievable for these highly dynamic RNP particles or requires trimming the complex to a stable, well-structured core that lacks the flexible, regulatory domains. Alternative techniques that can provide structural information for complexes in solution under native conditions, where they retain their natural dynamic properties, are needed. In this study, we explored the possibility of using a combination of NMR, biochemical data and molecular modeling to generate an accurate high-resolution model of RNP complexes. We applied this strategy to the ternary hPrp31 (human Prp31)–15.5K–U4 5′-SL (stem–loop) spliceosomal complex, which, due to its large size and instability and because of the difficulty in obtaining isotopically labeled hPrp31, is not amenable to complete structure determination by NMR. We designed a protocol where the protein–protein interaction surface is defined for 15.5K by NMR data, while the relative orientations of the U4 RNA and the hPrp31 protein are described by mutational and cross-linking data. Using these data in a restrained ensemble docking protocol, we obtained a model for the ternary complex that reveals a novel rationale for the hierarchical assembly of the complex. Comparison of the docking model with the crystal structure recently obtained for a trimmed version of the complex reveals the high accuracy of the docking model, even down to an atomic level. This work shows that the architecture of large RNP complexes is within reach by NMR investigation in solution even for those cases where a traditional structural determination cannot be performed.

Keywords: NMR; docking; spliceosome; RNP complex; intermolecular recognition

Introduction

Processing of eukaryotic pre-mRNAs involves excision of non-coding sequences (introns) and ligation of coding sequences (exons). The two transesterification reactions required for this process are catalyzed by the spliceosome, a complex ribonucleoprotein (RNP) machine.[1–3] The spliceosome is composed of small nuclear RNAs (snRNAs) and

**Corresponding author.* E-mail address: teresa.carlomagno@embl.de.

† P. Li and J. Kirkpatrick contributed equally to this work.

Abbreviations used: RNP, ribonucleoprotein; hPrp31, human Prp31; SL, stem–loop; K-turn, kink-turn; NOE, nuclear Overhauser enhancement; iRMSD, interface RMSD; SA–MD, simulated annealing–molecular dynamics.

proteins that associate to form five building blocks—the U1, U2, U4, U5 and U6 RNP particles.[4–6] None of these building blocks alone is able to sustain catalysis. Instead, the catalytically active particle is assembled through complicated pathways that include formation of different multiparticle complexes and large conformational rearrangements.[7,8]

In the catalytically active U2/U6·U5 complex, the U6 snRNA is base paired with the U2 snRNA and provides crucial residues for catalysis.[8] However, in an earlier stage during spliceosome assembly, the U6 snRNA is associated with the U4 snRNA in the U4/U6·U5 complex.[9,10] The U4 snRNA can be thought of as a chaperone that delivers the U6 snRNA to other spliceosomal particles in a repressed state by masking the catalytic residues. Thus, the assembly and disassembly of the U4/U6·U5 particle are of key importance for the regulation of spliceosomal activity.

The formation of the U4/U6 complex is initiated by the recognition of the 5′-stem–loop (SL) region of the U4 snRNA, located between the two U4/U6 base-paired regions (Fig. 1a), by the 15.5K protein.[13] This highly conserved protein recognizes and possibly stabilizes a particular sequence of the RNA, known as the kink-turn (K-turn) motif.[14] In the hierarchical assembly pathway of the U4/U6 particle, the 15.5K–U4 5′-SL complex provides a binding platform for the

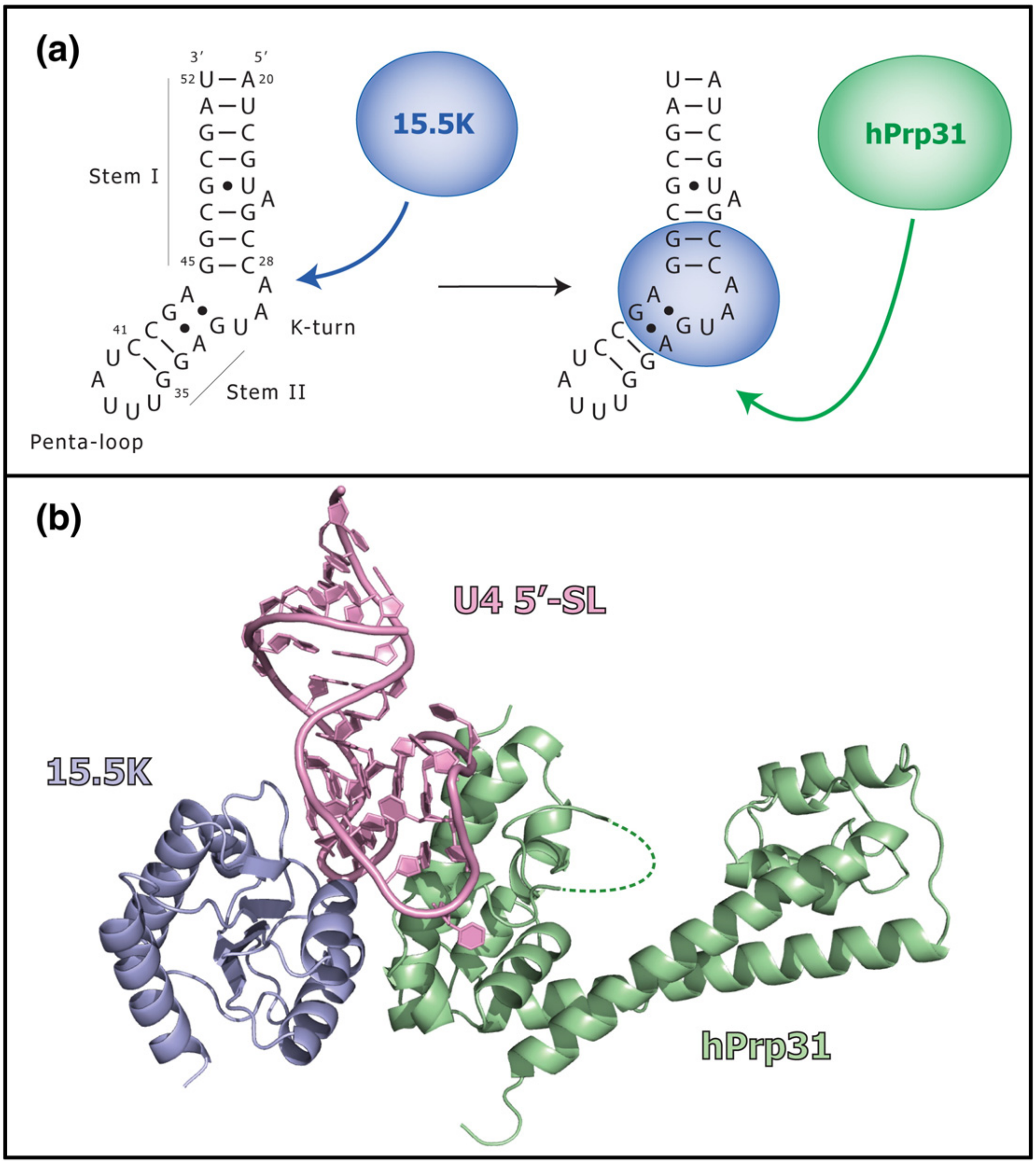

Fig. 1. (a) Schematic representation of the U4 5′-SL snRNA used in this study. Binding of 15.5K to the K-turn region and that of the secondary binding protein hPrp31 to the preformed 15.5K–U4 5′-SL RNP are indicated. (b) The crystal structure of the hPrp31$^{78–333}$–15.5K–U4 snRNA complex (green, hPrp31; dashed line in hPrp31, unstructured loop; blue, 15.5K; pink, U4 snRNA).[11] The hPrp31$^{78–333}$–15.5K–U4 snRNA complex has a triangular architecture with each molecule contacting both other molecules in the complex. The figures in this article were partly prepared with the PyMOL molecular graphics package.[12]

human Prp31 (hPrp31) (U4/U6-61K) protein.[15] In the major spliceosome (minor spliceosome), the hPrp31–15.5K–U4 5′-SL (hPrp31–15.5K–U4$_{atac}$ 5′-SL) ternary complex[16] subsequently recruits the hPrp3–hPrp4–CypH (U4/U6-90K–60K–20K) ternary complex to the U4/U6 di-snRNP. Visualizing the three-dimensional architecture of the U4/U6 particle is indispensable for understanding the mechanisms of stabilization and destabilization of the U4/U6·U5 complex at various stages during spliceosome assembly.[17]

The structure of the hPrp31$^{78–333}$–15.5K–U4 snRNA complex has recently been determined using both X-ray crystallography and NMR spectroscopy data.[11] Recent developments in NMR spectroscopy[18–21] have yielded a dramatic improvement in the sensitivity of NMR spectra of large macromolecules.[22] In the case of the hPrp31$^{78–333}$–15.5K–U4 snRNA complex, a detailed structural investigation by NMR is limited by its poor solubility (maximum stable concentration of 0.15–0.2 mM), by poor yields in the expression of recombinant hPrp31, which hinders ^{13}C/^{15}N/^{2}D labeling, and by the rapid degradation of hPrp31 in solution. Despite all these difficulties, we were able to map the interaction surface between the hPrp31 protein and the 15.5K protein[11] by chemical shift perturbation mapping and saturation transfer experiments[23] using the amide resonances of perdeuterated 15.5K. Crystallographic structural analysis (Fig. 1b) revealed the atomic details of the intermolecular interactions in the ternary complex.[11] The RNA is sandwiched between 15.5K and hPrp31, and the interaction surface between the hPrp31 protein and the binary RNP is composed of approximately equal contributions from the U4 5′-SL RNA and the 15.5K protein. Both the NMR experiments and the crystallographic analysis identified the hPrp31$^{78–333}$ fragment as a genuine RNP recognition domain.

While a detailed description of side-chain interactions in multimeric assemblies of the size of the hPrp31–15.5K–U4 snRNA complex (~80 kDa) is best achieved by crystallographic methods, the difficulties encountered in the crystallization of particles containing flexible regions or inhomogeneous, rapidly degrading components are often a serious obstacle. Moreover, the presence of crystal packing forces and the need to trim the single components of the multimeric particle to crystallographically well-behaved units call for complementation of the crystallographic results with a second approach where the full-length native complex can be investigated in solution. Here, we introduce an integrated strategy that uses NMR spectroscopy and biochemical data in combination with molecular modeling to obtain an accurate model of ternary RNP complexes. In this approach, NMR analysis is used to define the protein–protein interaction surface in the hPrp31–15.5K–U4 snRNA complex, while cross-linking and mutational data define the RNA–protein contacts. Subsequently, molecular modeling is employed to assemble the ternary hPrp31–15.5K–U4 snRNA complex from single-component structural models, including the crystallographic structure of the 15.5K–U4 snRNA binary complex[14] and a homology model for hPrp31. We demonstrate that this procedure leads to a ternary complex model that differs at the intermolecular interface by only 2.3 Å [root-mean-square deviation (RMSD)] from the crystallographic structure of the ternary complex.[11] This procedure is based on a few rapid and easily accessible NMR and biochemical experiments and constitutes a valid alternative to structural analysis by X-ray diffraction when crystallization fails, as, for example, with inhomogeneous, dynamic, rapidly degrading or even partially unfolded systems. Alternatively, when crystallographic data are available on trimmed constructs of the complex, the methodology presented here allows validation of the crystal structure in an environment close to physiological conditions and using full-length proteins.

Results

Visualization of the three-dimensional arrangement of single components in spliceosomal RNP complexes and definition of the intermolecular interfaces are indispensable for understanding the mechanisms of the dynamic spliceosome assembly process. However, crystal structures are not easily accessible for these highly dynamic RNP complexes. Alternative techniques that can provide structural information for full-length complexes in solution, namely in an environment where the complex components retain their dynamic properties, are desirable. In this work, we explored the possibility of generating an accurate docking model for the ternary hPrp31–15.5K–U4 5′-SL spliceosomal complex from a few biochemical and NMR data. The protein–protein interaction surface was defined for 15.5K by NMR data, while hPrp31–RNA contacts were identified from mutational data and cross-linking experiments. These few highly ambiguous pieces of information were used as restraints in a docking protocol to generate a model for the hPrp31–15.5K–U4 5′-SL ternary complex. We used ensemble models of the 15.5K–U4 5′-SL-33nt binary complex generated from the crystallographic coordinates of the 15.5K–U4 5′-SL-22nt complex[14] and ensemble homology models for the N-terminal domain of hPrp31 (see Materials and Methods for detailed descriptions of the generation of these two models) as starting structures for the single components in the complex. The accuracy of the ternary complex docking model was evaluated by comparison with the recently solved crystal structure of the hPrp31$^{78–333}$–15.5K–U4 5′-SL complex.[11]

NMR analysis of the protein–protein interface

The existence and nature of the 15.5K–hPrp31 interface were probed by NMR experiments. All NMR experiments were conducted with both full-length hPrp31 and the hPrp31$^{78–333}$ fragment in order to ascertain whether any of the observed effects might be elicited by the flexible C-terminus of hPrp31 (residues 334–499). The chemical shift changes induced on the 15.5K protein by addition of either full-length hPrp31 or the hPrp31$^{78-333}$ fragment were monitored in ^{15}N transverse relaxation-optimized spectroscopy heteronuclear single-quantum coherence spectra.[18] These changes localized primarily on the α2 and 3_{10} helices of 15.5K (Fig. 2a and Fig. S1a) for both full-length hPrp31 and hPrp31$^{78-333}$. Residues K37, A39–T45 and N47 of helix α2 exhibited the most pronounced changes ($\delta_{NH}>0.06$ ppm).

To explore whether such chemical shift changes are induced by direct protein–protein contacts or by the restructuring of the U4 5′-SL RNA upon hPrp31 binding, we performed cross-saturation experiments[23] on the ternary complex assembled from ^{2}H/^{15}N uniformly labeled 15.5K and unlabeled hPrp31 or hPrp31$^{78-333}$. In these experiments, the methyl protons of hPrp31 are saturated using a decoupling pulse train and the saturation of magnetization is transferred via cross-relaxation from hPrp31 to those amide protons of 15.5K that are in close proximity to hPrp31 protons. Due to both the high deuteration level (>95%) of 15.5K and the absence of RNA resonances in the methyl region of the spectrum, the saturation transfer is specific for the protein–protein interaction surface.

Large intensity changes (I_{change}) were observed for residues in the α2, α3 and 3_{10} helices of 15.5K in the hPrp31–15.5K–U4 5′-SL complex (Fig. 2b and Fig. S1b), with the most pronounced changes for R36,

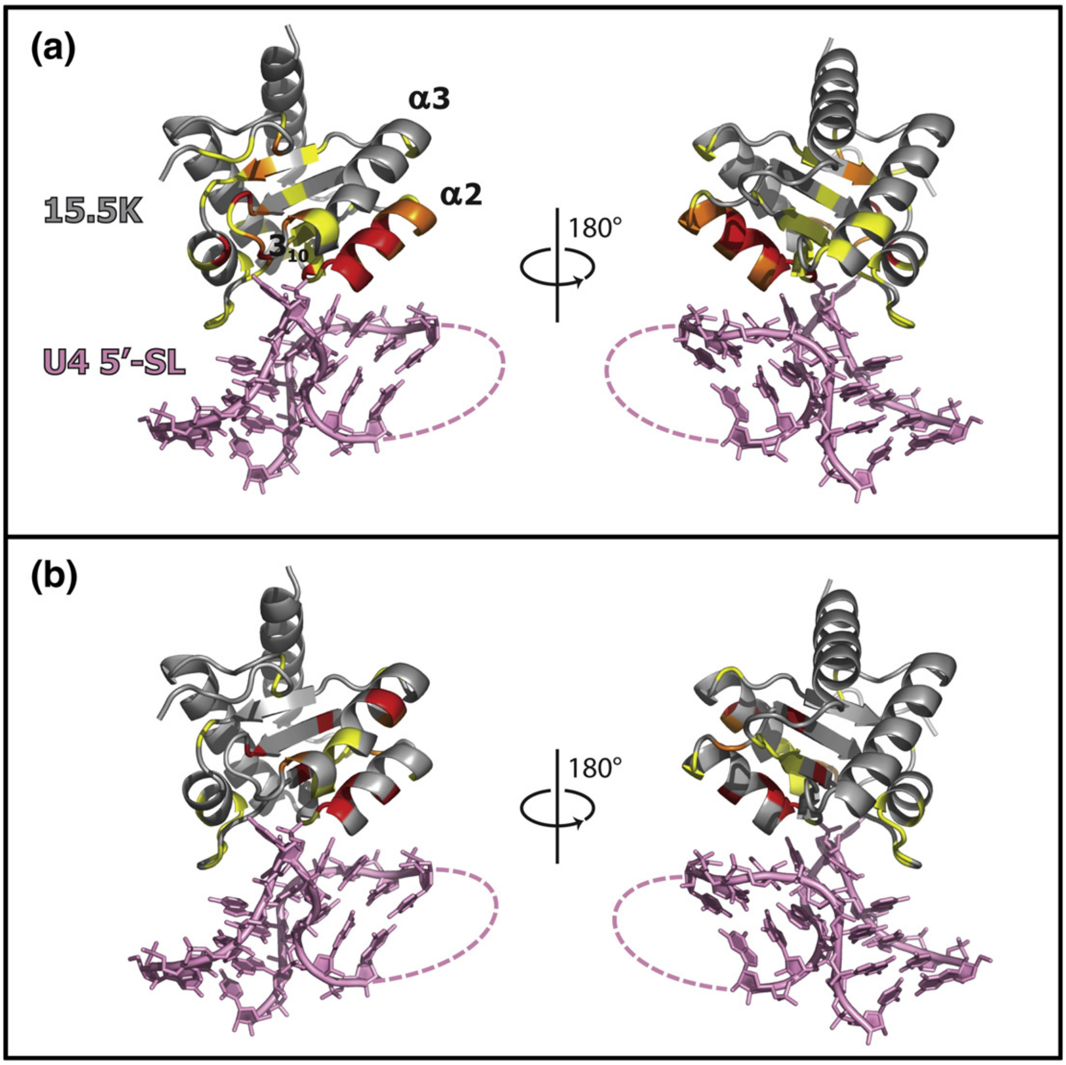

Fig. 2. (a) Chemical shift changes (δ_{NH}) of 15.5K in the 15.5K–U4 5′-SL RNP observed upon binding of both hPrp31 and hPrp31$^{78-333}$ are plotted on the crystal structure of the 15.5K–U4 5′-SL binary complex.[14] Red indicates $\delta_{NH}>0.06$ ppm; orange, $0.04<\delta_{NH}<0.06$ ppm; and yellow, $0.02<\delta_{NH}<0.04$ ppm. (b) Signal intensity changes observed for the 15.5K amide resonances in the hPrp31–^{2}D/^{15}N-labeled 15.5K–U4 5′-SL ternary complex upon saturation of the methyl resonances of hPrp31 are plotted on the crystal structure of the 15.5K–U4 5′-SL binary complex. The I_{change} values observed for the ^{2}D/^{15}N-labeled 15.5K–U4 5′-SL binary complex upon saturation of the residual methyl protons of ^{2}D/^{15}N-labeled 15.5K were subtracted, after scaling, from those observed for the ternary complex to give $I_{change,norm}$. $I_{change,norm}>0.5$ values are plotted: red indicates $I_{change,norm}>2$; orange, $1<I_{change,norm}<2$; and yellow, $0.5<I_{change,norm}<1$ (gray, 15.5K; pink, RNA; dashed line, disordered penta-loop).

A39, A42 and K44 of α2 and for L63, L71 and L72 of the 3_{10} and α3 helices. A very similar picture is obtained for the hPrp31$^{78-333}$–15.5K–U4 5′-SL complex. The control experiment performed for the [^{2}H,^{15}N]15.5K–U4 5′-SL complex in the absence of hPrp31 or hPrp31$^{78-333}$ showed much smaller intensity changes, localized to the hydrophobic core of 15.5K rather than to the α2 and α3 helices. These effects stem from the incomplete deuteration of the methyl groups of 15.5K (deuteration level of ∼95%). The intensity changes observed for 15.5K in the 15.5K–U4 5′-SL binary complex were subtracted, after scaling, from those observed in the hPrp31–15.5K–U4 5′-SL and hPrp31$^{78-333}$–15.5K–U4 5′-SL ternary complexes in order to ensure that the values plotted in Fig. 2b and Fig. S1b are specific for the interaction surface between the two proteins. The intensity changes produced by saturation of the hPrp31 aliphatic protons and the chemical shift perturbations both map onto the α2, 3_{10} and α3 helices of 15.5K, identifying this region as the protein–protein interaction surface in the RNP complex. Furthermore, the strong similarity between the results from the saturation transfer experiments on the hPrp31–15.5K–U4 5′-SL and hPrp31$^{78-333}$–15.5K–U4 5′-SL complexes indicates that the interface with 15.5K is entirely contained within the hPrp31$^{78-333}$ fragment.

In principle, the location of the protein–protein interaction surface on hPrp31 could be identified by a similar approach using ^{2}H,^{15}N-labeled hPrp31 and unlabeled 15.5K. However, the low expression levels of hPrp31 in *Escherichia coli* did not permit the production of this protein with either ^{13}C,^{15}N labeling for assignment or ^{2}H,^{15}N labeling for the saturation transfer experiments. Thus, the residues of hPrp31 interacting with 15.5K are left highly ambiguous in the docking calculations and are localized on the basis of geometrical considerations, as explained below.

Docking of the ternary complex

The docking protocol used models for the 15.5K–U4 5′-SL-33nt binary complex and hPrp31$^{188-332}$ as starting structures. A detailed description of the generation of the starting structures is given in Materials and Methods. Briefly, models for the 15.5K–U4 5′-SL-33nt binary complex were generated using the crystal structure of the 15.5K–U4 5′-SL-22nt complex as a template.[14] 150 structures differing in the conformation of the RNA penta-loop were generated. Models of hPrp31$^{188-332}$ were obtained by comparative modeling with the structure of the *Archaeoglobus fulgidus* Nop5p[24] (sequence alignment shown in Fig. S2). 150 differing in the conformation of loop residues 253–272 were generated. We chose to use an ensemble of starting structures for both the hPrp31 protein and the 15.5K–U4 5′-SL-33nt complex, differing in the conformation of the flexible regions (loop 253–272 in hPrp31 and the RNA penta-loop), rather than one single structure per component. These regions were missing in the respective crystal structures, indicating that they adopt a range of conformations. By generating structural ensembles for these flexible loops, we sought to realistically represent the molecular recognition process, wherein the bound structure is selected from the range of conformations sampled by the free components in solution.

With the models of 15.5K–U4 5′-SL-33nt binary complex and hPrp31$^{188-332}$ in hand, we set out to build the model for the ternary complex using HADDOCK (High Ambiguity-Driven biomolecular DOCKing) 2.0.[25,26] In the following discussion, we refer to the helical structures of hPrp31 using the numbers of the analogous structural elements in Nop5p, which are different from those used by Liu *et al.*[11] Defining the ambiguous intermolecular contacts between 15.5K and hPrp31$^{188-332}$ is essential for the docking procedure. While the contribution of 15.5K to the protein–protein contact surface was well defined by the NMR chemical shift perturbation and saturation transfer experiments, the interfacial surface of hPrp31 remained to be determined. Our strategy to locate the hPrp31 protein–protein contact surface was based on the application of geometrical restrictions to define the relative orientations of the U4 5′-SL RNA and hPrp31, which can be obtained from a combination of (i) electrostatic complementarities, (ii) mutagenesis data and (iii) cross-linking experiments.

The electrostatic potential calculation for hPrp31$^{188-332}$ revealed a strongly electropositive surface and an electronegative surface (Fig. 3a). A large portion of the electropositive surface is the roughly flat surface formed by the helices corresponding to α8, α9, α11 and α12 of the Nop domain of Nop5p. This surface is likely to bind to the highly negatively charged surface of the 15.5K–U4 5′-SL binary complex, defined by the major groove of the stem II and K-turn regions of the U4 5′-SL RNA, as well as the N-terminal part of helix α2 of 15.5K (Fig. 3b). Helix α3 of 15.5K, which shows mainly neutral charges, is directly adjacent to helix α2 and hence is also located on this electronegative surface. Furthermore, the observation that the 15.5K–hPrp31 interactions detected by NMR are identical for the full-length hPrp31–15.5K–U4 5′-SL complex and the hPrp31$^{78-333}$–15.5K–U4 5′-SL complex excludes the possibility that the C-terminal domain of hPrp31 contacts 15.5K.

The involvement of the positively charged surface of hPrp31 in binding to the 15.5K–U4 5′-SL complex was confirmed by both mutagenesis and UV cross-linking experiments. The mutagenesis information used in this docking study was defined for an archaeal complex that shares a similar architecture with the spliceosomal hPrp31–15.5K–U4 5′-SL complex. The archaeal complex is constituted by the L7Ae protein (the archaeal orthologue of 15.5K), which recognizes the K-turn of the box C/D RNA.[27,28] The structure of the L7Ae–box C/D binary RNP is very closely related to that of the 15.5K–U4 5′-SL binary complex.[14,29] The L7Ae–box C/D RNP recruits the Nop5p and fibrillarin proteins to form a tetrameric complex whose function is the methylation of

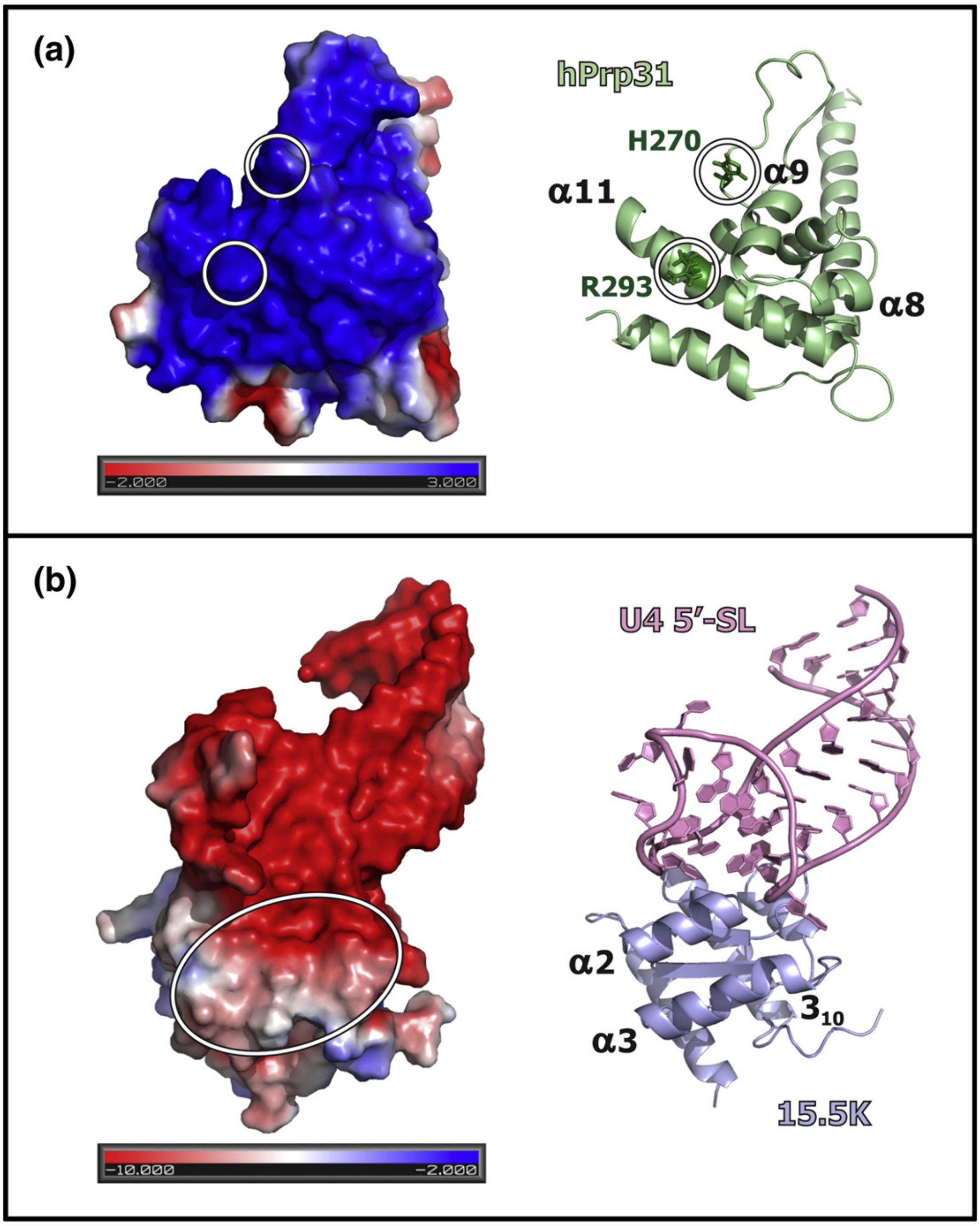

Fig. 3. (a) hPrp31[188–332] possesses a strongly electropositive surface. H270 and R293, the two residues of hPrp31 that have been identified to be in contact with the 15.5K–U4 5′-SL complex by UV cross-linking and mutagenesis experiments, are shown in white circles. R293 is located unambiguously on the electropositive surface, while H270 is part of the loop region between helices α9 and α10. (b) The electrostatic surface of the 15.5K–U4 5′-SL complex. The stem II and K-turn regions of the RNA constitute a highly electronegative surface. Helices α2, 3_{10} and α3 of 15.5K, which have been identified to be in contact with hPrp31 by NMR experiments, are shown in the white circle.

rRNA.[30–32] The Nop5p protein shows strong similarity to the Nop domain of hPrp31 (residues 215–333) and was therefore used as a template for the comparative modeling. Mutational studies conducted for the fibrillarin–Nop5p–L7Ae–box C/D complex showed that R224 of Nop5p, which corresponds to R293 of hPrp31, is essential for the association of the fibrillarin–Nop5p dimeric complex with the L7ae–box C/D sRNP.[24] This arginine is located on the positively charged surface of hPrp31 (Fig. 3a) and is highly conserved throughout the Nop family of proteins, as well as between the same proteins from different species. Thus, it is reasonable to assume that, as with R224 in the fibrillarin–Nop5p–L7Ae–box C/D complex, R293 of hPrp31 is involved in the interaction with the 15.5K–U4 5′-SL-33nt complex and furthermore that it interacts with the highly negatively charged U4 RNA.

The second anchor point was derived from a previous UV cross-linking and mass spectrometric analysis, which demonstrated that residue H270 of hPrp31 is in close contact with U44 of the U4$_{atac}$ 5′-SL RNA in the highly homologous hPrp31–15.5K–U4$_{atac}$ 5′-SL RNP complex of the minor spliceosome.[15] U44 corresponds to A39 in the hPrp31–15.5K–U4 5′-SL complex of the major spliceosome (Fig. S3). H270 is not unambiguously located on the same electropositive surface as R293 as it is part of the disordered 253–272 loop region, but it is free to orient toward this surface.

The mutagenesis and UV cross-linking experiments, together with the charge complementarities, define the positively charged surface of hPrp31 that contains residues H270 and R293 (Fig. 3a) as being in contact with the negatively charged surface of the 15.5K–U4 5′-SL complex that consists of both the stem II, K-turn and loop regions of the U4 RNA and the α2, 3_{10} and α3 helices of 15.5K. The relative orientations of the two charge complementary surfaces of hPrp31$^{188-332}$ and the 15.5K–U4 5′-SL complex were defined as follows. The interaction partner of H270 has been uniquely located to loop residue A39 by UV cross-linking experiments.[15] The mutagenesis data available for R293 do not reveal its interaction partner on the U4 5′-SL RNA; however, H270 and R293 of hPrp31 are separated by approximately 15 Å, which spans the distance between loop residue A39 of the U4 RNA and the tip of the K-turn (Fig. S4). Therefore, we assigned ambiguous nuclear Overhauser enhancements (NOEs) between R293 and the stem II and K-turn regions of the RNA (see Materials and Methods for a detailed list), while well-defined NOEs were imposed between hPrp31-H270 and U4-A39, according to the UV cross-linking data.[15] These constraints define the relative position of the U4 RNA to hPrp31.

The protein–protein interaction surface on 15.5K is principally located on the α2, 3_{10} and α3 helices. The hPrp31 surface residues located at the interface with helices α2, 3_{10} and α3 of 15.5K can be assigned on the basis of geometrical considerations. Following the relative position of the RNA to hPrp31, the protein–protein interaction surface on hPrp31$^{188-332}$ must be located on helices α8, α9 and α11, whose orientations are then approximately perpendicular to those of the α2, α3 and 3_{10} helices of 15.5K. The hPrp31 surface was divided into three zones (Fig. S4b): T239, N240, V305 and F308 on the very N-terminal end of helix α8 and the very C-terminal end of α11 (upper zone, red in Fig. S4b) were given contacts to the 3_{10} helix and the N-terminal end of helix α3 of 15.5K; residues 245–248 on helix α9 and A297 and K298 on the N-terminal region of helix α11 (lower zone, blue in Fig. S4b) were given contacts to α2 of 15.5K; and K243 on the C-terminal end of α8 and T300, L301 and R304 on the C-terminal half of α11 of hPrp31 reside in the middle region of the interaction surface (magenta in Fig. S4b) and were given ambiguous contacts to the α2, 3_{10} and α3 helices of 15.5K. The contacting residues on 15.5K are the surface residues, namely N40, T43, N47 and R48 of helix α2 (red in Fig. S4a) and L63, I65, I66, H68 and L71 of the 3_{10} and α3 helices (blue in Fig. S4a).

The docking protocol described in Materials and Methods was repeated four times to assess the reproducibility of the results. Each of the four repeat calculations generated 500 water-refined structures.

Analysis of the docking results

Choice of scoring function

The first step in the analysis is to rank the solutions according to an empirical score. For the HADDOCK program,[25] the current default scoring function is defined as:

$$\text{Score (Score 1)} = E_{\text{vdW}} + 0.2\ E_{\text{elec}} + 0.1\ E_{\text{AIR}} + E_{\text{desolv}}$$

where E_{vdW} and E_{elec} represent the contribution to the energy from the van der Waals and electrostatic interactions, respectively; E_{AIR} reflects the agreement with the ambiguous interaction restraints; and E_{desolv} represents the penalty for surface desolvation upon complex formation. This score has been optimized for application to protein–protein complexes. However, we expect that the interface between hPrp31 and the 15.5K–U4 RNP will include a significant contribution from the U4 RNA. The highly electrostatic nature of RNA–protein interactions suggests that a different weighting of the electrostatic energy in the scoring function may be appropriate. A previous study on protein–DNA docking used a weight of 1 for the electrostatic energy in the scoring function.[33] Therefore, in this work, we compared two scores for the purposes of structure selection: score 1 is the classic HADDOCK score as defined above, while the weight of the electrostatic term is increased in score 2:

$$\text{Score 2} = E_{\text{vdW}} + E_{\text{elec}} + 0.1\ E_{\text{AIR}} + E_{\text{desolv}}$$

Selection of the best docking solution

The selection of the best docking solution proceeds via plotting the scores of the solutions against the interface RMSD (iRMSD) to the lowest scoring structure. The plots for the two scores are shown for the first run in Fig. 4 (see Fig. S5 for all four runs). The correlation between the score and the iRMSD is stronger for score 1 (classic HADDOCK score) than for score 2 (increased electrostatic weight), indicating that there is closer structural similarity between similarly scoring structures for score 1, while structures with a similar score 2 can be more structurally diverse. This finding is not unexpected given that the RNA–protein interface, whose influence on the score is increased in score 2 relative to score 1 by the increased electrostatic weight, is formed in large part by interaction between the flexible U4 penta-loop and the flexible loop in hPrp31. The flexibility of these two loops was represented by performing ensemble docking using multiple conformations of the loop regions. Such ensemble docking has been shown to improve the probability of finding near-

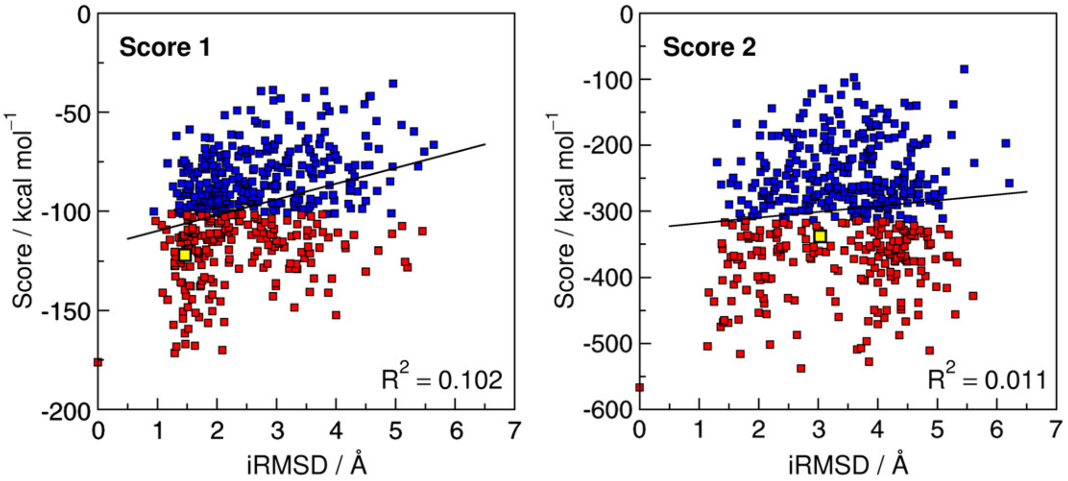

Fig. 4. Plots for the first run of score *versus* iRMSD to the lowest scoring structure for score 1 (left) and score 2 (right). The square of the correlation coefficient is shown at the bottom right corner of each plot. The 200 lowest scoring structures selected for the representative ensemble from each run are shown in red, and the closest-to-the-mean structures of these ensembles are highlighted in yellow. The correlation is stronger when the structures are ordered by score 1, and, consequently, the closest-to-the-mean structure for the ensemble of 200 lowest scoring structures is also closer to the lowest scoring structure.

native complex structures among the docking solutions but at the same time generates low-energy structures that are further from the true complex.[34]

In terms of selecting structures for the representative ensemble, the first score may be deemed more appropriate, as it yields less structural diversity within the ensemble. On the other hand, we expect that score 2, which better reflects the highly electrostatic nature of the interface, will give higher rankings for solutions close to the true structure.

It is common practice when analyzing structures generated by docking protocols to group the structures into clusters, where each cluster contains a collection of structurally similar solutions. Such clustering of the solutions is appropriate for situations where the structures are naturally grouped into families that exhibit significant structural differences. However, when all the members of the ensemble are structurally similar, clustering these solutions into families is no longer appropriate. The structures resulting from our docking calculations did not group into structurally distinct families and instead constituted a single cluster. Hence, the solutions were analyzed using a procedure similar to that traditionally employed with solution NMR structural ensembles, whereby the lowest scoring

Table 1. Summary of iRMSDs (in angstrom) for ensembles containing the top 20, top 50, top 100, top 200 and all 500 water-refined structures as ranked by scores 1 and 2

	Top 20	Top 50	Top 100	Top 200	All 500
Score 1					
Run 1	1.27 (1.46) 2.71	1.47 (1.81) 2.53	1.46 (1.94) 2.22	*1.46 (1.93)* *2.22*	1.37 (2.09) 2.41
Run 2	1.41 (1.45) 2.60	1.40 (1.59) 2.44	1.40 (1.80) 2.44	*1.40 (1.81)* *2.44*	2.41 (2.11) 2.09
Run 3	1.63 (1.39) 2.63	1.18 (1.53) 2.62	1.45 (1.62) 2.23	*1.45 (1.85)* *2.23*	1.85 (2.03) 2.20
Run 4	1.19 (1.63) 2.85	1.19 (1.78) 2.85	1.03 (1.82) 2.75	*1.50 (1.89)* *2.40*	2.05 (2.13) 2.60
Score 2					
Run 1	1.69 (1.88) 1.88	3.20 (2.10) 2.47	3.20 (2.03) 2.47	*3.03 (1.98)* *2.22*	3.17 (2.09) 2.41
Run 2	1.95 (1.99) 1.53	2.44 (2.03) 1.98	2.44 (1.91) 2.56	*2.54 (1.92)* *2.48*	2.03 (2.11) 2.09
Run 3	3.11 (1.58) 2.41	3.11 (1.89) 2.41	2.43 (1.93) 1.89	*3.07 (1.87)* *2.23*	2.73 (2.03) 2.20
Run 4	1.93 (1.88) 1.44	1.93 (1.95) 1.44	2.54 (1.84) 1.99	*3.19 (1.98)* *2.57*	2.88 (2.13) 2.60

The first figure in each box is the iRMSD between the structure closest to the mean and the lowest scoring structure; the second figure (in parentheses), the average iRMSD to the mean structure; and the third figure, the iRMSD between the structure closest to the mean and the crystal structure. The ensemble of 200 structures performs best in terms of reproducibility of the results in the four docking runs and was therefore chosen for the final analysis.

Table 2. Average pairwise iRMSDs (in angstrom) between closest-to-the-mean structures from the four runs for ensembles containing the top 20, top 50, top 100, top 200 and all 500 water-refined structures as ranked by scores 1 and 2

	Score 1	Score 2
Top 20	1.01	1.68
Top 50	1.03	1.49
Top 100	1.03	1.22
Top 200	*0.92*	*0.96*
All 500	1.11	1.11

Selecting 200 structures as the representative ensemble leads to the lowest average pairwise iRMSD between the four closest-to-the-mean structures for both scores, and both values are less than 1 Å, demonstrating the excellent reproducibility of the results. The ensemble of 200 structures performs best in terms of reproducibility of the results in the four docking runs and was therefore chosen for the final analysis.

structures are selected as the representative ensemble and the structure closest to the mean structure of the ensemble is the representative single structure.

In order to determine the number of structures to select for the representative ensemble, we considered the iRMSD between the structure closest to the mean and the lowest scoring structure, with the criterion of obtaining similar iRMSDs for the four docking runs (Table 1). This criterion ensures that the chosen ensemble of complex structures can be stably reproduced in different docking runs and is independent of the instabilities in the lowest-scoring solutions. These instabilities are expected for ensemble docking, namely when flexible interfaces are represented by an ensemble of input conformations, as in the present case for the flexible U4 penta-loop and the flexible loop in hPrp31.

When the 200 lowest scoring structures ranked by score 1 were selected, there was agreement between the four runs to an accuracy of 0.1 Å, with an iRMSD of approximately 1.5 Å between the structure closest to the mean and the lowest scoring structure (Table 1). When ranking the structures according to score 2, there was a larger average iRMSD between the closest-to-the-mean and lowest scoring structures, and more variability among the runs, as shown in Fig. 4 and Fig. S5. However, selecting 200 structures gave a much improved agreement between the runs compared to ensembles with fewer structures. The agreement between the runs deteriorates again when all 500 structures are considered.

In addition to the iRMSD between the closest-to-the-mean and lowest scoring structures for each run, the pairwise iRMSDs between the closest-to-the-mean structures from the four runs can be used to assess the number of structures required to obtain good reproducibility between the runs (Table 2). For both scores, there is a decrease in the average pairwise iRMSD between the closest-to-the-mean structures as the number of structures increases; as expected, the reduction is more pronounced for score 2. When 200 structures are selected, both scores yield close agreement between the runs, with average iRMSDs between the closest-to-the-mean structures of less than 1 Å, demonstrating the excellent reproducibility of the docking protocol. The average pairwise iRMSD increases when all 500 structures are considered.

Thus, both criteria suggest using 200 structures as the representative ensemble for both scores.

Assessment of docking performance

As the structure of the hPrp31$^{188–332}$–15.5K–U4 complex has recently been solved by X-ray diffraction, we were in the fortunate position to be able both to assess the performance of the docking protocol by comparing the docking solutions with

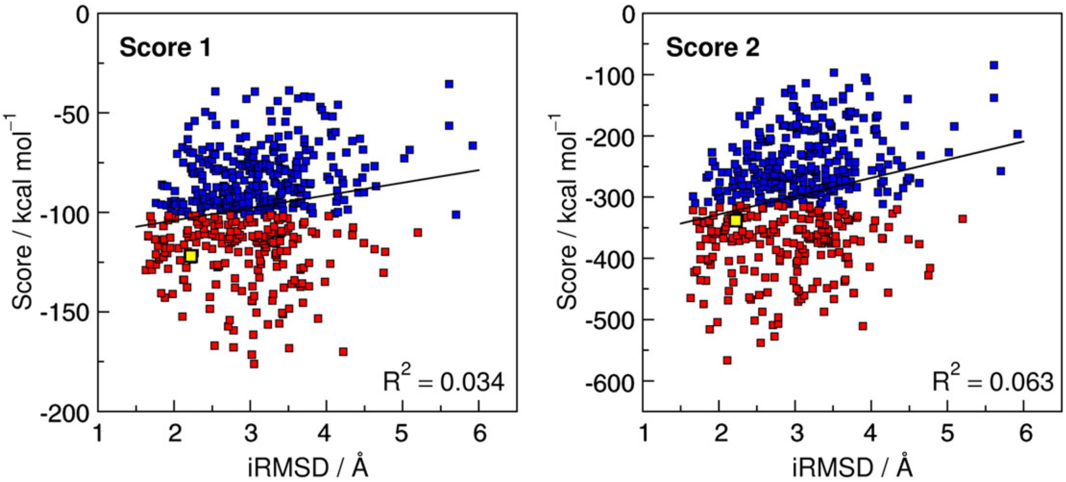

Fig. 5. Plots for the first run of score *versus* iRMSD to the crystal structure for score 1 (left) and score 2 (right). The square of the correlation coefficient is shown at the bottom right corner of each plot. The 200 lowest scoring structures selected for the representative ensemble from each run are shown in red, and the closest-to-the-mean structures of these ensembles are highlighted in yellow. Score 2 yields the stronger correlation with the iRMSD to the crystal structure, reflecting the importance of electrostatic interactions in the formation of the RNP. However, the closest-to-the-mean structure (yellow), which is representative of the ensemble, is the same for both scores.

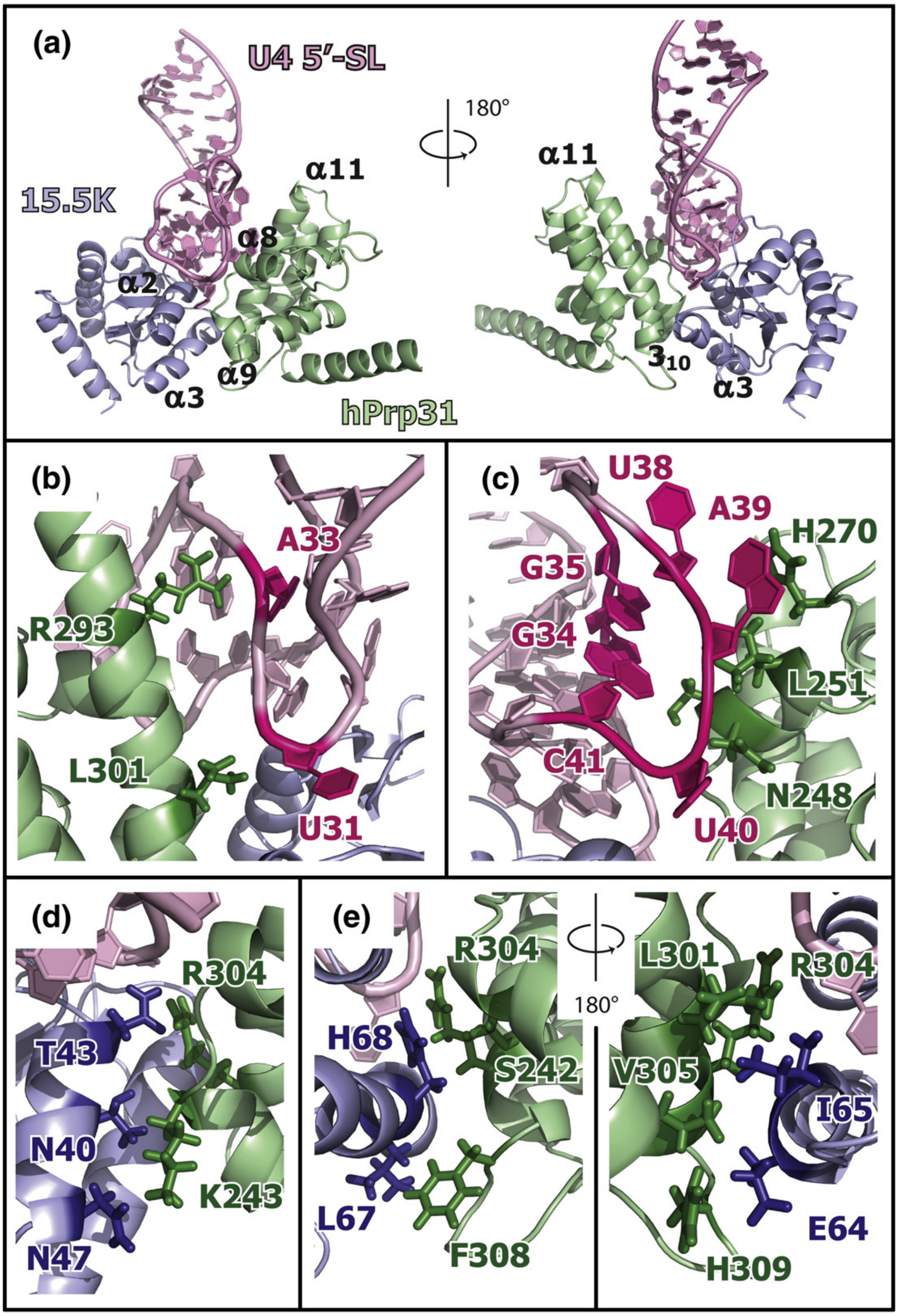

Fig. 6. (a) Overview of the docking model for the hPrp31$^{188–332}$–15.5K–U4 5′-SL complex (structure closest to the mean for first docking run). hPrp31$^{188–332}$ is shown in green; 15.5K, in blue; and RNA, in pink. The hPrp31$^{188–332}$ has extensive contacts with both the RNA and the 15.5K. (b and c) Close-up views of the interaction of hPrp31$^{188–332}$ with the RNA. In particular, R293 and L301 of hPrp31 are close to the phosphate backbones of A33 and U31, respectively. H270 is interacting closely with A39, and residues C247, N248 and L251 of helix α9 are packed against the major groove of stem II and the penta-loop. (d) hPrp31 shows several electrostatic interactions with the α2 helix of 15.5K, involving in particular salt bridges between R304 and N40 and between K243 and N47. (e) The interactions between hPrp31 and the 3_{10} and α3 helices of 15.5K are mostly of a hydrophobic nature: F308 stacks below the side chain of L67 and H68 has contacts to R304 and S242, while I65 and E64 are interacting with several residues at the C-terminal end of helix α11.

the crystal structure and to evaluate the two scoring functions with respect to their ability to select those solutions closest to the true structure. Plots of score *versus* iRMSD to the crystal structure for both scores are shown in Fig. 5 (first run; for all four runs, see Fig. S6). The correlation for score 2 is higher than that for score 1, revealing the importance of the electrostatic contribution to the energetics of complex formation in this RNP. This is reflected in the iRMSDs between the closest-to-the-mean structures and the crystal structure for representative ensembles with fewer structures (Table 1)—these iRMSDs are lower for score 2 than for score 1 (averages of 2.7 and 1.8 Å over the four runs when taking 20 structures for scores 1 and 2, respectively). However, when 200 structures are considered, the two scores give very similar results, with average iRMSDs of ~2.3 Å for both scores, demonstrating both the accuracy and the reliability of the docking.

Overall, these results indicate that selecting 200 of the 500 water-refined structures leads to highly reproducible and accurate results, largely removing the dependency on the exact scoring function employed and giving closest-to-the-mean structures for multiple runs that exhibit iRMSDs of less than 1 Å to one another and less than 2.5 Å to the crystal structure.

Comparison of docking structures with the crystal structure

The representative structure from the first run (which is the same for both scores when 200 structures are considered) is shown in Fig. 6a. Here, the RNA is sandwiched between the 15.5K

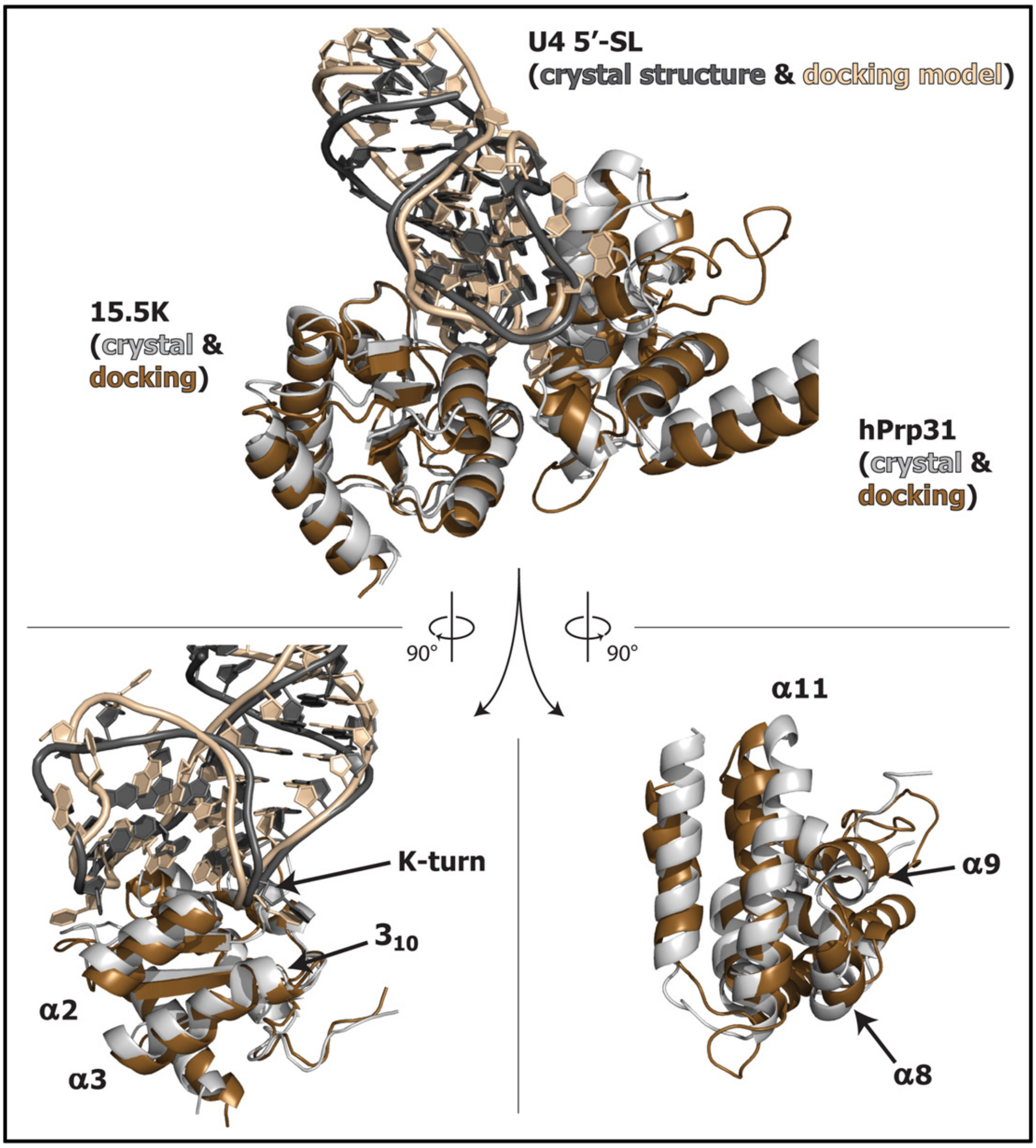

Fig. 7. Overlap of the representative docking model of the hPrp31$^{188–332}$–15.5K–U4 5′-SL complex from the first docking run (dark brown, proteins; pale brown, RNA) with the crystal structure (light gray, proteins; dark gray, RNA). The agreement of the two complex architectures is excellent.

and the hPrp31. The hPrp31 surface buried by formation of the complex covers an area of 1150 Å^2, distributed almost equally between the RNA and the 15.5K (570 and 580 Å^2, respectively). R293 of hPrp31 contacts the phosphate backbone of A33 (Fig. 6b), in agreement with the mutational data, underlining the importance of R293 in binding the binary RNP. hPrp31 residue L301 is in contact with U31 of the RNA K-turn, and H270 is interacting closely with A39 of the penta-loop (Fig. 6c). The hPrp31–15.5K interprotein interactions are also extensive (Fig. 6d and e). The α2 helix of 15.5K shows primarily electrostatic contacts with hPrp31, involving two salt bridges between hPrp31-R304 and 15.5K-N40 and between hPrp31-K243 and 15.5K-N47. E64, I65, L67 and H68 in the 3_{10} and α3 helices have significant van der Waals interactions with several hPrp31 residues at the C-terminal end of helix α11 (Fig. 6e). Thus, the docking model of the ternary complex identifies the Nop domain of hPrp31 as a true RNP recognition domain that binds to a composite platform composed of both the U4 RNA and 15.5K with extensive protein–protein and protein–RNA contacts.

This docking structure also shows excellent agreement with the hPrp31$^{78-333}$–15.5K–U4 5′-SL crystal structure (iRMSD = 2.22 Å), faithfully reproducing the three-dimensional architecture of the complex (Fig. 7). Closer inspection of the interfacial regions reveals that the overlap between the docking model and the crystal structure is slightly better for the 15.5K–U4 component (iRMSD = 1.71 Å) than for the hPrp31 (iRMSD = 2.72 Å). In particular, there is a slight tilt of helix α11 and a small translational offset of helices α8 and α9. However, despite these small differences, the protein–protein contacts observed in the crystal structure[11] are very well reproduced in the docking model, even at an atomic level, featuring the same salt bridges between hPrp31-R304 and 15.5K-N40 and between hPrp31-K243 and 15.5K-N47 and the same van der Waals contacts between hPrp31-L301/F308 and 15.5K-I65/L67 (Fig. 6).

Discussion

Many cellular complexes are of a dynamic nature as the plasticity of intermolecular interactions is essential for functional regulation. Such plasticity often implies that macromolecules are constituted by well-folded domains in combination with large unstructured regions. The presence of unfolded regions renders crystallization a tedious task that requires trimming of the macromolecules down to their folded core. This procedure, however, carries the intrinsic risk that fragment constructs may exhibit non-native interactions. As an alternative to crystallization of complexes consisting of fragment components, the native complex can be studied in solution by NMR. NMR is an exquisitely powerful technique to study both folded and unfolded biomolecules in an aqueous environment. The limitation here lies in the size of the macromolecular assembly. A complete high-resolution structure determination by NMR is limited to species with a molecular mass of less than 50 kDa,[35,36] which is smaller than most cellular complexes of interest. However, recent technological advances[18–21] have allowed visualization of species as large as 900 kDa.[22] The structural information that can be obtained in these cases is limited and requires prior knowledge of the spectroscopic chemical shifts. Nonetheless, it is conceivable that a complex of the size of 100 kDa or more can be constructed through a stepwise procedure that includes (i) detailed NMR investigations of the single, smaller components of the complex; (ii) acquisition of structural information on the intermolecular contacts in the native complex by NMR; and (iii) a restraints-driven docking of the single components to assemble the macromolecular complex. The characterization of intermolecular interfaces by NMR is a feasible task, even for large complexes, as it takes advantage of the chemical shift assignments obtained for the single-molecule components.

We have applied this procedure to the ternary RNP complex constituted by the U4 5′-SL RNA and the 15.5K and hPrp31 proteins. This example is particularly challenging due to the fact that the hPrp31 expresses poorly in *E. coli* and hence cannot be obtained with the ^{15}N, ^{13}C and ^{2}D labeling necessary for its visualization by NMR. Thus, the NMR-based intermolecular restraints available in this case are limited to the identification of the surface of 15.5K in contact with hPrp31, as observed from the amide resonances of 15.5K. These data are supplemented with cross-linking and mutagenesis data, which define two anchor points connecting the U4 5′-SL RNA to hPrp31. In summary, the protein–protein interaction interface in the ternary complex is defined by NMR data, while the hPrp31–RNA interface is defined by cross-linking and mutagenesis data. This highly ambiguous set of restraints, obtained with a combination of techniques, all applied to native complexes in solution, is used in a restrained docking protocol (HADDOCK)[25,26] using the crystal structure of the binary 15.5K–U4 5′-SL complex and a homology model of the hPrp31$^{188–332}$ fragment as the building blocks. The protocol successfully converges to one cluster of complex structures, which can be represented by a subset of the lowest scoring structures.

Ranking of structures obtained by docking protocols is a recognized problem, and the most appropriate scoring function depends on the nature of the complex. Here, we evaluated two scoring functions that differed in the contribution of the electrostatic energy to the total score. The score in which the electrostatic energy was weighted more strongly gave more structural variability between similarly scoring structures, but the lowest scoring structures were in closer agreement to the crystal structure. Selecting a small subset of the lowest scoring structures gave a closest-to-the-mean structure in very good agreement with the crystal structure, but with some variability between duplicate runs. Therefore, we selected 200 of 500 water-refined structures as the representative ensemble, which, while slightly increasing the iRMSD of the closest-to-the-mean structure to the crystal structure, leads to excellent stability between different runs. The stability of the representative model structure across different runs is a good measure of the reproducibility of the results and therefore of the uniqueness of the solution.

Comparison of the docking model obtained for the hPrp31–15.5K–U4 5′-SL complex with the recently determined crystal structure reveals that the docking model is accurate and reproduces all the features of the three-dimensional architecture of the complex (Fig. 7). Furthermore, the atomic details of the protein–protein interaction surface, defined by the NMR data in the docking model, also show excellent agreement to the crystal structure (Fig. 6). Both electrostatic and van der Waals contacts between the 15.5K and hPrp31 proteins are correctly predicted by the model, underlining that exceptional accuracy can be obtained at an atomic level even when using sparse and highly ambiguous NMR restraints.

A critical point in the optimization of the docking protocol was the use of ensemble starting structures for both the 15.5K–U4 5′-SL binary complex and the hPrp31 protein. The involvement of two flexible regions (RNA penta-loop and the 253–272 loop of hPrp31) in the intermolecular interface calls for the description of the conformational plasticity at these sites. When the docking calculations are performed using single conformers for both the 15.5K–U4 5′-SL complex and the hPrp31 protein, the results are less accurate and poorly reproducible as they depend on the precise structure of the chosen conformer. Much improved results can be achieved by sampling the conformational space of the RNA penta-loop and the hPrp31 253–272 loop in the starting structures. This observation supports the "conformer selection" theory, according to which binding-competent conformers can be found in ensemble structures of the unbound components and can be selected on the basis of their low binding energy. However, while the use of ensemble starting structures provides access to accurate complex models, it also complicates the procedure of selecting the representative structure, as it generates an inhomogeneous pool of low-scoring structures (Fig. 4).[34] We have shown that such difficulties can be overcome by averaging over a larger number (200) of low-scoring structures.

To the best of our knowledge, this is the first time that HADDOCK has been used to obtain a model of a ternary RNP. Given the paucity and the highly ambiguous nature of the restraints, together with the involvement of flexible regions at the intermolecular interface, the protocol performs exceptionally well (Fig. 7). The results shown here prove the feasibility of assembling large macromolecular complexes using a few NMR data, easily accessible even for large complexes at low concentrations, complemented by biochemical data. The protocol presented here is of broad applicability and provides a reliable route to the determination of the three-dimensional architecture of multimeric complexes in a native environment.

The docking model for the hPrp31–15.5K–U4 5′-SL complex provides a rationale for the hierarchical assembly of the RNP. The hPrp31 recognizes the U4 5′-SL RNA only in the presence of 15.5K.[15] Two mechanisms could explain the sequential assembly of the ternary complex: (i) the 15.5K folds the RNA into a conformation that is competent for hPrp31 binding but does not itself contact the hPrp31 protein and (ii) the 15.5K shares an interaction surface with the hPrp31 protein, thus directly contributing to the binding energy. The first mechanism is the most commonly found in RNP assembly pathways.[37] Nevertheless, the NMR experiments disprove this mechanism in the case of the hPrp31–15.5K–U4 5′-SL complex and demonstrate that the two proteins interact with each other through an extensive surface involving helices α2, 3_{10} and α3 of 15.5K (Fig. 2). The docking model shows that hPrp31 contacts 15.5K with a surface of approximately the same area as that in contact with the RNA, ultimately identifying the hPrp31$^{188–332}$ fragment as a true RNP recognition domain.

At present, no structural information is available for the C-terminal domain of hPrp31. This part of the protein does not show any homology with known folded domains and hindered crystallization of the full-length hPrp31–15.5K–U4 5′-SL complex.[11] Both these observations point to the presence of a largely unfolded amino acid sequence. The identity of the protein–protein interactions detected by NMR for the hPrp31–15.5K–U4 5′-SL and hPrp31$^{78–333}$–15.5K–U4 5′-SL complexes excludes the possibility that the C-terminal domain of hPrp31 contacts 15.5K. Thus, it may be hypothesized that this part of hPrp31 interacts with the RNA. Indeed, a stem I containing eight rather than four base pairs has previously been found necessary in the 15.5K–U4 5′-SL complex for high-affinity binding of hPrp31.[38] In agreement with this hypothesis, residues G26–C28 in stem I have been shown to be protected from hydroxyl radical digestion upon binding of hPrp31 but not of hPrp31$^{78–333}$.[38] In the docking model of the ternary complex, the C-terminal tail of hPrp31$^{188–332}$ is located close to the K-turn on the side of the minor groove of stem I and is therefore in a favorable position to interact with this region of the U4 RNA. Interestingly, a sequence of 15 basic residues (^{351}RKKRGGRRYRKMKER^{365}) is located downstream of the hPrp31$^{188–332}$ fragment, which could easily interact with the negatively charged RNA stem I (Fig. S7). Cross-saturation experiments on the full-length hPrp31–15.5K–U4 5′-SL complex using ^{13}C/^{15}N-labeled RNA and unlabeled hPrp31 and 15.5K proteins are in progress in our laboratory to verify this hypothesis.

In conclusion, we have demonstrated a robust protocol, based on a combination of NMR and biochemical data, that provides access to the three-dimensional structure of complexes with size of the order of 100 kDa. The protocol consists of restrained docking of the single components of the complex using information on the intermolecular interaction surfaces derived by NMR. Such information is accessible from simple N–H^N correlations and can be obtained for very large complexes at low concentrations (~0.1 mM). The NMR-derived restraints are complemented with mutagenesis and cross-linking data. We applied this approach to the hPrp31–15.5K–U4 5′-SL spliceosomal complex and were able to obtain a very accurate docking model, which differed by only 2.2 Å from the crystal structure of the hPrp31$^{78–333}$–15.5K–U4 5′-SL complex in the interfacial region.

The work presented here demonstrates that NMR investigation in solution can be successfully applied to large, weakly concentrated complexes and opens the way to the definition of the structural basis for hierarchical complex assembly even in the absence of crystal structures.

Materials and Methods

Protein expression, purification and *in vitro* reconstitution of RNPs

The uniform labeling of glutathione *S*-transferase–15.5K with ^{2}H and ^{15}N was carried out using minimal medium prepared in 99.9% ^2H_2O (Eurisotop) with 10% v/v of *E. coli* OD2 DN growth medium (Silantes). Full-length hPrp31 and hPrp31$^{78–333}$ constructs were expressed and purified in unlabeled forms. Expression and purification of the proteins mentioned above have been described in detail by Liu *et al.*[11] The purified unlabeled U4 5′-SL-33nt oligonucleotide was purchased from IBA BioTAGnology. The RNA was dialyzed into the buffer (10 mM Tris–HCl, pH 7.6, 120 mM NaCl and 2 mM DTT) before annealing at 65 °C for 90 s. hPrp31–15.5K–U4 5′-SL and hPrp31$^{78–333}$–15.5K–U4 5′-SL complexes were reconstituted *in vitro* as described by Liu *et al.*[11]

NMR experiments on hPrp31–15.5K–U4 5′-SL and hPrp31$^{78–333}$–15.5K–U4 5′-SL complexes

All samples were prepared in buffer containing 10 mM Tris–HCl, pH 7.6, 120 mM NaCl and 2 mM DTT. [^{15}N,^{1}H] transverse relaxation-optimized spectroscopy[18] experiments on 15.5K–U4 5′-SL, hPrp31–15.5K–U4 5′-SL and hPrp31$^{78–333}$–15.5K–U4 5′-SL complexes were carried out at 308 or 303 K on a Bruker 900-MHz spectrometer equipped with a cryo-probe. The sample concentrations were 0.1 mM for the ternary complex containing the full-length hPrp31 and 0.3 mM for the complex containing hPrp31$^{78–333}$. The total experiment time was 12 h. Chemical shift changes were calculated according to $\delta_{NH} = \surd((\delta_N/5)^2 + (\delta_H)^2)/2)$. For the cross-saturation experiments[23] on the same complexes, a buffer containing 40% ^1H_2O and 60% ^2H_2O was used to minimize spin diffusion mediated by ^1H_2O. The experimental time was 4 days. Saturation of the methyl region of hPrp31 (−0.1 ppm) was achieved by application of WURST-2 decoupling for 800 ms with a maximum radiofrequency amplitude of 6 kHz and a pulse length of 40 ms. Two experiments were recorded in an interleaved manner: a reference experiment without saturation and a second experiment with the use of the saturation scheme. Peaks were integrated in both saturated and unsaturated (reference) spectra, and the intensity changes were calculated according to $I_{\mathrm{change}} = I_{\mathrm{unsat}}/I_{\mathrm{sat}}$. As small saturation transfer effects can also be mediated by residual protons in the 15.5K itself, a control experiment was performed using the 15.5K–U4 binary complex in the absence of hPrp31. The intensity changes in the control experiment were scaled by a multiplicative factor defined by the ratio of the average intensity changes in the ternary complexes (with full-length hPrp31 and hPrp31$^{78–333}$) to those in the binary complex: $I_{\mathrm{change,mean}}(\mathrm{ternary})/I_{\mathrm{change,mean}}(\mathrm{binary})$. These factors account for the different size ratios between the ternary complexes and the 15.5K–U4 binary complex, which result in different rates of spin diffusion. The scaled control intensity changes were then subtracted from the intensity changes for the ternary complexes to give the final normalized intensity changes, $I_{\mathrm{change,norm}}$.

Generation of the 15.5K–U4 5′-SL-33nt model

The U4 5′-SL-33nt construct has previously been identified as containing the minimal binding site for hPrp31[15] and was therefore used in the NMR structural analysis. To generate a structural model of this binary complex to be used in the docking to hPrp31, we used the crystal structure of the 15.5K–U4 5′-SL-22nt complex as a template.[14] Besides elongation of stem I (Fig. 1), the modeling included the generation of structures for the penta-loop, which is not present in the crystallographic coordinates but has been shown to form direct contacts with hPrp31.[38] Models of the binary 15.5K–U4 5′-SL-33nt complex were generated by first adding the elongated stem I and the penta-loop, followed by randomization of the penta-loop conformation and subsequent water refinement of the entire complex. Nucleotides of the penta-loop (36–40) and of the missing part of stem I (20–25 and 48–52) were built as three separate A-form strands (Insight II, Accelrys). These were then manually incorporated at approximately the correct positions in the 15.5K–U4-22nt crystal structure, and the resulting assembly was saved as a single Protein Data Bank file. An ensemble of 150 15.5K–U4-33nt structures for docking was then generated from this starting point in two stages. In the first stage, the modeled stem I strands and the penta-loop were connected to the 22-nt U4 RNA fragment using a simulated annealing–molecular dynamics (SA-MD) protocol (implemented in the XPLOR-NIH simulation program)[39,40] in which pseudo-NOEs were given at the junctions of the modeled and 22-nt RNA fragments. This protocol was repeated with different initial velocities to yield 150 structures with randomized penta-loop conformations. In the second stage, these structures were water refined using a slow-cooling SA-MD protocol (XPLOR-NIH). During both stages, the U4 RNA was bound to 15.5K using a set of pseudo-NOEs derived from the crystal structure of the binary 15.5K–U4 5′-SL-22nt complex.[14] In a similar manner, the structure of the 22-nt U4 fragment was restrained using pseudo-NOEs measured from the crystal structure in combination with appropriate planarity and hydrogen-bond restraints. In addition, the complete stem I was restrained to a standard A-form helical structure using dihedral restraints.

Comparative modeling of hPrp31$^{188–332}$

Before the crystal structure of the ternary hPrp31$^{78–333}$–15.5K–U4 5′-SL complex was solved, the only available template for modeling of the human hPrp31 was the structure of the *A. fulgidus* Nop5p.[24] As the goal of this study was to propose an alternative method to assemble multimeric complexes when crystallization fails, we did not use any structural information derived from the structure of the hPrp31$^{78–333}$–15.5K–U4 5′-SL complex[11] in the docking. Therefore, a model of the hPrp31$^{188–332}$ fragment was obtained by comparative modeling with the Nop5p structure. The sequences of hPrp31 and Nop5p were aligned using ClustalX v.1.83.[41] Amino acids 186–334 in hPrp31, containing the Nop domain (residues 215–333), and 112–261 in Nop5p showed excellent alignment (28% sequence identity) (Fig. S2). Secondary structure prediction for both proteins, performed using the PredictProtein server,[29] showed strikingly similar structural motifs (data not shown). The disordered loop region in Nop5p, comprising residues 182–201 (KSLYKAFARMKKGKKAKIPK), was aligned to residues 256–270 (RKTLSGFSSTSVLPH) in hPrp31$^{188–332}$. Comparative modeling was carried out using the "first-approach mode" provided by the SWISS-MODEL server.[42] After a short minimization, the model was subjected to an SA-MD protocol (XPLOR-NIH) consisting of heating to 3000 K (300 steps of 1 fs), 40,000 steps (2 fs) at 3000 K and cooling to 25 K in 300 steps (1 fs). All atomic positions of the hPrp31$^{188–332}$ model were restrained, except for loop amino acids 253–272 (GAQRKTLSGFSSTSV-LPHTG). One hundred fifty conformations of the 253–272 loop region were generated using different initial velocities. Each of the 150 models consists of eight helices, with the helices corresponding to α8, α9, α11 and α12 of the Nop5p Nop domain, forming a roughly flat surface. The 150 structures were then refined in the presence of water. The refinement protocol consisted of heating to 300 K in 100 steps (3 fs), followed by 3000 steps (4 fs) at 300 K.

Docking using HADDOCK 2.0

Electrostatic surfaces of hPrp31$^{188–332}$ and 15.5K–U4 5′-SL-33nt were calculated from the Protein Data Bank files using the PDB2PQR server[43] and the Adaptive Poisson–Boltzmann Solver software.[44] The ambiguous restraints between the two components of the ternary complex were assigned by considering the surface charge complementarity and the available biochemical and NMR data. A detailed description of the rationale behind the choice of the restraints is provided in Results. In summary, the ambiguous interaction restraints were given as follows: R293 of hPrp31$^{188–332}$ to nucleotides 32–35 and 41–43 (stem II and K-turn) of U4 5′-SL; residues N40, T43, N47 and R48 of the α2 helix of 15.5K to A297, K298, T300, L301, R304 and K243 of hPrp31$^{188–332}$; N40 and T43 of 15.5K to residues 245–248 of hPrp31$^{188–332}$; and residues L63, I65, I66, L68 and L71 of the α3 helix of 15.5K to T239, N240, K243, T300, L301, R304, V305 and F308 of hPrp31$^{188–332}$. H270 was given NOE restraints to nucleotide A39 of the penta-loop. HADDOCK 2.0[25,26] was used to generate the hPrp31$^{188–332}$–15.5K–U4 5′-SL-33nt complex model.

The 150 models of hPrp31$^{188–332}$ differing in the conformation of residues 253–272 and the 150 models of the 15.5K–U4 5′-SL-33nt complex with randomized penta-loop conformations were used as the initial structures for the two components. The semi-flexible interface of hPrp31$^{188–332}$ was defined as residues 237–250 and 291–310, and that of 15.5K–U4 5′-SL-33nt consisted of residues 38–50 and 61–74 in the 15.5K and nucleotides 32–35 and 41–43 in U4. Fully flexible segments were defined as residues 252–272 for hPrp31$^{188–332}$ and nucleotides of the RNA penta-loop (36–40) for 15.5K–U4 5′-SL-33nt. Other parameters were assigned the default HADDOCK values. A total of 22,500 structures corresponding to all combinations of the 150 hPrp31$^{188–332}$ and 150 15.5K–U4 starting structures was calculated in the rigid-body stage. The 1000 lowest scoring structures were refined in the semi-flexible annealing stage. The 500 lowest scoring structures after semi-flexible annealing were subjected to a final water refinement stage. The scoring functions for ordering the structures after the rigid-body and semi-flexible stages were the HADDOCK defaults. The scoring functions for ordering the water-refined structures are described in the main text. The docking protocol was repeated four times to assess the reproducibility of the resulting structures.

Acknowledgements

This work was supported by the Max Planck Institute and the European Molecular Biology Laboratory. We thank Devanathan Raghunathan and Marcel Reese for their help with setting up the modeling protocol. We also thank Reinhard Lührmann and Markus Wahl of the Max Planck Institute for stimulating discussions.

Supplementary Data

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.jmb.2009.03.001

References

1. Burge, C. B., Tuschl, T. & Sharp, P. A. (1999). *The RNA World.* Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY.
2. Will, C. L. & Lührmann, R. (2006). *The RNA World.* Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY.
3. Nilsen, T. W. (2003). The spliceosome: the most complex macromolecular machine in the cell? *BioEssays*, **25**, 1147–1149.
4. Gornemann, J., Kotovic, K. M., Hujer, K. & Neugebauer, K. M. (2005). Cotranscriptional spliceosome assembly occurs in a stepwise fashion and requires the cap binding complex. *Mol. Cell*, **19**, 53–63.
5. Kambach, C., Walke, S. & Nagai, K. (1999). Structure and assembly of the spliceosomal small nuclear ribonucleoprotein particles. *Curr. Opin. Struct. Biol.* **9**, 222–230.
6. Reed, R. & Palandjian, L. (1997). *Eukaryotic mRNA Processing.* IRL Press, Oxford, UK.
7. Brow, D. A. (2002). Allosteric cascade of spliceosome activation. *Annu. Rev. Genet.* **36**, 333–360.
8. Staley, J. P. & Guthrie, C. (1998). Mechanical devices of the spliceosome: motors, clocks, springs, and things. *Cell*, **92**, 315–326.
9. Bringmann, P., Appel, B., Rinke, J., Reuter, R., Theissen, H. & Luhrmann, R. (1984). Evidence for the existence of snRNAs U4 and U6 in a single ribonucleoprotein complex and for their association by intermolecular base-pairing. *EMBO J.* **3**, 1357–1363.
10. Hashimoto, C. & Steitz, J. A. (1984). U4 and U6 RNAs coexist in a single small nuclear ribonucleoprotein particle. *Nucleic Acids Res.* **12**, 3283–3293.
11. Liu, S., Li, P., Dybkov, O., Nottrott, S., Hartmuth, K., Luhrmann, R. *et al.* (2007). Binding of the human Prp31 Nop domain to a composite RNA–protein platform in U4 snRNP. *Science*, **316**, 115–120.

12. DeLano, W. L. (2002). *The PyMOL Molecular Graphics System.* DeLano Scientific, Palo Alto, CA.
13. Nottrott, S., Hartmuth, K., Fabrizio, P., Urlaub, H., Vidovic, I., Ficner, R. & Luhrmann, R. (1999). Functional interaction of a novel 15.5kD [U4/U6·U5] tri-snRNP protein with the 5′ stem–loop of U4 snRNA. *EMBO J.* **18**, 6119–6133.
14. Vidovic, I., Nottrott, S., Hartmuth, K., Luhrmann, R. & Ficner, R. (2000). Crystal structure of the spliceosomal 15.5kD protein bound to a U4 snRNA fragment. *Mol. Cell*, **6**, 1331–1342.
15. Nottrott, S., Urlaub, H. & Luhrmann, R. (2002). Hierarchical, clustered protein interactions with U4/U6 snRNA: a biochemical role for U4/U6 proteins. *EMBO J.* **21**, 5527–5538.
16. Schneider, C., Will, C. L., Makarova, O. V., Makarov, E. M. & Luhrmann, R. (2002). Human U4/U6·U5 and $U4_{atac}/U6_{atac}$·U5 tri-snRNPs exhibit similar protein compositions. *Mol. Cell. Biol.* **22**, 3219–3229.
17. Liu, S. B., Rauhut, R., Vornlocher, H. P. & Luhrmann, R. (2006). The network of protein–protein interactions within the human U4/U6·U5 tri-snRNP. *RNA*, **12**, 1418–1430.
18. Pervushin, K., Riek, R., Wider, G. & Wuthrich, K. (1997). Attenuated T-2 relaxation by mutual cancellation of dipole–dipole coupling and chemical shift anisotropy indicates an avenue to NMR structures of very large biological macromolecules in solution. *Proc. Natl Acad. Sci. USA*, **94**, 12366–12371.
19. Riek, R., Wider, G., Pervushin, K. & Wuthrich, K. (1999). Polarization transfer by cross-correlated relaxation in solution NMR with very large molecules. *Proc. Natl Acad. Sci. USA*, **96**, 4918–4923.
20. Rosen, M. K., Gardner, K. H., Willis, R. C., Parris, W. E., Pawson, T. & Kay, L. E. (1996). Selective methyl group protonation of perdeuterated proteins. *J. Mol. Biol.* **263**, 627–636.
21. Tugarinov, V., Hwang, P. M., Ollerenshaw, J. E. & Kay, L. E. (2003). Cross-correlated relaxation enhanced H-1–C-13 NMR spectroscopy of methyl groups in very high molecular weight proteins and protein complexes. *J. Am. Chem. Soc.* **125**, 10420–10428.
22. Fiaux, J., Bertelsen, E. B., Horwich, A. L. & Wuthrich, K. (2002). NMR analysis of a 900K GroEL–GroES complex. *Nature*, **418**, 207–211.
23. Takahashi, H., Nakanishi, T., Kami, K., Arata, Y. & Shimada, I. (2000). A novel NMR method for determining the interfaces of large protein–protein complexes. *Nat. Struct. Biol.* **7**, 220–223.
24. Aittaleb, M., Rashid, R., Chen, Q., Palmer, J. R., Daniels, C. J. & Li, H. (2003). Structure and function of archaeal box C/D sRNP core proteins. *Nat. Struct. Biol.* **10**, 256–263.
25. De Vries, S. J., van Dijk, A. D. J., Krzeminski, M., van Dijk, M., Thureau, A., Hsu, V. *et al.* (2007). HADDOCK *versus* HADDOCK: new features and performance of HADDOCK 2.0 on the CAPRI targets. *Proteins: Struct., Funct., Bioinf.* **69**, 726–733.
26. Dominguez, C., Boelens, R. & Bonvin, A. (2003). HADDOCK: a protein–protein docking approach based on biochemical or biophysical information. *J. Am. Chem. Soc.* **125**, 1731–1737.
27. Kuhn, J. F., Tran, E. J. & Maxwell, E. S. (2002). Archaeal ribosomal protein L7 is a functional homolog of the eukaryotic 15.5kD/Snu13p snoRNP core protein. *Nucleic Acids Res.* **30**, 931–941.
28. Rozhdestvensky, T. S., Tang, T. H., Tchirkova, I. V., Brosius, J., Bachellerie, J. P. & Huttenhofer, A. (2003). Binding of L7Ae protein to the K–turn of archaeal snoRNAs: a shared RNA binding motif for C/D and H/ACA box snoRNAs in archaea. *Nucleic Acids Res.* **31**, 869–877.
29. Rost, B., Yachdav, G. & Liu, J. F. (2004). The PredictProtein Server. *Nucleic Acids Res.* **32**, W321–W326.
30. Dennis, P. P. & Omer, A. (2005). Small non-coding RNAs in archaea. *Curr. Opin. Microbiol.* **8**, 685–694.
31. Omer, A. D., Ziesche, S., Decatur, W. A., Fournier, M. J. & Dennis, P. P. (2003). RNA-modifying machines in archaea. *Mol. Microbiol.* **48**, 617–629.
32. Omer, A. D., Ziesche, S., Ebhardt, H. & Dennis, P. P. (2002). *In vitro* reconstitution and activity of a C/D box methylation guide ribonucleoprotein complex. *Proc. Natl Acad. Sci. USA*, **99**, 5289–5294.
33. van Dijk, M., van Dijk, A. D. J., Hsu, V., Boelens, R. & Bonvin, A. (2006). Information-driven protein–DNA docking using HADDOCK: it is a matter of flexibility. *Nucleic Acids Res.* **34**, 3317–3325.
34. Chaudhury, S. & Gray, J. J. (2008). Conformer selection and induced fit in flexible backbone protein–protein docking using computational and NMR ensembles. *J. Mol. Biol.* **381**, 1068–1087.
35. Choy, W. Y., Tollinger, M., Mueller, G. A. & Kay, L. E. (2001). Direct structure refinement of high molecular weight proteins against residual dipolar couplings and carbonyl chemical shift changes upon alignment: an application to maltose binding protein. *J. Biomol. NMR*, **21**, 31–40.
36. Gardner, K. H., Zhang, X. C., Gehring, K. & Kay, L. E. (1998). Solution NMR studies of a 42 KDa *Escherichia coli* maltose binding protein beta-cyclodextrin complex: chemical shift assignments and analysis. *J. Am. Chem. Soc.* **120**, 11738–11748.
37. Agalarov, S. C., Prasad, G. S., Funke, P. M., Stout, C. D. & Williamson, J. R. (2000). Structure of the S15,S18–rRNA complex: assembly of the 30S ribosome central domain. *Science*, **288**, 107–112.
38. Schultz, A., Nottrott, S., Hartmuth, K. & Luhrmann, R. (2006). RNA structural requirements for the association of the spliceosomal hPrp31 protein with the U4 and $U4_{atac}$ small nuclear ribonucleoproteins. *J. Biol. Chem.* **281**, 28278–28286.
39. Schwieters, C. D., Kuszewski, J. J. & Clore, G. M. (2006). Using Xplor-NIH for NMR molecular structure determination. *Prog. Nucl. Mag. Res. Sp.* **48**, 47–62.
40. Schwieters, C. D., Kuszewski, J. J., Tjandra, N. & Clore, G. M. (2003). The Xplor-NIH NMR molecular structure determination package. *J. Magn. Reson.* **160**, 65–73.
41. Jeanmougin, F., Thompson, J. D., Gouy, M., Higgins, D. G. & Gibson, T. J. (1998). Multiple sequence alignment with ClustalX. *Trends Biochem. Sci.* **23**, 403–405.
42. Schwede, T., Kopp, J., Guex, N. & Peitsch, M. C. (2003). SWISS-MODEL: an automated protein homology-modeling server. *Nucleic Acids Res.* **31**, 3381–3385.
43. Dolinsky, T. J., Nielsen, J. E., McCammon, J. A. & Baker, N. A. (2004). PDB2PQR: an automated pipeline for the setup of Poisson–Boltzmann electrostatics calculations. *Nucleic Acids Res.* **32**, W665–W667.
44. Baker, N. A., Sept, D., Joseph, S., Holst, M. J. & McCammon, J. A. (2001). Electrostatics of nanosystems: application to microtubules and the ribosome. *Proc. Natl Acad. Sci. USA*, **98**, 10037–10041.