Description of the model and predictions of Φ-values for RNA

Assignment of coarse-grained structural model and energy parameters for basepairing, base-stacking, and hydrophobic interactions

We have two possibilities to calculate the energy of a RNA structure.

In the first case we have adapted the coarse-grained RNA model elaborated in Dokholyan's group [1] for description of RNA structure and energy parameters. To simplify calculations, those authors considered a full-scale atomic RNA model in which three beads correspond to each nucleotide. Beads P and S are positioned at the center of the mass of the corresponding phosphate group and the five-atom ring sugar, the base bead (B) is positioned at the center of the six-atom ring for both purines and pyrimidines (Fig. 1).

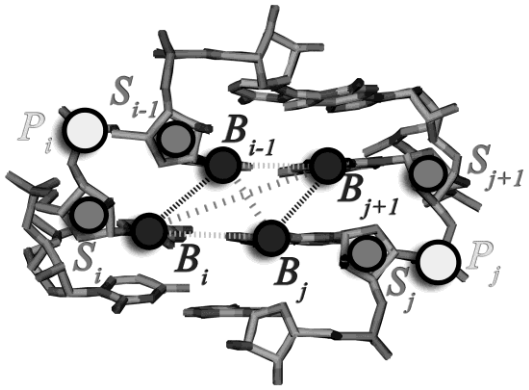


Fig. 1. Coarse-grained RNA structural model. Beads in the RNA: sugar (S), phosphate (P), and base (B). Distances vary depending on the type of the nucleic base in the nucleotide. Hydrogen bonds upon interactions of the bases and pairing contacts are shown between bases B_{i-1}:B_{j+1} and B_i:B_j.

The nonbonded interactions are crucial to model the process of RNA folding. In our adapted model [1], we included the base pairing (only A–U, G–C, and U–G pairs are involved in hydrogen bonding), base-stacking, and hydrophobic interactions. The basic energies of the hydrogen-bonding interactions are ε_{HB} = -0.5 for A–U, ε_{HB} = -1.2 for G–C, and ε_{HB} = -0.5 kcal/mol for U–G, respectively. If the distance between bases B_i and B_j is within the limits of d_{min}-d_{max}, then the hydrogen bond energy is calculated. The hydrogen bond energy depends on three distances (see Table 1) between bases and sugars: B_iB_j, S_iB_j, and B_iS_j (Fig. 1). The distances between S_iB_j and S_jB_i, define the orientations between the two nucleotides. If the distances satisfy the predetermined range, we allow the hydrogen bond to be formed, and forbid its formation otherwise. If all these distances are within the limits of $d_1 < R < d_{max}$, then coefficient 3 is given to the hydrogen bond energy ($3\varepsilon_{HB}$). In the case of further reduction of the distance (i.e. in the case of the interval $d_0 < R < d_1$), $0.5\varepsilon_{HB}$ (kcal/mol) is added to each pair (B_iB_j, or S_iB_j, or B_iS_j). Thus, when all three distances fall within $d_{min} < R < d_0$, then E_{HB} = 0 kcal/mol. When a branched hydrogen bond is formed, the energy value is divided by two.

In our model the energy of base-stacking and hydrophobic interactions is considered as follows: if two bases are at distance $r < 4.65$ Å for purines, $r < 4.60$ Å for pyrimidines, and $r < 3.8$ Å for purine–pyrimidines as well as for all modified bases, then $E_{\text{Stack}} = -0.6$ kcal/mol; if base pairs are closer to each other than 6.5 Å but no stacking is formed, then the energy of hydrophobic interactions $E_{\text{Hydrophobic}} = -0.4$ kcal/mol is attributed to them. We have calculated average free energy values for these interactions from the work [1]. The considered parameters for all possible neighboring base pairs have been determined from the experimentally tabulated energy [2]. It should be noted that for noncanonical pairs, the energy of stacking and hydrophobic interactions was also considered.

In the second approach the contacts in a RNA structure are calculated in a similar way as in a protein case. We use all three parameters for interactions between atoms: heavy atoms (non-hydrogen) interacting with heavy atoms, heavy atoms interacting with hydrogen atoms and hydrogen atoms interacting with hydrogen atoms. However, to describe stacking and close hydrophobic interactions in RNA structure we calculate interactions between neighbor nucleotides, in opposite to the protein case, where we omit interactions between closest residues in a chain sequence. We also not include any interaction with phosphate groups (atoms with conventional PDB names P, OP1, OP2) because they usually interact with water in folded and unfolded states.

Network of folding/unfolding pathways and the point of thermodynamic equilibrium

Why do we investigate RNA (protein) unfolding rather than folding? Simulation of unfolding is simpler than that of folding (since one can avoid exploring numerous high-energy dead-ends), while, according to the detailed balance principle [3], the pathways for folding and unfolding must coincide when both processes take place under the same conditions. Hence, we are interested in conditions close to that of thermodynamic equilibrium between the native and the coil states. Calculation algorithms are already elaborated for prediction of protein unfolding pathways [4].

The spatial structure of RNA in the native state is taken from the database of PDB or NDB. The RNA folding/unfolding process is modeled as reversible unfolding of its native structure by the dynamic programming technique [4]. We consider the network of unfolding pathways in which each pathway is a simplified virtual consecutive RNA unfolding (Fig. 2), i.e. the artificial exclusion of one or another nucleotide from all interactions within the molecule. The removed nucleotide gains the unfolded state entropy with the exception of the entropy spent to close disordered loops protruding from the remaining structure. It is assumed that the other nucleotides keep their native positions and that the unfolded regions do not fold into another, non-native structure. To use dynamic programming in searching for transition state(s) at a large

network of folding-unfolding pathways, we have to restrict this network by ~10^7 intermediates. Therefore we consider only the intermediates with no more than two closed loops in the middle of the strand plus the N- and C-terminal disordered tails. To the same end we use "strand links" consisting of a few nucleotides: of two for RNA with less than 80 nucleotide residues, and of four (or three) for larger RNAs.

Estimation of free energy and calculation of folding nuclei

The process of consecutive folding/unfolding of the native structure of a nucleotide strand consisting of U nucleotide links is shown in Fig. 2. This strand has a completely folded native state S_0, completely unfolded state S_U, and multiple intermediate partially unfolded structures S_v including v disordered links and the native-like globular part of $U - v$ links ($v = 0$ for native state S_O, $v = U$ for completely unfolded state S_U, $v = 1, \ldots, U - 1$ for partially unfolded structures).

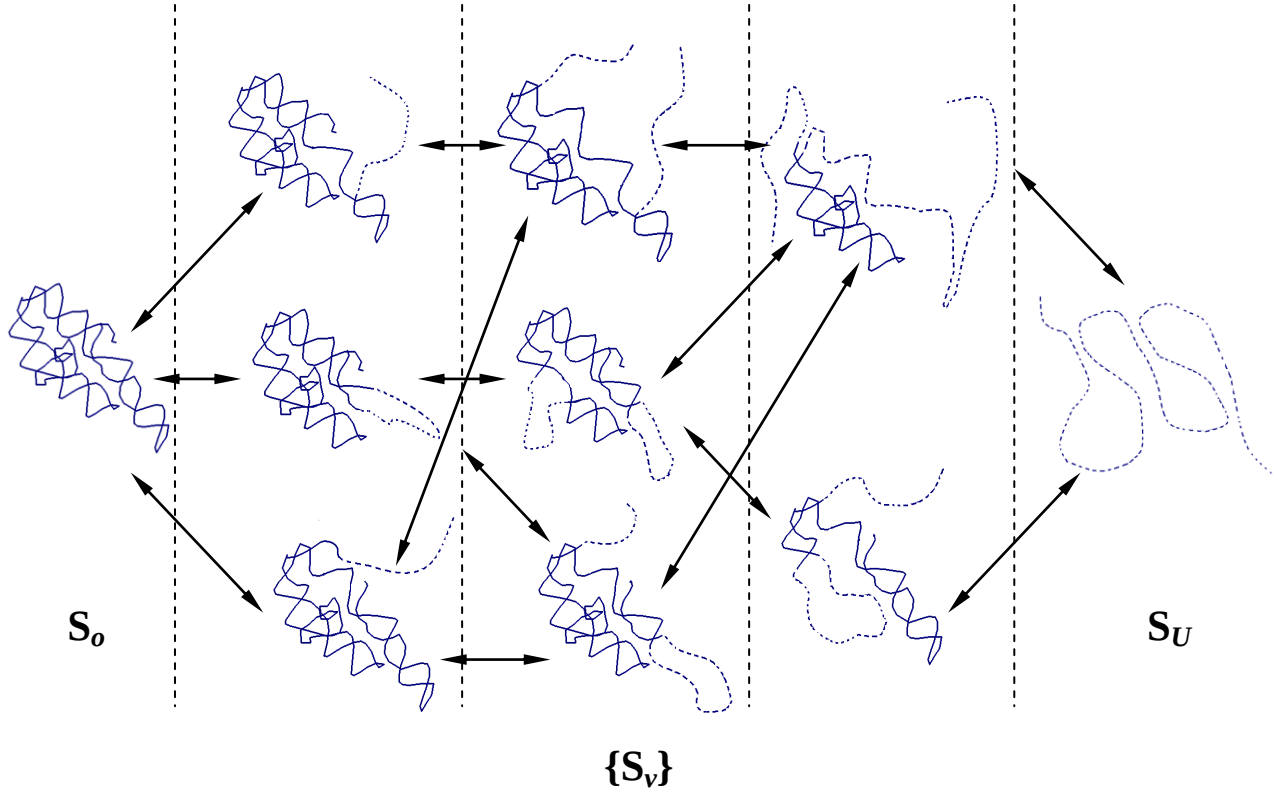


Fig. 2. Scheme of folding and unfolding pathways in native spatial structure S_0. S_U is a completely unfolded state in which all U of nucleotide strand links are unfolded (this figure shows the structure of domain P4-P6 from the *Tetrahymena thermophila* ribozyme first group intron). In each partially unfolded structure (type S_v), v links are unfolded (dotted line), while the other $U - v$ links retain their native position and conformation (continuous line). Vertical dotted lines separate microstates with a different number v of unfolded links in the chain. The central structure in the bottom row represents the microstate with v unfolded links forming one closed disordered loop and one unfolded tail; the central structure in the central row is the microstate in which v unfolded links form two closed disordered loops. The pathway networks used in calculations are much more extensive than in this scheme: they include millions of partially unfolded microstates.

All free energy calculations given in this work relate to the point of thermodynamic equilibrium between native structure S_0 and random coil S_U. The free energy of an intermediate state of an RNA molecule is calculated using the equation:

$$F(S_v) = E_{sum}(S_v) - RT[\sigma N_{free.nucl} + S_{loop}]. \quad (1)$$

Total energy $E_{sum}(S_v)$ is taken from all nucleotides of RNA structure S_v and is calculated as the sum of energies of base-pairing (energies of hydrogen bond, E_{HB}), base-stacking (E_{Stack}), and hydrophobic interactions ($E_{Hydrophobic}$) of each of nucleotides described by the coarse-grained model:

$$E_{sum}(S_v) = E_{HB} + E_{Stack} + E_{Hydrophobic}. \quad (2)$$

The main designations are as follows: T is the temperature in Kelvin (350 K); R is the universal gas constant; σ is the difference in entropy upon transition of one nucleotide residue from the unfolded to the structured part of the molecule in R units; $N_{free.nucl}$ is the number of nucleotide residues in unfolded part of the molecule; S_{loop} is the Flory entropy (cost for locking the loops leaving the globule between residues k and l). The loop entropy is calculated using the formula:

$$S_{loop} = -\frac{5}{2} R \ln|k - l| \quad (3)$$

We have shown in the course of protein structure modeling that the term responsible for the persistent length does not make a large contribution upon calculation of the loop entropy (persistent length 20 Å) [4]. The persistent length for RNA molecules is ~10-20 Å [5], and therefore we ignore it in this work.

Special attention should be given to calculation of σ, the entropy difference between random coil and native states of a nucleotide residue that can be calculated if the RNA structure is at the point of thermodynamic equilibrium between native and random coil phases $F(S_0) = F(S_U)$, i.e. $E_{sum}(S_0)$ and σ obey the ratio $E_{sum}(S_0) = -RTN_{all.nucl}\sigma$, where $N_{all.nucl}$ is the number of nucleotides in the native RNA structure.

A complete analysis of pathways passing through these "semi-unfolded" structures is carried out using the dynamic programming technique [4].

The value of ratio $\Phi = \Delta\Delta F_{\#-U}/\Delta\Delta F_{N-U}$ is the measure of the involvement of the amino acid residue in the transition state structure formation [6,7]. The $\Delta\Delta F_{N-U}$ value is the difference of free energies between the folded and unfolded states of the wild type RNA and mutant, $\Delta\Delta F_{\#-U}$ is the difference between free energies of transient and denatured states. If $\Phi = 1$, then contacts that define native state at the moment of the transition state have been already formed; this means that this residue is incorporated into the folding nucleus. If $\Phi = 0$, then these contacts

evolve at the last moment of protein folding, after overcoming the free energy barrier. It is very difficult to interpret intermediate Φ values because they depend on many factors. Such values may show both that these contacts at the moment of the transient state were formed partially, and that weak interactions between pairs could be formed or not at the moment of the transient state.

Φ-values for a definite nucleotide (Φ_n) are calculated using the formula:

$$\Phi_n = \frac{\sum_{S^{\#} \in TS} \Delta_n E(S^{\#}) \cdot P(S^{\#})}{\Delta_n E_N}, \quad (4)$$

where summation is carried out using the ensemble of transition states constructed by the dynamic programming technique upon construction of a complete folding/unfolding network, $\Delta_n E(S^{\#})$ is the change in the energy of interactions upon removal of the assigned nucleotide (n) in transition state $S^{\#}$, the words "nucleotide removal" mean exclusion of the latter from all interactions (this is similar to a particular amino acid residue replacement by glycine in proteins [1]; $\Delta_n E_N$ is the change in the interaction energy in the native state in response to the removal of nucleotide n. It is supposed that in the unfolded state the nucleotides form no contacts, i.e. they are not involved in any interaction.

To average the values in a set of transient states ($S^{\#}$), Boltzmann weights are used:

$$P(S^{\#}) = \frac{\exp(-F^{\#}(S^{\#})/RT)}{\sum_{S^{\#} \in TS} \exp(-F^{\#}(S\#)/RT)}, \quad (5)$$

where ($S^{\#}$) is the transient state from a set of all structures in this state. These Φ–values have the same sense as the Φ_f values derived from the protein/RNA engineering experiments. They are compared to see the correlation of the theory and experiment.

One of the advantages of our method is that it allows us to do prediction of the folding nucleus for nontrivial RNA motifs, such as pseudoknots and tRNAs if their spatial structures are in the PDB or NDB databases (see Table 2).

Table 3 presents the Φ-value profiles for four structures determined in the free unbound states: $tRNA^{Phe}$ (NDB file: 1EHZ), $tRNA^{Lys}$ (1FIR), $tRNA^{fMet}$ (3CW5), and $tRNA^{Asp}$ (3TRA). One can see that these profiles look similar: nucleotides corresponding to the D and T loops have the lowest Φ-values compared to different regions of the tRNA molecule, whereas the anticodon helix has the highest Φ-values, which correlates with the experimental data [8,9]. The Φ-value profile for $tRNA^{Lys}$ is slightly different, but this is due to poor resolution for this structure (3.3 Å). We obtained values for each component of interaction energy, number of hydrogen bonds, numbers of pairs of stacking and hydrophobic interactions for the four considered structures (Table 3).

REFERENCES

1. Ding, F., Sharma, S., Chalasani, P., Demidov, V.V., Broude, N.E., and Dokholyan, N.V. (2008) Ab initio RNA folding by discrete molecular dynamics: from structure prediction to folding mechanisms. *RNA*, **14**, 1164-1173.

2. Mathews, D.H., Burkard, M.E., Freier, S.M., Wyatt, J.R., Turner, D.H. (1999) Predicting oligonucleotide affinity to nucleic acid targets. *RNA*, **5**, 1458-1469.

3. Landsberg, P.T. (1971) *Problems in Thermodynamics and Statistical Physics*. PION, London.

4. Galzitskaya, O.V., and Finkelstein, A.V. (1999) A theoretical search for folding/unfolding nuclei in three-dimensional protein structures. *Proc. Natl. Acad. Sci. USA*, **96**, 11299-11304.

5. Caliskan, G., Hyeon, C., Perez-Salas, U., Briber, R.M., Woodson, S.A., Thirumalai, D. (2005) Persistence length changes dramatically as RNA folds. *Phys. Rev. Lett*., **95**, 268303.

6. Matouschek, J.T., Kellis, Jr., Serrano, L., and Fersht, A.R. (1989) Mapping the transition state and pathway of protein folding by protein engineering. *Nature*, **340**, 122–126.

7. Matouschek, J.T., Kellis, Jr., Serrano, L., Bycroft, M., and Fersht, A.R. (1990) Transient folding intermediates characterized by protein engineering. *Nature*, **346**, 440–445.

8. Maglott, E.J., Goodwin, J.T., and Glick, G.D. (1999) Probing the structure of an RNA tertiary unfolding transition state. *J. Am. Chem. Soc.*, **121**, 7461-7462.

9. Pereyaslavets, L.B., Baranov, M.V., Leonova, E.I., Galzitskaya, O.V. (2011) Prediction of folding nuclei in tRNA molecules. *Biochemistry (Mosc)*, **76**, 236-244.

Table 1. Distances between bases (C, G, U, A) and sugars (S)

Nucleotide pair and its components	d_{min}	d_0	d_1	d_{max}
C_i G_j	5.20 Å	5.46 Å	5.62 Å	5.74 Å
S_i G_j	7.70 Å	8.08 Å	8.63 Å	9.00 Å
C_i S_j	9.74 Å	9.74 Å	10.53 Å	10.82 Å
A_i U_j	5.00 Å	5.25 Å	5.68 Å	5.84 Å
S_i U_j	9.76 Å	9.94 Å	10.50 Å	10.76 Å
A_i S_j	7.72 Å	7.92 Å	8.82 Å	9.00 Å
U_i G_j	5.10 Å	5.65 Å	6.10 Å	6.25 Å
S_i G_j	7.00 Å	7.44 Å	8.24 Å	8.70 Å
U_i S_j	9.50 Å	10.25 Å	10.80 Å	11.35 Å

Table 2. Profiles of Φ-values of the pseudoknots and hairpins. Yellow color corresponds to nucleotides with high probability to be in the folding nucleus.

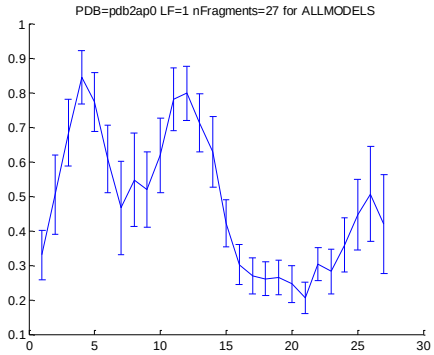
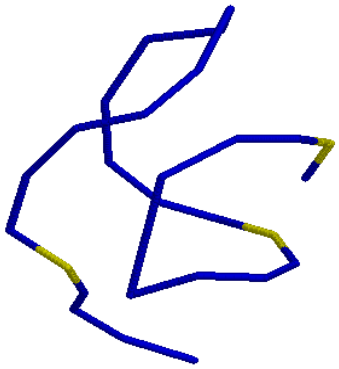
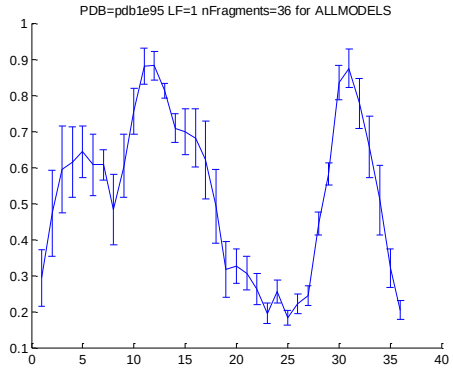
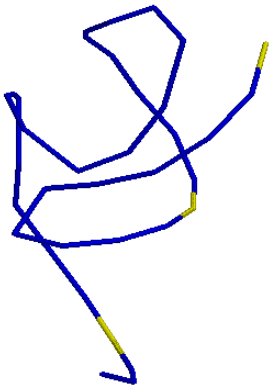
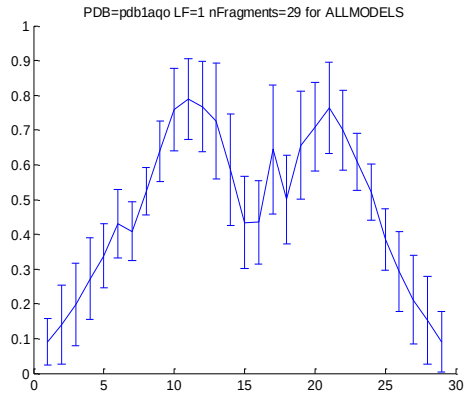
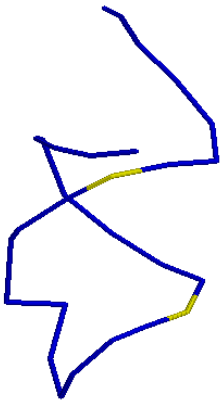
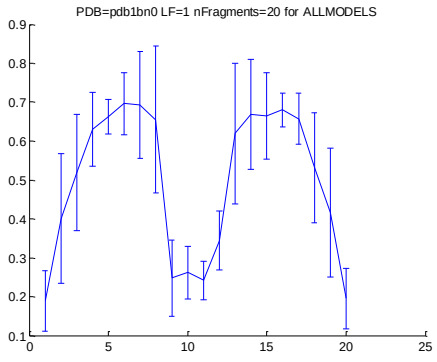
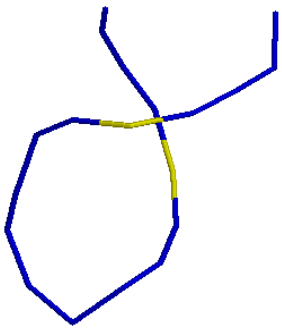
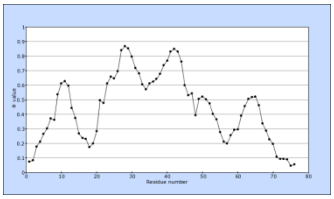
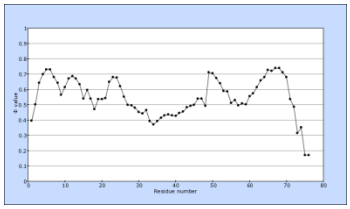
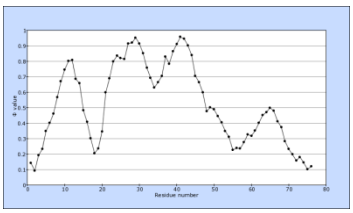
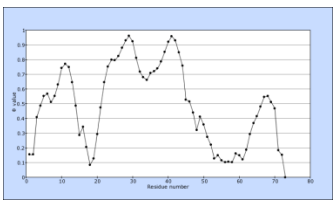
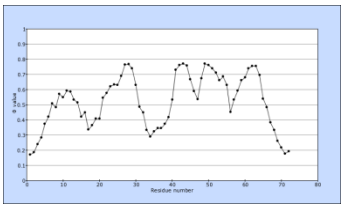
PDB entry	Name	Profile Φ-values	3D structure of molecule
2ap0 (NMR)	C27A SUGARCANE YELLOW LEAF VIRUS RNA PSEUDOKNOT		
1e95 (NMR)	SOLUTION STRUCTURE OF THE PSEUDOKNOT OF SRV-1 RNA, INVOLVED IN RIBOSOMAL FRAMESHIFTING		
1aqo (NMR)	IRON RESPONSIVE ELEMENT RNA HAIRPIN		
1bn0 (NMR)	SL3 HAIRPIN FROM THE PACKAGING SIGNAL OF HIV-1		

Table 3. Calculated energy characteristics of tRNA molecules by Dokholyan's method

PDB code (resolution), Name and origin	Agreement with experiment	Energy components (kcal/mol)				Number of interactions		
		Complete energy of molecule	*Hydrogen bonds*	*Stacking interactions*	*Hydrophobic interactions*	*Number of hydrogen bonds*	*Number of stacking interactions*	*Number of hydrophobic interactions*
1EHZ (1.93 Å) Yeast tRNA^{Phe}	 Yes	-127.2	-31.4	-59.4	-36.4	22	99	90
1FIR (3.3 Å) Bovine tRNA^{Lys}	 No	-116.62	-21.82	-60.0	-34.8	20	100	86
3CW5 (3.1 Å) *E. Coli* tRNA^{fMet}	 Yes	-116.75	-17.75	-67.8	-31.2	19	113	77
3L0U (3 Å) *E. Coli* tRNA^{Phe}	 Yes	-116.1	-29.9	-58.2	-28.0	22	97	69
3TRA (3 Å) Yeast tRNA^{Asp}	 Yes	-116.43	-25.43	-57.0	-34.0	23	95	84