RNA sampling and crystallographic refinement using

\textit{Rappertk} *

Swanand Gore and Tom Blundell
\{swanand,tom\}@cryst.bioc.cam.ac.uk
Department of Biochemistry, University of Cambridge
Cambridge CB2 1GA England

Abstract

\textbf{Background} Dramatic increases in RNA structural data have made it possible to recognize its conformational preferences much better than a decade ago. This has created an opportunity to use discrete restraint-based conformational sampling for modelling RNA and automating its crystallographic refinement.

\textbf{Results} All-atom sampling of entire RNA chains, termini and loops is achieved using the Richardson RNA backbone rotamer library and an unbiased distribution for glycosidic dihedral angle. Sampling behaviour of \textit{Rappertk} on a diverse dataset of RNA chains under varying spatial restraints is benchmarked. The iterative composite crystallographic refinement protocol developed here is demonstrated to outperform CNS-only refinement on parts of tRNA\textsuperscript{Asp} structure.

\textbf{Conclusion} This work opens exciting possibilities for further work in RNA modelling and crystallography.

\*This document is very similar to a chapter in SG’s thesis submitted in Sept. 2007 to the University of Cambridge, England.
1 Introduction

1.1 Role of RNA

RNA is involved in many important biochemical functions involving genetic information, such as its storage (viral RNA), communication (mRNA) and modulation (snRNA, microRNA). RNA also performs protein-like functions like enzymatic catalysis (ribosomal peptide bond formation - tRNA) and specific binding (amino-acid-specific tRNA) etc. It is believed to have played a major role in the early evolution of cellular life because it is functionally intermediate to proteins and DNA, exhibiting enzymatic activity as well as information storage and transfer (Voet and Voet (1995)). There is an increasing recognition of RNA’s importance in cellular life (Schlick (2006)) and attempts to organize available experimental information as RNA ontology (Leontis et al. (2006)).

1.2 RNA structure

RNA is simpler than proteins in the sequence space due to a much smaller alphabet, but structurally it is more complicated. A typical nucleotide contains at least thrice as many non-hydrogen atoms as an amino acid residue. The most prominent parts of polynucleotide structures are nucleotide bases which are purines or pyrimidines. Purines adenosine and guanine are 5,6 aromatic rings and resemble tryptophan’s sidechain. Pyrimidines uracil and cytosine are aromatic 6-rings which resemble phenylalanine and tyrosine sidechains. The bases can undergo a variety of post-transcriptional modifications, increasing the effective number of base types (Dunin-Horkawicz et al. (2006)). A striking feature of DNA and (most) RNA structures is the common Watson-Crick pairing of purines with pyrimidines, and the associated base stacking. But in RNA structures, there are other non-canonical base interactions which contribute to stabilization of various RNA motifs (Leontis and Westhof (2003)). Bases are linked to 5-membered ribose sugar rings through glycosidic linkages. The \( \chi \) torsion angle, which describes base rotation with respect to sugar, is distributed around \(-120^\circ\) or diametrically opposite to it, around \(-60^\circ\) (Schneider et al. (2004)). Sugar ring connects bases to the backbone, and occurs only in two conformations, \( C3' \)-endo or \( C2' \)-endo. The phosphate-sugar backbone has six torsion angles (\( \alpha, \beta, \gamma, \delta, \epsilon, \zeta \)) and much greater freedom than the protein mainchain. But conformational correlations in that space have been recognized recently (Duarte and Pyle (1998), Murray et al. (2003), Schneider et al. (2004)).

Despite chemical differences, protein and RNA chains are logically similar. RNA backbone and protein mainchain are the unbranched chains in both polymers and show clear preferences for parts of their dihedral spaces. In proteins, mainchain completely determines \( C_\beta \) coordinate
and similarly, RNA backbone almost completely determines the sugar coordinates. Bases are similar to sidechains, because both are rotameric and confer chemical characteristics to respective polymers. Thus, RNA backbone, sugar and bases are analogous respectively to protein mainchain, $C_\beta$ atom and sidechains.

### 1.3 RNA structure prediction

Like other biopolymers, sequence data for RNA is far greater than 3D structural data. RNA crystals generally do not diffract as well as proteins because RNA is harder to purify and crystallize, possibly due to size and flexibility. Hence structure prediction methods are important to bridge the sequence-structure gap. RNA structure prediction is done at two levels - secondary and 3D. Secondary structure prediction is important because it can help identify a variety of motifs like stem, hairpin loop, internal loop, junction loop, bulges and pseudo-knots. These predictions can prove to be important restraints to guide further 3D structure prediction. 3D structure prediction is important to locate interesting sites and tertiary interactions, but it has so far been dependent on secondary structure prediction (Shapiro et al. (2007)).

Secondary structure prediction estimates the base pairings given a sequence. Due to standard Watson-Crick base-pairing, RNA commonly exhibits helical stem regions. The sequence that connects the two strands in a stem is called a loop. Stem and loop arrangement can develop in a hierarchical fashion, giving rise to a structure that can be represented like a tree. Dynamic programming based algorithms like Mfold (Zuker et al. (1999)), Sfold (Mathews et al. (1999)), RNAstructure (Mathews et al. (2004)) assign secondary structure in such a way as to minimize the free energy for the sequence. Optimal and highly-ranked suboptimal solutions are very likely to contain the correct secondary structure. Suboptimal solutions can be filtered using Boltzmann sampling (Ding et al. (2004)) or abstract shape analysis (Steffen et al. (2006)) to enrich the solutions of dynamic programming algorithms. In addition to dynamic programming, various other approaches have also been utilized such as genetic algorithms (Shapiro et al. (2001)) and Monte-Carlo sampling (Xyaphoummine et al. (2005)). All approaches can be further enhanced by using multiple sequence alignments, based on the information-theoretic principle that MSAs improve the signal to noise ratio.

Tree-like simplicity of RNA secondary structure is lost when pseudoloops are formed by base-pairing of a stretch in loop with another strand. Pseudoloops are known to occur in many more complicated ways than the simplest H-type. They reduce flexibility of the

---

1. But base rotamerility is weaker and not used in this work.
2. Free energies used here are experimentally determined as a function of host secondary structure type and base-pairing.
structure because often the stems involved in a pseudoloop are coaxially stacked. Dynamic programming algorithms which include general pseudoknots scale poorly but simple H-type pseudoknots can be incorporated without loss of efficiency (Shapiro et al. 2007).

Fully-automated 3D structure prediction procedures are yet to be devised for RNA. This is perhaps due to the complexity of RNA structure and relatively less structural information as RNA is studied more often from a non-structural perspective. Present approaches encoded in programs like ERNA-3D (Zwieb and Muller 1997), RNA2D3D (Yingling and Shapiro 2006) and S2S (Jossinet and Westhof 2005) are focused on assisting the 3D model building exercise interactively. The inputs are a combination of known/predicted secondary structure, features derived from 3D structural data and available experimental restraints. The interactively assembled model is generally subjected to molecular dynamics refinement and minimization (Shapiro et al. 2007).

Recurrent 3D motifs in RNA structure are short sequence-dependent combinations of backbone conformations and base interactions. A complex set of noncovalent interactions stabilize them. Motif identification has not matured enough to be usable in 3D structure prediction (Leontis and Westhof 2003).

1.4 RNA crystallographic refinement

RNA crystallography is harder than protein crystallography because nucleotides are bigger and more flexible than amino acid residues. RNA crystals rarely diffract better than 2Å. Due to many high-quality protein structures, their statistical preferences can be used effectively to solve more protein structures. This critical mass effect is yet to be achieved for RNA as there are not enough structures for confident identification of backbone preferences and 3D motifs. Apart from their stand-alone utility, high-quality single-chain RNA structures are also essential for docking into low-resolution EM data of large complexes containing RNA chains.

Temperature factors suggest that RNA flexibility is the least for paired bases and the highest for phosphates. Yet phosphates are also easy to detect due to greater electron density. Hence RNA crystallographer identifies bases and phosphates of RNA chain in the initial map and then iteratively completes and refines the structure. Due to lack of structural preferences, this process is manual, tedious and laborious. Methods and progress in RNA crystallography have been reviewed by Holbrook and Kim (1999) and Holbrook (2005).

1.5 This work

This work is inspired by the success of RAPPER’s protein sampling which proved effective in loop sampling, comparative modelling and automation of crystallographic refinement.
(de Bakker et al. 2003, DePristo et al. 2005, Furnham et al. 2006). It is the last task that would be very useful to the crystallographer if replicated for RNA. In protein crystallography, approximate locations of $C_\alpha$ atoms and sidechains identified by the crystallographer are sufficient for Rapper to reach an almost refined structure. It is expected that a similar approach would work for RNA chains too, given the approximate locations of phosphates and bases visually identifiable in the electron density. Apart from crystallographic use, a generalized restraint-based all-atom sampler of RNA would be useful for generating decoy structures useful for benchmarking of energy functions. It would also allow generation of models with a prescribed sequence and secondary structure, and serve as a tool for generating 3D models of RNA motifs.

In this work, we show that Rapper’s GABB (genetic algorithm using branch-and-bound technique) algorithm can be extended to RNA structures to sample it accurately and efficiently under a variety of positional restraints on backbone and bases. We also demonstrate the all-atom iterative crystallographic refinement of parts of a tRNAAsp structure.

2 RNA tracing

These benchmarks assess the utility of RNA sampling for the intended application of crystallographic refinement, hence the restraints chosen here reflect the kind of information a crystallographer can provide. Spherical positional restraints are used for phosphates ($P$ atoms) and $C4'$ atoms. Base planes are restrained using a union-of-spheres restraint (base-plane restraint). This restraint is satisfied when the sampled set of coordinates lie within the union of given spheres.

As described in Gore et al. (2007), Rappertk uses the Richardson rotamer library (Murray et al. 2003) for RNA backbone sampling. Sugar-phosphate backbone consists of six dihedral angles: $\alpha$ ($O3i'_{i-1} - Pi - O5i'_{i} - C5i'_{i}$), $\beta$ ($Pi - O5i'_{i} - C5i'_{i} - C4i'_{i}$), $\gamma$ ($O5i'_{i} - C5i'_{i} - C4i'_{i} - C3i'_{i}$), $\delta$ ($C5i'_{i} - C4i'_{i} - C3i'_{i} - O3i'_{i}$), $\epsilon$ ($C4i'_{i} - C3i'_{i} - O3i'_{i} - Pi+1$) and $\zeta$ ($C3i'_{i} - O3i'_{i} - Pi+1 - O5i'_{i+1}$). Murray et al. (2003) define the RNA backbone suite as a set of seven dihedral angles $\{\delta^i, \epsilon^i, \zeta^i, \alpha^{i+1}, \beta^{i+1}, \gamma^{i+1}, \delta^{i+1}\}$ and identify 42 distinct rotamers. In a recent effort, this library has been extended to 46 rotamers, with standard deviations specified for each cluster (J. M. Richardson, personal communication). Glycosidic dihedral $\chi$ is defined over $O4i'_{i} - C1i'_{i} - N1i'_{i} - C2i'_{i}$ for pyrimidines $O4i'_{i} - C1i'_{i} - N9i'_{i} - C4i'_{i}$ for purines. $\chi$ preferences have not been rigorously analyzed, hence in this work it is randomly sampled between $-180^\circ$ and $+180^\circ$ in steps of $10^\circ$.

The basic operation in RNA chain extension is building next or previous backbone suite by sampling a backbone suite rotamer and building sugar/base of present nucleotide using a
random sample for glycosidic linkage. Various styles of sampling use this building block in different ways.

2.1 Sampling styles

For iterative crystallographic refinement, basic operations over the RNA chain are rebuilding the whole chain, or its terminal (5’ or 3’) or an intermediate fragment (loop):

- **Forward sampling (5’ → 3’)** is performed using the default RNA builder as described in [Gore et al. (2007)](Gore et al. 2007). This builder depends on atoms \((C5'_i, C4'_i, C3'_i)\) and yields \((O3'_i, P_{i+1}, O5'_{i+1}, C5'_{i+1}, C4'_{i+1}, C3'_{i+1})\) atoms (see [Gore et al. (2007)](Gore et al. 2007) for figure). It also builds the sugar and base of \(i^{th}\) nucleotide.

- **Bootstrapping required for sampling the whole chain** is explained in [Gore et al. (2007)](Gore et al. 2007). It involves approximate positioning of \((P, O1P, O2P, O5', C5', C4', C3')\) atoms of the first nucleotide.

- **Backward sampling (3’ → 5’)** is performed by slightly changing the forward builder. The same backbone rotamers are sampled, but the builder depends on atoms \((O3'_i, C3'_i, C4'_i)\) to calculate coordinates for atoms \((C5'_i, O5'_i, P_{i-1}, O3'_{i-1}, C3'_{i-1}, C4'_{i-1})\) (see Fig.1). Sugar and base for \(i^{th}\) nucleotide are also built.

- **Loop sampling uses forward sampling.** Nucleotides between and including start and end indices are rebuilt. Base of \((\text{start} - 1)^{th}\) nucleotide is resampled within 2Å positional restraints. Approximate loop closure is achieved by partial sampling of \((\text{end} + 1)^{th}\) nucleotide’s \((P, C5', C4', C3')\) atoms under similar restraints. Loop closure restraint is back-propagated by enforcing a spherical positional restraint centered at \(P_{\text{end}+1}\) atom with radius \(7 \times (\text{end} + 1 - i)\) Å on \(P^i\) atom and also forcing it to remain 5Å away from the \(P_{\text{end}+1}\) atom.

2.2 Initial observations

For benchmarking of RNA tracing capabilities, we have used a set of diverse RNA chains compiled by [Duarte and Pyle (1998)](Duarte and Pyle 1998) for their virtual dihedral analysis (summarized in Table 1). In the first exercise, we restrained \(P, C4'\) atoms to 2Å positional restraints and sampled only the backbones of the chains. But a model could be generated for only 12 of the 48 chains. This suggested that the Richardson rotamer set consisting of only 46 states was too coarse-grained for the task being attempted. Indeed, 46 is a small number for capturing
Figure 1: Reverse RNA builder
Table 1: Dataset of RNA chains used for the tracing exercise.

| PDB id | Size\(^a\) | Filtered Size\(^b\) | PDB id | Size\(^a\) | Filtered Size\(^b\) |
|--------|-----------|----------------|--------|-----------|----------------|
| 1i6u   | 37        | 11             | 1jj2   | 121       | 18             |
| 1kh6   | 27        | 10             | 1l9a   | 125       | 15             |
| 1l2x   | 27        | 10             | 1n78   | 75        | 24             |
| 2fmt   | 77        | 17             | 361d   | 19        | 1              |
| 1mzp   | 55        | 18             | 1k8w   | 21        | 10             |
| 1duh   | 44        | 19             | 1kq2   | 6         | 0              |
| 1cx0   | 71        | 18             | 1f7y   | 56        | 9              |
| 1ec6   | 19        | 6              | 1c0a   | 77        | 37             |
| 1m5k   | 91        | 20             | 1kxk   | 69        | 11             |
| 1b7f   | 12        | 3              | 1hq1   | 48        | 35             |
| 1f1t   | 37        | 4              | 1ivs   | 75        | 6              |
| 1cvj   | 8         | 1              | 1gid   | 158       | 29             |
| 1b23   | 73        | 13             | 1qtq   | 73        | 28             |
| 1ddy   | 35        | 4              | 1lng   | 96        | 8              |
| 1e7x   | 16        | 16             | 1f27   | 18        | 11             |
| 1et4   | 35        | 9              | 1jbs   | 28        | 11             |
| 1g1x   | 39        | 14             | 1ehz   | 76        | 26             |
| 1f7u   | 75        | 23             | 1hr2   | 156       | 19             |
| 1hmh   | 34        | 11             | 1mji   | 33        | 7              |
| 1jbt   | 28        | 7              | 1ffy   | 74        | 11             |
| 1qf6   | 76        | 25             | 1e7k   | 9         | 4              |
| 1m8x   | 8         | 3              | 1ntb   | 21        | 4              |
| 1ddl   | 7         | 0              | 1ser   | 64        | 9              |
| 1h4s   | 67        | 12             | 429d   | 12        | 6              |

\(^a\)Size is the number of nucleotides in the chain as in the deposited PDB structure.

\(^b\)Filtered size is the size of the largest contiguous segment of rotameric backbone suites in the given chain. The Richardson RNA backbone rotamer set consists of 46 7-dihedral tuples, along with standard deviations for all dihedrals in each tuple. A backbone suite is rotameric if the largest single-dihedral difference between the suite and the closest Richardson backbone rotamer is < 30° or < 3σ of that dihedral angle.
preferences of a flexible backbone consisting of 6 dihedral angles. Hence it was decided to supplement sampling by perturbation - after a rotamer is sampled, a random noise within 1σ of the respective dihedrals is added to them. Standard deviations were kindly provided by J. M. Richardson (personal communication). When the latest exercise was repeated with perturbed sampling, at least one model could be generated for 47 of 48 examples. When perturbed backbone-only sampling was performed within tighter 1Å restraints on P,C4' atoms, this dropped to 36 of 48 chains. These failures could be traced to the non-rotameric backbone suites present in the chains. Then the longest stretch of good backbone suites was identified within every chain. A good suite was defined to be the one for which the largest single angle difference from the closest rotamer was within 30° or 3σ for that angle. As seen from Table 1 such good fragments are fairly small as compared to whole chain. 45 of 46 such fragments could be sampled successfully under the same restraints. On these fragments, all-atom sampling was also possible within 1Å restraints on P,C4' atoms and 5Å baseplane restraints. By dropping the C4' from these restraints, an increase in sampling time was observed, accompanied by a reduction in number of examples for which 10 models could be built (33 of 45). This is due to population dilution, which in this case is the reduction in number of members which will satisfy restraints for the base of next nucleotide. As expected, using stricter base restraint of 3Å made the matters worse due to greater base restraint violations and no propagation of base restraints onto the backbone. Base restraint used here is hard to satisfy closely because a small error close to sugar amplifies towards the far end of the planar base. This problem can be addressed if given base restraint can be propagated onto C4' atom, but it is unclear at present how to achieve this.

2.3 Sampling performance

Two characteristics are desirable in a sampling process: (a) given tight restraints, sampling should be efficient and (b) given loose restraints, sampling should produce native-like conformations owing to the knowledge of native structure incorporated in it. In other words, sampling cost should be directly proportional to length of the sampled fragment and inversely proportional to the restraint strictness. Sampling accuracy should be directly proportional to restraint radius.

To check conformity with these expected traits, we carried out backbone-only sampling of filtered fragments under positional restraint of 1, 2 and 3Å on phosphorous atom. Note that fragments may be the entire chains or at either terminus of the RNA chain or in between, hence this also tests corresponding sampling styles. All-atom sampling exercises were carried out under the same restraints on P atom and baseplane restraint of 5Å on bases. 10 modelling attempts were made in each sampling exercise. A modelling attempt fails if it cannot produce
a model in 5 trials. Each trial uses backtracking, i.e. if sampling fails at a nucleotide, it is restarted from a position 3 nucleotides before it in the sampling order. In all-atom sampling, glycosidic linkage (χ dihedral) is sampled uniformly over the entire range at 10° intervals. van der Waals radii of base and sugar atoms are reduced by 50%. Sampling performance is quantified by measuring the average RMSD of models and average time taken to produce a model as functions of fragment size, restraint radius and whether bases are modelled.

The time plots (Fig.2) suggest a linear correlation between fragment size and sampling time for both backbone-only and all-atom models, hence lines of best fit have also been plotted. Regression coefficients of these lines are informative. In both cases, regression coefficients suggest that the time needed for sampling with P restraint radius of 2Å is twice as much as that with 3Å and four times as much with 1Å as with 2Å. Comparison of the regression coefficients in backbone and all-atom cases suggest that latter is nearly ten times costlier than the former.

The RMSD plots (Fig.3) suggest a weak correlation between the RMSD and fragment sizes, i.e. RMSD is lower for smaller fragments with the same restraint radius. This prompted the fitting of a log curve. RMSD falls with the size of P restraint. For each restraint size, all-atom RMSD is more than backbone RMSD. Interestingly, the backbone RMSD in all-atom case is better than that in backbone-only case, indicating the influence of base restraint in guiding the backbone.

Thus, the main sampling trends are: (a) smaller restraint radius leads to greater sampling time, (b) all-atom sampling is costlier than backbone-only, but leads to backbones with less RMSD (c) sampling time is proportional to fragment size and (d) RMSD tends to be smaller for smaller fragments. These trends are expected from previous experience with protein sampling exercises. But there are significant differences too, due to differences in restraint density. In protein Cα tracing, backbone sampling models 3N atoms under N positional restraints (ignoring carbonyl oxygen), hence positional restraint density is 1/3. In the RNA backbone tracing exercise carried out here, this density is 1/6 (ignoring phosphate oxygens O1P, O2P). This is reflected in the backbone RMSD: in the case of proteins, backbone RMSD is generally lower than the Cα restraint radius but it is generally higher for RNA than the P restraint radius. Another difference is rotamericity of protein sidechains and lack of it in glycosidic linkages. This is indicated by lower all-atom RMSD for proteins than RNA chains under similar restraints. To sum up, trends observed in RNA sampling are expected and satisfactory enough to attempt application to the crystallographic scenario.
Figure 2: Variation in sampling time with RNA fragment size. The following scatter plots (above backbone-only, below all-atom) indicate a linear correlation between average sampling times and fragment sizes. Regression coefficients of lines of best fit suggest that sampling time nearly doubles from 3Å to 2Å and almost quadruples from 2Å to 1Å. Similarly, all-atom sampling is roughly a magnitude costlier than backbone-only case. Note that 3 outliers have not been considered for the 1Å backbone-only plot and 4 fragments did not yield any model during 1Å all-atom sampling.
Figure 3: Variation in sampling accuracy with RNA fragment size. The following plots (top-1Å, middle-2Å, bottom-3Å) show relationship between average RMSD of models of fragments and fragment lengths. There is a weak tendency to have lower RMSD for lower lengths, hence a log curve was fitted for each scatter. In general, at a given $P$ restraint radius, all-atom models have better backbone RMSD than backbone-only models. All-atom RMSD is slightly greater than backbone RMSD in all-atom models. RMSD increases as restraint radius increases.
3 Foray into crystallographic refinement

3.1 About tRNA structure

Transfer RNAs are classic structures from the 1970s. Till mid-90s, structures of tRNA \cite{Hingerty1978, Sussman1978, Westhof1988, Westhof1986} were the only large RNA structures in PDB \cite{Shi2000}, making them remarkable achievements of crystallography techniques of that decade. tRNA is a cloverleaf-shaped molecule in its secondary structure representation and has a L-shaped 3D form (Fig. 4). tRNA is an essential cog in the translational machinery of the cell which incrementally translates the transcripted mRNA into peptide chain one residue at a time. Ribosome finds a tRNA with a 3-nucleotide anticodon complementary to current mRNA codon. This tRNA has an amino acid attached to its 5’ end, which the ribosome then attaches to the growing polypeptide.

tRNA structures are attractive for demonstrating crystallographic utility of discrete restraint-based RNA sampling because they are neither too small nor too large, are structurally well-studied and have 3 loop regions (anticodon loop, $T\psi C$ loop, D loop) with non-Watson-Crick base pairing. For this work, tRNA$^{Asp}$ structure was used, solved at 3Å by \cite{Westhof1988}. This structure (PDB 2tra) refines to $R/R_{free}$ of 0.2552/0.3063 with CNS starting from deposited structure and data.

3.2 Composite refinement protocol

Similar to composite refinement protocols used earlier in the thesis, this work also uses perturbed starting structures and rebuilds them with the aim of improving $R_{free}$. In brief, RAPPERTk identifies the ill-fit nucleotides by calculating the correlation coefficient between $F_c$ map and $\sigma_A$-weighted CNS omit map for regions around the backbone, sugar/base and entire nucleotide. Low ($<0.9$) correlation coefficient indicates nucleotide stretches to rebuild, which are then built incrementally using GABB algorithm. Ten times more children are generated as the population size, and top 10% are retained based on their electron density occupation score, leading to an enriched population. Resampled nucleotides get a $B$-factor of 30 assigned to all of their atoms. Non-RNA atoms (ligands and waters) are not used during sampling. Best member of population (according to density occupation) is written out as the new model along with non-RNA atoms appended to it. The coordinates and $B$-factors of non-RNA atoms are copied from the previous refinement iteration. This model is refined with CNS (2 rounds of MDSA starting at 5000K, intervened by a 200-step minimization). This procedure is repeated for 10 iterations. It is expected that RNA models generated with rotameric backbone states
Figure 4: tRNA structure: the schematic diagram shows the typical secondary structure of tRNA. 3D representation below it shows all-atom and cartoon representation of tRNA$^{Asp}$ as in PDB entry 2tra (Westhof et al. (1988)).
to obey given positional restraints, positive electron density restraints and excluded volume restraints would be within the convergence radius of CNS, i.e. such models can be used to assist CNS in finding well-refined structures, starting from ill-fitting ones.

### 3.3 Refining a helical fragment

CNS refinement was performed initially on the anticodon loop (nucleotides 33 – 37) and the $T\psi C$ loop (nucleotides 54 – 60), starting from models where the loops were perturbed by RNA tracing within tight positional restraints ($P$, baseplane restraints of 2Å, 5Å respectively). In both cases we observed that CNS was able to correct the errors introduced in the native structure. This was in contrast to proteins where similar positional restraints on $C_\alpha$ and sidechains generally result in unsatisfactory CNS-only refinement. But removal of baseplane restraints from the RNA trace deteriorated the refinement quality. This suggested that CNS convergence radius is larger for RNA structures than proteins, and Rappertk sampling may be of value only in cases where spatial information about the structure is highly uncertain.

In order to use a simple example to begin with, a fragment in RNA duplex (nucleotides 23 – 27) was chosen, with clear base densities. Initial perturbation was carried out with 2Å $P$ restraints and no base restraints to generate 5 models. The perturbed models were subjected to CNS refinement only. In 4 of 5 cases, CNS refinement was unsatisfactory. 3 such cases are shown in Fig.5. When the composite refinement protocol was applied to the same region with the same starting models, all trajectories resulted in well-refined structures (Fig.6). The mean of best $R_{free}$ values in CNS-only trajectories was 0.311 as compared to 0.304 for the composite protocol trajectories. It is interesting to note that $R_{free}$ does not strongly reflect the salient differences in the refinement trajectories indicated by Fig.5 and Fig.6.

### 3.4 Refining the $T\psi C$ loop

The same exercise was repeated for nucleotides 54 – 60, the $T\psi C$ loop. The native density for this loop is not as good as the helical fragment (see Fig.7). CNS-only refinement resulted in mean best $R_{free}$ of 0.316 over the 5 refinement attempts, whereas the same for composite refinement was 0.303. Visual inspection of these models shows the greater variability in the CNS models and that each attempt was stuck in a local minimum. 3 of 5 composite models refined to a structure very similar to native, but the rest were trapped in a local minima. Close observation of these 2 cases revealed that spurious density appearing elsewhere led Rappertk sampling away from the native.
Figure 5: CNS-only refinement of tRNA\textsuperscript{Asp} (PDB 2tra) can be unsatisfactory. Starting models were generated by perturbing a 5 nucleotide fragment (23-27) with 2Å P restraints and no base restraints. Top-left panel shows the native structure of the fragment with its omit map contoured at 2σ. Other 3 panels show the best $R_{\text{free}}$ structures in 3 different CNS-only refinement trajectories, with respective omit maps also contoured at 2σ. This suggests that CNS can get trapped in local minima in case of high initial structural uncertainty. Note that the CNS refinement here is with minimal restraints, i.e. hydrogen-bonding restraints between base-pairs were not provided to CNS.
Figure 6: Composite CNS/Rappertk refinement of a helical fragment from tRNA^{Asp}. Spherical positional restraints of radius 2Å were used around $P$ atoms of the 5-nucleotide (23-27) fragment from PDB 2tra. No restraints were imposed on bases. The CNS/Rappertk refinement resulted in satisfactory refinement in all 5 attempts. Best $R_{free}$ models in each trajectory are shown in magenta, with the deposited structure in sticks representation.
Figure 7: Composite cns/Rappertk refinement of the 7-nucleotide $T\psi C$ loop (54-60) from tRNA$_{Asp}$ with 2Å $P$ restraints and no base restraints. The top panel shows the native fragment with its omit map density. Middle panel shows the best $R_{free}$ models of cns-only refinement. Bottom panel shows the same for the composite refinement. Native structure (green) is shown for reference in middle and bottom panels.
3.5 Refining the anticodon loop

Anticodon loop spans nucleotides 33 - 37, of which G - 34, U - 35 and C - 36 have two equally occupied states in the 2tra structure. Initial attempts to repeat the previous exercise on this loop were unsatisfactory because Rappertk tried to fit a single conformation to these heterogenous nucleotides. Due to this, we created artificial diffraction data at the same resolution by considering only the first conformation of each nucleotide and assigning full occupancy to it. This significantly changed the refinement trajectories and a similar trend as previous two exercises could be observed. For five CNS-only and composite refinements, mean best $R_{free}$ values were 0.254 and 0.215 respectively, indicating a much improved refinement with the composite protocol. Fig.8 shows that the composite protocol yields almost identical structures and CNS-only refinement gets trapped in different local minima.

3.6 Typical problems with refinement protocols

There are two main reasons for suboptimal CNS-only refinement, identifiable from successive structures in the refinement trajectories (Fig.9). Firstly, if a base is very far away from its native-like location, CNS refinement does not restore it. Secondly, a base may get trapped into densities of phosphate, sugar or another base, in which case even if the base is not too far away, it is difficult to restore it. This is reminiscent of bulky misplaced sidechains in protein crystallographic refinement.

Structure trajectories suggest that improved refinement with CNS/Rappertk protocol must be due to relocation of bases by RNA sampling, which is brought about by the electron-density based enrichment of incremental building using rotameric backbones. A typical corrective rebuilding step is shown in Fig.10. The obvious mistakes in base placement are corrected with Rappertk whereas CNS carries out small corrections to take the conformations towards the optimal.

There are some imperfections in the Rappertk sampling scheme which may sometimes lead to incorrect final structures: (a) RNA is very flexible and population size of 300 and enrichment factor of 10 may not be sufficient (b) Lack of $\chi$ preferences means that selective pressure due to bases is low - it is further weakened in case of weak base density (c) Collateral damage may be caused by CNS refinement of a defective loop, e.g. perturbations in nearby regions of structure or symmetry-related copies do not get repaired during Rappertk rebuilding step leading to higher $R_{free}$ (d) Scoring scheme based on maximizing the electron density occupation may promote occupation of sharp peaks like waters and phosphates although there are obvious dissimilarities between such peaks and the shape of a base.
Figure 8: Composite cns/Rappertk refinement of the 5-nucleotide anticodon loop (33-37) with 2Å P restraints and no base restraints. The top panel shows the native fragment with its omit map density. Middle panel shows the best $R_{free}$ models of cns-only refinement. Bottom panel shows the same for the composite refinement. Native structure (green) is shown for reference in middle and bottom panels.
Figure 9: CNS refinement of RNA can get trapped in local minima. An instance of CNS-only refinement yields a structure (green) very different from native (magenta). Corresponding $2F_o - F_c$ omit map is shown contoured at 1\(\sigma\) around a three nucleotide stretch (green sticks, nucleotides 54-56) only for clarity. Corresponding nucleotides in the two structures are shown with arrows. CNS model has bases too far away from native locations and also occupying the wrong density. Also note the appearance of density around wrong bases.
Figure 10: A typical corrective step carried out by Rappertk RNA sampling with rotameric backbone and density enrichment. Blue is the native structure, green is a perturbed and CNS-refined model. Magenta is the Rappertk model found using the positional restraints and omit map of the green model. Note that Rappertk model removes gross errors yet small errors may remain in comparison to native.
4 Conclusion

This work suggests that knowledge-based sampling can be applied efficiently and productively to RNA structures. GABB algorithm was extended to sampling of RNA chains at 5'-end, 3'-end and intermediate regions. Modified nucleotides were incorporated in addition to standard ones for all-atom sampling. Using a 48-chains dataset, we showed that sampling performance is along expected lines and suggests its suitability for real-world applications like crystallography. Then we demonstrated that a helical strand, the TψC loop and the anticodon loop in the tRNA_{Asp} structure can be automatically sampled and iteratively refined using crystallographic data. It was found that the composite CNS/Rappertk protocol yields structures better refined than those by the CNS-only protocol. Shortcomings of both protocols were discussed. This work shows that automated crystallographic refinement of RNA chains is possible given the approximate trajectory of phosphates. This is a promising result for reducing manual effort and allowing exploration of multiple conformations.

Yet some concerns remain and must be addressed in future work. Sampling preferences themselves are imperfect. A-form conformation is adopted by more than 50% RNA suites but population frequencies for rest of the backbone rotamers are unclear. Hence we have used equal weights for all suite rotamers. For similar reasons, we have not used the weakly bimodal nature of the glycosidic linkage. A careful analysis of available structural data will be required before incorporating such preferences reliably, because sampling preferences are meant to bias the conformational search and not restrict it. Another improvement necessary for quicker sampling is the propagation of phosphate and base restraint onto the backbone (e.g. on C4') so that base restraint satisfaction becomes more likely. At present this is a sampling bottleneck.

There are a few promising ways to extend this work. Firstly, whole-chain crystallographic refinement of RNA structures can be performed for low resolution structures to reduce the number of non-rotameric suites. Secondly, RNA sampling can be used to generate 3D all-atom conformations for secondary structures or motifs by expressing the base pairing/stacking interactions as distance restraints. All conformations sampled to satisfy these restraints will be useful in 3D structure prediction which is, as noted before, a process of assembling 3D coordinates of predicted secondary structures. Finally, protein and RNA sampling can be combined together for automating the crystallography of protein-RNA complexes, especially the very large ones like ribosomes so that human attention will be required only in the early and late stages of refinement.
References

de Bakker, P.I.W., DePristo, M.A., Burke, D.F. and Blundell, T.L. (2003) Ab Initio Construction of Polypeptide Fragments: Accuracy of Loop Decoy Discrimination by an All-Atom Statistical Potential and the AMBER Force Field With the Generalized Born Solvation Model. Proteins: Struct., Func., and Genet., 51, 21–40.

DePristo, M.A., de Bakker, P.I., Johnson, R.J. and Blundell, T.L. (2005) Crystallographic refinement by knowledge-based exploration of complex energy landscapes. Structure, 13 (9), 1311–1319.

Ding, Y., Chan, C.Y. and Lawrence, C.E. (2004) Sfold web server for statistical folding and rational design of nucleic acids. Nucleic Acids Res, 32, W135–W141.

Duarte, C.M. and Pyle, A.M. (1998) Stepping through an RNA structure: A novel approach to conformational analysis. J Mol Biol, 284, 1465–1478.

Dunin-Horkawicz, S., Czerwoniec, A., Gajda, M.J., Feder, M., Grosjean, H. and Bujnicki, J.M. (2006) MODOMICS: a database of RNA modification pathways. Nucleic Acids Res, 34, D145–D149.

Furnham, N., Dore, A.S., Chirgadze, D.Y., de Bakker, P.I.W., Depristo, M. and Blundell, T.L. (2006) Knowledge-Based Real-Space Exporations for Low-Resolution Structure Determination. Structure, 14 (8), 1313–1320.

Gore, S.P., Karmali, A.M. and Blundell, T.L. (2007) Rappertk: a versatile engine for discrete restraint-based conformational sampling of macromolecules. BMC Structural Biology, 7:13, doi:10.1186/1472-6807-7-13.

Hingert, B., Brown, R.S. and Jack, A. (1978) Further refinement of the structure of yeast tRNA. J Mol Biol, 124, 523–534.

Holbrook, S. (2005) RNA structure: the long and the short of it. Curr Op Struct Biol, 15, 302–308.

Holbrook, S.R. and Kim, S.H. (1999) RNA crystallography. Biopolymers, 44, 3–21.

Jossinet and Westhof, E. (2005) Sequence to Structure (S2S): display, manipulate and interconnect RNA data from sequence to structure. Bioinformatics, 21, 3320–3321.

Leontis, N.B., Altman, R.B., Berman, H.M., Brenner, S.E., Brown, J.W., Engelke, D.R., Harvey, S.C., Holbrook, S.R., Jossinet, F., Lewis, S.E., Major, F., Mathews, D.H., Richardson, J.,
Williamson, J.R. and Westhof, E. (2006) The RNA Ontology Consortium: An open invitation to the RNA community. *RNA*, 12, 533–541.

Leontis, N.B. and Westhof, E. (2003) Analysis of RNA motifs. *Curr Op Struct Biol*, 13, 300–308.

Mathews, D., Sabina, J., Zuker, M. and Turner, H. (1999) Expanded Sequence Dependence of Thermodynamic Parameters Provides Robust Prediction of RNA Secondary Structure. *J Mol Biol*, 288, 911–940.

Mathews, D.H., Disney, M.D., Childs, J.L., Schroeder, S.J., Zuker, M. and Turner, D.H. (2004) Incorporating chemical modification constraints into a dynamic programming algorithm for prediction of RNA secondary structure. *Proc Nat Acad Sci*, 101, 7287–7292.

Murray, L.J.W., III, W.B.A., Richardson, D.C. and Richardson, J.S. (2003) RNA backbone is rotameric. *PNAS*, 100, 13904–13909.

Schlick, T. (2006) RNA: The Cousin Left Behind Becomes a Star. In *Computational Studies of DNA and RNA*, (Spomer, J. and Lankas, F., eds),, Springer Verlag, Dordrecht, The Netherlands pp. 259–281.

Schneider, B., Moravek, Z. and Berman, H.M. (2004) RNA conformational classes. *Nucleic Acids Res*, 32, 1666–1677.

Shapiro, B.A., Wu, J.C., Bengali, D. and Potts, M.J. (2001) The massively parallel genetic algorithm for RNA folding: MIMD implementation and population variation. *Bioinformatics*, 17, 137–148.

Shapiro, B.A., Yingling, Y.G., Kasprzak, W. and Bindewald, E. (2007) Bridging the gap in RNA structure prediction. *Curr Op Struct Biol*, 17, 157–165.

Shi, H. and Moore, P.B. (2000) The crystal structure of yeast phenylalanine tRNA at 1.93 Å resolution: A classic structure revisited. *RNA*, 6, 1091–1105.

Steffen, P., Vob, B., Rehmsmeier, M., Reeder, J. and Giegerich, R. (2006) RNAshapes: an integrated RNA analysis package based on abstract shapes. *Bioinformatics*, 22, 500–503.

Sussman, J.L., Holbrook, S.R., Warrant, R.W., Church, G.M. and Kim, S.H. (1978) Crystal structure of yeast phenylalanine transfer RNA I. Crystallographic refinement. *J Mol Biol*, 123, 607–630.

Voet, D. and Voet, J. (1995) *Biochemistry*. John Wiley & Sons, Inc.
Westhof, E., Dumas, P. and Moras, D. (1988) Restrained refinement of two crystalline forms of yeast aspartic acid and phenylalanine transfer RNA crystals. *Acta Cryst.*, **44**, 112–123.

Westhof, E. and Sundaralingam, M. (1986) Restrained refinement of the monoclinic form of yeast phenylalanine transfer RNA: Temperature factors and dynamics, coordinated waters, and base-pair propeller twist angles. *Biochemistry*, **25**, 4868–4878.

Xayaphoummine, A., Bucher, T. and Isambert, H. (2005) Kinefold web server for RNA/DNA folding path and structure prediction including pseudoknots and knots. *Nucleic Acid Res.*, **33**, 605–610.

Yingling, Y. and Shapiro, B. (2006) The prediction of the wild-type telomerase RNA pseudoknot structure and the pivotal role of the bulge in its formation. *J Mol Graph Model*, **25**, 261–274.

Zuker, M., Mathews, D. and Turner, D. (1999) Algorithms and Thermodynamics for RNA Secondary Structure Prediction. In *A Practical Guide in RNA Biochemistry and Biotechnology*, (Barciszewski, J. and Clark, B., eds.), NATO ASI Series, Kluwer Academic Publishers.

Zwieb, C. and Muller, F. (1997) Three-dimensional comparative modeling of RNA. *Nucleic Acids Symp Ser*, **36**, 69–71.