Entomological Research

Entomological Research **40** (2010) 104–112

RESEARCH PAPER

A three-dimensional model of the U1 small nuclear ribonucleoprotein particle

Jason A. SOMARELLI[1], Annia MESA[1], Ambrish ROY[2], Yang ZHANG[2] and Rene J. HERRERA[1]

1 Florida International University, College of Medicine, University Park, Miami, Florida
2 University of Kansas, Center for Bioinformatics, Lawrence, Kansas, USA

Correspondence
Rene J. Herrera, Florida International University, Human & Molecular Geneties College of Medicine, University Park, OE 304, Miami, FL 33199, USA. Email: herrerar@fiu.edu

Received 27 October 2009; accepted 13 January 2010.

doi: 10.1111/j.1748-5967.2010.00266.x

Abstract

Most of the pre-mRNAs in the eukaryotic cell are comprised of protein-coding exons and non-protein-coding introns. The introns are removed and the exons are ligated together, or spliced, by a large, macromolecular complex known as the spliceosome. This RNA-protein assembly is made up of five uridine-rich small nuclear RNAs (U1-, U2-, U4-, U5- and U6-snRNA) as well over 300 proteins, which form small nuclear ribonucleoprotein particles (snRNPs). Initial recognition of the 5′ exon/intron splice site is mediated by the U1 snRNP, which is composed of the U1 snRNA as well as at least ten proteins.

By combining structural informatics tools with the available biochemical and crystallographic data, we attempted to simulate a complete, three dimensional U1 snRNP from the silk moth, *Bombyx mori*. Comparison of our model with empirically derived crystal structures and electron micrographs pinpoints both the strengths and weaknesses in the *in silico* determination of macromolecular complexes. One of the most striking differences between our model and experimentally generated structures is in the positioning of the U1 snRNA stem-loops. This highlights the continuing difficulties in generating reliable, complex RNA structures; however, three-dimensional modeling of individual protein subunits by threading provided models of biological significance and the use of both automated and manual docking strategies generated a complex that closely reflects the assembly found in nature. Yet, without utilizing experimentally-derived contacts to select the most likely docking scenario, *ab initio* docking would fall short of providing a reliable model. Our work shows that the combination of experimental data with structural informatics tools can result in generation of near-native macromolecular complexes.

Key words: 3D protein modeling, *Bombyx mori*, I-TASSER, pre-mRNA splicing.

Introduction

The spliceosome is a large, ribonucleoprotein complex responsible for the removal of non-protein-coding introns and ligation of protein-coding exons of pre-mRNA to form mature mRNA (Moore & Sharp 1993). By including or excluding specific exons in a spatio-temporal manner, the spliceosome generates the incredible diversity of proteins observed in the eukaryotic cell. Five uridine-rich small nuclear RNAs (U1, U2, U4, U5 and U6) make up the scaffold of the macromolecule, onto which a multitude of protein splicing factors bind, forming small nuclear ribonucleoprotein particles (snRNPs) (Raker *et al.* 1999; Jurica *et al.* 2002; Mesa *et al.* 2008). The initial recognition of the 5′ exon/intron splice junction is made possible, in large part, by the U1 snRNP (Tatei *et al.* 1987; Eperon *et al.* 1989;

Kohtz *et al.* 1994), which is comprised of a core set of ten proteins: the heptameric Sm core, shared by U1, U2, U4 and U5, as well as the U1-specific proteins U1A, U1C and U1-70K (Yuo & Weiner 1989). The Sm core, made up of the SmB (or SmB′), SmD1, SmD2, SmD3, SmE, SmF and SmG proteins, forms a seven membered ring through which a highly conserved, single stranded region of the snRNA (known as the Sm binding site) passes (Mura *et al.* 2001). Formation of the Sm core around the snRNAs is necessary for hypermethylation of the 7-methyl-guanosine (m^7G) cap to a 2,2,7-trimethyl-guanosine (m3G) cap (Mattaj 1986; Kambach *et al.* 1999). Both the m3G cap and the Sm core are required for nuclear import of the snRNAs (Hamm *et al.* 1990; Fischer *et al.* 1993; Plessel *et al.* 1994; Kambach *et al.* 1999). Once inside the nucleus, the Sm core facilitates association of the U1-specific proteins, U1A, U1-70K and U1C, with the U1 snRNA (Nelissen *et al.* 1994). The U1A protein interacts directly with stem-loop II of U1 snRNA (Scherly *et al.* 1989). Although U1A is known to participate in the splicing reaction (Tang & Rosbash 1996), little is known about its exact functional role in this process. U1-70K binds to stem-loop I of U1 snRNA and is responsible for anchoring the U1 snRNP to the 5′ splice site through interactions between its arginine-serine rich (RS) domain and the SR proteins as well as other regulatory polypeptides (Salz *et al.* 2004). Although U1C does not associate directly with the U1 snRNA, the N-terminal 97 amino acids of U1-70K as well as the SmB protein allow the binding of U1C to the U1 snRNP (Muto *et al.* 2004). U1C, along with U1-70K, mediates binding of the U1 snRNP to the 5′ splice site (Heinrichs *et al.* 1990; Muto *et al.* 2004) and increases formation of the early spliceosomal complex (complex E) (Will *et al.* 1996).

Numerous experimental strategies have elucidated distinct protein-RNA and protein-protein contacts within the U1 snRNP, including biochemical (Scherly *et al.* 1989, 1991; Heinrichs *et al.* 1990; Lutz-freyermuth *et al.* 1990; Jensen *et al.* 1991; Nelissen *et al.* 1991, 1994; Tang & Rosbash 1996; Raker *et al.* 1999; Katsamba *et al.* 2001), X-ray crystallographic and nuclear magnetic resonance (NMR) (Hall 1994; Howe *et al.* 1994; Oubridge *et al.* 1994; Kambach *et al.* 1999; Mura *et al.* 2001; Nagai *et al.* 2001; Toro *et al.* 2001; Muto *et al.* 2004; Pommeranz krummel *et al.* 2009) and electron microscopy (Stark *et al.* 2001) technologies. By utilizing these empirically determined interaction points as a guide, we report here a complete three dimensional model of the U1 snRNP, generated with RNA/protein modeling and docking simulations. It is noteworthy that, overall, our model is similar in shape and size to a cryo-electron micrograph (EM) of the human U1 snRNP (Stark *et al.* 2001) as well as the crystal structure (Pommeranz krummel *et al.* 2009). The parallelism between the empirically assessed topology and the simulated model of the U1 snRNP is indicative of the power and utility in coupling biochemical experimentation with structural informatics. This work represents the first attempt to create and evaluate an *in silico* macromolecular assembly and presents the scientific community with a new application to the field of structural informatics.

Materials and methods

Identification of U1 snRNP proteins in the 6X genome of Bombyx mori

Proteins corresponding to *Bombyx mori* SmB, SmD1, SmD2, SmE, SmF, SmG and U1A were downloaded from the National Center for Biotechnology Information (NCBI) protein database (http://www.ncbi.nlm.nih.gov/) (GenBank accession numbers **NP_001091757**, **NP_001040404**, **NP_001077096**, **NP_001040370**, **NP_001093276,** **NP_001040405** and **NP_001037384**, respectively). To identify the remaining U1 proteins (SmD3, U1-70K and U1C) in the *Bombyx mori* genome, the corresponding *Drosophila melanogaster* proteins were used as queries to search the *Bombyx mori* 6X whole genome shotgun (WGS) database using a translated nucleotide BLAST (tBLASTn) (http://blast.ncbi.nlm.nih.gov/Blast.cgi). Genomic scaffolds containing each protein were downloaded from the NCBI nucleotide database (http://www.ncbi.nlm.nih.gov/) and translated in six frames using BioEdit version 7.0.5.3 (Hall 2004). Translated amino acid sequences were chosen from the *Bombyx mori* scaffolds that corresponded to each *Drosophila melanogaster* protein and checked for completeness using multiple alignments with corresponding proteins from *Homo sapiens*, *Apis mellifera*, *Aedes aegypti*, *Drosophila melanogaster*, *Drosophila pseudoobscura*, *Nasonia vitripens* and *Tribolium castaneum*.

Three dimensional modeling of the U1 snRNP proteins

Tertiary structures for each of the ten proteins of the U1 snRNP were generated using the I-TASSER server (http://zhang.bioinformatics.ku.edu/I-TASSER/), which was ranked as the best 3D structure prediction server in the recent CASP7 experiment (Zhang 2006; Battey *et al.* 2007). I-TASSER generates structure predictions by first threading query sequences through the Protein Data Bank (PDB) library and then reassembling the continuous fragments from templates into full-length atomic models (Zhang & Skolnick 2004a,b; Wu *et al.* 2007; Zhang 2008). The final models are ranked by the size of the structural clusters (i.e. the frequency of occurrence of the conformation appearing in the I-TASSER reassembly simulations) (Zhang & Skolnick 2004a,b). Finally, the top-ranked models for each subunit generated by I-TASSER were selected for subse-

quent docking simulations with the exception of U1-70K. For U1-70K, the first I-TASSER model exhibits a structure highly analogous to Properdin, one of the templates used in the modeling (PDB ID: **1W0R**), which adopts a conformation that is highly similar to the partial crystal structure that was reported during the preparation of this manuscript (Pommeranz krummel *et al.* 2009); however, the top model of U1-70K contains an RNA recognition motif in which the two β-strands (residues 147–150 and 172–175) are in the same plane as the alpha helix of the RRM, rather than being folded over and packed against the helix, as previous empirical data indicates (Howe *et al.* 1994). Careful examination of the model revealed that the RRM of U1-70K needed to be remodeled using the template from the second I-TASSER model, which more closely reflects the known folding properties of an RRM (Howe *et al.* 1994; Katsamba *et al.* 2001). As a result, the U1-70K tertiary structure used in assembling the U1 snRNP represents a combination of the U1-70K topology modeled using the template from the top model with the RRM modeled with the template from model two.

Modeling of structures with multiple protein chains is an arduous task using current computational resources and modeling techniques. As an approach to alleviate this problem, the subunits of the Sm heptamer were first modeled separately and then assembled in an hierarchical fashion. PDB ID: **1I8F** was selected as a template for the assembly of the Sm complex because it had the best sequence-profile match and structural similarity with the modeled subunits.

To combine the subunits into a heptameric Sm core, we first split the modeled subunits into structured core-region and unstructured tail regions. The core region of the modeled subunits is then superimposed on the template subunits utilizing the structure alignment program TM-align (Zhang & Skolnick 2005). Structurally aligned residue pairs in TM-align are considered as "conserved residues" in the Sm core. Since the initial complex structure assembled from the structure superposition had a number of steric clashes, we made a quick Metropolis Monte Carlo (MC) simulation to remove these steric clashes. The energy for the MC movement was defined as the Relative Mean Standard Deviation (RMSD) of subunit models to the template along the "conserved residue pairs" plus the reciprocal of the number of steric clashes between subunit cores. The MC movement includes random rotation and translation of each of the subunit models. The models with the best similarity to the template and free of clashes are selected. In the second step, we remodel the unstructured tails by I-TASSER simulations. The new I-TASSER models of the tails were then docked onto the Sm core complex using a Monte Carlo procedure similar to that described above, while the core structures were kept fixed. The complex model having maximum global similarity with the template of **1I8F** and free of clashes was then selected for subsequent docking onto the U1 snRNA.

The reason for docking the structures in a two-step manner is that the subunits have been modeled by I-TASSER independently and the orientation of the unstructured tails/loops (relative to each other) may be wrong, which render a compact rigid-body docking of subunit models impossible. The two-step docking procedure allowed us to remodel the tail orientation, while considering inter-subunit interactions, to allow us to pack the complex in a compact structure but with the maximum similarity with the template and the I-TASSER individual models.

Generation of U1 snRNA and pre-mRNA

A model representing the three dimensional conformation of the U1 snRNA from humans was obtained online (http://www-ibmc.u-strasbg.fr/upr9002/westhof/).

Assembly of the Sm core and creation of the U1 snRNP complex by step-wise docking

To create a model of the U1 snRNP, individual RNA and protein molecules were successively docked in an iterative fashion using the Patchdock server (http://bioinfo3d.cs.tau.ac.il/PatchDock/index.html). Docking constraints were added in each step by creating receptor and ligand binding site files using empirical strategies that provide intermolecular contacts in the U1 snRNP, including biochemical (Scherly *et al.* 1989, 1991; Heinrichs *et al.* 1990; Lutz-freyermuth *et al.* 1990; Jensen *et al.* 1991; Nelissen *et al.* 1991, 1994; Tang & Rosbash 1996; Raker *et al.* 1999; Katsamba *et al.* 2001), X-ray crystallographic and nuclear magnetic resonance (NMR) (Hall 1994; Howe *et al.* 1994; Oubridge *et al.* 1994; Kambach *et al.* 1999; Mura *et al.* 2001; Nagai *et al.* 2001; Toro *et al.* 2001; Muto *et al.* 2004; Pommeranz krummel *et al.* 2009) and electron microscopy techniques (Stark *et al.* 2001). First, the U1A protein was docked to the U1 snRNA. Next, the Sm binding site of U1 snRNA was manually threaded through the heptameric ring of the Sm core using the DeepView/SwissPDBViewer (version 3.7). U1-70K was then manually docked with DeepView/SwissPDBViewer (version 3.7) to stem-loop I using empirically determined contacts as a guide. U1C was docked last because it associates with the U1 snRNP by way of its interaction with the Sm core and U1-70K. Subunits of the U1 snRNP were docked in this order in an effort to minimize clashes between the Sm core and the other subunits. All models were visualized using PyMOL (version 0.99 beta14). A supplementary Protein Databank (.pdb) file (Supplementary File 1) has been included as part of this manuscript. This file can be opened and manipulated with any available .pdb viewer.

Results and discussion

U1-70K Interacts with U1 stem-loop I, SmB and U1C

U1-70K is made up of three functional regions, including the U1C binding region at the N-terminus, the central RNA recognition motif (RRM) and the C-terminal RS domain (indicated in yellow, pink and light blue, respectively, in Supplementary Fig. 1a). The RRM of U1-70K forms the classic β1-αA-β2-β3-αB-β4 configuration typical of RRMs (Katsamba *et al.* 2001), with the β-sheets packed flat against the backbone of stem-loop I of the U1 snRNA (Supplementary Fig. 1b). The first 97 amino acids in U1-70K (delineated in yellow in Supplementary Fig. 1c), to which U1C is known to bind (Nelissen *et al.* 1994), forms a long stretch of amino acids with no secondary structure, which passes across the Sm core and appears to interact directly with SmB (gray in Supplementary Fig. 1c) and U1C (red in Supplementary Fig. 1c). Similarly, the C-terminal RS domain of U1-70K (indicated in light blue in Supplementary Fig. 1a) maintains a flexible architecture, which facilitates binding to SR splicing factors, including, but not limited to SC35 and SF2/ASF (Wu & Maniatis 1993; Kohtz *et al.* 1994). This region of U1-70K is displaced from the U1 snRNP complex and contacts stem-loop III of U1 snRNA in the model. While it is difficult to accurately place the low complexity C-terminus in three dimensional space, the position of this region in the model simultaneously minimizes steric repulsion and may allow for a large degree of flexibility in the protein, perhaps as a mechanism to associate with one or more SR splicing factors.

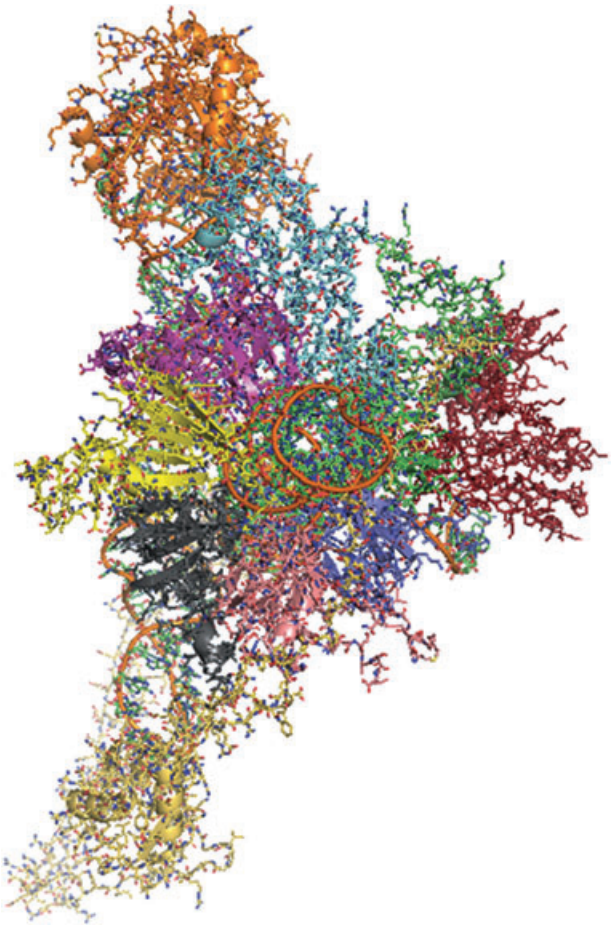

Figure 1 A three dimensional model of the U1 snRNP. Tertiary structural modeling coupled with a step-wise docking approach was employed to generate a simulated structure of the U1 snRNP. The positions of proteins within the model fit closely with the available biochemical and crystallographic data. The majority of protein-protein and protein-RNA contacts within the simulated U1 snRNP appear to be between just a few residues or nucleotides, with overall stabilization of the macromolecule resulting from many weak associations among members of the complex. U1 snRNA is depicted with an orange backbone and green nucleotides. U1A is delineated in orange, U1-70K in yellow, SmB in light blue, SmD3 in green, SmD1 in purple, SmD2 in yellow, SmE in pink, SmF in gray, SmG in lavender and U1C in red.

The RRM domains of U1A bind to U1 stem-loop II

U1A contains two RRMs from residues 29–84 and 151–193, with a 67 amino acid linker between them. The U1 snRNP model depicts the RRMs forming a clamp-like structure around stem-loop II of U1, which is extremely similar to the interaction between nucleolin and an RNA stem-loop (PDB ID: **1fje**) (reviewed by Maris *et al.* 2005). The N-terminal RRM (the left side of U1A in Supplementary Fig. 2) binds to the top of loop II, which contains free (non-base paired) nucleotides. The C-terminal RRM (to the right in Supplementary Fig. 2) is located more toward the stem, with seemingly no capacity to hydrogen bond to U1. This prediction fits well with empirical data, which suggests that the N-terminal RRM of U1A alone is sufficient to bind U1 snRNA (Scherly *et al.* 1989; Lutz-freyermuth *et al.* 1990; Katsamba *et al.* 2001). It is also remarkable that the region from amino acids 87–108 appears to contribute to stabilizing the U1 snRNA-U1A protein complex by extending over the top of stem-loop II and forming a pocket for the hairpin to bind (Supplementary Fig. 2). This interaction is supported by previous findings from several investigators, in which deletions or truncations within this region greatly reduce the binding capacity of U1A for U1 snRNA (Scherly *et al.* 1989, Boelens *et al.* 1991; Jensen *et al.* 1991; Hall 1994). Although the docking simulation predicts similar overall RNA-protein contacts when compared to the crystal structure (Oubridge *et al.* 1994), U1A in complex with U1 stem-loop II is rotated downward and differs by as much as 19.7 Å when superimposed to the empirically determined 3D con-

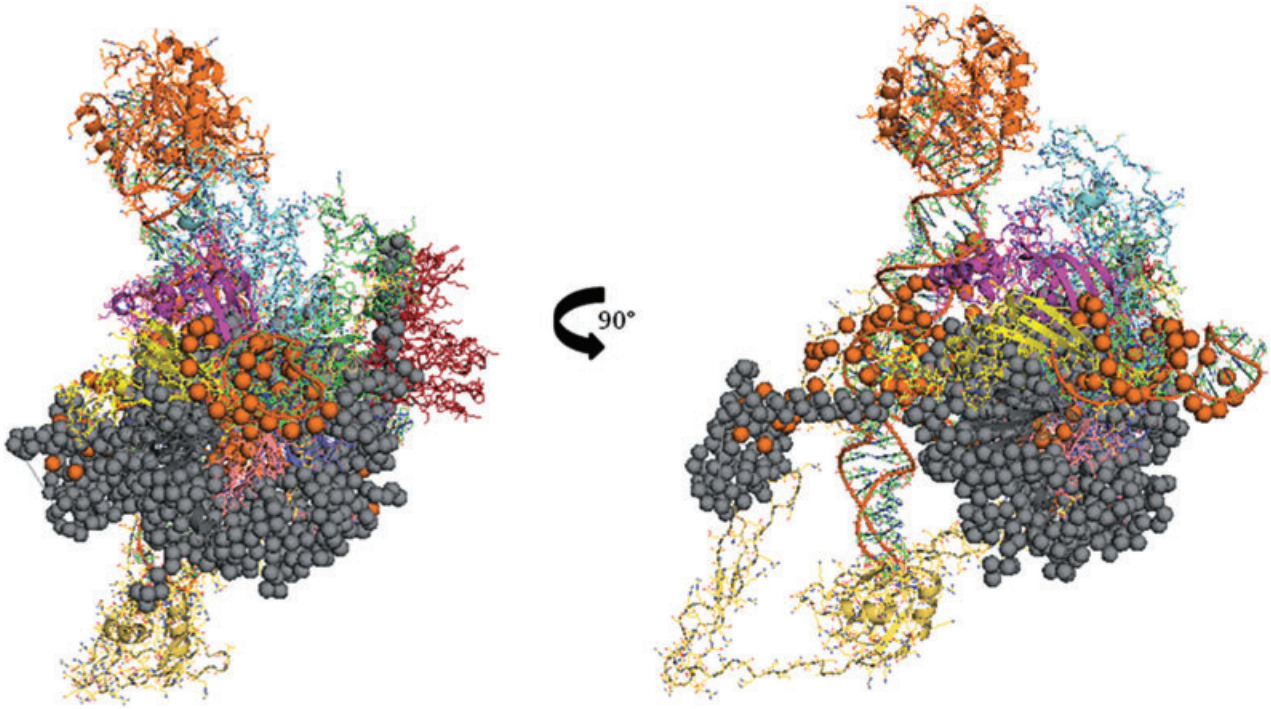

Figure 2 Superposition of the *in silico* U1 snRNP with the U1 snRNP crystal structure. The entire U1 snRNP model (multi-colored ribbon diagram) superimposes with the crystal structure of the U1 snRNP (gray and orange dots) with a resolution of 14.05 Å. The largest differences are in the placement of U1 snRNA stem-loop I, which is bent away from stem-loop III in the model, but 20 Å closer to hair-pin III in the crystal structure. Despite these disparities, the shape and relative placement of protein subunits within the model is supported by the X-ray crystallography data.

formation of the N-terminal RRM of U1A bound to a fragment of RNA (Oubridge *et al.* 1994) (PDB ID: **1URN**) (data not shown). This difference may be attributed to variables in the docking simulation and the X-ray crystallography study.

The Sm core

The heptameric Sm core forms a ring-like conformation through which the conserved Sm binding site ($RAU_{4\text{-}6}GR$) of U1 and other snRNAs pass. The Sm core model presented here was assembled manually following the previously determined biochemical interactions among monomers (Kambach *et al.* 1999; Raker *et al.* 1999; Nagai *et al.* 2001). Overlapping loop regions were adjusted to minimize steric hindrance and yet maintain an energetically favorable packing of the extended loops relative to the central Sm region of each monomer.

The smallest internal diameter of the opening in the core is 11.9 Å, which is 3.2 Å larger than a crystallized archaeal Sm-like ring structure (Mura *et al.* 2001). This difference is likely the result of the inclusion of the entire, full-length Sm proteins in the U1 snRNP model, as opposed to partial Sm-like proteins used in the crystal structure (Mura *et al.* 2001). The external diameter and width of the modeled core are also very close to the size of the Sm core described by Mura *et al.* (2001) (~65 Å and ~34 Å, respectively). The entire Archaeal Sm core (Mura *et al.* 2001) superimposes over the *Bombyx mori* Sm core model with an RMSD of 12.49 Å.

SmB, SmD1 and SmD3 all possess additional loops of 104, 55 and 86 amino acids, respectively (colored orange, yellow and red in Supplementary Fig. 3, respectively). The amino acids generating these loops were not included in any of the previously determined crystal structures of Sm proteins (Kambach *et al.* 1999; Mura *et al.* 2001; Toro *et al.* 2001). In addition, the α-carbon backbone within these regions in the simulations has no distinct tertiary conformation and, as a result, can fill any possible three dimensional space. Therefore, it is not known where these areas of the Sm core should be placed. Yet, despite these issues, the location of the loop regions of SmB within the model corresponds well with empirical data, which suggests that SmB contacts both the U1-70K and U1C proteins (Nelissen *et al.* 1994). Residues 1 through 66 of U1-70K are thought to interact with the Sm complex (Nelissen *et al.* 1994). In the model presented here, amino acids 15, 16, 30, 31 and 32 of U1-70K (U1-70K is rendered in yellow in Fig. 1) are in close prox-

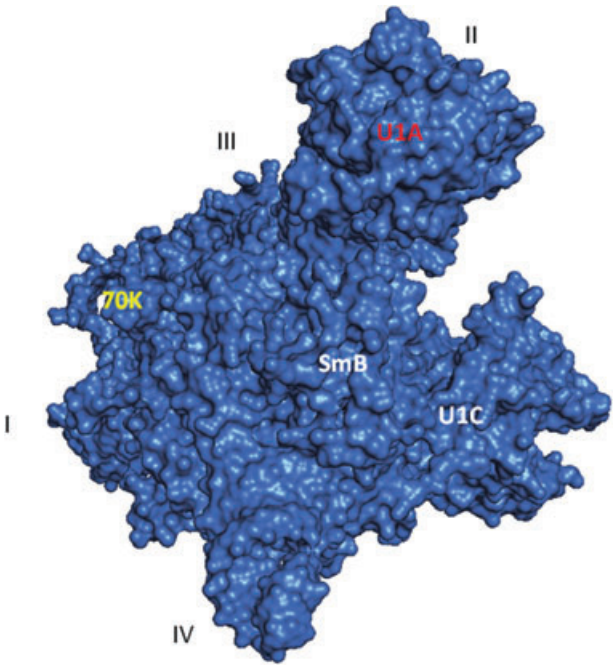

Figure 3 Comparison of the *in silico* generated model with an electron micrograph of the U1 snRNP. The overall shape of the simulated macromolecule forms a main body with two large projections and shares many similarities with an independently generated EM of the U1 snRNP (see: http://www.gpmolbio.uni-goettingen.de/faculty/f_luehrmann.html) (Stark *et al.* 2001). In addition, the placement of proteins within the molecule follows the positions estimated by Stark and coworkers (2001). The largest protrusions at the top left and top right correspond to U1-70K (in bold, yellow type) and U1A (in bold, red type), respectively. The two smaller regions at the bottom and bottom right are U1C and the C-terminal loop regions of SmB, respectively indicated in bold, white type).

imity to SmB (amino acids 37–76) (SmB is depicted in light blue at the top of the Sm core in Fig. 1). The loops comprised of residues 118–122 and 146–149 of U1C (delineated in red in Fig. 1) are also tightly associated with the β-4 and β-3 strands of SmD3 (in green at the top of the Sm core in Fig. 1) (Supplementary File 1). In addition, residues 49–64 of U1C interact with the SmD3 α-helix at amino acids 94–103.

U1C associates with the U1 snRNP via U1-70K and the Sm core

The U1C protein adopts two main folds (amino acids 1–58, which corresponds to a zinc finger motif) and a set of seven loops at the C-terminus that are stacked tightly at approximately 90° to the angle of the zinc finger. As indicated above, U1C interacts with U1-70K within its C-terminal loops and at its zinc finger motif, respectively; however, U1C (depicted in red in Fig. 1) does not seem to associate directly with U1 snRNA. This is in direct agreement with previous studies, which established that U1C is bound to the U1 snRNP only by way of its connections to U1-70K and the Sm core (Nelissen *et al.* 1991, 1994; Nagai *et al.* 2001).

Comparison of the U1 snRNP model with empirically derived structures

A detailed comparison of the *in silico* generated U1 snRNP with a recently reported crystal structure of much of the U1 snRNP (Pommeranz krummel *et al.* 2009) reveals a number of similarities. Superposition of the U1 snRNP model (depicted with ribbon diagrams in Fig. 2) with that of the crystal structure (rendered with dots in Fig. 2) indicates that the two complexes overlap with a RMSD of 14.05 Å (Fig. 2). The positioning of all protein components in the *in silico* determined snRNP mirrors that in the crystal structure; however, one of the most striking differences in the overall topology of the *in silico* U1 snRNP as compared to the crystal structure is the relative placement of stem-loop I in the U1 snRNP (Fig. 2). At their greatest distance, the hairpin I loop regions of the crystal and the model are 50.7 Å apart (Fig. 2). While the modeled U1 snRNP contains a conformation in which the four stem-loops are nearly perpendicular to one another, with a distance of 71.2 Å from stem-loop I to stem-loop III, the crystal structure reveals that the angle of stem-loop I relative to stem-loop III is approximately 75°, at a distance of just 51.2 Å. The difference between the modeled U1 snRNA and the crystal structure represents the largest disparity between our model and the crystal structure and highlights the continuing difficulties in modeling complex RNA conformations without prior experimental data (Shapiro *et al.* 2007; Chai 2008).

Another difficult step in generating the U1 snRNP is in the determination of the conformations of the low complexity N- and C-termini of U1-70K and their placement in the U1 snRNP. Although the position of the first 97 residues of U1-70K relative to the Sm core and U1C in the U1 snRNP model is strikingly similar to that observed in the crystal structure, amino acids 180–352 are not available in the crystal structure. Furthermore, residues 180–352 are made up of multiple arginine and serine repeats, which form a long chain of almost no secondary structure and can be placed in any three dimensional space. This makes the assignment of both the folding properties and the positioning of this portion of the model within the complex extremely difficult. Yet, despite these challenges, it is attractive to postulate that the proposed topology of the C-terminal RS motif of U1-70K may enable the protein to have the flexibility to interact with SR proteins that are bound to cis-acting elements within pre-mRNAs. The structural plasticity that may exist in U1-70K could contribute to the ability of U1-70K to recognize multiple SR factors in a variety of contexts on the

pre-mRNA, thereby contributing to recognition of the 5′ exon-intron junction in multiple possible states.

It is noteworthy that the U1 snRNP model also shares distinct similarities with an independently derived cryo-electron micrograph (EM) of the U1 snRNP (An image of the EM can be found at: http://www.gpmolbio.uni-goettingen.de/faculty/f_luehrmann.html) (Stark *et al.* 2001) (Fig. 3). The overall topology of the U1 snRNP derived from both techniques reveals a main circular body comprised of the Sm core and U1C along with two protruding segments at the upper right and upper left comprised of U1A and U1-70K, respectively (Fig. 3). The main body of the complex is approximately 75Å across, which is in close agreement with the EM, in which the main body is 70–80 Å in diameter (Stark *et al.* 2001). In addition, the intersection of the four U1 snRNA stem-loops in both the model and the EM connects immediately below U1A (Fig. 3) (Stark *et al.* 2001). SmD2 and SmB are also in close contact with U1-70K, as indicated in the EM (Stark *et al.* 2001).

Despite the parallelisms provided by the two lines of evidence, there are also differences between the model and the EM. As with the crystal structure, the most profound difference between the model and the EM generated by Stark *et al.* (2001) is in the positioning of U1 snRNA stem-loop I relative to the rest of the snRNP. In the model, stem-loop I is bent toward stem-loop IV, while in the EM and the crystal structure, stem-loop I is closer to stem-loop III. Some discrepancies in the U1 RNA are expected given the unique challenges of developing reliable RNA tertiary folds, especially considering the lack of solved RNA three dimensional conformations and the inherent computational difficulties with developing models of large and complex RNA molecules (Shapiro *et al.* 2007; Chai 2008). In addition, the relative positions of SmB and U1C are shifted downward and to the left in both the crystal structure and the simulated complex when compared to the EM (Stark *et al.* 2001).

Conclusion

The present work employs structural informatics to illuminate the complex RNA-protein and protein-protein interactions within the U1 snRNP. Utilizing the available biochemical and crystallographic data, a complete model of the *Bombyx mori* U1 snRNP was generated *in silico*. Quantitative superposition and comparison with empirically derived structures enabled us to evaluate the robustness of the modeling effort. While the size of the simulated complex, its overall topology and the relative positions of most of the proteins within the U1 snRNP superimpose closely to within 14 Å of the crystal structure (Pommeranz krummel *et al.* 2009), there are also disparities between the model and the experimentally-determined topologies. In particular, the positioning of U1 snRNA stem-loop I and the folding properties of the low complexity regions of U1-70K were challenging. Despite these differences, generation of the U1 snRNP by computer-based methods represents an unique application to the field of structural modeling. It is important to mention that the assembly of the U1 snRNP by *in silico* techniques was largely dependent upon close attention to prior biochemical characterization of interaction sites within the complex and, in some cases manual docking of protein subunits rather than *ab initio* reconstruction of the U1 snRNP. Nonetheless, the research described here represents one of the first attempts to develop a macromolecular assembly *in silico* and demonstrates that detailed analyses of available empirical data can aid in creation of macromolecular complexes that have biological relevance.

Acknowledgments

JAS acknowledges EPA STAR Fellowship number FP-91670801-3. AM acknowledges NIH/NIGMS R25 GM061347. RJH acknowledges NIH award 1SG1 GM083685-01. YZ acknowledges the support from the Alfred P. Sloan Foundation, NSF Career Award 0746198 and NIH grant GM-083107.

References

Battey JN, Kopp J, Bordoli L, Read RJ, Clarke ND, Schwede T (2007) Automated server predictions in CASP7. *Proteins* **69**: 68–82.

Boelens W, Scherly D, Jansen EJ, Kolen K, Mattaj IW, van Venrooij WJ (1991) Analysis of in vitro binding of U1-A protein mutants to U1 snRNA. *Nucleic Acids Research* **19**: 4611–4618.

Chai D (2008) RNA structure and modeling: progress and techniques. *Progress in Nucleic Acid Research and Molecular Biology* **82**: 72–93.

Eperon IC, Makarova OV, Mayeda A, Munroe SH, Cáceres JF, Hayward DG *et al.* (1989) Selection of alternative 5′ splice sites: role of U1 snRNP and models for the antagonistic effects of SF2/ASF and hnRNP A1. *Molecular and Cellular Biology* **20**: 8303–8318.

Fischer U, Sumpter V, Sekine M, Satoh T, Lührmann R (1993) Nucleo-cytoplasmic transport of U snRNPs: definition of a nuclear location signal in the Sm core domain that binds a transport receptor independently of the m3G cap. *The EMBO Journal* **12**: 573–583.

Hall KB (1994) Interaction of RNA hairpins with the human U1A N-terminal RNA binding domain. *Biochemistry* **33**: 10076–10088.

Hall TA (2004) BioEdit, a user-friendly biological sequence alignment editor and analysis program for Windows 95/98/NT. *Nucleic Acids Symposium Series* **41**: 95–98.

Hamm J, Darzynkiewicz E, Tahara SM, Mattaj IW (1990) The trimethylguanosine cap structure of U1 snRNA is a component of a bipartite nuclear targeting signal. *Cell* **62**: 569–577.

Heinrichs V, Bach M, Winkelmann G, Lührmann R (1990) U1-specific protein C needed for efficient complex formation of U1 snRNP with a 5′ splice site. *Science* **247**: 69–72.

Howe PW, Nagai K, Neuhaus D, Varani G (1994) NMR studies of U1 snRNA recognition by the N-terminal RNP domain of the human U1A protein. *The EMBO Journal* **13**: 3873–3881.

Jensen TH, Oubridge C, Teo CH, Pritchard C, Nagai K (1991) Identification of molecular contacts between the U1A small nuclear ribonucleoprotein and U1 RNA. *The EMBO Journal* **10**: 3447–3456.

Jurica MS, Licklider LJ, Gygi SR, Grigorieff N, Moore MJ (2002) Purification and characterization of native spliceosomes suitable for three-dimensional structural analysis. *RNA* **8**: 426–439.

Kambach C, Walke S, Young R, Avis JM, de la Fortelle E, Raker VA *et al.* (1999) Crystal structures of two Sm protein complexes and their implications for the assembly of the spliceosomal snRNPs. *Cell* **96**: 375–387.

Katsamba PS, Myszka DG, Laird-Offringa IA (2001) Two functionally distinct steps mediate high affinity binding of U1A protein to U1 hairpin II RNA. *Journal of Biological Chemistry* **276**: 21476–21481.

Kohtz JD, Jamison SF, Will CL, Zuo P, Lührmann R, Garcia-Blanco MA *et al.* (1994) Protein-protein interactions and 5′-splice-site recognition in mammalian mRNA precursors. *Nature* **368**: 119–124.

Lutz-Freyermuth C, Query CC, Keene JD (1990) Quantitative determination that one of two potential RNA-binding domains of the A protein component of the U1 small nuclear ribonucleoprotein complex binds with high affinity to stem-loop II of U1 RNA. *Proceedings of the National Academy of Sciences of the United States of America* **87**: 6393–6397.

Maris C, Dominguez C, Allain FH (2005) The RNA recognition motif, a plastic RNA-binding platform to regulate post-transcriptional gene expression. *The FEBS Journal* **272**: 2118–2131.

Mattaj IW (1986) Cap trimethylation of U snRNA is cytoplasmic and dependent on U snRNP protein binding. *Cell* **46**: 905–911.

Mesa A, Somarelli JA, Herrera RJ (2008) Spliceosomal immunophilins. *FEBS Letters* **582**: 2345–2351.

Moore MJ, Sharp PA (1993) Evidence for two active sites in the splicesome provided by stereochemistry of pre-mRNA. *Nature* **23**: 364–368.

Mura C, Cascio D, Sawaya MR, Eisenberg DS (2001) The crystal structure of a heptameric archaeal Sm protein, Implications for the eukaryotic snRNP core. *Proceedings of the National Academy of Sciences* **98**: 5532–5537.

Muto Y, Pomeranz Krummel D, Oubridge C, Hernandez H, Robinson CV, Neuhuas D *et al.* (2004) The structure and biochemical properties of the human spliceosomal protein U1C. *Journal of Molecular Biology* **341**: 185–198.

Nagai K, Muto Y, Pomeranz Krummel DA, Kambach C, Ignjatovic T, Walke S *et al.* (2001) Structure and assembly of the spliceosomal snRNPs. *Biochemical Society Transactions* **29**: 15–26.

Nelissen RL, Heinrichs V, Habets WJ, Simons F, Lührmann R, van Venrooij WJ (1991) Zinc finger-like structure in U1-specific protein C is essential for specific binding to U1 snRNP. *Nucleic Acids Research* **11**: 449–454.

Nelissen RLH, Will CL, van Venrooij WJ, Lührmann R (1994) The association of the U1-specific 70K and C proteins with U1 snRNPs is mediated in part by common U snRNP proteins. *The EMBO Journal* **13**: 4113–4125.

Oubridge C, Ito N, Evans PR, Teo CH, Nagai K (1994) Crystal structure at 192 A resolution of the RNA-binding domain of the U1A spliceosomal protein complexed with an RNA hairpin. *Nature* **372**: 432–438.

Plessel G, Fischer U, Lührmann R (1994) M3G cap hypermethylation of U1 small nuclear ribonucleoprotein snRNP in vitro, evidence that the U1 small nuclear RNA-guanosine-N2-methyl-transferase is a non-snRNP cytoplasmic protein that requires a binding site on the Sm core domain. *Molecular and Cellular Biology* **14**: 4160–4172.

Pommeranz Krummel DA, Oubridge C, Leung AK, Li J, Nagai K (2009) Crystal structure of human spliceosomal U1 snRNP at 55 Å resolution. *Nature* **458**: 475–480.

Raker VA, Hartmuth K, Kastner B, Lührmann R (1999) Spliceosomal U snRNP core assembly, Sm proteins assemble onto an Sm site RNA nonanucleotide in a specific and thermodynamically stable manner. *Molecular and Cell Biology* **19**: 6554–6565.

Salz HK, Mancebo RSY, Nagengast AA, Speck O, Psotka M, Mount SM (2004) The Drosophila U1-70K protein is required for viability but its arginine-rich domain is dispensable. *Genetics* **168**: 2059–2065.

Scherly D, Boelens W, van Venrooij WJ, Dathan NA, Hamm J, Mattaj W (1989) Identification of the RNA binding segment of human U1A protein and definition of its binding site on U1 snRNA. *The EMBO Journal* **8**: 4163–4170.

Scherly D, Kambach C, Boelens W, van Venrooij WJ, Mattaj IW (1991) Conserved amino acid residues within and outside of the N-terminal ribonucleoprotein involved in U1 RNA binding. *Journal of Molecular Biology* **219**: 577–584.

Shapiro BA, Yingling YG, Kasprzak W, Bindewald E (2007) Bridging the gap in RNA structure prediction. *Current Opinion in Structural Biology* **17**: 157–165.

Stark H, Dube P, Lührmann R, Kastner B (2001) Arrangement of RNA and proteins in the spliceosomal U1 small nuclear ribonucleoprotein particle. *Nature* **409**: 539–542.

Tang J, Rosbash M (1996) Characterization of yeast U1 snRNP A protein, identification of the N-terminal RNA binding domain RBD binding site and evidence that the C-terminal RBD functions in splicing. *RNA* **2**: 1058–1070.

Tatei K, Takemura K, Tanaka H, Masaki T, Ohshima Y (1987) Recognition of 5′ and 3′ splice site sequences in pre-mRNA studied with a filter binding technique. *Journal of Biological Chemistry* **262**: 11667–11674.

Toro I, Thore S, Mayer C, Basquin J, Seraphin B, Suck D (2001) RNA binding in an Sm core domain, X-ray structure and functional analysis of an archaeal Sm protein complex. *The EMBO Journal* **20**: 2293–2303.

Will CL, Rümpler S, Gunnewiek JK, van Venrooij WJ, Lührmann R (1996) In vitro reconstitution of mammalian U1 snRNPs active in splicing, the U1-C protein enhances the formation of early E spliceosomal complexes. *Nucleic Acids Research* **24**: 4614–4623.

Wu JY, Maniatis T (1993) Specific interactions between proteins implicated in splice site selection and regulated alternative splicing. *Cell* **75**: 1061–1070.

Wu S, Skolnick J, Zhang Y (2007) Ab initio modeling of small proteins by iterative TASSER simulations. *BMC Biology* **5**: 17.

Yuo Y, Weiner AM (1989) A U1 small nuclear ribonucleoprotein particle with altered specificity induces alternative splicing of an adenovirus E1A mRNA precursor. *Molecular and Cellular Biology* **9**: 3429–3437.

Zhang Y (2006) *Protein structure prediction by I-TASSER at CASP7*. Invited talk given at CASP7 conference, November 26–30, 2006, Asilomar Conference Center, Pacific Grove, CA.

Zhang Y (2008) I-TASSER server for protein 3D structure prediction. *BMC Bioinformatics* **9**: 40.

Zhang Y, Skolnick J (2004a) Automated structure prediction of weakly homologous proteins on a genomic scale. *Proceedings of the National Academy of Sciences of the United States of America* **101**: 7594–7599.

Zhang Y, Skolnick J (2004b) SPICKER, A clustering approach to identify near-native protein folds. *Journal of Computational Chemistry* **25**: 865–871.

Zhang Y, Skolnick J (2005) TM-align, a protein structure alignment algorithm based on the TM-score. *Nucleic Acids Research* **33**: 2302–2309.

Supporting information

Additional Supporting Information may be found in the online version of this article:

Figure S1 (a) U1-70K makes contacts with multiple U1 snRNP subunits. U1-70K associates with U1C within its first N-terminal 97 amino acids colored yellow). The central RNA recognition motif RRM) mediates binding of U1-70K to U1 snRNA stem-loop I delineated in pink). The C-terminal Arg/Ser rich RS) domain light blue) is responsible for interacting with SR proteins, thereby enabling recognition of the 5′ splice site by the U1 snRNP. (b) U1-70K binds stem-loop I of U1 snRNA. The central RRM of U1-70K associates with U1 snRNA stem-loop I, with the β-sheets packed flat against the loop and α-helices on the outside. (c) U1-70K interacts with U1C red). The N-terminal region of U1-70K indicated in yellow) associates with the zinc finger domain of U1C depicted in red). The SmB protein is delineated in gray with a space filling model.

Figure S2 (U1A yellow) associates with U1 snRNA stem-loop II via its N-terminal RRM. U1A contains two RRMs located at the termini of the protein. The N-terminal RRM to the left of stem-loop II) interacts with U1 snRNA hairpin II, while the C-terminal RRM to the right of stem-loop II) is twisted away from the RNA. The C-terminal RRM is necessary for splicing and may bind to other RNAs or proteins in the spliceosome.

Figure S3 The Sm core forms a heptameric ring. The Sm core, comprised of SmB or SmB′) orange), SmD1 yellow), SmD2 green), SmE blue), SmF light blue) and SmG purple) and SmD3 red) assembles into a ring-like structure, through which the Sm binding site of U1 and other snRNAs pass.

Please note: Wiley-Blackwell are not responsible for the content or functionality of any supporting materials supplied by the authors. Any queries (other than missing material) should be directed to the corresponding author for the article.