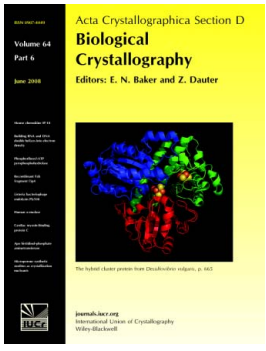
Acta Crystallographica Section D
Biological Crystallography
ISSN 0907-4449
Editors: **E. N. Baker and Z. Dauter**

Building of RNA and DNA double helices into electron density

Frantisek Pavelcik and Bohdan Schneider

Acta Cryst. (2008). D**64**, 620–626



Acta Crystallographica Section D: Biological Crystallography welcomes the submission of papers covering any aspect of structural biology, with a particular emphasis on the structures of biological macromolecules and the methods used to determine them. Reports on new protein structures are particularly encouraged, as are structure–function papers that could include crystallographic binding studies, or structural analysis of mutants or other modified forms of a known protein structure. The key criterion is that such papers should present new insights into biology, chemistry or structure. Papers on crystallographic methods should be oriented towards biological crystallography, and may include new approaches to any aspect of structure determination or analysis.

Acta Crystallographica Section D
Biological Crystallography
ISSN 0907-4449

Building of RNA and DNA double helices into electron density

Frantisek Pavelcik[a,b]* and Bohdan Schneider[c]*

[a]Department of Chemical Drugs, Faculty of Pharmacy, University of Veterinary and Pharmaceutical Sciences in Brno, CZ-61242 Brno, Czech Republic, [b]Department of Inorganic Chemistry, Faculty of Natural Sciences, Comenius University in Bratislava, SK-84215 Bratislava, Slovak Republic, and [c]Institute of Organic Chemistry and Biochemistry, Academy of Sciences of the Czech Republic, Fleming Square 2, CZ-16610 Prague, Czech Republic

Correspondence e-mail: pavelcikf@vfu.cz, bohdan@uochb.cas.cz

Received 7 November 2007
Accepted 13 March 2008

A method has been developed that automatically fits double-helical regions into the electron density of nucleic acid structures. Rigid fragments consisting of two Watson–Crick base pairs and three pairs of phosphate groups in the A-type or B-type conformation are positioned into the electron density by phased rotation and translation functions. The position and orientation of the localized double-helical fragments are determined by phased refinement. The method has been tested by building double-helical regions of nine RNA structures of variable crystallographic resolution and polynucleotide length and is available for free use.

1. Introduction

The automatic building of molecular models into electron density is an open challenge in macromolecular crystallography. The problem has been partially solved for proteins (Lamzin & Wilson, 1993; Perrakis *et al.*, 1999; Levitt, 2001; Terwilliger, 2001; Oldfield, 2003; Pavelcik, 2003, 2004), but the automatic building of nucleic acids has been virtually unexplored. This is in stark contrast to recent developments in the structural science of nucleic acids, especially of RNA. The last several years have witnessed unprecedented growth in the crystallography of large biological RNA molecules such as ribozymes and riboswitches and in particular ribosomes, so that the database of known nucleotide structures has multiplied in quantity as well as in the diversity of structural motifs. The multiple functions of RNA molecules in the essential processes of transcription and translation guarantee that new nucleic acid structures will continue emerging. At the same time, the diverse and complicated folds of RNA molecules make the fitting of electron densities a challenging task, especially when the relatively low resolution of most crystal structures containing RNA is considered. Of around 700 structures released by the NDB (Berman *et al.*, 1992), less than half (335) have a resolution better than 2.5 Å and a mere 40 have a resolution better than 1.5 Å.

This work reports an attempt to automate the fitting of the prevailing building block of nucleic acid structures, the double helix in the A- and B-forms; emphasis was placed on the main motif of RNA architecture, the A-form double helix. We utilize the advanced methodology of the so-called 'phased rotation, conformation and translation function' (PRCTF) as recently developed and described for protein fitting (Pavelcik *et al.*, 2002; Pavelcik, 2006). The method was tested on a set of nine RNA structures with crystallographic resolution varying between 1.5 and 3.1 Å and of very different sizes, from a moderate-size sarcin/ricin loop structure (Correll *et al.*, 1999)

to the structure of the large ribosomal subunit (Ban *et al.*, 2000). The work demonstrates that rigid double-helical tetranucleotide fragments with two Watson–Crick base pairs can be successfully used to build double-helical stems with Watson–Crick base pairs for phased maps at fairly low resolution (3.5 Å or even lower) in both small and large RNA structures.

2. Methods

2.1. Molecular fragments for fitting

The most prevalent and structurally conservative structural elements of nucleic acids are double helices, A-form in RNA and B-form in DNA, and a double-helical segment is therefore an obvious choice of fragment for fitting into electron density. We used a short double helix with two base pairs in the Watson–Crick arrangement and three pairs of phosphate groups (Fig. 1). The size of the fragment should be sufficient for fitting of electron densities into maps of lower resolution, certainly worse than 2.5 Å. A longer double helix was not used because it may lead to artificial smoothing of the density and oversimplification of the resulting model.

The double-helical fragments (Fig. 1) contain the backbone atoms of two full nucleotides plus the 3′-end phosphate; they start at atom O3′($i-1$) and end at O5′($i+2$) and contain 54 backbone and 36 base atoms: [O3′-PO_2-O5′-C5′-C4′(O4′)-C3′(C2′-C1′-B1)-O3′-PO_2-O5′-C5′-C4′(O4′)-C3′(C2′-C1′-B2)-O3′-PO_2-O5′]$_2$. B1 and B2 are bases. The fragments were built in the conformations of the two principal double-helical forms, A and BI (Schneider *et al.*, 1997), so that searches are possible for both RNA and DNA molecules. Bases are modelled as generalized purine (R) and pyrimidine (Y). To capture the basic sequence features of the electron density, fragments with all three combinations of R and Y were built, *i.e.* RR, RY and YR. A-RNA fragments are called NA_RY, NA_RR and NA_YR and have a radius of 11.2 Å; B-DNA fragments are correspondingly NB_RY, NB_RR and NB_YR with a radius of 11.8 Å. Fragments NA_YY and NB_YY are identical to the corresponding RR fragments owing to the Watson–Crick base pairing. The geometry of all fragments is kept rigid during fitting.

To test the robustness of the method and the possibility of a simplified procedure for building double helices, we also performed searches using only one fragment to fit all sequences; these searches were made using the fragment NA_YR. The resulting models, further referred to as NAhelix, did not discriminate between different sequences and only traced the sugar-phosphate backbones; their base atoms were therefore removed after the fitting.

2.2. Virtual geometry parameters for double-helical fragments

The fragments that were fitted into the electron density needed to be connected into longer chains. The orientation of the neighbouring fragments and their possible connection was determined by the calculation of virtual bonds and virtual bond angles: a virtual bond (VB) is the distance between the geometrical centres of two successive fragments and a virtual angle is defined as the angle between the geometrical centres of three successive fragments; for computational reasons, the virtual angle is represented by the distance (VT) between the end points of two successive virtual bonds. The above virtual parameters and their 'critical values' (see §2.3) were estimated from the structure of the canonical A-RNA double helix; the mean value is VB = 3.9 Å with a standard deviation of 0.4 Å.

2.3. Fragment overlap

A fragment positioned in the electron density (a refined peak of the PRCTF) is called a peak. Connecting two peaks is directed by their virtual distance and angle, as described above, and mutual overlap. The overlap was only calculated for peaks that were within the critical values defined by the virtual distances and was measured as the root-mean-square difference between the positions of the four corresponding P atoms (*e.g.* M–B, N–C, Q–D and R–E in Fig. 2),

$$\mathrm{DPP} = \left(\frac{\sum d^2}{4}\right)^{1/2}, \qquad (1)$$

where d is the distance of the overlapping P atoms. Calculation of the overlap of two double-helical fragments is more complicated than in proteins (Pavelcik, 2004) or single-stranded nucleotides because double-helical fragments have exact or approximate twofold rotation symmetry with equivocal 'upper' and 'lower' ends, so that four relative positions of two peaks have to be considered; this is shown in Fig. 2. All four combinations are calculated and the best overlap is selected. If the r.m.s.d. of the peak overlap [DPP in (1)] is less than 1.5 Å then the two peaks are accepted as being

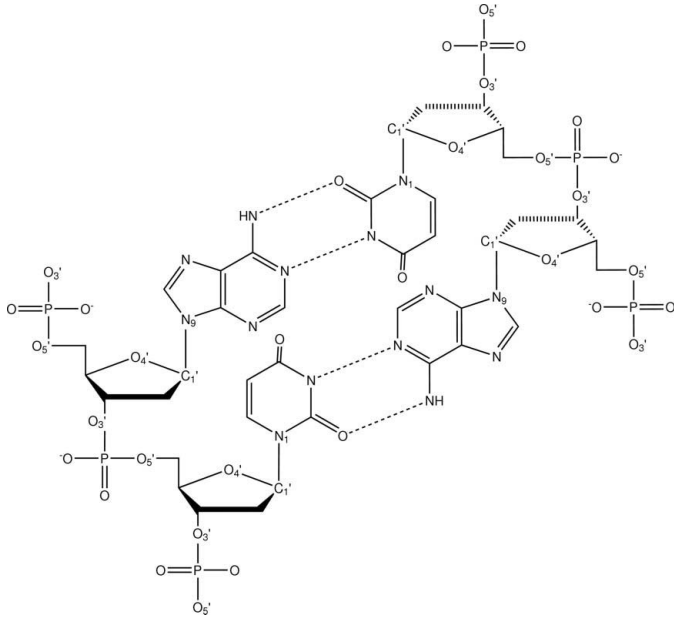


Figure 1
Definition of the double-helical fragments. Note that the fragment has the third phosphate group attached to the 3′-end of the dinucleotide. The chemical diagram shown is for NA_RY or NB_RY.

connected. The best overlap determines the relative position and orientation of two fragments. Information about the overlap is used for construction of the polynucleotide strand and for sorting peaks.

Another parameter resulting from the overlap is FIT. FIT is used as a qualitative parameter reflecting base overlap. If two fragments overlap by the same type of base, purine over purine and pyrimidine over pyrimidine, then FIT = 0.2DPP; for the overlap of different bases FIT = 0.6DPP [DPP is defined in (1)]. Possible overlaps between two peaks have a distribution that is characterized by a weight w calculated by

$$w = \exp\left[-\frac{\mathrm{DPP}}{K\sigma(\mathrm{DPP})}\right]\exp\left[-\frac{\mathrm{FIT}}{K\sigma(\mathrm{FIT})}\right], \qquad (2)$$

where the standard estimated deviations σ(DPP) and σ(FIT) are estimated from all overlaps. K is an empirical parameter set to $K = 4$. The weight w is calculated for all peaks and is used to sort them before they are connected. (2) is empirical and was adopted from protein-building methods (Pavelcik, 2004).

2.4. Connecting peaks

The algorithm for connecting peaks (nucleotide fragments) was adopted from the procedure developed for amino acids (Pavelcik, 2004). Chain building starts at the peak with the highest score by positioning the double-helical fragment (A-B-C/D-E-F in Fig. 2); its orientation fixes the 5′- and 3′-ends of the built double helix. Other peaks are added to both ends of the first peak according to the relative positions and orientations of overlapping PRCTF peaks. Only fragments that have VB in the interval 3.5–4.3 Å are accepted for model building and the VT between three successive fragments is not allowed to be shorter than 6.0 Å. An acceptable value of the virtual angle (the distance between the first and last geometrical centres of three successive fragments) is an important criterion because it prevents building strands in the wrong direction. Which peak is attached is decided by the weight [quality of overlap, calculated by (2)], connectivity and virtual angle. Four possible connections of two fragments are shown in Fig. 2. Peaks are connected until no further peaks can be added and building of the chain is then terminated. Building of another chain is then initiated by selecting the highest unconnected peak. All symmetry-equivalent positions are considered for all peaks in order to find the longest possible double-helical fragments, but the assignment of separate chains to one molecule in a single asymmetric unit cannot be guaranteed before the model has been almost finalized. A molecular envelope is not calculated and the chain building uses a purely crystallographic approach; nonbonding interactions are not currently considered in the procedure.

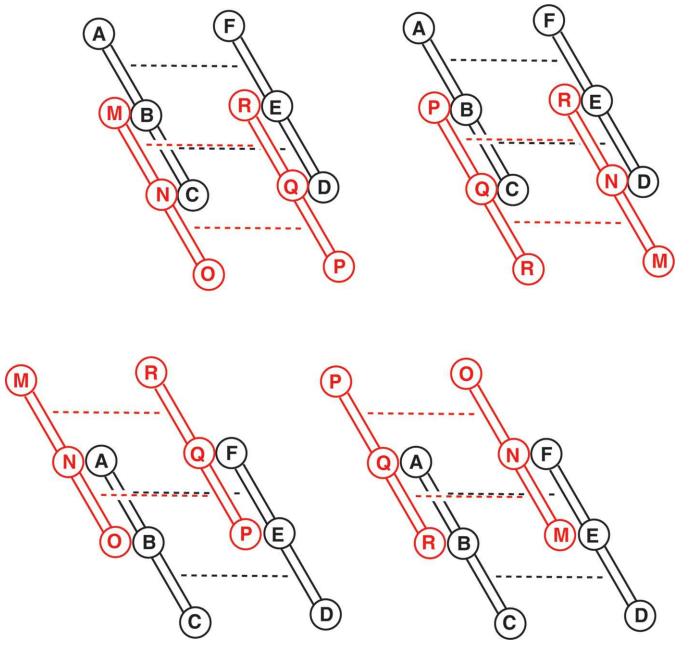


Figure 2
Four possible ways to connect two double-helical fragments into a polynucleotide chain. The first fragment has phosphates 5′-ABC-3′ and 5′-DEF-3′ and the second 5′-MNO-3′ and 5′-PQR-3′; base pairs are indicated by dashed lines.

2.5. Building of the nitrogenous bases

The present algorithm recognizes only purine, R (G or A), and pyrimidine, Y (C or U/T), base types for initial positioning into the electron density; it does not distinguish between A–U and G–C base pairs. Searches considering R/Y sequences require three fragments reflecting three dinucleotide sequences: YR, RY and RR. The chain can, in principle, be extended at both ends by three fragments of sequence YR, RY and RR; selection of the fragment and consequently the chain sequence is based directly on the result of the connecting algorithm. Model building with one fragment (NAhelix), ignoring the R/Y sequence, is simpler but the resulting fit is less discriminative.

2.6. The main steps of building

The basic steps of RNA or DNA building are similar to those of protein building (Pavelcik, 2004). Two separate programs, *NUT* (Pavelcik, 2006) and *DHL* (Pavelcik, unpublished work), are used for electron-density fitting. *NUT* is a program for calculation of the phased translation and rotation function (PRCTF) and *DHL* is a program for fragment connecting and for assembling the coordinate file of the model. It should be noted that neither protein chains nor water molecules or counter-ions are modelled in the current version of the *DHL* program. For testing purposes, the results are analyzed by the program *CMP* (Pavelcik, unpublished work), which compares the coordinates of the model with the coordinates of the fully refined structure. The input file is the same for the *NUT*, *DHL* and *CMP* programs. The process of model building is described in greater detail below.

2.6.1. Preparation of input. The basic parameters are unit-cell parameters, space-group symmetry and structure-factor amplitudes and phases. The sequence does not have to be specified; the number of nucleotides in the asymmetric unit is sufficient.

2.6.2. Electron-density expansion. The radius of the electron-density expansion was fixed at 12.5 Å for building with double-helical fragments. The step in the fast Fourier transform of the translation function is between 0.6 and 1.0 Å and is typically 0.7 Å. The maximal indices for spherical harmonics and Bessel functions are specified by 'quantum numbers' n_{max} and l_{max}. These parameters reflect a compromise between the accuracy of building, the speed of calculations and the disk space required; their maximum values are set to 6 and 8, respectively. The above-mentioned parameters are suitable for routine building of most structures with resolution better than 3.5 Å; more examples of input parameters are shown in the program manual.

2.6.3. Fragment overlap. The positions and orientations of the fragments in the electron density are determined by the PRCTF; the expected number of peaks is one half the number of nucleotides. The positions and orientation of the peaks are refined as geometrically rigid objects and stored in a file for use by the *DHL* and *CMP* programs.

2.6.4. Building a backbone model. Refined peak positions are sorted based on quality of the fit to the electron density, quality of the peak overlap and peak connectivity. Peaks connected to other peaks at both ends (*i.e.* those having two low values of DPP and FIT) are given higher priority in sorting because they form (longer) double helices; double-helical fragments of at least two connected peaks are saved in a PDB-style coordinate file and peaks with no connectivity are deleted. Building of double helices is based on virtual distances and fragment overlap between peaks as explained in §§2.3 and 2.4.

2.7. Validation of the model

To test the success of the method, fitted models of RNA structures were compared with the final refined structure as deposited in the PDB (Berman *et al.*, 2002). A root-mean-square deviation (r.m.s.d.) was calculated between the positions of corresponding atoms of the fitted model and the refined structure, r.m.s.d. = $[(\sum d^2)/n]^{1/2}$, where d are the distances between the related atoms and n is the number of atoms compared. Because the base types were uncertain, the r.m.s.d. was only calculated between the backbone atoms and $n = 54$ ($n = 49$ when the 5′-end nucleotide without the phosphate group was fitted). The overall quality of the fit was assessed by the calculation of an R factor and correlation coefficient for the fitted model using the program *REFMAC* (Murshudov *et al.*, 1999).

2.8. Selection of structures for testing

Testing was performed on medium to large RNA structures that had measured structure factors deposited in the PDB (Berman *et al.*, 2002); the basic crystallographic data are given in Table 1. Experimental phases were available for four structures: three publicly deposited CIF files, 1ffk (Ban *et al.*, 2000), 1j5e (Wimberly *et al.*, 2000) and 1mme (Scott *et al.*, 1995), and for the structure eden, phase set $\varphi2$ directly from the authors (Ennifar & Dumas, unpublished work). For the remaining structures, phases were calculated from the refined PDB-deposited coordinates. The structures for testing were also selected with their resolution in mind. Structures with near-atomic resolution were considered (sarcin-ricin loop of rRNA, PDB code 480d; Correll *et al.*, 1999) as well as structures with moderate resolution below 3.0 Å (hammerhead ribozyme; PDB code 1mme; Scott *et al.*, 1995) and the structure of a small (30S) ribosomal subunit (PDB code 1j5e; Wimberly *et al.*, 2000); the molecular weight or size of the models varies between the relatively small sarcin-ricin loop structure of 27 nucleotides to the very large structures of ribosomal subunits.

Table 1
Tested nucleic acid structures.

'Code' identifies the structure (the PDB code for deposited structures), 'Resol' is the crystallographic resolution in angstroms, 'SG' is the space group, 'Phases' describes what phases were used ('Exper' are experimental phases provided by the authors, 'Calc' are phases calculated from the refined atomic coordinates as deposited in the PDB). Observed structure-factor moduli ($|F_o|$) were used in all calculations.

Code	Resol	SG	Phases	Reference
480d	1.5	$P4_3$	Calc	Correll *et al.* (1999)
Eden $\varphi1$	1.6	$C2$	Calc	Ennifar & Dumas (unpublished)
Eden $\varphi2$	1.6	C2	Exper	Ennifar & Dumas (unpublished)
1nlc	1.9	$P3_121$	Calc	Ennifar *et al.* (2003)
1ehz	1.9	$P2_1$	Calc	Shi & Moore (2000)
1dk1	2.8	$P6_422$	Calc	Nikulin *et al.* (2000)
1u9s	2.9	$C222_1$	Calc	Krasilnikov *et al.* (2004)
1mme	3.1	$P3_121$	Exper	Scott *et al.* (1995)
1ffk	2.4	$C222_1$	Exper	Ban *et al.* (2000)
1j5e	3.1	$P4_12_12$	Exper	Wimberly *et al.* (2000)

All calculations were carried out on a 2.4 GHz Intel Core2 Duo CPU with 4 GB of RAM under Windows XP with the program compiled by the Compaq (Digital) Fortran90 compiler (parallel tests took place on a 2.4 GHz Intel P4 CPU with 1 GB of RAM under Linux Fedora with the program compiled with Intel Fortran 8.0). The times of density fitting varied between seconds for small structures (25 s CPU time for 480d) to a few hours for extremely large structures (7321 s CPU time for 1ffk).

3. Results and discussion

Table 2 summarizes the results of automated fitting of the three double-helical dinucleotide fragments NA_YR, NA_RY and NA_RR to the electron densities of the nine structures and Table 3 demonstrates how the fit of these fragments to the electron density of structure 480d depends on the crystallographic resolution.

3.1. Comparison between fitted models and fully refined structures

When the coordinates of the connected peaks are compared with the atomic coordinates of the refined PDB file, the number of residues correctly positioned by the PRCTF into the density is the main criterion for the evaluation of the method. An individual peak is considered as correctly positioned if each of its six P atoms are inside the sphere defined

Table 2
Fitting of the double-helical fragments NA_RY, NA_RR and NA_YR into the electron density of RNA structures.

n_{NTs} is the number of nucleotides in the asymmetric unit of the refined structure, n_{WC} is the number of nucleotides forming Watson–Crick pairs in double helices of the refined structure, n_{NUT} is the number of nucleotides located by the procedure, %fit is $100 \times n_{NUT}/n_{WC}$, ⟨R.m.s.d.⟩ is the mean root-mean-square deviation for n_{NUT} nucleotides in angstroms, n_{DHL} is the number of peaks connected by the program *DHL*, n_{ch} is the number of connected chains, *R* is the *R* factor calculated by *REFMAC*5 for the fitted model and Correl is the correlation coefficient from *REFMAC*5 calculated for the model.

Structure	n_{NTs}	n_{WC}	n_{NUT}	%fit	⟨R.m.s.d.⟩	n_{DHL}	n_{ch}	*R* (%)	Correl (%)
480d	27	10	16	160	0.85	6	1	46	75
Eden φ1	46	40	42	105	0.80	21	3	37	85
Eden φ2	46	40	39	97	0.90	14	3	43	79
1nlc	46	40	36	90	0.49	13	3	39	84
1ehz	76	42	44	105	0.77	20	4	49	73
1dk1	57	32	40	125	0.88	17	5	41	71
1u9s	155	94	95	101	0.90	44	10	39	76
1mme	82	52	42	81	1.16	14	6	46	73
1ffk	2833	1542	1782	116	0.81	789	184	—	—
1j5e	1494	792	825	104	0.80	319	76	—	—

Table 3
Quality of the electron-density fit as a function of crystallographic resolution.

One double-helical fragment, NA_YR, was fitted into the electron density of the RNA structure 480d at various resolutions. 'd_{min} is the cutoff resolution limit, n_{max} and l_{max} are the parameters of electron-density expansion (spherical harmonics and Bessel functions) used in the input and n_{Ref} is the number of reflections. n_{DHL}, n_{ch}, n_{NUT} and ⟨R.m.s.d.⟩ are as defined in Table 2.

d_{min}	n_{max}	l_{max}	n_{Ref}	n_{DHL}	n_{ch}	n_{NUT}	⟨R.m.s.d.⟩
3.0	6	8	1349	8	2	16	0.89
3.5	6	8	852	8	2	16	0.88
4.0	6	8	576	6	2	12	0.83
4.0	8	10	576	6	1	12	0.73
4.0	12	12	576	8	2	16	0.88
4.5	6	8	400	3	1	12	0.83
4.5	8	10	400	5	1	16	0.99
4.5	12	12	400	6	2	12	0.74
4.8	6	8	328	2	1	12	0.90
4.8	8	10	328	5	1	16	0.99
4.8	12	12	328	5	2	12	0.74
5.0	6	8	292	4	2	10	0.87
5.0	8	10	292	4	2	10	0.83
5.0	12	12	292	4	2	14	1.03
5.5	12	12	220	2	1	10	0.87

by the four phosphate O atoms in the refined structure and, at the same time, the root-mean-square difference (r.m.s.d.) between the corresponding atoms of the fitted model and the refined structure is smaller than 1.7 Å. The overall accuracy of the fit is assessed primarily by the mean r.m.s.d. (⟨r.m.s.d.⟩) calculated as the average for all positioned peaks; further criteria are (i) the number of peaks connected by the *DHL* program (n_{DHL} in Table 2) and (ii) the number of formed chains (n_{ch} in Table 2).

The average r.m.s.d. is between 0.5 and 1.0 Å; this is slightly larger than the value obtained for the fitting of small protein fragments using the program *PROTF* (Pavelcik, 2004). These relatively large r.m.s.d. values reflect the inaccuracy of density fitting but also the conformation differences between the rigid search fragment and real flexible double helices. However, the typical value of the r.m.s.d. of about 0.8 Å, which is relatively independent of resolution, is better than the r.m.s.d. typically obtained for automated protein fitting (DiMaio *et al.*, 2007) and represents an objective measure of the fitting success. Comparison of the number of fitted residues and nucleotides in Watson–Crick pairs (n_{NUT} and n_{WC} in Table 2) shows that most double-helical regions were localized. Double-helical regions are even overfitted to some extent as indicated by values larger than 100% in the %fit column of Table 2. This is a consequence of the fact that Watson–Crick and several non-Watson–Crick ('mismatched') pairs are isosteric (Leontis *et al.*, 2002) and are likely to have similar electron-density envelopes. The isosteric mismatched pairs can therefore fit double-helical regions relatively well. Overfitting is possible, especially at the ends of double helices, where only half of the search fragment is fitted to the last Watson–Crick pair while the second half is fitted into electron density potentially outside the helix. Low-ranking peaks usually represent partial fits to nucleic acid structure and surrounding solvent, protein residues *etc.* and the correlation coefficients between these peaks and the crystal electron density are usually low. More reliable models could be built when the first and the last residue of the chain were deleted. A more detailed comparison of the peak properties in the centre and at the ends of fitted double-helical regions is required to eliminate some or most of the overfitting.

The fragment fitting was also verified by the program *REFMAC* (Murshudov *et al.*, 1999). The overall *R* factor calculated for the models is relatively high, around 40–45%, but the correlation coefficient shows strong correlation between the models and the 'experimental' electron densities (70–85%) as summarized in Table 2. These refinement statistics are typical for initial phases of refinement and confirm the validity of the automatic fitting.

The fit of a double helix into the experimentally phased electron density of structure eden φ2 (Ennifar & Dumas, unpublished; Table 1) is shown in Fig. 3. The highest density contour on the left side of the figure in pink shows the Se atom used for phasing. The phosphate group outside the density at the bottom of the figure illustrates problems with fitting 5′-end nucleotides that do not have phosphates.

3.2. Quality of the fit as a function of various factors

The influence of the nucleotide sequence on the fitting was tested by using only one fragment, NA_YR. The results of this fitting protocol are only slightly but consistently worse than fitting discriminating the purine and pyrimidine bases. Acceleration of the calculation by a factor of up to three does not seem to justify the loss of sequence information and we prefer fitting by three fragments.

Crystallographic resolution is not a critical parameter for the quality of the fit, as can be seen in Table 2 and especially in Table 3. Structures with higher resolution are fitted better but the difference is not significant. Table 3 summarizes fitting into the electron density of the structure 480d (Correll *et al.*, 1999) as a function of resolution. The quality of the fit does not

deteriorate dramatically to a relatively low resolution, where it is important to change the default parameters of the program *NUT* to higher 'quantum' numbers n_{max} and l_{max}. If correctly treated, fitted models are still reliable below 4 Å. Owing to the size (radius) of the double-helical models, about 11 Å, we estimate that the quality of the fitting deteriorates significantly below 4.8 Å (Table 3). The proposed method can therefore serve for initial fitting of structures of moderate to low crystallographic resolution.

The method scaled well with the size of the fitted RNA molecules. Small to medium RNA molecules with 27–155 nucleotides fitted similarly well; additionally, multiple copies of the biological assembly in the asymmetric unit, *e.g.* the two biological assemblies in the structure 1mme, did not impair the quality of the resulting model to a measurable degree. The proposed method is capable of fitting double-helical regions in large and complicated ribosomal RNA (1ffk and 1j5e). The overall accuracy of the fit to the electron density of these large structures measured by ⟨r.m.s.d.⟩ and the percentage of located residues are comparable to the other tested structures. The fitting of double-helical fragments to the density of a small ribosomal subunit, 1j5e, shows comparable results to the fitting of the much smaller structure of hammerhead ribozyme, 1mme, at the same resolution. The presence of other molecules of high molecular weight, mainly proteins (as in structure 1dk1), did not worsen the fit, probably because the double-helical fragments are too large to be fitted to the protein region. However, development of a method of fitting both nucleic acid and protein fragments into electron density is clearly a logical subsequent step, if only for the sake of the convenience of a method that would be able to build a model for protein/nucleic acid complexes in one step. A quick search with one helical fragment ignoring the nucleotide sequence can help to delineate the envelope of the nucleic acid double-helical regions and may be followed by detailed protein and nucleotide searches.

How the quality of the phases influences the fit of the double-helical fragments into electron densities was more quantitatively estimated using two different sets of phases for an unpublished structure of an RNA 23-mer labelled eden in Tables 2 and 3 (Ennifar & Dumas, unpublished work). The phases in eden $\varphi1$ are the best available phases calculated from the final refined atomic model, while the phases in eden $\varphi2$ are the initial experimental phases obtained directly by MAD phasing. The r.m.s. difference between the two phase sets is 69.8°. Evidently, the best fit is obtained with the phases calculated from the final coordinate model, eden $\varphi1$, but the initial phases from the MAD experiment generate a fit of comparable quality.

3.3. Availability

The programs *NUT* and *DHL*, the manual, examples and double-helical fragments for fitting the A- and B-forms are freely available for academic use at http://vfu-www.vfu.cz/3900/software/XFP/XFP.html.

4. Conclusions

Short double-helical fragments in the A-RNA conformation with two Watson–Crick base pairs and three phosphates as defined in Fig. 1 fit a significant portion of double helices in the nine tested X-ray RNA structures (Table 1). The presence of three phosphate groups on each strand seems to be important for the evaluation of reliable overlaps and for validation of the built model. A higher quality fit was achieved when the fitting procedure considered three purine/pyrimidine (R/Y) sequences, NA_RR, NA_RY and NA_YR, but acceptable results were obtained with only one fragment, effectively ignoring the oligonucleotide sequence. All fragments are kept rigid during the fitting procedure. This limitation is unlikely to seriously bias the resulting model because the A-RNA conformation is known to be quite rigid; only a few variants of the 'canonical' A-RNA are known and in addition they are structurally very similar (Richardson *et al.*, 2008). The double-helical fragments can be located in electron densities of resolution as low as 4.8 Å. Fitting of rigid double-helical dinucleotide fragments can be a good first step in model building and phase improvement for the refinement of RNA structures with low as well as fairly high crystallographic resolution.

There are two principal limitations of the proposed method: (i) it only fits double-helical regions and (ii) it does not fit protein fragments. Double helices are the most

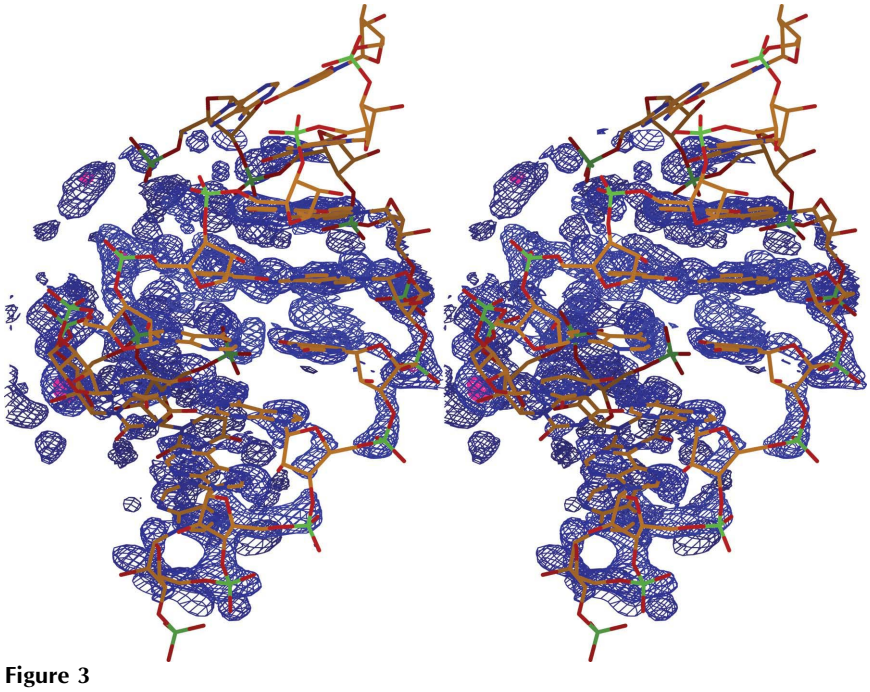


Figure 3
A stereoview of the fit of a double helix into the experimentally phased electron density of structure eden $\varphi2$ (Ennifar & Dumas, unpublished).

abundant structural units in the architecture of even the most complicated RNA molecules, but not all may be fitted by rigid double-stranded fragments in the idealized conformations of the A- and B-forms. A new procedure allowing the building of irregular and conformationally more complex loops, bulges and non-Watson–Crick double-helical segments of RNA molecules would require the use of a library of flexible single-stranded RNA fragments. Such work based on previously published RNA fragments of approximately dinucleotide length (Schneider *et al.*, 2004) is under development. Integration of the fitting of both protein and nucleic acid fragments is another direction for future development of the method.

A typical root-mean-square deviation (r.m.s.d.) between the corresponding atoms of the fitted model and the fully refined PDB structure is about 0.8 Å. This value is almost independent of crystallographic resolution and is similar to the corresponding values obtained for automated protein fitting. However, only serious testing by crystallographers will determine whether the proposed methodology for the automation of fitting of RNA fragments into electron density is useful or not.

The authors are grateful to Drs E. Ennifar and P. Dumas from Institute de Biologie Moleculaire et Cellulaire, Strasbourg, France for atomic coordinates and experimentally phased structure factors of their unpublished structure 'eden' of an RNA 23-mer. This research was supported by grants 204/06/1007 from the Grant Agency of the Czech Republic, 1/2333/05 from the VEGA grant agency of the Slovak Republic and DBI 0110076 of the NSF to the Nucleic Acid Database. BS is grateful for support by grant LC512 from the Ministry of Education of the Czech Republic.

References

Ban, N., Nissen, P., Hansen, J., Moore, P. B. & Steitz, T. A. (2000). *Science*, **289**, 905–920.
Berman, H. M. *et al.* (2002). *Acta Cryst.* D**58**, 899–907.
Berman, H. M., Olson, W. K., Beveridge, D. L., Westbrook, J., Gelbin, A., Demeny, T., Hsieh, S.-H., Srinivasan, A. R. & Schneider, B. (1992). *Biophys. J.* **63**, 751–759.
Correll, C. C., Wool, I. G. & Munishkin, A. (1999). *J. Mol. Biol.* **292**, 275–287.
DiMaio, F., Kondrashov, D. A., Bitto, E., Soni, A., Bingman, C. A., Phillips, G. N. Jr & Shavlik, J. W. (2007). *Bioinformatics*, **23**, 2851–2858.
Ennifar, E., Walter, P. & Dumas, P. (2003). *Nucleic Acids Res.* **31**, 2671–2682.
Krasilnikov, A. S., Xiao, Y., Pan, T. & Mondragon, A. (2004). *Science*, **306**, 104–107.
Lamzin, V. S. & Wilson, K. S. (1993). *Acta Cryst.* D**49**, 129–147.
Leontis, N. B., Stombaugh, J. & Westhof, E. (2002). *Nucleic Acids Res.* **30**, 3497–3531.
Levitt, D. G. (2001). *Acta Cryst.* D**57**, 1013–1019.
Murshudov, G. N., Vagin, A. A., Lebedev, A., Wilson, K. S. & Dodson, E. J. (1999). *Acta Cryst.* D**55**, 247–255.
Nikulin, A., Serganov, A., Ennifar, E., Tishchenko, S., Nevskaya, N., Shepard, W., Portier, C., Garber, M., Ehresmann, B., Ehresmann, C., Nikonov, S. & Dumas, P. (2000). *Nature Struct. Biol.* **7**, 273–277.
Oldfield, T. J. (2003). *Acta Cryst.* D**59**, 483–491.
Pavelcik, F. (2003). *Acta Cryst.* A**59**, 487–494.
Pavelcik, F. (2004). *Acta Cryst.* D**60**, 1535–1544.
Pavelcik, F. (2006). *J. Appl. Cryst.* **39**, 483–486.
Pavelcik, F., Zelinka, J. & Otwinowski, Z. (2002). *Acta Cryst.* D**58**, 275–283.
Perrakis, A., Morris, R. & Lamzin, V. S. (1999). *Nature Struct. Biol.* **6**, 458–463.
Richardson, J. S., Schneider, B., Murray, L. W., Kapral, G. J., Immormino, R. M., Headd, J. J., Richardson, D. C., Ham, D., Hershkowits, E., Williams, L. D., Keating, K. S., Pyle, A.-M., Micallef, D., Westbrook, J. D. & Berman, H. M. (2008). *RNA*, **14**, 465–481.
Schneider, B., Moravek, Z. & Berman, H. M. (2004). *Nucleic Acids Res.* **32**, 1666–1677.
Schneider, B., Neidle, S. & Berman, H. M. (1997). *Biopolymers*, **42**, 113–124.
Scott, W. G., Finch, J. T. & Klug, A. (1995). *Cell*, **81**, 991–1002.
Shi, H. & Moore, P. B. (2000). *RNA*, **6**, 1091–1105.
Terwilliger, T. C. (2001). *Acta Cryst.* D**57**, 1755–1762.
Wimberly, B. T., Brodersen, D. E., Clemons, W. M. Jr, Morgan-Warren, R. J., Carter, A. P., Vonrhein, C., Hartsch, T. & Ramakrishnan, V. (2000). *Nature (London)*, **407**, 327–339.