Acta Crystallographica Section D
Biological Crystallography
ISSN 0907-4449

Inclusion of weak high-resolution X-ray data for improvement of a group II intron structure

Jimin Wang

Department of Molecular Biophysics and Biochemistry, Yale University, New Haven, CT 06520, USA

Correspondence e-mail: jimin.wang@yale.edu

Received 18 February 2010
Accepted 27 July 2010

PDB Reference: group II intron, 3g78.

It is common to report the resolution of a macromolecular structure with the highest resolution shell having an averaged $I/\sigma(I) \geq 2$. Data beyond the resolution thus defined are weak and often poorly measured. The exclusion of these weak data may improve the apparent statistics and also leads to claims of lower resolutions that give some leniency in the acceptable quality of refined models. However, the inclusion of these data can provide additional strong constraints on atomic models during structure refinement and thus help to correct errors in the original models, as has recently been demonstrated for a protein structure. Here, an improved group II intron structure is reported arising from the inclusion of these data, which helped to define more accurate solvent models for density modification during experimental phasing steps. With the improved resolution and accuracy of the experimental phases, extensive revisions were made to the original models such that the correct tertiary interactions of the group II intron that are essential for understanding the chemistry of this ribozyme could be described.

1. Introduction

Both the quality of X-ray sources and measurement technologies have been greatly improved in the past few decades so that we can routinely and accurately measure very weak high-resolution diffraction data. In contrast to X-ray analysis of small molecules, where there is little or no radiation damage, macromolecular crystals often suffer from severe radiation damage even at liquid-nitrogen temperatures. The highest resolution data often disappear first as a consequence of global radiation damage or change as a consequence of local radiation damage, which, for example, can reduce the binding of heavy atoms on extended irradiation at their peak-absorption wavelengths. It is challenging to collect a complete data set at the highest possible diffraction limit from radiation-sensitive macromolecular crystals. For example, long exposures reduce statistical counting errors and improve the quality of weak high-resolution data, but can lead to an incomplete data set. The highest possible resolution must be traded off against the completeness of the data. How to define the actual resolution of macromolecular structures thus becomes an issue and how to process data to the highest possible resolution become a challenge.

It is common to report the resolution of a macromolecular structure with an averaged $I/\sigma(I)$ of 2 for the highest resolution shell so that the resolution of one structure may be compared with the next. Merging statistics for this resolution shell are typically around 50% or slightly higher, provided that there are no major problems in the crystals. Some crystallographers prefer to use an averaged $F/\sigma(F)$ of 2 as a resolution-cutoff criterion. This is the default value for reflection selection in refinement programs such as *X-PLOR* (Brünger, 1993). This criterion corresponds to an averaged $I/\sigma(I)$ of 1. Using this criterion, the highest resolution shell may have merging statistics greater than 100%, *i.e.* the measurement errors for most reflections could actually exceed their measured mean intensities, making their absolute values unreliable. However, the fact that they are very weak and near zero (but not uniformly zero) is very reliable.

The question addressed in this study is whether there is any advantage to including weak high-resolution data during experimental density modification as well as during structure refinement. We previously demonstrated for a protein structure that the benefits that result from the inclusion of these weak data in structure refinement far exceed those resulting from their exclusion (Wang & Boisvert, 2003). Their inclusion increased the number of X-ray observations by 40% when the resolution was extended from 2.2 Å using an $F/\sigma(F)$ = 2.0 cutoff criterion to 2.0 Å using $F/\sigma(F)$ = 1.16. More importantly, it improved the quality of the refined model based on refinement statistics such as the free *R* factor, which fell by 8 percentage units from that at the original resolution of 2.2 Å.

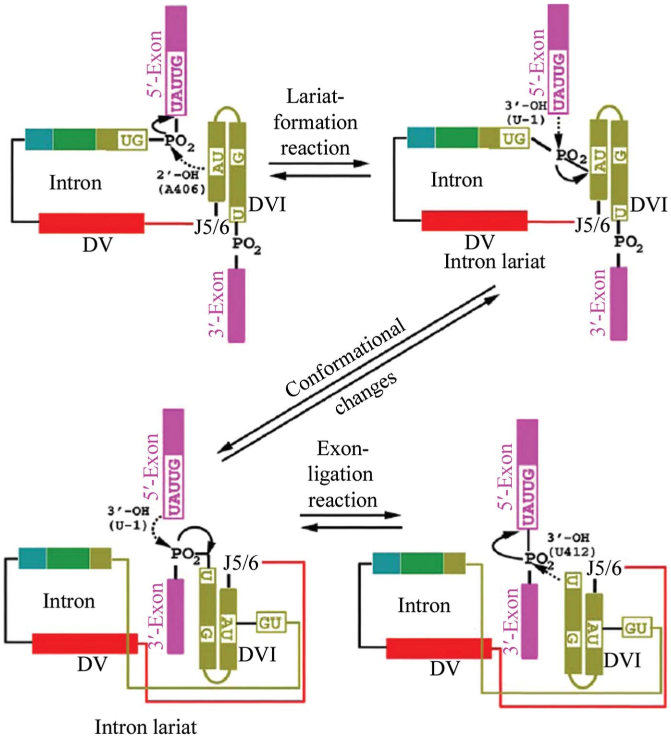


Figure 1
Two self-splicing reactions for the *O. iheyensis* group II intron. Schematic drawing of two sequential transesterification reactions, the first splicing reaction or lariat-formation reaction (first row) and the second splicing reaction or exon-ligation reaction (second row), coupled with proposed large conformational changes (diagonal) hinged at domain junction J5/6 between DV and DVI. The 5′-exon is numbered using minus signs and the 3′-exon is numbered using plus signs relative to the respective scissile phosphates. The intron is numbered without signs. Selection of the branch site A406 is defined by a preceding U·G wobble base pair. Dotted arrows indicate the attacking nucleophiles A406, U412 and U−1; solid arrows indicate bond-breaking processes.

The physical basis for these benefits is that the inclusion of these weak reflections in refinement ensures that the refined models also have weak calculated structure factors for these reflections; otherwise, incorrect models may result, particularly with incorrect atomic *B*-factor distributions, when the calculated structure factors for these reflections are unconstrained. In this study, we report that the inclusion of weak high-resolution data during experimental density modification improved the experimental electron-density maps in a recently reported RNA structure, the self-splicing group II intron from *Oceanobacillus iheyensis*, with original PDB code 3bwp (Toor, Keating *et al.*, 2008). This structure has previously been revised once, without the inclusion of additional high-resolution data (PDB codes 3eoh and 3eog; Toor, Rajashankar *et al.*, 2008). By including weak high-resolution data, we obtained an improved atomic model with new PDB code 3g78 and provided an accurate description of its tertiary interactions (PDB entry 3igi; Toor *et al.*, 2010). The coordinates 3igi were derived from the current coordinates 3g78 after the data used for refinement were truncated back to 3.1 Å in line with the commonly acceptable definition of resolution within the macromolecular crystallographic community. These PDB codes are used throughout the text to refer to the corresponding structures.

Self-splicing group II introns have six structurally conserved domains designated DI–DVI encoding the entire catalytic apparatus for two distinct phosphodiester-transfer or self-splicing reactions (Lambowitz & Zimmerly, 2004; Pyle & Lambowitz, 2006; Toor, Keating *et al.*, 2008): lariat formation (the exchange of a 2′–5′ phosphodiester for a 3′–5′ phosphodiester) and exon ligation (the exchange of one 3′–5′ phosphodiester for another) (Fig. 1). They are ideal model systems for exploring the more complex mechanisms of RNA splicing mediated by spliceosomes, which have the same architecture as the group II introns (Ceck, 1986; Sharp, 1994; Lambowitz & Zimmerly, 2004; Pyle & Lambowitz, 2006). Within the six domains of group II introns many long-range tertiary interactions have been established, for example in the γ–γ' or κ–κ' pairs (Jacquier & Michel, 1990; Boudvillain & Pyle, 1998). Biochemical studies have shown that in the catalytic centers of group II introns and the spliceosome as well as group I introns there are two metal ions that are critical for catalyzing phosphodiester-transfer reactions (Steitz & Steitz, 1993; Moore & Sharp, 1993; Gordon & Piccirilli, 2001; Gordon *et al.*, 2007).

2. Reprocessing of X-ray diffraction data for native and derivative crystals

In order to determine whether we could extend the resolution of the native group II intron structure and how far we could extend it, we systematically reprocessed the X-ray diffraction data from 3.2 to 3.1, 3.0, 2.9 and 2.8 Å resolution with a 0.1 Å resolution increment using the program *HKL*-2000 (Otwinowski & Minor, 1997). We increased a stringent criterion for rejecting radiation-damaged images according to their relative scaling *B* factor with each resolution increment. For the highest targeted resolution of 2.8 Å, any images with an

Table 1
Overall statistics of data reprocessing.

Values in parentheses are for the highest resolution shell.

		Yb		Ir		
Data set	Native (Mg/K)	Peak	Inflection	Peak	Inflection	Remote
Unit-cell parameters (Å)						
a	89.11	88.71	88.73	88.44	88.40	88.50
b	94.97	95.16	95.53	94.99	95.10	95.06
c	226.0	225.0	224.2	225.1	224.8	225.2
Wavelength (Å)	0.9795	1.3861	1.3865	1.1055	1.1058	1.0957
Resolution (Å)	40–2.80 (2.90–2.80)	40–3.30 (3.42–3.30)	40–3.30 (3.42–3.30)	40–3.20 (3.31–3.20)	40–3.40 (3.52–3.40)	40–3.20 (3.31–3.20)
No. of reflections†	47323	49762/2	45876/2	57567/2	42471/2	55967/2
Redundancy	6.3	6.6	6.5	6.6	6.9	6.6
$I/\sigma(I)$	20.7 (0.38)	26.8 (1.67)	27.8 (0.50)	25.6 (2.04)	25.0 (2.72)	25.5 (2.26)
Completeness (%)	98.9 (92.1)	89.8 (22.4)	79.4 (9.3)	94.8 (81.6)	86.5 (84.4)	95.3 (87.2)
$R_{\rm merge}$ (%)	7.2 (>100)	7.9 (62.8)	9.3 (>100)	9.4 (85.2)	10.1 (83.8)	8.9 (85.8)
No. of metal-ion sites	[76]	5		17		

† For the heavy-atom derivative data sets the option 'scale anomalous' was used during scaling, which yielded a number of reflections that was approximately twice that of unique reflections with all Friedel mates. The space group was $P2_12_12_1$ for all data sets.

overall B factor greater than 7 Å^2 or a χ^2 greater than 2 were rejected. Data from individual native crystals were first processed to establish a baseline for acceptable statistics before attempting to merge their images together to give an increased redundancy. In general, highly redundant averaging of data through merging data from multiple crystals can substantially reduce counting statistical errors and thus improve the quality of weak data only if all data represent exactly the same structure, *i.e.* strictly isomorphous crystals with no radiation damage. However, in the presence of the non-isomorphism problems among group II native crystals and severe radiation damage in individual crystals such merging could lead to an averaged structure with relatively poor quality. We monitored the improvement of density-modified experimental maps as well as structure refinement to justify any resolution extension, during which we had to make a trade-off between the highest possible resolution and the highest possible redundancy. We also sharpened the native amplitudes systematically by $B = -20, -40, -60$ or -80 Å^2.

In order to properly assess the effectiveness of resolution extension and amplitude sharpening, we recalculated anomalous difference Fourier maps using the peak-wavelength data from both Ir and Yb derivatives with new experimental phases from density modification or new calculated phases from

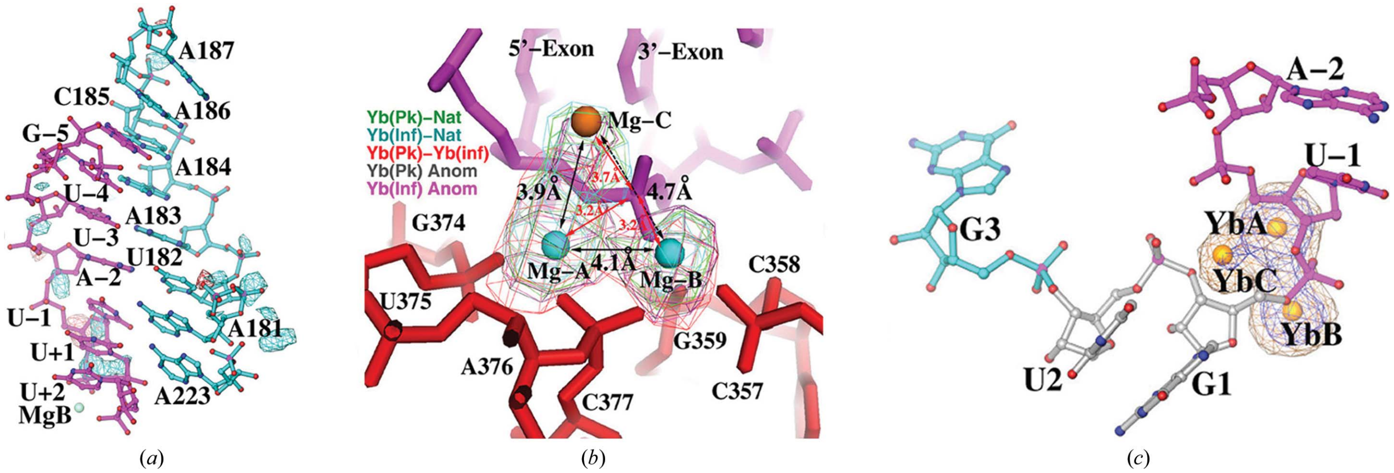


Figure 2
Major revisions of cocrystallized intron structures and heavy-atom structures. (*a*) An isomorphous difference Fourier map between the second and first group II intron structures (Toor, Rajashankar *et al.*, 2008; Toor, Keating *et al.*, 2008), contoured at +3.5σ (cyan) and −3.5σ (red), superimposed onto the catalytic site (cyan) with the bound RNA product (magenta, see below) shows that there is no negative difference at the G−5 position to support the displacement of the originally bound RNA product containing G−5 by a smaller oligonucleotide lacking G−5. (*b*) The revised heavy-atom structures show that there are three, not two, Yb-binding sites near the catalytic site in the Yb derivative. A third site (MgC) is about 3.7 Å from the P atom of the scissile phosphate of the RNA product bound in the catalytic site, while the two original sites (MgA and MgB) are 3.2 Å away from it. These sites have been verified using all possible combinations of experimental difference Fourier maps: (i) isomorphous, (ii) dispersive and (iii) anomalous difference Fourier maps as indicated. Maps are contoured at 25σ. (*c*) Relationship between the three metal ions and an unspliced substrate. These metal ions are shown in completely unbiased anomalous difference Fourier maps using phases from the partially refined intron model contoured at 25σ and 30σ. The unspliced substrate joins its 5′-exon (magenta) to the 3′-end of the intron (cyan), where G1 and U2 (silver) have been hypothetically modeled in the native intron structure.

partially refined atomic models to monitor peak heights for known heavy-atom sites. With improved phases, the peak heights should increase. Indeed, we observed a noticeable increase when the resolution of the native data was extended from 3.2 to 3.0 Å followed by a complete iteration of the experimental phasing procedures. The peak heights continued to increase with calculated phases from partially refined models when the resolution was extended from 3.0 to 2.8 Å for refinement, justifying the extension of the resolution to 2.8 Å. Peak heights also increased when the native amplitudes were sharpened from 0 to −40 Å^2, but rapidly decreased when the sharpening went beyond −40 Å^2. A disadvantage of this criterion was that there were no anomalous signals beyond 3.2 Å in these derivative data and that any direct improvement in the highest resolution shell was not detectable. An advantage of this criterion was that we could objectively monitor the improvement of phases in medium- and low-resolution shells when high-resolution data were added. Using this criterion, we obtained the best-quality experimental maps as well as the best model in structure refinement when we included all data to 2.8 Å resolution with a sharpening of $B = -40$ Å^2. Unsuccessful attempts were also made to extend the data beyond 2.8 Å.

The best native data set was processed mainly from a single native crystal with the highest possible 2.8 Å resolution and included only about a third of the usable diffraction images from the crystal. All other images with noticeable radiation damage were discarded. Images from other native crystals were also discarded for this reprocessing. Here, we chose the highest possible resolution for the native data over the highest possible redundancy by not averaging data that had suffered from substantial radiation damage. At 2.8 Å resolution, the reprocessed native data (3g78) had a merging *R* factor of 7.2% with an overall redundancy of about 6, compared with the 14.9% at 3.1 Å resolution with a redundancy of 15 reported for the original structure 3bwp (Toor, Keating *et al.*, 2008). The overall statistics of data reprocessing are summarized in Table 1 and details per resolution shell for the native data set are given in Table 2. This native data set had 95% completeness and was used for heavy-atom structure determination, heavy-atom parameter refinement and experimental phase calculation. For density-modification procedures, we used the complete data set, which included the remaining 5% of the data at medium and low resolution from other native crystals using a filling-in procedure according to the smallest possible non-isomorphous amplitude differences. There were large non-isomorphous differences between different native crystals, which were as high as 13.6% at 3.1 Å resolution. These native crystals did not contribute any useful information at a resolution beyond 3.1 Å.

With the reprocessed native data, we observed an amplitude difference of 11.6% between the first structure (3bwp, 3eoh and then 3g78) and the second structure (3eog), which was reported to contain a cocrystallized oligonucleotide that had displaced the originally copurified bound RNA product (Toor, Rajashankar *et al.*, 2008). This amplitude difference is actually smaller than the non-isomorphous differences of 13.6% observed among the native data used in the original structure determination (Toor, Keating *et al.*, 2008). With the improved experimental phases, we can unambiguously determine any differences between the first (3bwp, 3eoh and reprocessed 3g78) and second (3eog) structures using observed difference Fourier maps. Particularly, we would expect to see strong negative peaks if the smaller cocrystallized oligonucleotide were to replace the originally bound larger RNA product, for example in the G−5 position, as this G nucleotide found in the RNA product was not present in the oligonucleotide. In contrast, the observed difference Fourier maps showed that the first and second structures were identical to one another and that the oligonucleotide used in cocrystallization failed to displace the copurified RNA product bound in the catalytic site (Fig. 2*a*).

Table 2
Detailed statistics of native data reprocessing.

Resolution shell (Å)	$I/\sigma(I)$	$R_{\rm merge}$ (%)	Completeness (%)
40.00–6.03	135.2	3.6	97.0
6.03–4.79	28.7	5.3	99.9
4.79–4.18	23.3	7.1	100.0
4.18–3.80	20.3	9.8	100.0
3.80–3.53	14.8	14.4	100.0
3.53–3.32	9.87	21.9	100.0
3.32–3.15	6.36	32.6	100.0
3.15–3.02	3.29	61.2	100.0
3.02–2.90	0.82	>100	99.8
2.90–2.80	0.38	>100	92.1
40.00–2.80	20.71	7.2	98.9

We also systematically reprocessed all derivative data with merging *R* factors in the range 7.9–10.1% (Table 1). These merging *R* factors are much smaller than those reported originally, which were as high as 17.9% (Toor, Rajashankar *et al.*, 2008). Again, we chose the highest possible quality and resolution for derivatives in data reprocessing over the highest possible redundancy. As judged by these merging statistics, the quality of all reprocessed data has been greatly improved. Consistent with the improved merging statistics, the peak heights for known heavy-atom sites greatly increased in anomalous difference Fourier maps calculated using density-modified experimental phases when radiation-damaged images were discarded during data processing. With the reprocessed derivative data, the heavy-atom signal-to-noise ratios for experimental phasing were increased.

3. Redetermination of heavy-atom structures for experimental phasing

Heavy-atom substructures were iteratively redetermined using a combination of anomalous, dispersive and isomorphous difference Fourier methods with reprocessed X-ray diffraction data and experimental phases from previous runs of density-modification procedures as described elsewhere (Cura *et al.*, 1992; Rould *et al.*, 1992). The heavy-atom structures were further verified by reasonable physical interactions with the RNA structure. The parameters of the heavy-atom structures were re-refined using the program *MLPHARE*

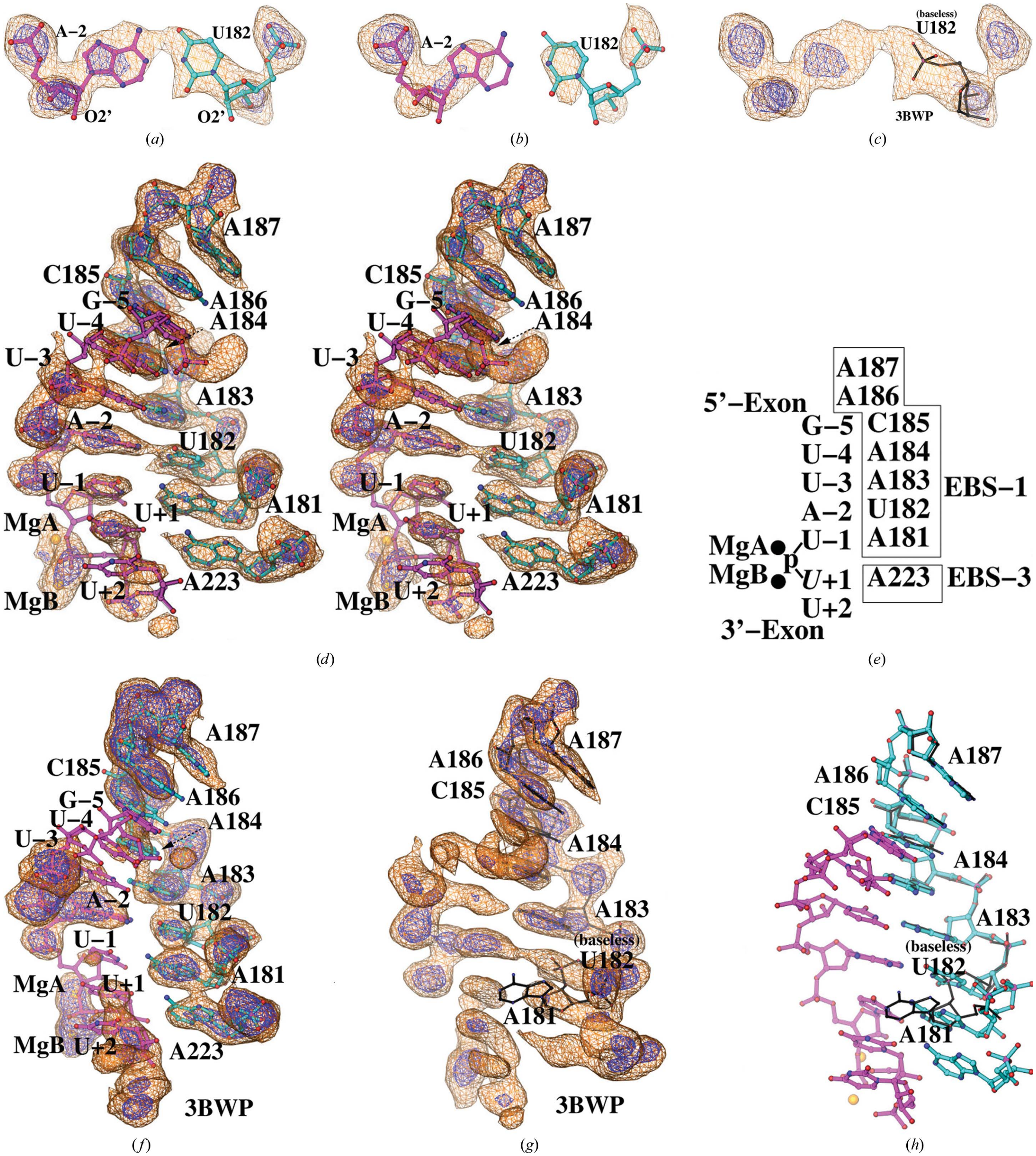
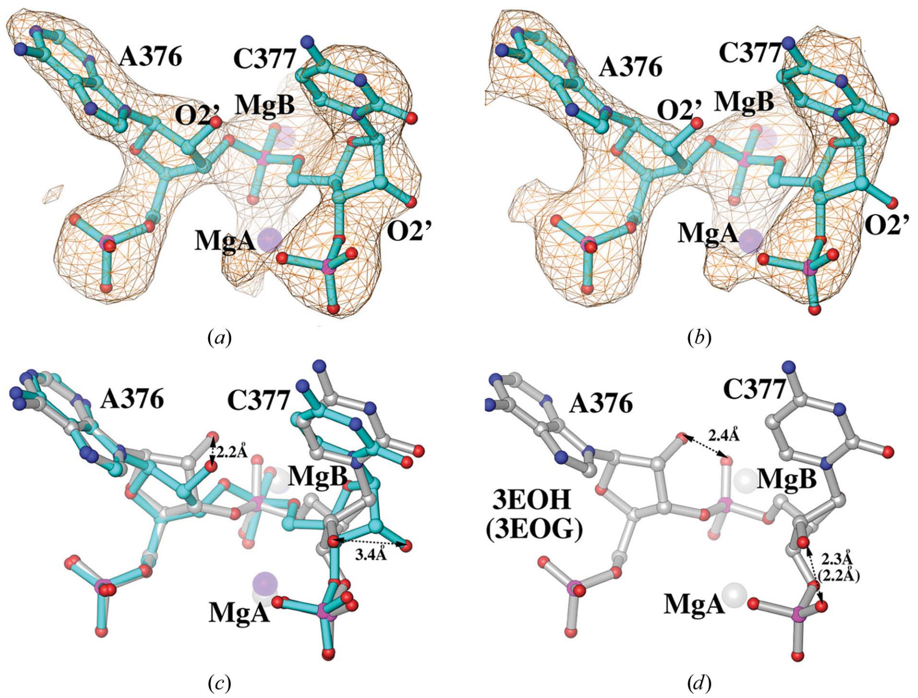


Figure 3
The binding of an RNA product in the catalytic site. (*a*) Our new experimental map superimposed onto our new model for the A−2·U182 Watson–Crick base pair (3g78). (*b*) The original map 3bwp superimposed onto our model 3g78. (*c*) Our experimental map 3g78 superimposed onto the original model 3bwp, which included several 'baseless' nucleotide residues that were built without bases, such as U182. (*d*) Stereodiagram of our experimental map superimposed onto our model for the RNA product and product-binding sites (3g78). (*e*) Schematic drawing of the product–EBS-1 (boxed) and product–EBS-3 (boxed) interactions. Non-Watson–Crick interaction pairs are indicated in italics for U+1. Metal ions A and B are also shown in relationship to the scissile phosphate that connects U−1 of the 5′-exon to U+1 of the 3′-exon. (*f*) The original experimental map 3bwp superimposed onto our model 3g78. (*g*) Our experimental map 3g78 superimposed onto the original model 3bwp. (*h*) Comparison of our model 3g78 with the original model 3bwp. Experimental maps were contoured at 1σ (golden) and 3σ (blue) in (*a*), (*b*), (*c*), (*d*) and (*g*) and 0.5σ (golden) and 1.5σ (blue) in (*f*). The RNA product is shown in magenta, EBS-1/EBS-3 in cyan and the original model in gray.



(*a*)

(*b*)

(*c*)

(*d*)

Figure 4
Catalytic bulge A376/C377. (*a*) Our new experimental map superimposed onto our model (3g78) with visible bumps for O2′. (*b*) The original experimental map 3bwp superimposed onto our model (3g78). (*c*) Superposition of our model (cyan, 3g78) and an early revised model (gray, 3eoh) with large differences for O2′ indicated. (*d*) The original models (3eoh and 3eog) had poor stereochemistry with very short unfavorable van der Waals contacts, as indicated. All maps were contoured at 1σ.

under an option for using external phases from previous phasing iterations (Otwinowski, 1991). The resultant experimental phases were further improved using density-modification procedures as implemented in *CNS* (Brünger *et al.*, 1998).

Improvement of the reprocessed data and experimental phases led to new heavy-atom structures. For example, the two previously reported Yb^{3+} sites (A and B; Toor, Keating *et al.*, 2008) have now been split into three well defined sites (A, B and C), all of which were within 5 Å of each other near the catalytic site of the intron (Fig. 2*b*). Without this splitting, the location of the original heavy-atom site (A) was distorted, being the weighted average of two split sites (A and C). The poor quality of the originally processed data (as apparent from the merging statistics in Toor, Keating *et al.*, 2008) could have made it difficult to resolve these two split sites. In this case, the resulting structure of the catalytic site was highly distorted, which may in part be responsible for its uninterpretability.

The presence of a third Yb^{3+} site (C) in the derivative structure is consistent with a newly observed RNA product bound in the ordered catalytic site in the intron, because this Yb^{3+} site can make reasonable interactions with the product (Fig. 2*b*). In fact, this site would interact even better with an unspliced substrate (*i.e.* the 5′-exon remains connected to the 3′-end of the intron) according to computer modeling (Fig. 2*c*). It is possible that the binding of Yb^{3+} at this site in the derivative structure has shifted the equilibrium to a state containing the unspliced substrate by directly stabilizing it. Indeed, we found that there were more connecting densities between the scissile phosphate and the first visible nucleotide G3 of the intron (through the modeled G1 and U2) in the Yb^{3+}-derivative structure when the amplitudes of this derivative data set were used for density modification. This observation further supports the shifted-equilibrium hypothesis. When different native data sets were used for density modification in parallel, the resulting experimental maps showed that there were more variations near the catalytic site than elsewhere. This observation suggests that these native structures may be in different equilibrium states along the reaction pathway. Improper merging of the data from such native crystals could lead to improper averaging of these states.

After systematically extending the resolution of the native and derivative data, we have improved the experimental maps to 2.8 Å from the original 3.1 Å resolution (Toor, Keating *et al.*, 2008). Our new experimental electron-density maps have unambiguously revealed a well resolved RNA product bound in the catalytic site of the intron (Fig. 3), the 2′-OH positions of many nucleotide residues in the active site (Fig. 4) and elsewhere (Fig. S1[1]). From these new maps, we observed proper configurations of bases in Hoogsteen and triplet base pairs (Fig. S1), a previously unnoticed single-nucleotide bulge (Fig. S2) and a new complexity of the κ–κ' tertiary interactions (Figs. 5 and S3). This revision provides new insights into the catalytic mechanism of the self-splicing reaction (see below) and gives an accurate description of the tertiary interactions that are essential for understanding the folding of this ribozyme (Toor *et al.*, 2010).

4. Interpretations of new experimental maps

Our new experimental maps have revealed an A-like duplex at the catalytic site formed between exon-binding site 1 (EBS-1) and its reciprocal intron-binding site (IBS) (Fig. 3). Electron densities for the bound RNA product persist at equally high contour levels as for the intron itself (Figs. 3*a* and 3*d*), suggesting that EBS-1 is fully occupied. This is consistent with the fact that EBS-1 binds this site during both steps of the self-splicing reaction (Lambowitz & Zimmerly, 2004). Electron densities for EBS-3 are relatively weak, which is consistent with the fact that the second site binds different RNAs in the

[1] Supplementary material has been deposited in the IUCr electronic archive (Reference: DZ5201). Services for accessing this material are described at the back of the journal. Throughout this article the supplementary figures are numbered using the prefix S.

two steps of the reaction. The occupancy of different RNAs at the second site is likely to be determined by the equilibrium between all possible substrates and products of the self-splicing reactions. In retrospect, the newly built RNA product also fits well into the original experimental maps (Figs. 3*b* and 3*f*), in which this unrecognized RNA was incorrectly interpreted as part of the intron structure itself (Toor, Keating *et al.*, 2008; Figs. 3*c* and 3*g*). Besides the currently modeled product, the exon-binding site may also be partially occupied by some other autodegradation products of DVI with relatively low occupancies, as observed in other native crystals (Toor *et al.*, 2010).

With the improved quality of the experimental maps after resolution extension, we began to see small bumps corresponding to the location of 2′-OH for many nucleotides, including those in the catalytic metal-ion binding site (Fig. 4*a*). These features helped in the building of nucleotides with proper sugar puckers and configurations of bases. However, these features were largely absent from the original experimental maps (3bwp; Fig. 4*b*). With accurately positioned metal ions and revised sugar puckers in the catalytic site, which differ substantially from the original model (Figs. 4*c* and 4*d*), we have established the stereochemistry of the scissile phosphates of the bound RNA product as well as that of other surrounding phosphates in the binding pocket (see below).

Based on the positioning of O2′ and densities for bases from improved experimental maps, we can now recognize a number of new Hoogsteen A·U base pairs, including A110·U259, A50·U198, A268·U285 and A138·U150 (Figs. S1*a*–S1*d*). These maps also show a base triplet involving the Hoogsteen base G in G179·A220·U157 and wobble geometry for all G·U base pairs (Fig. S1*e* and S1*f*). In addition, the new maps helped to identify an extrahelical single-nucleotide bulge at G300 (Fig. S2) and to correct an out-of-register error in the original sequence assignment in this region (Toor, Keating *et al.*, 2008; Toor, Rajashankar *et al.*, 2008).

Perhaps the most perplexing issue in previous interpretations of intron structures (Toor, Keating *et al.*, 2008; Toor, Rajashankar *et al.*, 2008) was that the structures could not explain the biochemical observations as to how the κ and κ′ tertiary interactions are responsible for folding the catalytic bulge (Boudvillain & Pyle, 1998). An extensive revision was

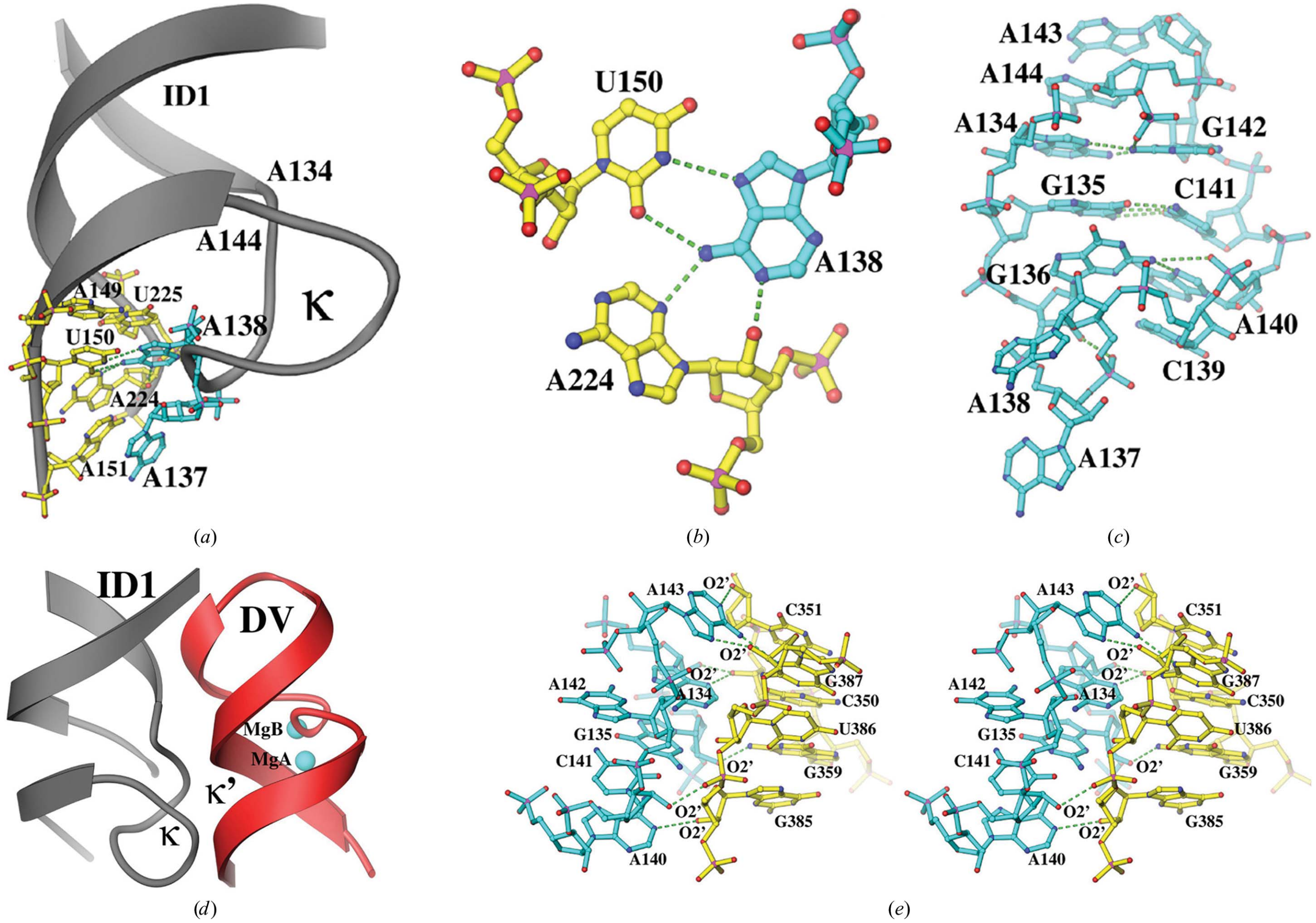


Figure 5
The κ–κ′ tertiary interactions. (*a*) A137 and A138 (cyan) are two splayed-out nucleotides at the tip of the κ stem-loop of ID1 and make extensive interactions with DV. (*b*) A close-up of the base triplet involving A138 and the A138·U150 Hoogsteen base pair. (*c*) Detailed base stacking within the duplex-like κ loop region with standard G135·C141 Watson–Crick base pairs and other non-Watson–Crick base pairs. (*d*) A ribbon representation of the κ–κ′ tertiary interactions. (*e*) Stereodiagram of the κ (cyan)–κ′ (yellow) interface with all interacting O2′s labeled.

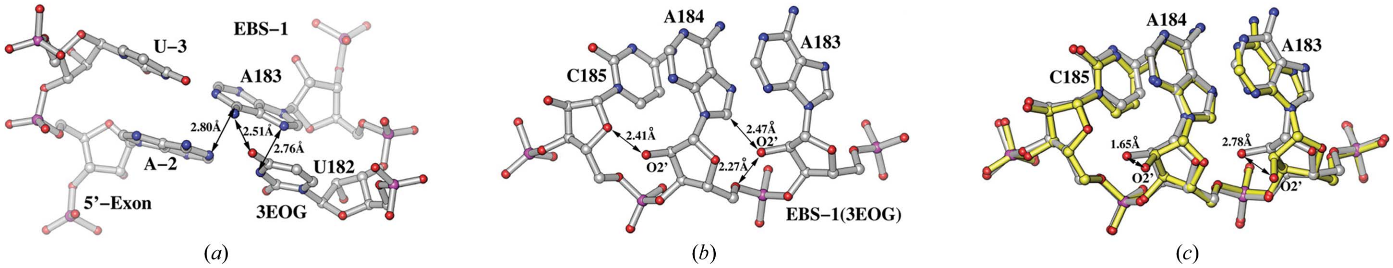


Figure 6
A-form helical RNA in the EBS-3 and IBS interaction. (*a*) Poorly stacked bases in the original model 3eog. (*b*) Poor stereochemistry for O2′ in the original model 3eog. (*c*) Comparison of our new idealized A-form-like duplex model (yellow, 3g78) with the original model (silver, 3eog) with repositioned O2′s indicated. The coordinate differences for O2′ were as large as 2.8 Å.

previously made in this region of the structure when the native data were reprocessed but without resolution extension (Toor, Rajashankar *et al.*, 2008). However, in both the original and the revised interpretations 3bwp and 3eoh the *κ*–*κ*′ ternary interactions were incorrect: (i) phosphate backbones were misplaced at locations where stacked bases should be, (ii) bases were present where phosphate backbones should be and (iii) stacked bases were improperly unstacked (Fig. S3). Similar misplacement errors also occurred in the catalytic site of the previous interpretations 3bwp and 3eoh (Figs. 3*c*, 3*g* and 3*h*). After the inclusion of high-resolution data in this revision, the *κ* stem can be recognized to be in a duplex with nearly all bases being stacked and forming inter-base hydrogen bonds (Figs. 5*a*, 5*b* and 5*c*). The *κ*–*κ*′ tertiary interactions are mediated by one inter-base and seven inter-domain O2′-mediated hydrogen bonds (Figs. 5*d* and 5*e*), the latter of which are a hallmark of all RNA structures. Further discussions on how the inclusion of high-resolution weak data helps in structure refinement and density modification can be found in the Supplementary Material. We would like to emphasize that amplitude-sharpening, like amplitude demodulation in correction for the lattice-translocation defects (Wang *et al.*, 2005), should affect unweighted statistics more than the weighted statistics such as *R* factors.

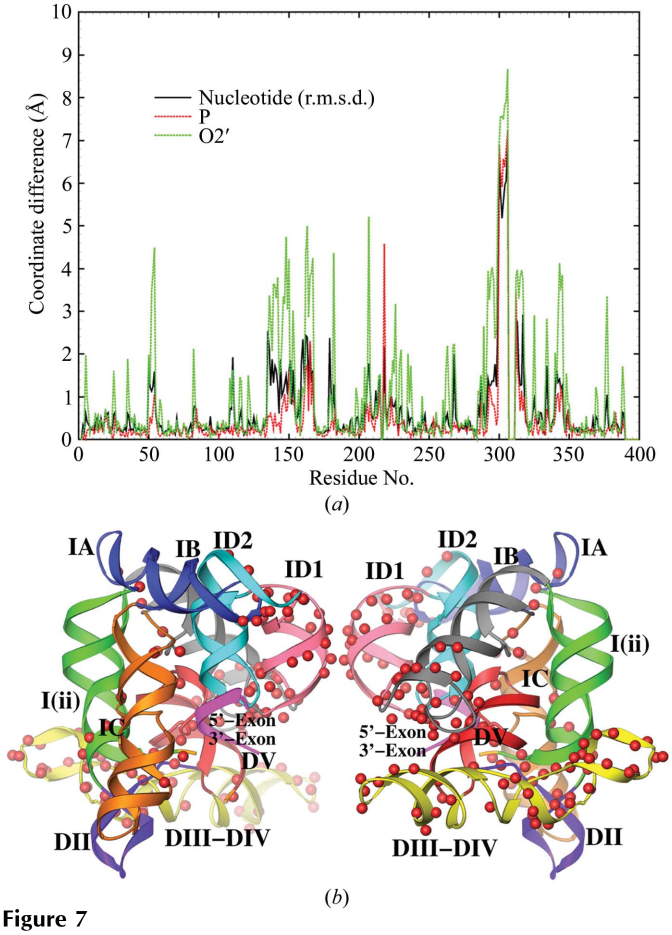


Figure 7
Distribution of revisions. (*a*) Coordinate revisions for P (red), O2′ (green) and the entire nucleotide (r.m.s.d., black) as a function of residue number. (*b*) Back and front views of the three-dimensional distribution of large coordinate revisions (>1.0 Å for O2′ or a nucleotide).

5. Structure refinement and model rebuilding

Although it was obvious that improved experimental electron-density maps greatly helped to correct many errors in the original structures (Figs. 3, 4, 5 and S1–S5), the issue was raised as to why previous structure refinement failed to do so (Toor, Keating *et al.*, 2008; Toor, Rajashankar *et al.*, 2008). Structure refinement is the most powerful method to bootstrap incomplete structures that cannot be interpreted from initial experimental maps and most crystallographers heavily rely on structure refinement to complete structure determination. Here, we hope to learn some valuable lessons by providing some clues to identify the major problems in the original structures that might have impaired structure refinement (Toor, Keating *et al.*, 2008; Toor, Rajashankar *et al.*, 2008). There were systematic errors in the interpretation of the O2′ positions and sugar puckers as well as in the sugar conformations of many nucleotides, even in many A-form helical regions (Figs. 6, S4 and S5). These led to large coordinate differences between this revision 3g78 and and the original structure 3bwp (Toor, Keating *et al.*, 2008), with a root-mean-square deviation (r.m.s.d.) for all atoms of about 1.6 Å and a maximal coordinate difference of 12.4 Å (Fig. 7), and between this revision 3g78 and the previously revised model 3eoh (Toor, Rajashankar *et al.*, 2008), with an r.m.s.d. for all atoms of about 1.4 Å and a maximal difference of 8.9 Å.

With the exception of a few misplaced or missing phosphate backbones, the placement of most phosphoryl atoms was largely accurate in the original structures 3bwp and 3eoh

Table 3
Refinement statistics of the revised intron structure.

Values in parentheses are for the highest resolution shell.

PDB code	3g78
Resolution (Å)	40–2.80 (2.88–2.80)
Unit-cell parameters (Å)	a = 89.11, b = 94.97, c = 226.0
No. of reflections	44197
Observation:parameter ratio	1.22
$R_{\rm work}$ (%)	19.6 (62.7)
$R_{\rm free}$ (%)	22.6 (69.5)
R.m.s.d. bonds (Å)	0.005
R.m.s.d. angles (°)	1.2
Total No. of atoms	9057
$\langle B \rangle$† (Å^2)	20.9
No. of intron nucleotides	389
No. of exon nucleotides	9
No. of Mg^{2+} ions	52
No. of K^{+} ions	24
No. of water molecules	435

† Average B factor corresponding to sharpened amplitudes.

(Toor, Keating *et al.*, 2008; Toor, Rajashankar *et al.*, 2008). However, owing to errors in the sugar pucker and the misplaced O2′, the orientations of the phosphate groups for many nucleotides were largely incorrect. Some errors in O2′ placement in the original structures may be attributable to difficulties in the interpretation of poor-quality regions of the original experimental maps, for example near the catalytic bulge residues A376 and C377 that bind two catalytic metal ions. In this bulge, O2′ was misplaced by 3.4 Å (Fig. 4*c*). However, it was puzzling how and why many C2′-*endo* form nucleotides were previously built even in A-form helical regions, where misplacement errors for O2′ were as high as 4.8 Å (Fig. S5*g*). In A-form helical regions the sugar pucker should always be C3′-*endo* (Rich, 2003). If one uses standard crystallographic software such as *Coot* (Emsley & Cowtan, 2004) to build an A-form duplex, the sugar pucker is always C3′-*endo*. The C3′-*endo* sugar pucker can be also re-enforced

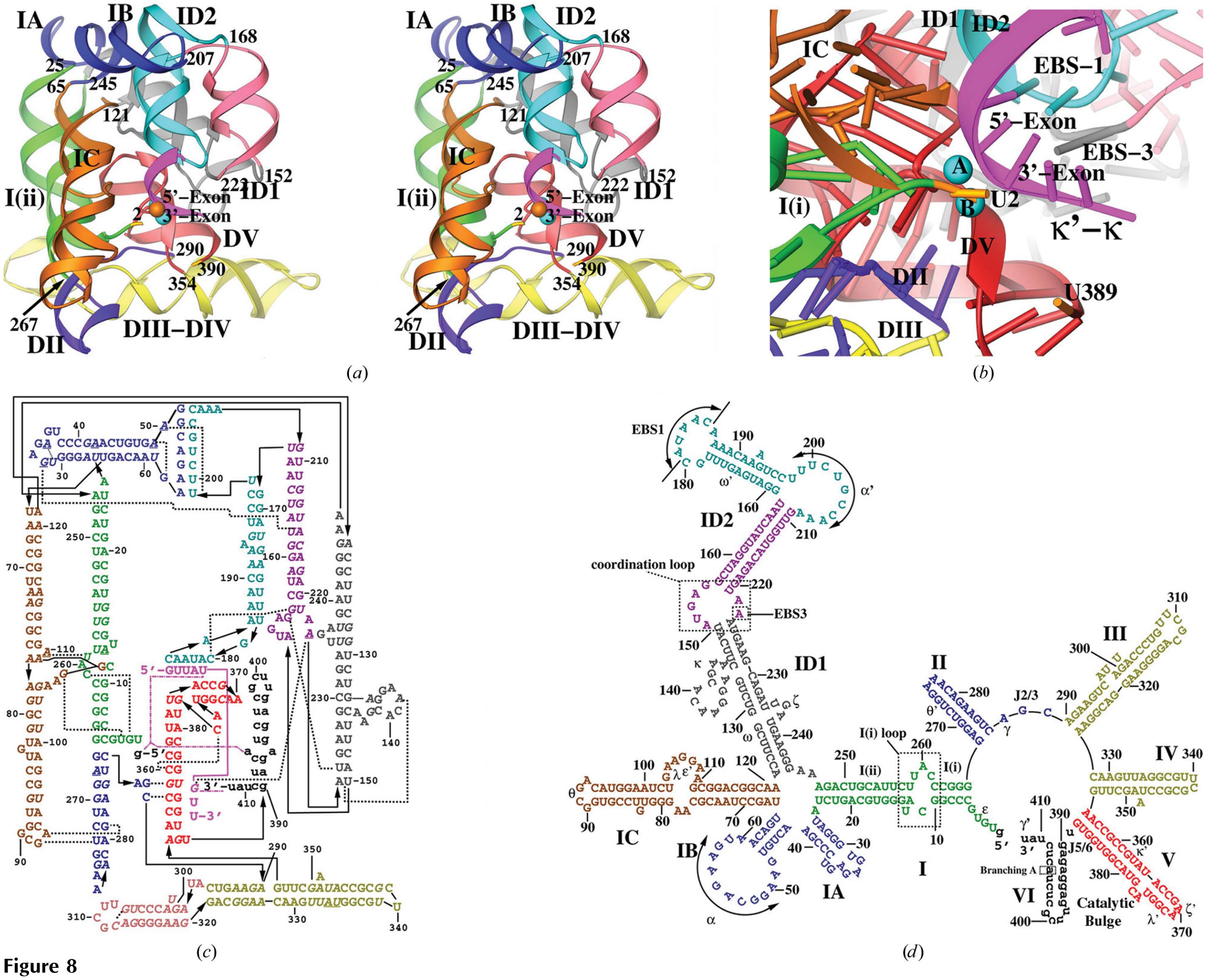


Figure 8
Revised secondary structures of the group II intron. (*a*) Stereodiagram of a ribbon representation of our revised structure (3g78) colored by domain, with residues at domain junctions numbered. (*b*) A close-up of the catalytic site. (*c*) A schematic drawing of secondary structures corresponding to (*a*). Non-Watson–Crick base pairs are in italics; underlined nucleotides are in Hoogsteen base pairs. Long-range tertiary interactions are shown as dotted lines and connections as solid arrows. (*d*) Classic drawing of the secondary structures with all known important long-range interactions labeled.

during refinement using *CNS* (Brünger *et al.*, 1998) even when a non-C3′-*endo* pucker is initially built. Other errors included a systematic building of non-wobble base pairs for nearly all wobble G·U base pairs in the original structure (Fig. S1*f*), which were largely corrected in an initial revision (Toor, Keating *et al.*, 2008; Toor, Rajashankar *et al.*, 2008).

The reason why the previous refinement failed to correct the modeling errors in the initial studies was likely to be because the errors were extensive and systematically distributed throughout the entire structure (Fig. 7). At 3.1 Å resolution, the observation-to-parameter ratio was 1.1 in the original structure 3bwp (Toor, Keating *et al.*, 2008). With such a low observation-to-parameter ratio and poorly processed data, refinement would rapidly run into local minima owing to initial incorrect model bias.

In this structure refinement, we gradually extended the resolution to 2.8 Å with tight geometric constraints, including Watson–Crick base pairing and proper 3′-*endo* sugar puckers whenever known or where possible, a procedure that proved to be successful in structure refinement of the group I intron structure (Adams *et al.*, 2004). These constraints were applied when refinement was carried out using the program *CNS* during torsion-angle dynamics simulation (Brünger *et al.*, 1998), because at 2.8 Å resolution the observation-to-parameter ratio of 1.2 is still relatively low. Once the geometry had been maintained properly in *CNS* refinement and structure refinement became stable, these constraints were then removed as refinement of the structure continued using the program *REFMAC* (Murshudov *et al.*, 1997). The model was rebuilt using the programs *ONO* and *Coot* (Jones *et al.*, 1991; Emsley & Cowtan, 2004). In this revised intron model 3g78, the last visible nucleotide is U389 at the domain junction between DV and DVI or J5/6. Experimental densities for the second residue U2 were relatively weak, but were well defined for the third intron residue G3. The first visible nucleotide was U2 in another related native structure, resulting in a more complete model with no gaps between U2 and U389. DVI remains missing from the intron structure. Refinement statistics are summarized in Table 3.

6. Revised secondary structures of the group II intron and metal-ion binding

In addition to the correction of register errors in sequence assignment and of geometric errors in secondary structure, we also observed that there were 52 Mg^{2+} ions (Fig. S6). Five of the Mg^{2+} ions corresponded to Yb^{3+}-binding sites and 17 to Ir^{3+}-binding sites (Fig. S6). Additionally, there were 27 K^{+}-binding sites. All of these sites were identified using criteria as described elsewhere (Adams *et al.*, 2004; Klein *et al.*, 2004), *i.e.* with strong densities and reasonable geometric relationships

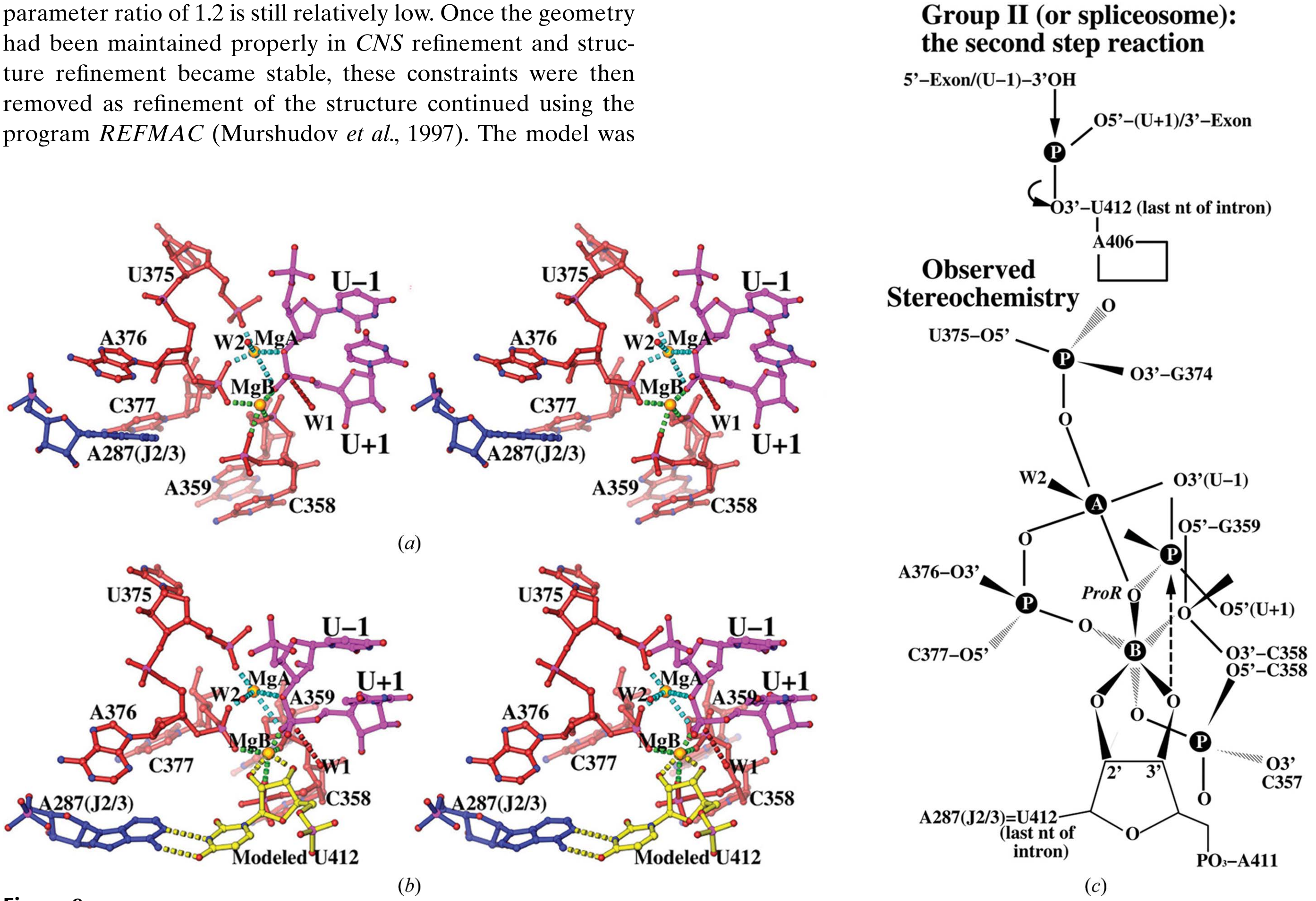


Figure 9

Stereochemistry of the exon-ligation reaction. (*a*) The observed catalytic site in stereo. An ordered solvent molecule W1 is near the attacking position. (*b*) The nucleophile U412 (yellow) was modeled into the catalytic site by making the U412·A287 (blue, at domain junction J2/3) Watson–Crick base pair. (*c*) Schematic drawing of the stereochemistry.

to the phosphates and bases of the intron. These metal ions are distributed throughout the entire intron structure (Fig. S6) and are known to play important stabilization roles in the folding of large RNA structures (Klein *et al.*, 2004). This revised structure also led to a new structure-based assignment of secondary structures for the group II intron (Fig. 8).

7. Stereochemistry of catalytic metal ions and scissile phosphates

In this revised structure of a group II intron, we observed a bound RNA product at the catalytic site that is relevant to the reversal of the second splicing step (Fig. 3), in which only one of the two nonbridging O atoms of the scissile phosphate interacts with the two metal ions A and B (Fig. 9). The two nonbridging O atoms are geometrically distinct and are named Pro-Sp or Pro-Rp according to the chirality of the scissile phosphate after phosphorothioate substitution. The nonbridging O atom that interacts with the two metal ions is Pro-Sp in the product in our structure (Fig. 9). This O atom is Pro-Rp in the substrate before the inversion of the phosphoryl center. This finding is consistent with the results of Gordon *et al.* (2000) for this step and with the predictions of the catalytic mechanism by Steitz & Steitz (1993).

This structure can also indirectly address the catalytic mechanism of the first self-splicing reaction, lariat formation (Fig. 1), even though it does not reveal the corresponding substrate-bound or product-bound state. However, we can model these states with reasonable confidence. We have observed a third noncatalytic metal ion C in the catalytic site in the Yb^{3+}-derivative structure (Fig. 2*b*). However, the presence of this metal ion could stabilize the unspliced substrate in which the 3′-end of the 5′-exon is joined to the 5′-end of the intron (Fig. 2*c*) or could slow the first self-splicing reaction, leading to the accumulation of the unspliced substrate in the derivative structure. Although the electron density did not unambiguously reveal where G1 was located in the model of the native structure, electron densities for the two flanking phosphodiesters were well defined, suggesting that the relative positioning of the 5′-exon and the 5′-end of the intron remains unchanged from an initial unspliced structure to the final spliced structure as observed here.

There was a gap about 12 Å between the scissile phosphate of the 5′-exon and the 5′-phosphate of G3 in the revised structure (U2 was modeled based on another native intron structure, but not as confidently as in the highest resolution intron structure). We can model the location of a single phosphodiester at the midpoint of this gap, about 6 Å away from the scissile phosphate (Fig. 2*c*), which bridges the two unobserved nucleotides G1 and U2 in the unspliced substrate. With this modeling, the stereochemistry of the scissile phosphate in the unspliced substrate is now defined as Pro-Rp. After the inversion of the phosphate center in the first splicing reaction, it is Pro-Sp in the product. Thus, our modeling shows that there is a rotation of 120° of the O5′ connection around the central axis of the pentacovalent phosphoryl intermediate, which alters the modeled O5′ path in the unspliced substrate to the observed continuous helical path of the observed exon-ligation product, as predicted previously by Steitz & Steitz (1993). This rotation also inverts the chirality of the scissile phosphate again so that it is Pro-Rp for the substrate of exon ligation. Thus, the revised structure is consistent with the observation that both of the forward splicing reactions are strongly inhibited by the Pro-Rp phosphorothioate diastereomer as substrate, but not by the Pro-Sp isomer (Moore & Sharp, 1993; Gordon *et al.*, 2000).

8. Two opposite orientations of DVI in two self-splicing reactions

We are still missing half of the catalytic site for both splicing reactions: (i) we observed the product of the exon-ligation reaction but not its substrate in the revised exon structure and (ii) we have deduced the substrate of the lariat-formation

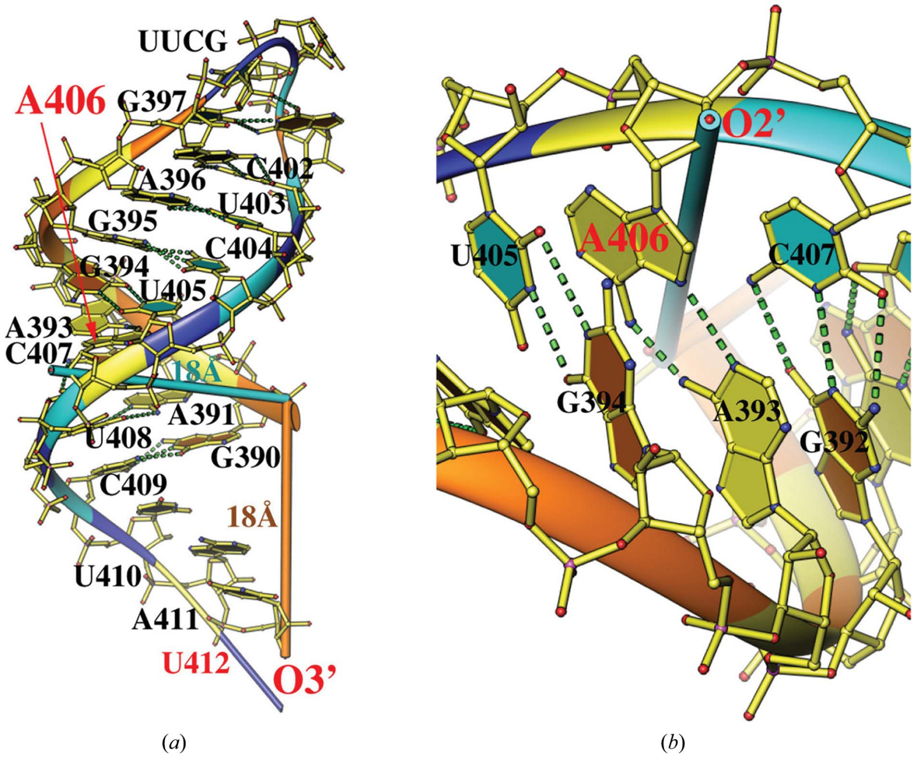


Figure 10
Docking of an isolated rigid DVI model onto the intron structure. (*a*) A rigid RNA duplex model for DVI with indicated distances (18 Å) from the nucleophiles A406 and U412 to the putative hinge point at the 5′-phosphate of G390. Backbone ribbons are colored by nucleotide type. (*b*) A close-up view of the branch-site A406 adjacent to a wobble U405·G394 base pair.

reaction but not yet its product. The missing product of the lariat-formation reaction is the 2′–5′ phophodiester-linked A406 and the missing substrate for the exon-ligation reaction is the 3′-OH of U412, both of which are unfortunately located on the missing DVI in the intron structure. There are at least three reasons why DVI might be invisible in the structure: (i) its truncation in the intron construct used for crystallization to remove the known η–η' interactions for proper orientation of DVI in the exon-ligation reaction (Toor, Keating *et al.*, 2008), (ii) its multiple orientations during the reaction cycle and (iii) its partial degradation during prolonged crystallization incubation (Toor *et al.*, 2010).

The missing DVI in the intron is a rigid helical stem with eight base pairs capped by a 398-UUCG-401 tetranucleotidyl loop and a 410-UAU-412 trinucleotidyl 3′-tail (Figs. 10*a* and 10*b*). Of the eight base pairs, six are Watson–Crick base pairs and flank two middle non-Watson–Crick base pairs: a U405·G394 wobble base pair that defines the branch-site A406 formation of an A406–A393 mispair (Chu *et al.*, 2001). As the product of the lariat-formation reaction, the product A406 (the nucleophile for the reverse reaction) is next to the scissile phosphate. As the substrate of the exon-ligation reaction, the substrate U412 (the nucleophile for the forward reaction) is next to the scissile phosphate. Because DVI is joined to DV by the J5/6 nucleotide U389 and because there is an equal distance of about 18 Å between the nucleophile A406 and the 3′-phosphate of U389 of the intron (*i.e.* the 5′-phosphate of G390 in the rigid DVI model; Fig. 10*a*) and between the nucleophile U412 and the 3′-phosphate of U389, J5/6 is likely to serve as the hinge point for the rotation of DVI to swap A406 or U412 out of and into the catalytic site.

By constraining either A406 or U412 next to the scissile phosphate and G390 of the rigid DVI model next to U389 of the intron, we can dock the rigid DVI model onto the intron in

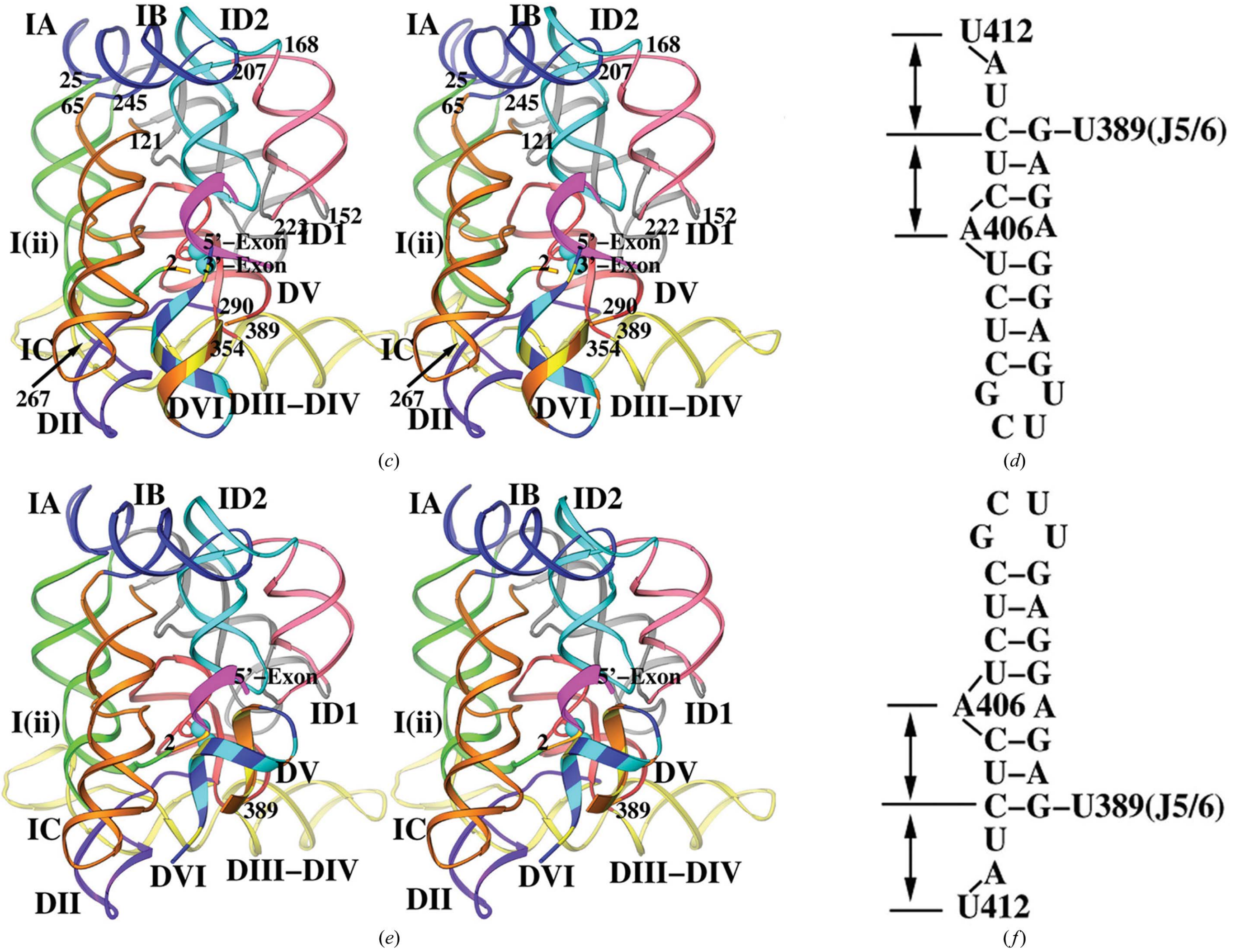


Figure 10 (continued)

(*c*) Docking of the DVI model [thick ribbons as in (*a*) colored by nucleotide] onto the intron structure (thin ribbons, colored by domain, with domain junctions numbered) with the ligated exon product (thick ribbons, magenta) bound in the catalytic site for reversal of the exon-ligation reaction. Two catalytic metal ions are shown as cyan spheres. (*d*) Schematic representation of (*c*) for the docked DVI in the exon-ligation reaction with its tetranucleotide loop in the down position. (*e*) Docking of DVI for the lariat-formation reaction with the 5′-exon connected to G1 of the intron, making a sharp turn. (*f*) Schematic representation of (*e*) for docked DVI with its tetranucleotide loop in an up position.

two distinct orientations, one with the tetraloop UUCG in the down position for the exon-ligation reaction (Figs. 10*c* and 10*d*) and the other with the tetraloop UUCG in the up position for the lariat-formation reaction (Figs. 10*e* and 10*f*). In the tetraloop-down position, U412 of DVI fits into the catalytic site better if the two nucleotides U410 and A411 form a bulge to loop themselves out. When U412 is in the catalytic site for the exon-ligation reaction, U412 forms Watson–Crick base pair with A287 at J2/3 through known γ–γ' interactions (Jacquier & Michel, 1990). With this constraint, one can see how both the 2′-OH and 3′-OH groups coordinate to the catalytic metal ion, while only the 3′-OH group is in the correct position for nucleophilic attack (Fig. 9*b*). The modeled 180° orientation of DVI in the two splicing reactions is consistent with the previous observation of a large conformational change between the two splicing reactions in group II introns (Chanfreau & Jacquier, 1994). In these two orientations of DVI, the relative positions of 2′-OH and 3′-OH and the polarity of the phosphate backbones that bear either A406 or U412 are also switched. This modeling is consistent with the observation of a switched preference between 2′-OH and 3′-OH as a nucleophile in the two splicing reactions in group II introns (Gordon *et al.*, 2000). Interestingly, two classes of tRNA synthetases also switch the preference between 2′-OH and 3′-OH for acylation, but by binding tRNAs in two opposite orientations to synthetases (Eriani *et al.*, 1990), *i.e.* by switching the polarity of the RNA strand that bears the nucleophile.

9. Concluding remarks

In this study as well as in an earlier study (Wang & Boisvert, 2003), we have demonstrated that weak high-resolution data with $\langle F/\sigma(F)\rangle = 1.0$ or $\langle I/\sigma(I)\rangle = 0.5$ in the highest resolution shell contain important structural information for accurate determination of macromolecular structures and should be included in structure refinement. The bottom line of this study is that useful weak high-resolution data should not be discarded in macromolecular structure determination. We should push the boundary of the resolution of macromolecular structures as far as possible during structure determination, where we learn more about the chemistry as well as the biology of these macromolecules.

I would like to acknowledge that Dr Pyle provided the original X-ray diffraction data corresponding to the structure of the group II intron published in *Science* (Toor, Keating *et al.*, 2008) for this analysis. Dr Pyle did not participate in the reinterpretation of the grounp II intron structure described in this paper. I thank Dr W. H. Konigsberg for proofreading and commenting on this manuscript.

References

Adams, P. L., Stahley, M. R., Kosek, A. B., Wang, J. & Strobel, S. A. (2004). *Nature (London)*, **430**, 45–50.
Boudvillain, M. & Pyle, A. M. (1998). *EMBO J.* **17**, 7091–7104.
Brünger, A. T. (1993). *X-PLOR Manual Version* 3.1. Yale University Press, New Haven, USA.
Brünger, A. T., Adams, P. D., Clore, G. M., DeLano, W. L., Gros, P., Grosse-Kunstleve, R. W., Jiang, J.-S., Kuszewski, J., Nilges, M., Pannu, N. S., Read, R. J., Rice, L. M., Simonson, T. & Warren, G. L. (1998). *Acta Cryst.* D**54**, 905–921.
Ceck, T. R. (1986). *Cell*, **44**, 207–210.
Chanfreau, G. & Jacquier, A. (1994). *Science*, **266**, 1383–1387.
Chu, V. T., Adamidi, C., Lie, Q., Perlman, P. S. & Pyle, A. M. (2001). *EMBO J.* **20**, 6866–6876.
Cura, V., Khrishnaswamy, S. & Podjarny, A. D. (1992). *Acta Cryst.* A**48**, 756–764.
Emsley, P. & Cowtan, K. (2004). *Acta Cryst.* D**60**, 2126–2132.
Eriani, G., Delarue, M., Poch, O., Gangloff, J. & Moras, D. (1990). *Nature (London)*, **347**, 203–206.
Gordon, P. M., Fong, R. & Piccirilli, J. A. (2007). *Chem. Biol.* **14**, 607–612.
Gordon, P. M. & Piccirilli, J. A. (2001). *Nature Struct. Biol.* **8**, 893–898.
Gordon, P. M., Sontheimer, E. J. & Piccirilli, J. A. (2000). *Biochemistry*, **39**, 12939–12952.
Jacquier, A. & Michel, F. (1990). *J. Mol. Biol.* **213**, 437–447.
Jones, T. A., Zou, J.-Y., Cowan, S. W. & Kjeldgaard, M. (1991). *Acta Cryst.* A**47**, 110–119.
Klein, D. J., Moore, P. B. & Steitz, T. A. (2004). *RNA*, **10**, 1366–1379.
Lambowitz, A. M. & Zimmerly, S. (2004). *Annu. Rev. Genet.* **38**, 1–35.
Moore, M. J. & Sharp, P. A. (1993). *Nature (London)*, **365**, 364–368.
Murshudov, G. N., Vagin, A. A. & Dodson, E. J. (1997). *Acta Cryst.* D**53**, 240–255.
Otwinowski, Z. (1991). *Proceedings of the CCP4 Study Weekend. Data Collection and Processing*, edited by W. Wolf, P. R. Evans & A. G. W. Leslie, pp. 80–86. Warrington: Daresbury Laboratory.
Otwinowski, Z. & Minor, W. (1997). *Methods Enzymol.* **276**, 307–326.
Pyle, A. M. & Lambowitz, A. M. (2006). In *The RNA World*, 3rd ed., edited by R. F. Gesteland, T. R. Cech & J. F. Atkins. New York: Cold Spring Harbor Laboratory Press.
Rich, A. (2003). *Nature Struct. Biol.* **10**, 247–249.
Rould, M. A., Perona, J. J. & Steitz, T. A. (1992). *Acta Cryst.* A**48**, 751–756.
Sharp, P. A. (1994). *Cell*, **77**, 805–815.
Steitz, T. A. & Steitz, J. A. (1993). *Proc. Natl Acad. Sci. USA*, **90**, 6498–6502.
Toor, N., Keating, K. S., Fedorova, O., Rajashankar, K., Wang, J. & Pyle, A. M. (2010). *RNA*, **16**, 57–69.
Toor, N., Keating, K. S., Taylor, S. D. & Pyle, A. M. (2008). *Science*, **320**, 77–82.
Toor, N., Rajashankar, K., Keating, K. S. & Pyle, A. M. (2008). *Nature Struct. Mol. Biol.* **15**, 1221–1222.
Wang, J. & Boisvert, D. C. (2003). *J. Mol. Biol.* **327**, 843–855.
Wang, J., Kamtekar, S., Berman, A. J. & Steitz, T. A. (2005). *Acta Cryst.* D**61**, 67–74.