Supplementary Information

1 Fragments, FARFAR motifs and junctions.

The non-redundant set of structures was taken from the BGSU RNA Site (http://rna.bgsu.edu/rna3dhub/nrlist), version 1.89, with a resolution of 3.0 Å or better.

The loops were identified as the internal and hairpin loops obtained with the RNA 3D Motif Atlas (rna.bgsu.edu/rna3dhub/motifs), using the version from April 4th, 2017. In this respect, the subindex was used internally to identify the loop from other ones belonging to the same structure. Table 1 contains the name of the loop and the number of mobile nucleotides. Table 2 contains the same information for the junctions.

2 Interaction parameters

In this section, we describe in detail the methods and parameters used for the construction of the tabulated interactions which define our model. The probability distributions were used for defining the interactions as expressed in Eqs. (3) and (4). They were defined by collecting a set of three-dimensional points from the structural database. In this version of the code, the Jacobian is considered as a constant. The points were analyzed with Mathematica 9

Structure	Mobile nts	Motif
$157D_1$	6	IL_157D_001
$1D4R_1$	8	IL_1D4R_001
$1JJ2_1$	15	IL_1JJ2_004
$1LNT_1$	12	IL_1LNT_002
$1Q9A_1$	6	HL_1Q9A_001
$1U9S_1$	7	HL_1U9S_001
$2GDI_1$	7	HL_2GDI_001
$2OIU_1$	14	IL_2OIU_004
$2R8S_1$	9	IL_2R8S_002
$2R8S_3$	13	IL_2R8S_003

Table 1: Details of FARFAR motifs.

Structure	Mobile nts	Name
$1GID_2$	14	J3_1GID_001
$2QBZ_3$	17	J3_2QBZ_001
$3R4F_2$	10	J3_3R4F_001
$4P8Z_2$	13	J3_4P8Z_001
$4P9R_2$	13	J3_4P9R_001

Table 2: Details of junctions.

using the `SmoothKernelDistribution` function, and evaluating it on a three-dimensional grid with an spacing of 0.2 Å. All the clouds obtained so were then truncated to zero when their value was smaller than a given threshold value. This value depends on the interaction, and in certain cases, it has been slightly adjusted in order to prevent the formation of isolated probability domains.

The strength of each interaction is defined as the minimum of the interaction energy of a certain kind between two nucleotides, and shifting it by its corresponding value of ϵ as defined in Eq. (2) in the paper, which is positive by construction. This yields the minimum energy of the potential well. In the case of the base-pairing interactions, we have taken the minimum value from all the possible base-pair combinations (including different faces and species) and assigned a value of $\epsilon = 0$. From here, we can determine the minimum value of the equivalent energy of a hydrogen bond, which yields 27.5 T_0. Stacking interaction and non-bonded base-phosphate strengths are normalized according to this value.

2.1 Stacking

The stacking interactions can be bonded or non-bonded, that is, between nucleotides which are neighbors in the same strand or any other case. In the first case, it can be distinguished between purines and pyrimidines, and whether the faces confronted are 3' and 5'. In all the cases, the threshold was of 0.001, with the exception of the case of two 5' faces interacting, on which the threshold was of 0.005. The points obtained were also filtered, considering only those whose z-coordinate lies in the range 2Å, 4.5Å or -5Å, -2 Å.

For the non-bonded case, due to the worse quality of the statistics, this interactions discerns over the face of each base, without treating the combinations separately. In all the cases, the threshold was of 0.01. Also, for the stacking interaction to be formed, it is required that the normal vectors to be aligned by an angle of 23^o or less. The points obtained were also filtered, considering only those whose z-coordinate lies in the range 2Å, 4.5Å or -5.6Å, -2 Å, with a distance in the x-y plane of 3.5 or less.

The strength of the stacking interactions was estimated as described in the main text. In more detail, we performed a series of tests on the GCAA tetraloop and the 255D duplex, to ensure their stability and obtain a first guess of the

stacking strength. Later on, we performed simulated annealing over a large set of internal and hairpin loops of a large ribosomal unit (pdb id: 1S72, hairpin loops (HL) 5, 7, 18, 24, 25, 41, 43, 45, 54 and 55; internal loops(IL) 1, 2, 7, 15, 16, 20, 21, 25, 26, 28, 32, 39, 41, 46, 57, 59, 60, 63, 65, 66, 69, 71, 72, 75-77, 79, 83, 84, 86, 90, 92-95, 98, and 100 according to FR3D, February 2017). These simulations were done by constraining the glycosidic bond angle and sugar puckers to the values found in the native structures. The result of this is a large set of structures (20 initial conditions per fragment) which are local energy minima. The idea is, from this set of decoys, to find the optimal parameters of stacking that will make that the structure with the largest INF^{st} will have the lowest energy of the set. In addition, we made some assumptions with regard to the strength of stacking between different species, and also made sure that at least the GCAA tetraloop and the test duplex were still folded. Therefore, we generated a table of different sets of ϵ^{st} values, using four parameters that we searched in a grid space: ΔRR, ΔRY, ΔYY and Δ_{NB}. These parameters are shifts that we apply on the set of stacking strengths, and they are specific for purines (RR), purines and pyrimidines (RY), and pyrimidines (YY), while there is also a shift which relates the stacking between bonded nucleotides with the non-bonded ones. After the optimal set of Δ shifts was found, we performed a similar approach for obtaining a guess of ϵ^{xp}. In this case, however, we also performed annealing simulations of the same set of fragments where all the sugar puckers and glycosidic bond angles were set to C3' endo and *anti*. By adjusting the values of ϵ^{xp}, searching manually over a grid of parameters, we were able to obtain a set of values which allows to favor energetically the native structures over the rest, or the structure with the lowest RMSD measured with respect to the native one.

This procedure could be repeated iteratively, and over a larger set of structures. However, given the statistical nature of our interactions, we preferred to stay with this first approach to avoid overfitting the parameters.

2.2 Backbone

In this case, the dataset is filtered using MolProbity, and removing all the points which do not have a suiteness index larger than 0.5. Later on, the pucker is classified according to the same software, and the glycosidic bond angle is classified as *anti* ($\chi < -120^o$ or $\chi > 155^o$), *high anti* ($120^o < \chi < -10^o$) or *syn* (purines : $35^o < \chi < 145^o$, pyrimidines $40^o < \chi < 145^o$). The interaction of the base with the phosphate group belonging to the same nucleotide had a threshold of 0.0002, and values ten times larger for *syn* conformations and the *anti* conformation of purines. The interaction of the base with its neighboring phosphate group had a threshold value of 0.0002, which was increased to 0.008 for the cases on which the statistics were poorer (less than 20 points).

The backbone interactions were multiplied by a prefactor as explained in the main text. The prefactors for the base-phosphate and angle interactions are of 13 and 3, respectively.

The sugar-phosphate-sugar angle is also treated according to the puckers in-

volved. Fig. 1 shows the distribution of this angle θ from the sampled database. The figure depicts the unrefined case and when both sugars have the same pucker, that is, C3' endo or C2' endo. The peaks of the distribution, for the C3' endo case matches nicely the angle found in an A-form of RNA. For the C2' pucker distribution, the maximum is observed at a larger angle and the distribution is softer. In a B-form of DNA, this angle is also larger than the one found in the A-form, but smaller than the distribution peak by 10^o.

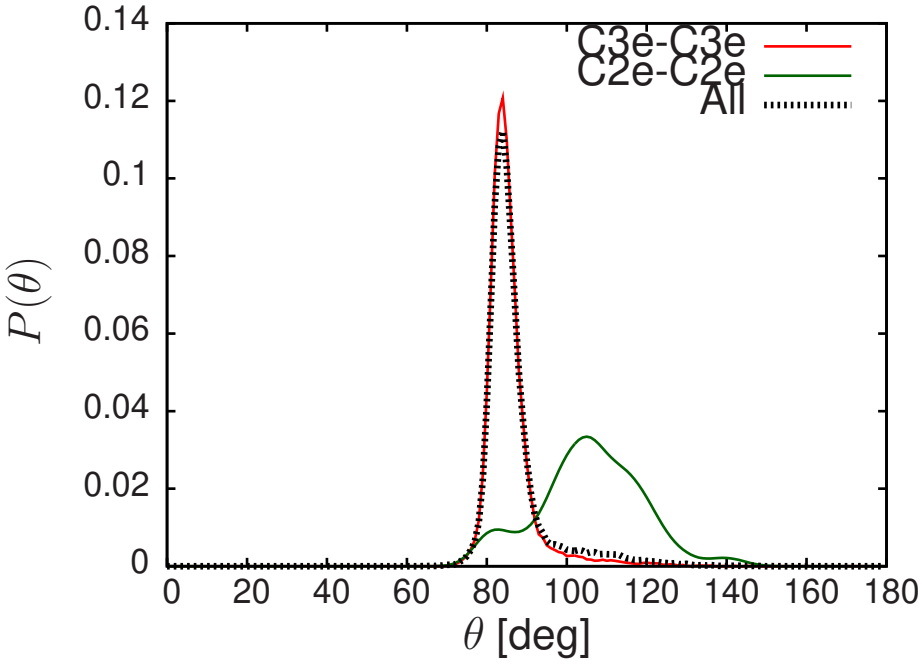

Figure 1: S-P-S angle distribution for two consecutive nucleotides. The cases shown correspond to two sugars in the C3' endo conformation, two sugars in the C2' endo conformation and the distribution without any refinement, on which the C3' endo conformation shades its counterpart.

2.3 Base-pairing

In this case, the default threshold was of 0.0007. However, in cases on which the volume of the cloud was too small, this value was tuned in such a way that a minimum volume of 10 $Å^3$ was reached.

For some species along certain particular faces, one can find more than one cloud of points. In order to be consistent with the annotation parameters, they must be treated in an exclusive manner. Fig. 2 exemplifies this along the cWW pairing of adenine-adenine. If one cloud is chosen for one base, the other

must be chosen for the paired base. Our code takes this into consideration, by sub-spanning the space for adenine-adenine (cSS, cWW, tSS), adenine-cytosine (cSS, cWW), adenine-guanine (cSS), cytosine-cytosine (cWW), cytosine-uracil (cWW) , guanine-guanine(cSS) and uracil-uracil(cWW) pairs.

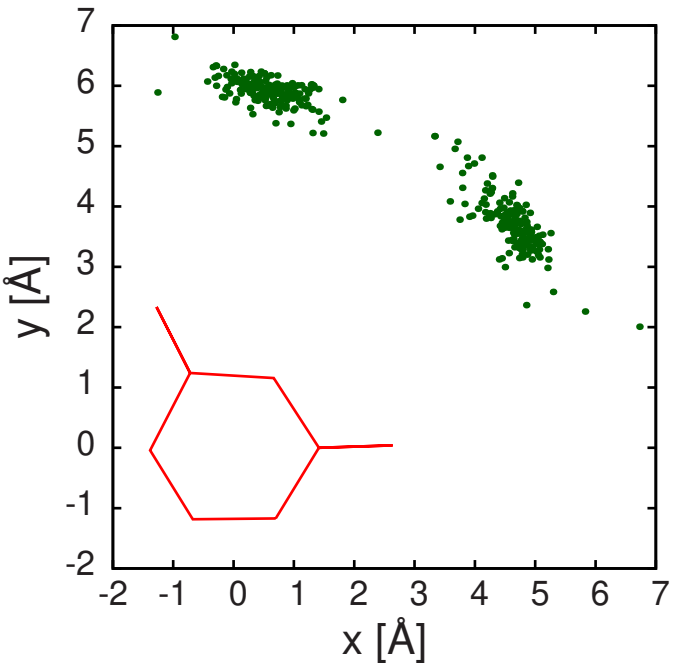

Figure 2: Cloud of cWW points collected from the database, for uracil-uracil (Uu and uU according to FR3D), projected on the x-y plane. The cloud has two domains well defined.

The base-base interaction is also exclusive on each face. That is, given two bases A and B, they will be paired through their corresponding faces only if B is the candidate to pair with A with the lowest energy, and viceversa. In this manner, no ambiguity nor excess of pairs is found.

It is also noted that from the 288 different combinations, some pairs are extremely scarce. Therefore, for the cases on which less than 10 ocurrences were found, we replaced them by more populated clouds with the smallest centroid distance between their points. This choice is quite reasonable, except in the cases on which the replacing cloud is formed by more than one domains. In such a case, the following cloud with the smallest centroid distance (usually with a lower RMSD with respect to the original), was used as the replacement. This approach was needed for 32 pairs, which are listed in Table 3.

It was also noted that the restriction of the interaction dependent on the dihedral defined by the z vector of each base and the vector that joins their center

Original	Replacement	$D_{centroids}$(Å)	RMSD(Å)
ACpSW	ACaSW	1.3	1.6
ACpWS	CAaWS	2.1	2.4
AGaSW	AApSW	1.4	1.6
AGpHS	UUpHW	1.1	1.6
AUpSH	AAaSH	0.6	1.2
AGaWS	CAaWS	1.9	2
AGaWH	GUpWS	0.7	1
AGaHH	UGaHS	0.3	1.1
AGpHW	AUaHW	0.6	1
AUpWS	AAaWS	2.4	2.5
AGpWH	UApWH	0.6	1.2
UApHS	UUpHW	3.6	3.6
UCaWS	GGaWH	3.5	3.6
UCpHH	ACpHS	2.2	2.3
UCpHS	UUpHS	2.6	2.9
UCpWS	GUpWS	0.3	0.8
UGpWW	GUaWW	0.6	0.9
UUaSS	CAaSS	2.8	3.5
UCpWW	CCpWW	0.5	0.8
GAaSW	GAaSH	0.4	1.3
GApSH	UUpSH	3.3	3.7
GAaSW	GAaSH	0.4	1.3
GUaHW	GGaHW	1.6	1.7
GAaWS	GCaWW	0.6	0.8
CApWS	CAaWS	0.9	1.3
CCaSS	CAaSW	2.4	2.8
CCpWH	UAaWH	0.8	1.3
CUaSS	CAaSS	2.1	2.3
CUaSW	AUaSH	1.5	1.8
CUpSH	GUaSH	1	1.3
CUpSW	AUaSH	0.6	1.1
CUpWW	CGaWW	0.4	0.7

Table 3: Pairs with less than ten points and their replacement. Nomenclature is XYqMN, where X and Y stand for the type of nucleotides and q and MN for the orientation and faces of the pair in the Leontis-Westhof nomenclature.

improved the results. Therefore, the pairing only takes place when such an angle lies between the minimum and maximum values observed in the structure database, in a specific manner for each kind of pair.

2.4 Base-phosphate

The non-bonded base-phosphate interaction must be treated with care. Due to the lack of internal structure of the phosphate group, it is very easy to produce false positive pairs. Therefore, we have callibrated it, resulting in energetic terms for these interactions corresponding to 0.58 times the energy of hydrogen bond in a base pair. This allows the formation of stems and duplexes without the necessity of a constraint.

BPh0 interactions are not included in the energy function. The interactions 7BP, 8BP and 9BP interact through the Hoogsteen face in cytosine, while 3BP, 4BP and 5BP appear over the Watson-Crick face in guanine. Still, there is a large overlap among the clouds of these interactions. Considering that the 4BP and 8BP are composed of two hydrogen bonds, we have treated them in a specific manner, and distinguish them inside a single face although their energy corresponds for the moment to one hydrogen bond, for simplicity.

A way of refining the base-phosphate interaction is by backmapping, in real time, the OP1 and OP2 atoms from the CG representation and check if it is possible to form a hydrogen bond with the base. The position of the oxygens can be defined with respect to the position of the phosphate group and the two neighboring sugar groups, as the average obtained over the base-phosphate pairs observed in the non-redundant set of structures used for the determination of the interactions. In the same way, the contact points in a base can be easily determined in the plane defined by the CG base, allowing to define virtually the position of a hydrogen and its donor group. When the angle and the distance between these components is suitable, then the base-phosphate interaction is formed. The values of the distance and angle between donor-hydrogen-acceptor were taken from Zirbel et al. [1] and later reduced manually in order reduce the formation of spurious interactions in the GCAA tetraloop.

2.5 Excluded Volume

The excluded volume interactions are treated in the same geometry as the base pairs. Between bases A and B, a cloud of points is built with all the bases that are separate at a distance smaller than a cutoff radius of 7Å. Later on, an ellipsoid is fitted through a Monte Carlo procedure to maximize the empty, interior volume of the cloud. For sugars and phosphates, the procedure is analogous.

3 Base flip

The *high-anti* conformation is practically reachable from the *anti* conformation, so we do not employ any trial move for this change. Nevertheless, when a

Nucleobase type	A3	H3	S3	A2	H2	S2
R	0	0.18	0.36	0.1	0.18	0.47
Y	0	0.36	-	0.36	0.44	-

Table 4: Values of $\epsilon^{\chi p}$ with respect to the hydrogen bond energy.

nucleobase is in this conformation, it has an additional contribution to its energy, which is also compensated by the different potentials that govern its backbone terms. In case the nucleotide is in a conformation which is not unambiguously defined between *anti* and *high anti*, the conformation is given by the one that has the lowest energy.

However, for changing to *syn* conformation and between different sugar puckers, the nucleotides must experience a successful Monte Carlo trial move. For this aim, we rotate and displace the base, and remap the sugar after this. The rotation matrix and displacement vector are obtained in the following way: We first consider a reference frame on the center of a base, from where we can estimate the most probable position of the neighboring phosphate groups by minimizing the energy between the base and them. This will give two points at each side of the plane defined by the base. Later on, we relate the two points belonging to a conformation (for example, C3' endo and *anti*) to the points of another conformation (for example, C3' endo and *syn*) by multiplying their positions by a rotation matrix and adding a displacement vector. In general, there is no such a matrix which can map directly two points of a conformation onto another arbitrary pair. Therefore, we parametrize our matrix and displacement and find the optimal values wich minimize the distance between the transformed coordinates of the phosphates in C3' endo, *anti* and the original coordinates of C3' endo, *syn*. For changing the state of the base, we apply the inverse of this operation on the base. Note that this move is reversible and has a unitary Jacobian.

The values of $\epsilon^{\chi p}$ are calculated with respect to the most common conformation, which is C3' endo for the sugar pucker and *anti* for the glycosidic bond angle. Due to their abundance, this conformation is the most favorable energetically, which is reflected in the value of its minimum, the lowest in comparison with the minimum of the rest. Considering the sum of energy of the base with its bonded phosphates, we obtain the minimum of this for each combination of χp conformations. With this information, we can add the $\epsilon^{\chi p}$ correspondingly. The values obtained after the parametrization, described briefly at the end of Section 2.1 are listed in Table 4, denoting the χp conformation by a letter and a number : the glycosidic bond angle conformation is denoted by A, H and S for *anti*, *high-anti* and *syn*, and the number is 3 or 2 for the sugar pucker C3' endo and C2' endo respectively.

Sequence	Structure	Nucleotides [chain|residue]	CG-RMSD[Å]
GAAA	1FJG	A|156 - A|165	1.5
GAGA	1FJG	A|294 - A|303	1.1
GCGA	1S72	0|574 - 0|583	1.5
GGAA	1FJG	A|1513 - A|1522	1.6
GGGA	1S72	0|2246 - 0|2255	0.9
GUAA	1U9S	A|97 - A|106	1.1
GUGA	1S72	A|1074 - A|1083	1.2

Table 5: Detail of PDB index and the nucleotides considered in the tetraloop fragments annealed, GNRA family.

4 Tetraloop structures

The UUCG tetraloop structure was taken from the PDB 2KOC, while the CUUG tetraloop was taken from the PDB 2L6I. Additionally, we tried all the possible sequences which formed a tetraloop structure with the sequence GNRA. For this, we identified the corresponding sequences in larger structures, when there was a reasonable long stem that kept it stable. We obtained thus a 10 nucleotide structure. The warmup and annealing procedure was done keeping frozen the last two pairs of the stem.

For GNRA, we tried the sequences GAAA, GAGA, GCGA, GGAA, GGGA, GUAA and GUGA. Table 5 displays the specific sequence of residues and the original structure from which the motif was extracted.

5 Figures from Results section

This section contains the figures of the motifs from the fragment from the structure 2GDI (Figs. 3 and 4). In this case, the base colored in black has a sugar pucker C2' endo. The auxiliar particles of each base are colored: X in red and Y in yellow, which indicate the orientation of the base. The sugar conformation allows the black nucleotide to form a U-U pair as in the native structure.

References

[1] Zirbel, C. L., Šponer, J. E., Šponer, J., Stombaugh, J., and Leontis, N. B. (2009) Classification and energetics of the base-phosphate interactions in RNA. *Nucl. Acids Res.*, **37**, 4898–4918.

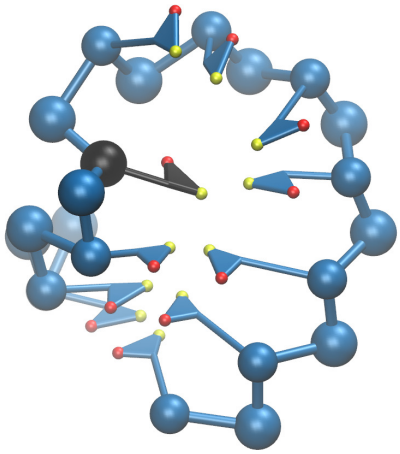

Figure 3: 2GDI after annealing with SPQR

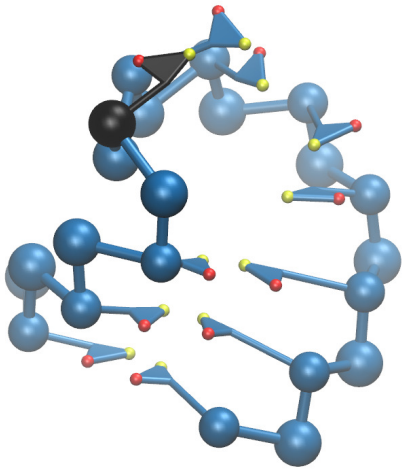

Figure 4: 2GDI after annealing with SPQR and χp constraint.