The MC-Fold | MC-Sym pipeline infers RNA structure from sequence data

Marc Parisien and François Major

Institute for Research in Immunology and Cancer (IRIC), Department of Computer Science and Operations Research, Université de Montréal, PO Box 6128, Downtown Station, Montréal, Québec H3C 3J7 CANADA

Methods

RNA-Select. Using a simple pair-wise Smith-Waterman sequence comparison, we grouped together the RNA 3-D structures that have similar sequences. The most recently solved structure for each group was selected to form RNA-Select (Tab. S2)

NCM database. The NCM database contains lone-pair loops up to six nucleotides (including the flanking lone base pair; see Fig. S1 "output1") and double-stranded NCMs up to eight nucleotides (including both flanking base pairs). For lone-pair loops, we use the syntax "L-<sequence>", where L is the length of the loop and <sequence> is the sequence. Therefore, the NCM database contains 4 types and 5440 different lone-pair loop NCMs: 64 3-loops (3-AAA, 3-AAC, ... 3-UUU); 256 4-loops (4-AAAA, 4-AAAC, ... 4-UUUU); 1024 5-loops; and, 4096 6-loops. For double-stranded NCMs, we use the syntax "L1_L2-<sequence>", where L1 is the length of the 5'-strand, L2 is the length of the 3'-strand, and <sequence> is the sequence. Therefore, the NCM-database contains 15 types and 407808 different double-stranded NCMs. The 2_2-<sequence> NCMs represent the 256 base pairing tandems: 2_2-AAAA, 2_2-AAAC, ... 2_2-UUUU. The 3_2-<sequence> represents 1024 5'-strand single-nucleotide bulges, and the 2_3-<sequence> the 1024 3'-strand single-nucleotide bulges. Similarly, the 4_2-<sequence> represents 4096 5'-strand double-nucleotide bulges, and so on; 2_4 (4096 NCMs), 5_2 and 2_5 (2 × 16384 = 32768 NCMs), 6_2 and 2_6 (2 × 65536 = 131072 NCMs), 3_3 (4096 NCMs), 3_4 and 4_3 (2 × 16384 = 32768 NCMs), 3_5 and 5_3 (2 × 65536 = 131072 NCMs), and 4_4 (65536 NCMs). Because there are so many NCMs, the database is built in a just in time fashion, i.e. instances of the NCMs are built as the MC-Fold | MC-Sym pipeline needs them.

NCM building. First, we build a database of RNA backbone templates for each NCM: the phosphate groups, riboses, and glycosidic bonds. These correspond to each of the 19 NCM types.

Second, we build a database of all possible base pairs: nucleobases and glycosidic bonds. Third, we align the four atoms of the glycosidic bonds of the base pairs with those of the backbone templates. A fit is found if the RMSD measured on the anchor points are within a user-defined precision in Å. Typically, we use values from 0.1 to 1.0 Å (for this study, we used 0.3 for the lone-pair loop and double-stranded NCMs).

MC-Fold structure enumeration. To generate the possible hairpins of a sequence, we first determine a list of initiation sites, which can be assigned lone-pair NCMs. Then, recursively, we match the rest of the sequence to double-stranded NCMs (see Fig. S10). Since we consider all possible positions for the initiation sites (even those of more than 6 nucleotides), this assignment process is in O(N^2), where N is the length of the sequence. For each possible hairpin loop, we must find an assignment of approximately N/2 NCMs for the rest of the sequence. Since we have 15 double-stranded NCM types, this process is exponential, in O($15^{N/2}$). This algorithm enumerates all possible NCM construction exhaustively. The various incompatibilities amongst NCM junctions limit the number of actual constructions, explaining why this algorithm works in practice (see Fig. S11).

For multi-branched structures, we use 4 indices: *i, j, k*, and *l, i* < *j* < *k* < *l*. We build stem-loops where the lone-pair of the hairpin is located at (*j, k*), and the last base pair in the stem at (*i, l*). We store them in a hyper-cube [(*i, j*) (*k, l*)]. We keep one (the best energy) stem-loop for each position, E[(*i, j*) (*k, l*)]. The time for filling the hyper-cube stays the same as described above, and the process results in a database of stem-loops, which we sort by the *i* indices.

We then fill a dynamic programming table using the following recurrence equation:

$$E(i,l) = \min\begin{cases} E(i+1,l) \\ E(i,l-1) \\ \min\limits_{i<j<k<l} E[(i,j)(k,l)] \\ \min\limits_{i<p<l}(E(i,p) + E(p+1,l)) \end{cases}$$

The value *E(1,N)* gives the best possible energy for an assembly of stem-loops. Note the similarity between these recurrence equations and those of Nussinov-Jacobson[1]. In the top equation, nucleotide *i* is free and in the second equation nucleotide *l* is free. The third equation is for considering a stem, whereas the last equation is for considering a multi-branch structure. This process is in $O(N^4)$ in time, due to the third equation, and does not consider pseudo-knotted structures. We do not mark the minimum value origins, as we do not need to reconstruct the minimum energy structure at this step.

The dynamic programming table is used to enumerate the sub-optimal solutions. We use the Waterman-Byers algorithm[2], which needs $E_{min} = E(1,N)$, as well as a fraction of the energy, Δ, that limits the sub-optimal solutions considered. The energy of a sub-optimal returned by the algorithm is E, $E_{min} \leq E \leq E_{min} + \Delta$, which is the Waterman-Byers condition.

We solve the problem by backtracking over the stem variables. We pick one, two, three, and so on stems from a list, L, generated *a priori*. In other words, we compute the Cartesian products, $\{L\}\times\{L\}$, $\{L\}\times\{L\}\times\{L\}$, and so on. We make sure that the selected stems are entirely embedded, i.e. $j < i' < l' < k$, as well as that they define distinct sequence regions, i.e. $(i' > l)$. Each time a new stem is added, the Waterman-Byers condition is verified. The current energy is added to the minimum energy of the remaining sequence, $E(j, k) + E(l, N \setminus \Omega)$, which are both available from the dynamic programming table. Ω is the set of the regions spanned by the previously selected stems (see Fig. S12). At anytime, if it is possible to build a structure that will respect the Waterman-Byers condition, then we continue the current construction; otherwise we try the next stem for the current variable or if no more stems are available, we backtrack to the previous stem-variable. This process is exponential and influenced greatly by the Δ value, which determines the probability of satisfying the condition. Haralick and Elliott developed a probabilistic time complexity model of backtracking algorithms in 1980[3].

For pseudo-knotted structures, we squeeze in an extra stem, B, in a complete secondary structure, such that B creates the ABAB configuration with another stem, A, previously selected in the structure. The ABAB pseudo-knot configuration constitutes the vast majority of pseudoknots (also called H-type)[4]. Several Aalberts and Hodas rules about pseudoknot stem lengths were implemented[4]. The total pseudo-knot energy is that of it's constituting stems, including the coaxial stacking contribution. Also, the Waterman-Byers condition must be relaxed to allow for the initial A-A- stem configuration (on which the ABAB pseudo-knot can form). This increases significantly the search space size, and thus computation time.

MC-Fold scoring function. MC-Fold generates a set of sub-optimal structures given a single input sequence. The structures are ranked by their probability of occurrence given the sequence. These scores are transformed in energies by assuming a Boltzmann distribution:

$$\Phi(structure \mid sequence) = -RT \ln \Psi(structure \mid sequence),$$

where RT has the value 0.606 kcal/mol.

The scoring function accounts for the probabilities of observing the NCMs given the sequence, their junctions, the base pairs in the context of the junctions, and the base pairs themselves, out of any context (Fig. S10). As a result, we obtain the following Master equation:

$$\Psi(structure \mid seq) = \Psi(NCMs \mid seq) \times \Psi(junctions \mid NCMs) \times \Psi(hinges \mid junctions) \times \Psi(pairs \mid hinges)$$

When a suite of NCMs is assigned to a sequence, each NCM, c_i, is mapped to a subsequence of the sequence, s_i. The sequence-NCM affinity is evaluated by the first term of the scoring function:

$$\Psi(NCMs \mid seq) = \prod_{i}^{cycles} \Psi(c_i \mid s_i),$$

which can be written as:

$$\Psi(c_i \mid s_i) = \frac{\Psi(s_i \mid c_i)\Psi(c_i)}{\Psi(s_i)}$$

using Bayes's theorem. Since $\Psi(c_i \mid s_i)$, the probability of c_i given s_i, cannot be computed directly, we compute $\Psi(s_i \mid c_i)$, the probability of observing s_i in c_i, $\Psi(s_i)$, the probability of s_i, and $\Psi(c_i)$, the probability of c_i. The probability of s_i, $\Psi(s_i)$, is the product of the occurrence probabilities of each nucleotide in s_i, or $\Psi_p(s_i)$.

Note that in the PDB we do not find every sequence within each NCM. To avoid null probabilities whenever a sequence cannot be found in a specific NCM, we accept sibling alternative sequences. Each nucleotide in the sequence is allowed the following IUPAC-IUB single-letter code lists: A:[A,R,M,N], C:[C,Y,M,N], G:[G,R,K,N], and U:[U,Y,K,N]. Consequently, a sequence of *n* nucleotides is represented by 4^n sequences. We call the generalized sequence, gs_i, the sequence that maximizes the ratio of the actual sequence probability within a given cycle on the *a priori* sequence probability:

$$\Psi(c_i \mid s_i) \propto \max_{g} \frac{\Psi(gs_i \mid c_i)}{\Psi_{apriori}(gs_i)}$$

Here, the maximization of the ratio prevents the over-generalization of the sequence into the degenerate N-only sequence.

For computation speedup, all sequence variations of each cycle were pre-calculated, and their worst probabilities, $\Psi(c_i \mid s_i)$, were arbitrarily assigned a maximum energy of +1.0 kcal/mol; the term $\Psi(s_i)$ has now been absorbed into the scaling of converting the probability into energy.

The second term evaluates the junction of two cycles, corresponding to a Markov chain of order 1:

$$\Psi(junctions \mid NCMs) = \prod_{(j,k)}^{junctions} \Psi(junction_{(j,k)} \mid NCM_j \wedge NCM_k)$$

where $\Psi(junction_{(j,k)} \mid NCM_j \wedge NCM_k)$ is the probability to observe a junction composed of NCM$_j$ followed by NCM$_k$. The maximum energy associated with the lowest junction probabilities was arbitrarily assigned to +1.0 kcal/mol.

When two NCMs are joined, the base pairing type of the common base pair depends not only on the sequence, but also on the two NCMs. For example, the flanking base pair of a tri-loop must accommodate the sharp turn of the RNA backbone. Thus, the hinge can be scored by:

$$\Psi(hinges \mid junctions) = \prod_{l}^{hinges} \Psi(hinge_l \mid junction_{(j,k)})$$

where $\Psi(hinge_l \mid junction_{(j,k)})$ is the probability of observing hinge$_l$ at junction$_{(j,k)}$. Let $\Psi(type_m \mid NCM_j^l)$ be the probability to observe base pairing type *m* in NCM$_j$ in hinge$_l$. To consider all base pairing types of the hinge, we must consider all common base-pairing types of NCM$_j$ and NCM$_k$:

$$\Psi(hinge_l \mid junction_{(j,k)}) = \sum_{m}^{j} \sum_{n}^{k} \delta_{m,n} \Psi(type_m \mid NCM_j^l) \Psi(type_n \mid NCM_k^l)$$

where δ is the Dirac delta function, which ensures that the joint probabilities are calculated for the common base pairing types only. This computation prevents the incorporation of an invalid base pair in the hinge (see Table S3).

Finally, once the hinge has been specified, we must quantify the specific nucleotide association of the base pair. Thus:

$$\Psi(pairs \mid hinges) = \prod_{p}^{pairs} \Psi(pair_p \mid hinge_l),$$

where $\Psi(pair_p \mid hinge_l)$ is the probability of observing pair$_p$ in the hinge$_l$. The maximum energy has been arbitrarily fixed to +1.0 kcal/mol.

Coaxial stacking energetic contributions. The coaxial stacking between two stems is scored accordingly to the creation of a new 2_2 NCM between the two stems. This NCM is similar to the others of its class, but lacks one phosphodiester linkage, which is substituted by a base stacking interaction. The total energetic contribution of coaxial stacking, therefore, comes from the new NCM itself, i.e. its fitness to the sequence, as well as from the two new junctions (-2.9 kcal/mol). An entropy cost of +2.5 kcal/mol is added for the loss of the phosphodiester linkage. This arbitrary value is a compromise between single and multi-branched structures: low costs favour multi-branched structures; high costs hairpins.

MC-Sym structure generation. Libraries of 3-D fragments corresponding to each NCM are built (see NCM building above). The NCM fusion in MC-Sym is conceptually equivalent to that of MC-Fold, i.e. all possible NCM 3-D fragments are systematically assigned to the sequence. However, since MC-Fold has already assigned a score, no scoring is necessary. The concatenation of two adjacent NCMs is done by optimal superimposition of the two copies of the common base pair in 3-D. Since there are many possible NCM 3-D fragments for each NCM, an exhaustive assignment is prohibitive. Instead, a Las Vegas algorithm is used to explore as many structures as possible in a given period of time, fixed for this study to 12h. The difference between the Las Vegas and the better-known Monte Carlo algorithms is that the former never gives an incorrect result, i.e. all 3-D structures generated by MC-Sym are consistent with the input constraints.

MC-Fold | MC-Sym pipeline. The pipeline is described in Fig. S1. Input 1 is a single sequence. MC-Fold performs the NCM fusion 1 and returns a sorted list of possible structures in dot-bracket notations (Fig. S13). An MC-Sym input script for any MC-Fold solution can be generated by providing it in the "mask" field of MC-Fold (see Fig. S14). This represents Input 2 in the pipeline diagram of Fig. S1. MC-Sym is invoked and run for 24 hours, producing atomic-precision 3-D models that satisfy the interactions specified in the script. An RMSD threshold for each NCM merge, an overall atomic clash constraint, a ribose construction threshold, an implicit phosphate restraint, a time limit or a maximum number of models, and a threshold RMSD amongst the models produced parameterize MC-Sym. These values can be edited in the script generated by MC-Fold. However, default values for these parameters are fixed, and the scripts generated by MC-Fold can be submitted to MC-Sym without editing. The output of MC-Sym is a set of 3-D structures in PDB format[5] (Fig. S15).

MC-Cons. The algorithm MC-Cons does not find a consensus structure deprived of many base pairs that fit all sequences of an RNA family. Instead, we assign to each sequence one of its suboptimal predictions that globally optimizes the sum of pair-wise similarities. In other words, we look for a global and structural consensus assignment (that may include more than one structures) rather than for a common structure. This is similar to the concept of RNA "shapes" proposed by Reeder and Giegerich[6]. First, a similarity score is computed for each pair of suboptimal solutions and stored in a similarity matrix. This score is largely biased towards structural alignment, rather than sequence alignment. Then, from the similarity matrix, the maximum sum is found by backtracking over all suboptimal solutions. As the sequence-structure space grows exponentially, a cyclic coordinate method[7], where the optimal structure of one sequence is searched while all others are fixed, is used as an optimization heuristic. We then apply hierarchical clustering to unveil the structural features of the consensus assignment.

RNA structure images. The 3-D structures were rendered using PyMOL. The secondary structure were rendered using a modified version of the CONTRAfold renderer[16].

Discussion

Arguments in favour of a new HIV-1 -1 frameshifting element. First, the double A bulge is conserved across all 753 sequences, suggesting a possible functional role. It can adopt the A-minor motif that can simultaneously kink the structure[8] and dock to any tandem of Watson-Crick base pairs[9]. In comparison, the GGA bulge is found in half of these sequences (Fig. S9), substituted by a GAA bulge in the other half. G and A have different chemical groups and, in general, cannot easily be substituted. Second, the flanking base pair above the bulge can either be GA or AA, which are frequent and stable at the end of double-helical stems[10]. Third, the model satisfies enzymatic probing data applied to the native sequence from two studies[11,12]. Fourth, the model applies to all HIV-1 subtypes, introducing three times less NC base pairs in only one rather than three sites in the NMR model[13]. Fifth, our has lower thermodynamic average energies than the NMR model (-23.0 vs. -21.3 kcal/mol; as computed by the RNAeval program of the Vienna package[14]). Sixth, the model corroborates with recent enzymatic cleavage data that indicate an unpaired nucleotide A45[15].

Tables

Table S1 | Comparison of the predictive power of three approaches. The predictions of three approaches are compared over 1968 base pairs (1665 Watson-Crick) in 264 hairpins extracted from 182 different PDB structures. Zipper implements a greedy algorithm that folds a sequence from bottom-up using exclusively tandems of base pairs. This gives us a lower bound on the predictive power. RNAsubopt implements the current thermodynamics model and enumerates exhaustively all suboptimal solutions. For each approach, the best predicted structures are analyzed. In each row, the best value is shown in bold. By increasing the number of sub-optimal solutions to 5, the Matthews coefficient ratios go up to 93.1 (99.1% of the canonical base pairs) and 87.7 (97.3% of the canonical base pairs), respectively for MC-Fold and RNAsubopt. Interestingly, MC-Fold's ratio reaches 92.2 when the top 2 solutions are analyzed (RNAsubopt 86.3).

Predicted base pairs (%)	*Zipper* *(Lower bound)*	RNAsubopt (Thermodynamics)	CONTRAfold (Machine learning)	MC-Fold (NCM)
False positives	*50.2*	**6.7**	7.5	17.9
False negatives	*25.9*	25.2	26.9	**10.1**
True Positives	*74.1*	74.8	73.1	**89.9**
Canonicals	*75.6*	*88.4*	*86.3*	***94.7***
Non-canonicals	*64.9*	*N/A*	*1.4*	***62.1***
Matthews = $\sqrt{\frac{\mathrm{TP}}{(\mathrm{TP+FN})}\frac{\mathrm{TP}}{(\mathrm{TP+FP})}}$	*66.5*	82.8	81.4	**86.6**

Table S2 | RNA-Select. The 531 PDB codes corresponding to the X-ray crystallographic and NMR structures.

104D	124D	157D	168D	170D	176D	17RA	1A34	1A4T	1A51
1A60	1A9N	1AFX	1AJF	1AJT	1AL5	1AM0	1APG	1ATO	1ATV
1ATW	1AUD	1AV6	1B23	1B36	1B7F	1BAU	1BGZ	1BJ2	1BMV
1BN0	1BR3	1BVJ	1BYJ	1BYX	1BZ2	1BZT	1C0A	1C0O	1C2Q
1C4L	1C9S	1CK5	1CQ5	1CSL	1CVJ	1CX0	1CX5	1D0T	1D0U
1D4R	1D6K	1D9H	1DDL	1DDY	1DFU	1DQF	1DRR	1DUH	1DUL
1DUQ	1DXN	1DZ5	1E4P	1E7K	1E95	1EBR	1EC6	1EFO	1EFS
1EFW	1EHZ	1EJZ	1EKA	1EKD	1EKZ	1ELH	1ESH	1ET4	1EUY
1EVP	1EXD	1EXY	1F27	1F5G	1F5U	1F6U	1F6X	1F6Z	1F7U
1F84	1F85	1F8V	1F9L	1FEQ	1FEU	1FG0	1FHK	1FIX	1FL8
1FMN	1FNX	1FQZ	1FUF	1FYO	1G1X	1G2E	1G2J	1G3A	1G4Q
1G70	1GKW	1GSG	1GTF	1GTN	1GUC	1H0Q	1H2C	1H2D	1H38
1H3E	1H4S	1HC8	1HJI	1HLX	1HO6	1HOQ	1HS1	1HS2	1HS3
1HS4	1HS8	1HWQ	1HYS	1I2X	1I2Y	1I3X	1I3Y	1I46	1I4B
1I5L	1I6U	1I7J	1I9F	1I9K	1I9V	1I9X	1ICG	1IDV	1IE1
1IK1	1IK5	1IKD	1IL2	1IVS	1J1U	1J4Y	1J6S	1J8G	1J9H
1JBR	1JBT	1JID	1JO7	1JOX	1JTJ	1JTW	1JU7	1JUR	1JZC
1JZV	1K1G	1K2G	1K4A	1K4B	1K5I	1K6G	1K6H	1K8S	1KAJ
1KD3	1KFO	1KH6	1KIS	1KKS	1KNZ	1KOC	1KOD	1KOS	1KP7
1KPD	1KPY	1KQ2	1KUO	1KUQ	1KXK	1L1C	1L1W	1L2X	1L3Z
1L8V	1L9A	1LDZ	1LMV	1LNT	1LPW	1LUU	1LUX	1LVJ	1M5K
1M5L	1M82	1M8V	1M8W	1M8X	1M8Y	1MDG	1ME0	1ME1	1MFJ
1MFK	1MFY	1MHK	1MHM	1MIS	1MJI	1MMS	1MNX	1MSY	1MT4
1MUV	1MV1	1MV6	1MWG	1MY9	1MZP	1N1H	1N35	1N38	1N53
1N66	1N77	1N7A	1N8X	1NA2	1NAO	1NB7	1NBK	1NBR	1NC0
1NEM	1NTA	1NTQ	1NTS	1NTT	1NUJ	1NXR	1NYB	1NZ1	1O15
1OKF	1OLN	1OO7	1OOA	1OQ0	1OSU	1OSW	1OW9	1P5M	1P5N
1P5O	1P79	1PBL	1PGL	1PJY	1PVO	1Q29	1Q75	1Q8N	1Q93
1Q96	1Q9A	1QBP	1QC0	1QC8	1QD3	1QES	1QET	1QF6	1QLN
1QU2	1QWB	1R2P	1R3E	1R3O	1R3X	1R4H	1R7W	1R7Z	1RAW
1RC7	1RFR	1RGO	1RKJ	1RLG	1RMV	1RNA	1RNG	1RNK	1ROQ
1RPU	1RXA	1S03	1S2F	1S76	1S9L	1SA9	1SAQ	1SDR	1SDS
1SER	1SI3	1SJ3	1SLP	1SYZ	1SZY	1T0D	1T0E	1T28	1T2R
1T4L	1T4X	1TFN	1TFW	1TJZ	1TLR	1TOB	1TTT	1TUT	1TXS
1U0B	1U2A	1U3K	1U6P	1U8D	1U9S	1ULL	1UTD	1UUD	1UUU
1UVJ	1UVK	1UVL	1UVN	1VFG	1VOP	1VQ7	1WKS	1WNE	1WPU
1WRQ	1WSU	1WTS	1WWD	1WWE	1WWF	1WWG	1XHP	1XJR	1XMQ
1XOK	1XP7	1XPE	1XPF	1XSG	1XSH	1XV0	1XV6	1XWP	1XWU
1Y26	1Y27	1Y39	1Y3O	1YFG	1YFV	1YG3	1YMO	1YN1	1YNC
1YNE	1YSV	1YTU	1YTY	1YVP	1YYK	1YYW	1YZ9	1Z2J	1Z30
1Z31	1Z43	1Z7F	1ZBI	1ZC5	1ZCI	1ZDJ	1ZDK	1ZE2	1ZEV
1ZFV	1ZIF	1ZIG	1ZIH	1ZJW	1ZL3	1ZX7	1ZZ5	205D	216D
219D	246D	247D	255D	259D	280D	283D	28SP	2A0P	2A1R
2A43	2A8V	2A9X	2AB4	2AD9	2ADC	2ADT	2AO5	2ASB	2ATW
2AU4	2AWE	2AWQ	2AZ0	2B3J	2B6G	2BBV	2BE0	2BGG	2BH2
2BJ6	2BNY	2BS0	2BS1	2BTE	2BX2	2C06	2C4Y	2C4Z	2C50
2C51	2CHJ	2CSX	2D17	2D18	2D1A	2ERR	2ES5	2ESI	2EUY
2EZ6	2F4X	2F88	2F8K	2FK6	2FMT	2FQN	2FRL	2FZ2	2G1W
2G8F	2G92	2GBH	2GM0	2TOB	2TPK	2TRA	310D	315D	332D
333D	353D	354D	361D	364D	377D	393D	397D	398D	3PHP
402D	404D	405D	409D	413D	418D	419D	420D	421D	422D
429D	430D	433D	435D	438D	439D	464D	466D	468D	469D
470D	471D	472D	479D	484D	485D	5MSF	6MSF	7MSF	8DRH
8PSH									

Table S3 | Hinge scoring. The score of a GA hinge, *l*, at the junction of NCMs *i* (4-GAGA) and *j* (2_2-CGAG) is 0.731: the sum of the products of the probabilities of appearance, Ψ, of the GA base pairing types, *m* and *n*, found in the instances of the two NCMs in RNA-Select, independently of the junction. The sheared GA base pair (S/H anti) validates the hinge created by the junction of the two cycles since it is the most frequent among all possible base pairs (probability of 0.730). The Watson-Crick/Hoogsteen is another valid option, but is less likely to appear in this context (probability of 0.001).

Number of occurences	Probabilities of appearance (%)	Base pairing type
5’ G-A base pair of NCM$_i$	$\Psi_{m\mid i}(type_m \mid cycle_i^l)$	
72	0.889	`S/H anti trans`
3	0.037	`S/W anti cis`
3	0.037	`S/W para trans`
2	0.025	`W/H anti trans`
3’ G-A base pair of NCM$_j$	$\Psi_{n\mid j}(type_n \mid cycle_j^l)$	
161	0.821	`S/H anti trans`
29	0.148	`W/W anti cis`
2	0.010	`H/W para cis`
2	0.010	`W/B anti cis`
2	0.010	`W/H anti trans`
G-A base pair of hinge$_l$	$\sum^{i}\sum^{j}\delta_{m,n}\Psi(type_m \mid cycle_i^l)\times\Psi(type_n \mid cycle_i^l)\cong 0.731$	
	$0.889\times 0.821 = 0.730$	`S/H anti trans`
	$0.037\times 0.000 = 0.000$	`S/W anti cis`
	$0.000\times 0.148 = 0.000$	`W/W anti cis`
	$0.025\times 0.010 \approx 0.001$	`W/H anti trans`

Figures

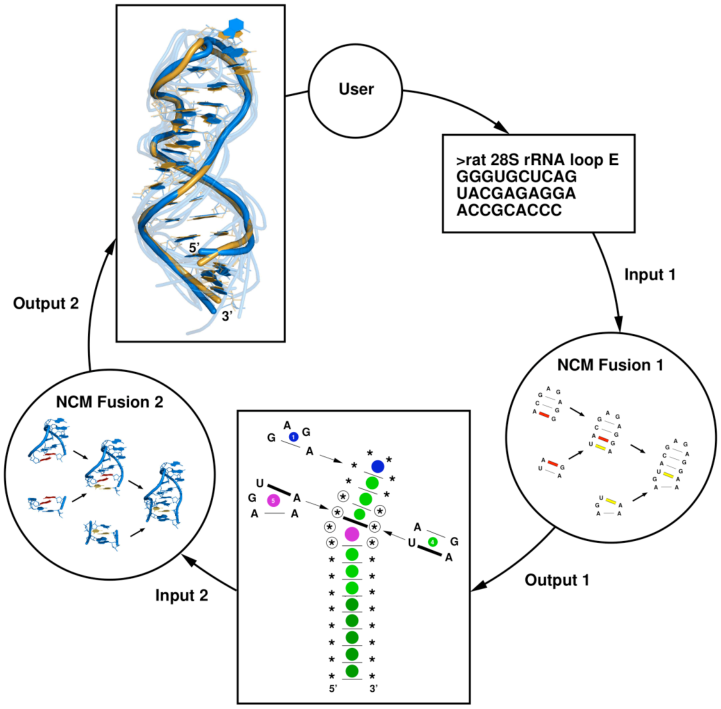


Figure S1 | The MC-Fold | MC-Sym pipeline applied to the rRNA loop E. Input 1: Sequence of the rat 28S rRNA loop E. NCM Fusion 1: MC-Fold. Two adjacent NCMs share a common hinge base pair (red and yellow). Output 1/Input 2: The optimal assignment contains 13 NCMs (circles), 14 base pairs (lines), and 29 nucleotides (stars). The three main NCM types are shown: blue) lone-pair loops (GAGA tetraloop; NCM #1); green) base pair tandems (dark green indicates canonical tandems); and, purple) bulge and interior loops, an extension of the base pair tandem. The NC UA hinge base pair (bold line) is common to NCMs #4 and #5, which combination forms the sarcin/ricin motif. Each stem-loop is one chain of NCMs. Since the output of MC-Fold can be a multi-branch or pseudo-knotted structure made of more than onoe hairpin, the output is a set of chains of NCMs. NCM Fusion 2: MC-Sym. Output 2: The closest prediction (blue) that shares 1.8 Å of RMSD and a representative sampling of structures (light blue) are shown optimally superimposed on the rat 28S rRNA X-ray crystallographic loop E structure (gold).

```
>tRNA ASP
GCCGUGAUAGUUUAAUGGUCAGAAUGGGCGCUUGUCGCGUGCCAGAUCGGGGUUCAAUUCCCCGUCGCGGCGC
.................xx.............xxxx.................x...................  High
...............xx..x................xx................xx.................  Medium
(((((((..(((((......)))))((((((.......))))))...((((((...)..))))))))))))..  native - 2TRA   RNK TP FP FN Mthw
(((((((((((((.......))))))(((((.......)))))....((((((...)..)))))))))))).. -59.75 ( -1.25)    1 18  6  6  75.0
(((((((.(((((.......)))))((((((.......))))))...((((((...)..)))))))))))).. -59.73 ( -0.84)    2 19  5  5  79.2
(((((((.(.(((.......))))(((((((.......)))))))..((((((...)..)))))))))))).. -59.47 ( -0.71)    3 19  5  5  79.2
(((((((..((((.......))))(((((((.......)))))))..((((((...)..)))))))))))).. -58.72 ( -0.79)    4 19  5  5  79.2
(((((((.(.(((.......))))(((((((.......)))))).).((((((...)..)))))))))))).. -58.69 ( -0.71)    5 19  5  5  79.2
(((((((..(((((......)))))((((((.......))))))...((((((...)..)))))))))))).. -58.46 ( -0.71)    6 24  0  0 100.0**
(((((((...(((.......)))((((((((.......)))))).))((((((...)..)))))))))))).. -58.25 ( -0.71)    7 19  5  5  79.2
(((((((...((((......))))(((((((.......)))))))..((((((...)..)))))))))))).. -58.23 ( -1.07)    8 23  1  1  95.8
(((((((.((((((......))))))(((((.......)))))....((((((...)..)))))))))))).. -58.02 ( -1.14)    9 23  1  1  95.8
(((((((...(((.......)))((((((((.......)))))))).((((((...)..)))))))))))).. -57.93 ( -0.71)   10 19  5  5  79.2
```

Figure S2 | MC-Fold predictions for the yeast tRNA^{ASP}. The top ten structures generated by MC-Fold for the Yeast tRNA^{ASP} under SHAPE constraints are shown. The native structure (PDB file 2TRA) ranks 4th (Matthews coefficient ratio of 100%). The numbers in parenthesis represent the energy contributions of the coaxial stacking. The nucleotides marked with a dot under the “High” SHAPE constraints have an 8 kcal/mol penalty if found paired; 4 kcal/mol for “Medium”. Nucleotide 47 is absent. Real time = 159 seconds.

a

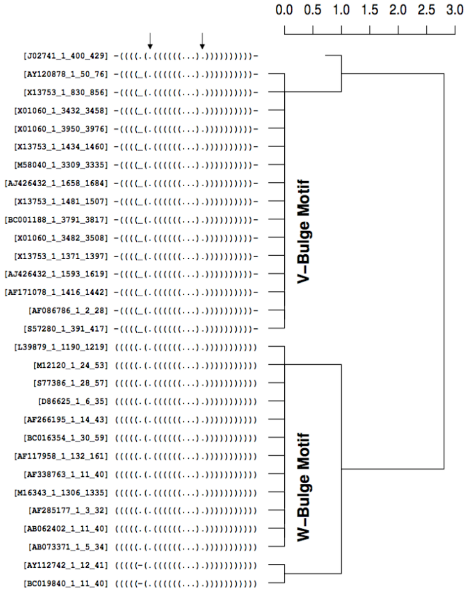
```
>E.coli 5
1   5   10   15   20   25   30   35   40   45   50   55   60   65   70   75   80   85   90   95  100  105  110  115  120
|   |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |    |
UGCCUGGCGGCCUUAGCGCGGUGGUCCCACCUGACCCCAUGCCGAACUCAGAAGUGAAACGCCGUAGCGCCGAUGGUAGUGUGGGGUCUCCCCAUGCGAGAGUAGGGAACUGCCAGGCAU
..............x.............................x...........................................................................  strong DMS
......................................x............xx.....x......x......................................................  moderate DMS
(((((((((((....((((((((....(((((((.............))))..)))...)))))).)).(((((((((((((((((...)))))))))))))))))..))))))))))).  native (2AW4)   RNK TP FP FN Mthw
(((((((((((((..(((((((((((..((((((((.(.....).))))))..))))).)))))).)).(((((((((((((((((...)))))))))))))))))))))))))))))). -110.44 ( -2.02)   1 42  8  1 90.6
(((((((((((((..(((((((((((..(((((((((......).))))))..))))).)))))).)).(((((((((((((((((...)))))))))))))))))))))))))))))). -110.08 ( -2.02)   2 42  8  1 90.6
(((((((((((.(..(((((((((((..((((((((.(.....).))))))..))))).)))))).))((((((((((((((((((...)))))))))))))))))))))))))))))). -109.35 ( +0.00)   3 42  8  1 90.6
((((((((((((...(((((((((((..((((((((.(.....).))))))..))))).)))))).))((((((((((((((((((...)))))))))))))))))))))))))))))). -109.29 ( -1.26)   4 42  8  1 90.6
(((((((((((.((.(((((((((((..((((((((.(.....).))))))..))))).)))))).)).(((((((((((((((((...)))))))))))))))))))))))))))))). -109.01 ( +0.00)   5 42  8  1 90.6
```

b

```
20   25   30   35   40   45   50   55   60   65
     |    |    |    |    |    |    |    |    |    |
aGCGCGGUGgUcCCacCUGAcccCAUGCCGaacUCAGaaGUGaAaCGCCGUAGCg
((((((((((((((((((((((((.....))))))))))).))))))))))).))) -45.05 ( +0.00)
```

Figure S3 | MC-Fold predictions for the *E. coli* 5S rRNA. a. The top 5 structures generated by MC-Fold for the *E. coli* 5S rRNA under DMS constraints are shown. The native structure (PDB file 2AW4) is not predicted (best Matthews coefficient ratio 90.6%). The numbers in parenthesis represent the energy contributions of the coaxial stacking. The nucleotides marked with a dot under the “strong” DMS reactivity have an 8 kcal/mol penalty if found paired; 4 kcal/mol for “moderate”. Real time = 2131 seconds. **b**. The optimal MC-Fold solution of the 16-69 *E. coli* 5S rRNA subsequence. The NC base pairs are shown using lowercase letters.

a.


0.0
0.5
1.0
1.5
2.0
2.5
3.0
[J02741_1_400_429] -((((.(.((((((...).))))))))))-
[AY120878_1_50_76] -((((_(.((((((...).))))))))))-
[X13753_1_830_856] -((((_(.((((((...).))))))))))-
[X01060_1_3432_3458] -((((_(.((((((...).))))))))))-
[X01060_1_3950_3976] -((((_(.((((((...).))))))))))-
[X13753_1_1434_1460] -((((_(.((((((...).))))))))))-
[M58040_1_3309_3335] -((((_(.((((((...).))))))))))-
[AJ426432_1_1658_1684] -((((_(.((((((...).))))))))))-
[X13753_1_1481_1507] -((((_(.((((((...).))))))))))-
[BC001188_1_3791_3817] -((((_(.((((((...).))))))))))-
[X01060_1_3482_3508] -((((_(.((((((...).))))))))))-
[X13753_1_1371_1397] -((((_(.((((((...).))))))))))-
[AJ426432_1_1593_1619] -((((_(.((((((...).))))))))))-
[AF171078_1_1416_1442] -((((_(.((((((...).))))))))))-
[AF086786_1_2_28] -((((_(.((((((...).))))))))))-
[S57280_1_391_417] -((((_(.((((((...).))))))))))-
[L39879_1_1190_1219] (((((.(.((((((...).)))))))))))
[M12120_1_24_53] (((((.(.((((((...).)))))))))))
[S77386_1_28_57] (((((.(.((((((...).)))))))))))
[D86625_1_6_35] (((((.(.((((((...).)))))))))))
[AF266195_1_14_43] (((((.(.((((((...).)))))))))))
[BC016354_1_30_59] (((((.(.((((((...).)))))))))))
[AF117958_1_132_161] (((((.(.((((((...).)))))))))))
[AF338763_1_11_40] (((((.(.((((((...).)))))))))))
[M16343_1_1306_1335] (((((.(.((((((...).)))))))))))
[AF285177_1_3_32] (((((.(.((((((...).)))))))))))
[AB062402_1_11_40] (((((.(.((((((...).)))))))))))
[AB073371_1_5_34] (((((.(.((((((...).)))))))))))
[AY112742_1_12_41] (((((-(.((((((...).)))))))))))
[BC019840_1_11_40] (((((-(.((((((...).)))))))))))
V–Bulge Motif
W–Bulge Motif

b.

EMBL Number	Sequence / Structure	Rank
Consensus	((((((.(((((((...).)))))))))))	
AY112742_1_12_41	GUcCUGCUUCAACAGUGCUUGAACGGaAC	(1st)
BC019840_1_11_40	GUcUUGCUUCAACAGUGUUUGAACGGaAC	(1st)
AF086786_1_2_28	-UcGUUCGUCCUCAGUGCAGGGCAACaG-	(1st)
S57280_1_391_417	-UcGUUCGUCCUCAGUGCAGGGCAAUaG-	(5th)
AF171078_1_1416_1442	-UgGUUCGUCCUCAGUGCAGGGCAACaG-	(1st)
AJ426432_1_1593_1619	-AUUAUCGGGAGCAGUGUCUUCCAUAAU-	(1st)
X13753_1_1371_1397	-AUUAUCGGGGGCAGUGUCUUCCAUAAU-	(3rd)
X01060_1_3482_3508	-AUUAUCGGAAGCAGUGCCUUCCAUAAU-	(1st)
BC001188_1_3791_3817	-AUUAUCGGGAACAGUGUUUCCCAUAAU-	(3rd)
X13753_1_1481_1507	-AUUAUCGGGGACAGUGUUUCCCAUAAU-	(3rd)
AJ426432_1_1658_1684	-UAUAUCGGAGACAGUGAUCUCCAUAUG-	(1st)
M58040_1_3309_3335	-UAUAUCGGAGaCAGUGAcCUCCAUAUG-	(1st)
X13753_1_1434_1460	-UAUAUCGGAGGCAGUGACCUCCAUAUG-	(1st)
X01060_1_3950_3976	-UGUAUCGGAGACAGUGAUCUCCAUAUG-	(1st)
X01060_1_3432_3458	-UUUAUCAGUGACAGAGUUCACUAUAAA-	(1st)
X13753_1_830_856	-UUUAUCAGUGACAGCGUUCACUAUAAA-	(1st)
AY120878_1_50_76	-GgUCGCGUCAACAGUGUUUGAUCGAaC-	(1st)
Consensus	(((((.(.(((((((...).)))))))))))	
AB062402_1_11_40	uUUCCUGCUUCAACAGUGCUUGGACGGAAc	(1st)
AB073371_1_5_34	uCUCCUGCUUCAACAGUGCUUGGACGGAGc	(1st)
AF285177_1_3_32	GUUCCUGCUUCAACAGUGCUUGGACGGAAC	(1st)
M16343_1_1306_1335	GUUCCUGCgUCAACAGUGCUUGGaCGGAAC	(2nd)
AF338763_1_11_40	UUaCCUGCUUCAACAGUGCUUGAACGGcAA	(1st)
AF117958_1_132_161	ucUCUUGUUUCAACAGUGUUUGGACGGAac	(21th)
BC016354_1_30_59	ucUCUUGCUUCAACAGUGUUUGGACGGAac	(2nd)
AF266195_1_14_43	AUUCUUGCUUCAACAGUGUUUGAACGGAAU	(1st)
D86625_1_6_35	GUUCUUGUUUCAACAGUGAUUGAACGGAAC	(1st)
S77386_1_28_57	GUUCUUGCUUCAACAGUGAUUGAACGGAAC	(1st)
M12120_1_24_53	GUUCUUGCUUCAACAGUGUUUGAACGGAAC	(1st)
L39879_1_1190_1219	GUaCUUGCUUCAACAGUGUUUGAACGGaAC	(1st)
J02741_1_400_429	-aUCUUGCUUCAACAGUGUUUGGACGGAa-	(1st)

Figure S4 | Clustering and aligned IRE sequences. a. The results of a hierarchical clustering of the predicted structures identified by MC-Cons using inputs from MC-Fold. Each sequence is identified by its EMBL identifier, and as found in the Rfam database. A structural distance of 0 indicates identical structures. The IRE sequences are clearly grouped in their respective structural class: the V-bulge (above) and the W bulge (below). The W bulge is recognized in the bracket notation by the typical "((.(.((", whereas the V bulge is recognized by "((.((". The arrows indicate the C involved in IRE function. MC-Cons determines the IRE consensus assignment in about 10 minutes. **b**. The alignment was made according to consensus structures identified by MC-Cons. The sequences are divided in two groups: the V-bulge (up) and the W-bulge (down). The non-canonical base pairs are highlighted using lowercase letters.

```
>tRNA-ASN
GACUCCAUGGCCAAGUUGGUUAAGGCGUGCGACUGUUAAUCGCAAGAUCGUGAGUUCAACCCUCACUGGGGUCGCCA
(((((((..((((.........))))((((((((...))))))))....((((((...)..))))))))))))....    (   2nd)
...............xx..xx.............xxx..........x.............................
 >tRNA-GLY
GCGCAAGUGGUUUAGUGGUAAAAUCCAACGUUGCCAUCGUUGGGCCCCGGUUCGAUUCCGGGCUUGCGCACCA
(((((((..(((((.....)))))((((((((...))))))))..((((((...)..))))))))))))....        (   2nd)
........x......x..x.............xxx......................................
 >tRNA-ILE
GGUCUCUUGGCCCAGUUGGUUAAGGCACCGUGCUAAUAACGCGGGGAUCAGCGGUUCGAUCCCGCUAGAGACCACCA
(((((((..(((((.......)))))((((((((...))))))))....((((((...)..))))))))))))....    (   5th)
........x......xx..xx.............xxx..........x.............................
 >tRNA-LYS
UCCUUGUUAGCUCAGUUGGUAGAGCGUUCGGCUUUUAACCGAAAUGUCAGGGGUUCGAGCCCCCUAUGAGGAGCCA
(((((((..(((((......)))))((((((((...))))))))....((((((...)..))))))))))))....     (   1st)
...............xx..x.............xxx..........x.............................
 >tRNA-MET
GCUUCAGUAGCUCAGUAGGAAGAGCGUCAGUCUCAUAAUCUGAAGGUCGAGAGUUCGAACCUCUCCUGGAGCACCA
(((((((..(((((......)))))((((((((...))))))))....((((((...)..))))))))))))....     (   5th)
...............x.................xxx..........x.............................
 >tRNA-THR
GCUUCUAUGGCCAAGUUGGUAAGGCGCCACACUAGUAAUGUGGAGAUCAUCGGUUCAAAUCCGAUUGGAAGCACCA
(((((((..((((........))))((((((((...))))))))....((((((...)..))))))))))))....     (   1st)
...............xx..x.............xxx..........x.............................
 >tRNA-TRP
GAAGCGGUGGCUCAAUGGUAGAGCUUUCGACUCCAAAUCGAAGGGUUGCAGGUUCAAUUCCUGUCCGUUUCACCA
(((((((..(((((.....)))))((((((((...))))))))....((((((...)..))))))))))))....      (   1st)
........x......x..x.............xxx..........x.............................
 >tRNA-ALA
GGGCGUGUGGCGUAGUCGGUAGCGCGCUCCCUUAGCAUGGGAGAGGUCUCCGGUUCGAUUCCGGACUCGUCCACCA
(((((((..(((((......)))))(((((((.....)))))))....((((((...)..))))))))))))....     (  60th)
........x......x...x.............xxxx.........x.............................
 >tRNA-ARG
UUCCUCGUGGCCCAAUGGUCACGGCGUCUGGCUACGAACCAGAAGAUUCCAGGUUCAAGUCCUGGCGGGGAAGCCA
(((((((..(((((......)))))((((((((...))))))))....((((((...)..))))))))))))....     (  48th)
........x......x..x..............xxx..........x.............................
 >tRNA-ASP
UCCGUGAUAGUUUAAUGGUCAGAAUGGGCGCUUGUCGCGUGCCAGAUCGGGGUUCAAUUCCCCGUCGCGGAGCCA
(((((((..(((((......)))))((((((.......))))))...((((((...)..))))))))))))....       2TRA   TP FP FN Mthw
(((((((..(((((......)))))(((((((.....)))))))...((((((...)..))))))))))))....      (  5th) 24  1  0 98.0
...............x..x..............xxxx......................................
>tRNA-GLU
UCCGAUAUAGUGUAACGGCUAUCACAUCACGCUUUCACCGUGGAGACCGGGGUUCGACUCCCCGUAUCGGAGCCA
(((((((..(((((......)))))((((((((...))))))))...((((((...)..))))))))))))....      (  55th)
...................x.............xxx.......................................
 >tRNA-HIS
GGCCAUCUUAGUAUAGUGGUUAGUACACAACAUUGUGGCUGUUGAAACCCUGGUUCGAUUCUAGGAGGUGGCACCA
((((((((..(((((......)))))(((((((.....)))))))...((((((...)..))))))))))))).)..     (  13th)
................x..xx.............xxxx......................................
 >tRNA-PHE
GCGGAUUUAGCUCAGUUGGGAGAGCGCCAGACUGAAGAUCUGGAGGUCCUGUGUUCGAUCCACAGAAUUCGCACCA
(((((((..((((........))))(((((((.....)))))))....((((((...)..))))))))))))....      4TRA    TP FP FN Mthw
(((((((..(((((......)))))((((((.(....)))))))....((((((...)..))))))))))))....     ( 336th) 23  2  1 93.9
...............xx................xxxx.......................................
>tRNA-VAL
GGUUUCGUGGUCUAGUCGGUUAUGGCAUCUGCUUAACACGCAGAACGUCCCCAGUUCGAUCCUGGGCGAAAUCACCA
(((((((..(((((.......)))))((((((((...))))))))....((((((...)..))))))))))))....    (  11th)
........x......x...xx.............xxx..........x.............................
```

Figure S5 | Consensus structural assignment for yeast tRNA sequences. The yeast non-mitochondrial tRNA sequences are from the September 2004 edition of the compilation of tRNA sequences and sequences of tRNA genes database. The modified nucleotides in MC-Fold are treated like their canonical counterparts. The modified nucleotides that cannot adopt the A-RNA helix are constrained. For each tRNA, the anticodon nucleotides are unpaired. The positions marked with 'x' are either modified nucleotides that cannot form the A-RNA helix (unpaired), or anticodon nucleotides. The average real time to fold each tRNA sequence is 223.6 sec. MC-Cons determines the consensus structural assignment in about 53 minutes.

```
>E.coli 1
UGCCUGGCGGCCGUAGCGCGGUGGUCCCACCUGACCCCAUGCCGAACUCAGAAGUGAAACGCCGUAGCGCCGAUGGUAGUGUGGGGUCUCCCCAUGCGAGAGUAGGGAACUGCCAGGCAU
((((((((((.....(((((((((((..((((((.............))))..))))).)))))).))((((((((((((((((((...))))))))))))))))))..)))))))))).     (   2nd)
>E.coli 2
UGCCUGGCGGCAGUAGCGCGGUGGUCCCACCUGACCCCAUGCCGAACUCAGAAGUGAAACGCCGUAGCGCCGAUGGUAGUGUGGGGUCUCCUCAUGCGAGAGUAGGGAACUGCCAGGCAU
((((((((((.....(((((((((((..((((((.............))))..))))).)))))).))((((((((((((((((((...))))))))))))))))))..)))))))))).     (   5th)
>E.coli 3
UGCCUGGCGGCAGUAGCGCGGUGGUCCCACCUGACCCCAUGCCGAACUCAGAAGUGAAACGCCGUAGCGCCGAUGGUAGUGUGGGGUCUCCCCAUGCGAGAGUAGGGAACUGCCAGGCAU
((((((((((.....(((((((((((..((((((.............))))..))))).)))))).))((((((((((((((((((...))))))))))))))))))..)))))))))).     (   3rd)
>E.coli 4
UGUCUGGCGGCAGUAGCGCGGUGGUCCCACCUGACCCCAUGCCGAACUCAGAAGUGAAACGCCGUAGCGCCGAUGGUAGUGUGGGGUCUCCCCAUGCGAGAGUAGGGAACUGCCAGACAU
((((((((((.....(((((((((((..((((((.............))))..))))).)))))).))((((((((((((((((((...))))))))))))))))))..)))))))))).     (   3rd)
>E.coli 5
UGCCUGGCGGCCUUAGCGCGGUGGUCCCACCUGACCCCAUGCCGAACUCAGAAGUGAAACGCCGUAGCGCCGAUGGUAGUGUGGGGUCUCCCCAUGCGAGAGUAGGGAACUGCCAGGCAU
..........xx..x.............................x...........................................................................     strong DMS
..................................x..xx............xx.....x......x......................................................     moderate DMS
(((((((((((....((((((((....(((((((.............))))..)))...)))))).)).(((((((((((((((((...)))))))))))))))))..))))))))))).     native (2AW4)  TP FP FN Mthw
((((((((((.....(((((((((((..((((((.............))))..))))).)))))).))((((((((((((((((((...))))))))))))))))))..)))))))))).     (   5th)       41  4  2 93.2
(((((((((((((((((((((((....(((((((.............))))..)))...)))))).))(((...))).((((((((...))))))))....)))))..)))))))))).     Mathews et al. 33  8 10 78.6
>E.coli 6
UGUCUGGCGGCAGUAGCGCGGUGGUCCCACCUGACCCCAUGCCGAACUCAGAAGUGAAACGCCGUAGCGCCGAUGGUAGUGUGGGGACUCCCCAUGCGAGAGUAGGGAACUGCCAGACAU
((((((((((.....(((((((((((..((((((.............))))..))))).)))))).))((((((((((((((((((...))))))))))))))))))..)))))))))).     (   3rd)
>E.coli 8
UGCCUGGCGGCCUUAGCGCGGUGGUCCCACCUGACCCCAUGCCGAACUCAGAAGUGAAACGCCGUAGCGCCGAUGGUAGUGUGGGGGCUCCCCAUGCGAGAGUAGGGAACUGCCAGGCAU
((((((((((.....(((((((((((..((((((.............))))..))))).)))))).))((((((((((((((((((...))))))))))))))))))..)))))))))).     (   5th)
>E.coli 10
UGUCUGGCGGCAGUAGCGCGGUGGUCCCACCUGACCCCAUGCCGAACUCAGAAGUGAAACGCCGUAGCGCCGAUGGUAGUGUGGGGUCUCCUCAUGCGAGAGUAGGGAACUGCCAUGCAU
((((((((((.....(((((((((((..((((((.............))))..))))).)))))).))((((((((((((((((((...))))))))))))))))))..)))))))))).     (   2nd)
>E.coli 11
UGCCUGGCGGCAGUAGCGCGGUGGUCCCACCUGACCCCAUGCCGAACUCAGAAGUGAAACGCCGUAGCGCCGAUGGUAGUGUGGGGUCUCCCCAUGCGAGAGUAGGGAACUGCCAGGCAUCA
((((((((((.....(((((((((((..((((((.............))))..))))).)))))).))((((((((((((((((((...))))))))))))))))))..))))))))))...   (   3rd)
>E.coli 14
UGCCUGGCGGCCGUAGCGCGGUGGUCCCACCUGACCCCAUGCCGAACUCAGAAGUGAAACGCCGUAGCGCCGAUGGUAGUGUGGGGUCUCCCCAUGCGAGAGUAGGGAACUGCCAGACAU
((((((((((.....(((((((((((..((((((.............))))..))))).)))))).))((((((((((((((((((...))))))))))))))))))..)))))))))).     (   2nd)
```

Figure S6 | MC-Cons consensus assignment for the *in vivo E. coli* 5S rRNA. The ten sequences were obtained from the 5S ribosomal RNA database. Each sequence was submitted to MC-Fold. The top 100 structures for each sequence were then submitted to MC-Cons. The *E. coli* sequence #5 is the same as used by Mathews and colleagues (*Proc. Natl Acad. Sci. U S A*. **101**, 7287-7292, 2004). For each consensus structure, the MC-Fold rank is shown in parenthesis. MC-Fold average real time = 925.6 sec. MC-Cons real time = 2151 sec.

```
>E.coli 1
UGCCUGGCGGCCGUAGCGCGGUGGUCCCACCUGACCCCAUGCCGAACUCAGAAGUGAAACGCCGUAGCGCCGAUGGUAGUGUGGGGUCUCCCCAUGCGAGAGUAGGGAACUGCCAGGCAU
((((((((((.(.(((((((((((((((((((((((((.....))))))))))).)))))))))).)))(((((((((((((((((...))))))))))))))))).)))))))))))).  (12th)
>E.coli 2
UGCCUGGCGGCAGUAGCGCGGUGGUCCCACCUGACCCCAUGCCGAACUCAGAAGUGAAACGCCGUAGCGCCGAUGGUAGUGUGGGGUCUCCUCAUGCGAGAGUAGGGAACUGCCAGGCAU
((((((((((.(.(((((((((((((((((((((((((.....))))))))))).)))))))))).)))(((((((((((((((((...))))))))))))))))).)))))))))))).  ( 1st)
>E.coli 3
UGCCUGGCGGCAGUAGCGCGGUGGUCCCACCUGACCCCAUGCCGAACUCAGAAGUGAAACGCCGUAGCGCCGAUGGUAGUGUGGGGUCUCCCCAUGCGAGAGUAGGGAACUGCCAGGCAU
((((((((((.(.(((((((((((((((((((((((((.....))))))))))).)))))))))).)))(((((((((((((((((...))))))))))))))))).)))))))))))).  ( 1st)
>E.coli 4
UGUCUGGCGGCAGUAGCGCGGUGGUCCCACCUGACCCCAUGCCGAACUCAGAAGUGAAACGCCGUAGCGCCGAUGGUAGUGUGGGGUCUCCCCAUGCGAGAGUAGGGAACUGCCAGACAU
((((((((((.(.(((((((((((((((((((((((((.....))))))))))).)))))))))).)))(((((((((((((((((...))))))))))))))))).)))))))))))).  ( 1st)
>E.coli 5
UGCCUGGCGGCCUUAGCGCGGUGGUCCCACCUGACCCCAUGCCGAACUCAGAAGUGAAACGCCGUAGCGCCGAUGGUAGUGUGGGGUCUCCCCAUGCGAGAGUAGGGAACUGCCAGGCAU
((((((((((.(.(((((((((((((((((((((((((.....))))))))))).)))))))))).)))(((((((((((((((((...))))))))))))))))).)))))))))))).  ( 6th)
>E.coli 6
UGUCUGGCGGCAGUAGCGCGGUGGUCCCACCUGACCCCAUGCCGAACUCAGAAGUGAAACGCCGUAGCGCCGAUGGUAGUGUGGGGACUCCCCAUGCGAGAGUAGGGAACUGCCAGACAU
((((((((((.(.(((((((((((((((((((((((((.....))))))))))).)))))))))).)))(((((((((((((((((...))))))))))))))))).)))))))))))).  ( 1st)
>E.coli 10
UGUCUGGCGGCAGUAGCGCGGUGGUCCCACCUGACCCCAUGCCGAACUCAGAAGUGAAACGCCGUAGCGCCGAUGGUAGUGUGGGGUCUCCUCAUGCGAGAGUAGGGAACUGCCAUGCAU
((((((((((.(.(((((((((((((((((((((((((.....))))))))))).)))))))))).)))(((((((((((((((((...))))))))))))))))).)))))))))))).  ( 1st)
>E.coli 11
UGCCUGGCGGCAGUAGCGCGGUGGUCCCACCUGACCCCAUGCCGAACUCAGAAGUGAAACGCCGUAGCGCCGAUGGUAGUGUGGGGUCUCCCCAUGCGAGAGUAGGGAACUGCCAGGCAUCA
((((((((((.(.(((((((((((((((((((((((((.....))))))))))).)))))))))).)))(((((((((((((((((...))))))))))))))))).))))))))))))...( 1st)
>E.coli 14
UGCCUGGCGGCCGUAGCGCGGUGGUCCCACCUGACCCCAUGCCGAACUCAGAAGUGAAACGCCGUAGCGCCGAUGGUAGUGUGGGGUCUCCCCAUGCGAGAGUAGGGAACUGCCAGACAU
((((((((((.(..((((((((((((((((((((((((.....))))))))))).)))))))))).)))(((((((((((((((((...)))))))))))))))))..))))))))))).  (64th)
```

Figure S7 | Unconstrained MC-Cons consensus assignment for the *E. coli* 5S rRNA. The nine sequences were obtained from the 5S ribosomal RNA database. Each sequence was submitted to MC-Fold. The top 100 structures for each sequence were then submitted to MC-Cons. The consensus structure resembles that deduced from structural probing in solution and computer modelling by Brunel et al. (*J. Mol. Biol*. **221**, 293-308, 1991). For each consensus structure, the MC-Fold rank is shown in parenthesis. MC-Cons real time = 508 sec.

```
>Se1
CCCAGAUGAUGGCUUCACUGCUUGAUGGG
((((...(((((((....)))))))))))   (   4th)
....x..........xx............
>Se3
CCCAGAUGAUGCUUUAUCAGGCGGAUGGG
((((...(((((((....)))))))))))   (   1st)
....x.........x..............
>Se5
CCCAGAUGAUAGUGAGGCGCGGCUUGAUGGG
((((...(((((((.....).)))))))))) (  14th)
....x.........xxx..............
>Se6
CCCAGAUGAUAGUAAGGCGCGGCUUGAUGGG
((((...(((((((.(...)))))))))))) (   3rd)
....x.........x................
>Se7
CCCAGAUGAUCCGACGCGCUUUGGUGAUGGG
((((...((((((((....)))))))))))) (   4th)
....x.............x............
```

Figure S8 | MC-Fold predictions for the SECIS element. Positions marked with 'x' have high reactivity to single-stranded enzymatic probing, and are penalized by 8 kcal/mol if they are found base-paired in MC-Fold solutions. The nucleotides that participate in the formation of the K-turn motif are shown in bold. MC-Cons real time = 23 seconds.

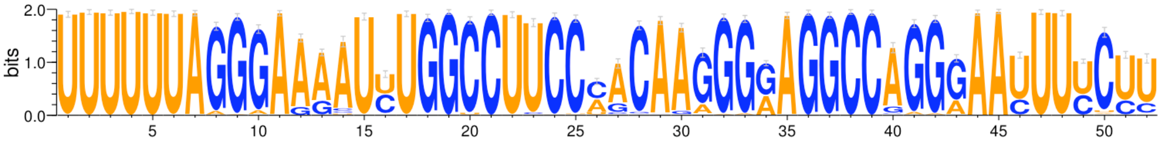


Figure S9 | The sequence variations observed in 753 HIV-1 frame-shifting elements. The sequences were obtained from Rfam. The slippery sequence is located in positions 1-7. The G(G|A)A bulge is the NMR model is located at positions 42-44. The AA bulge in our model is located at positions 44-45. The drawing was made with WebLogo (http://weblogo.berkeley.edu/logo.cgi).

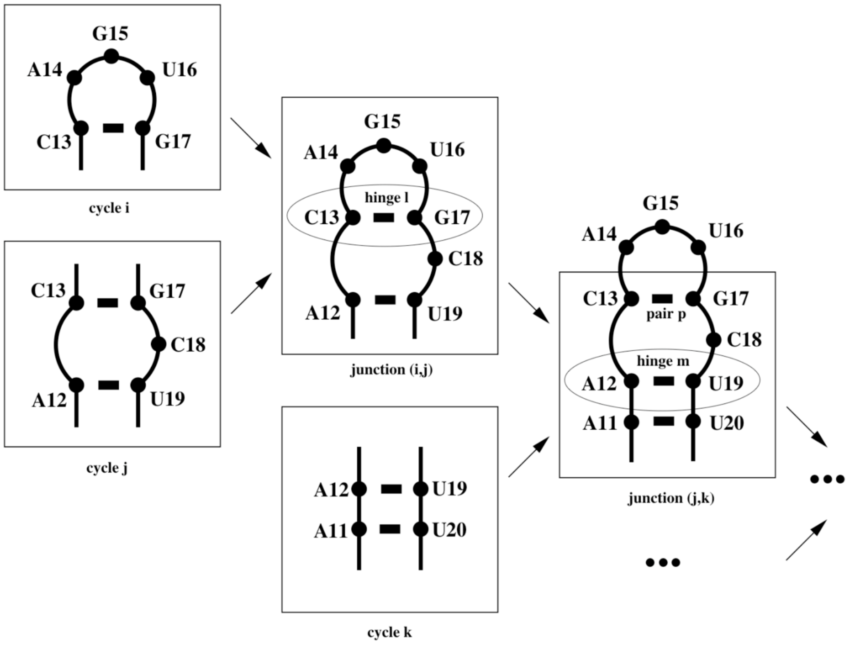


Figure S10 | Cycles, junctions, hinges, and base pairs. The dots represent nucleotides and the thick lines base pairs. Two NCMs (left), i and j, are joined, defining a junction (center above), (i, j), which includes a hinge (center above), l, and corresponding common pair (right), p. The junction (i, j) and the hinge l are valid, and thus a new NCM (center below), k, can be added. The arrows indicate the formation of junctions. The hinges are highlighted using ovals. The sum needed to compute the score resulting from this particular combination are:

1. $\Psi(5-NCM\,|\,"CAGUG")$, the probability of observing a 5-NCM given "ACGUU".
2. $\Psi(2\_3-NCM\,|\,"ACGUU")$, the score of assigning a 2_3-NCM to "ACGUU".
3. $\Psi(junction_{(i,j)}\,|\,5-ACGUU,2\_3-ACGUU)$, the probability of observing a junction between 5-ACGUU and 3_2-ACGUU.
4. $\Psi(CG\,|\,junction_{(i,j)})$, the probability of observing a CG base pair in junction$_{(i,j)}$.
5. $\Psi(CG\,|\,hinge_{l})$, the probability of observing a CG base pair given hinge$_{l}$.

Tab. S3 shows how the hinge probability is determined when a base pair tandem is added to a GNRA tetraloop.

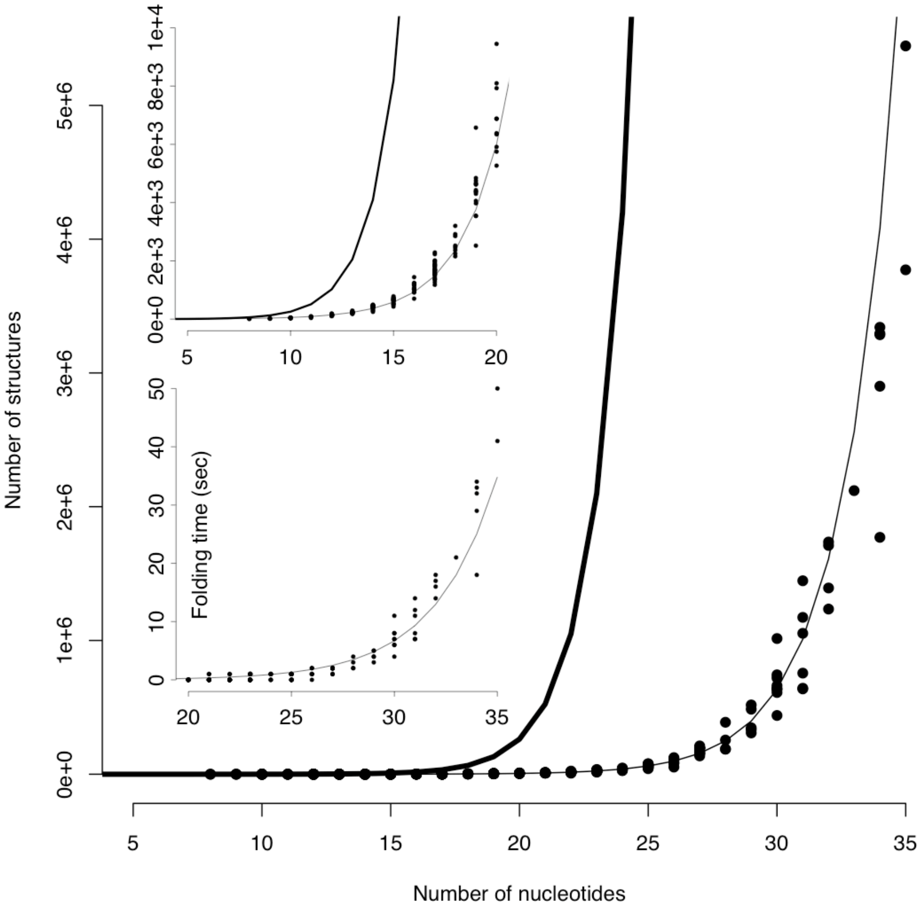


Figure S11 | Number of structure vs. sequence lenght. The number of secondary structures generated by MC-Fold versus the length of hairpin sequences. Each dot represents one hairpin. The curve for hairpins of 1 to 20 nucleotides is zoomed (inset above). The thick line shows the theoretical number of structures approximated by an exponential least square fit. The time required to compute the hairpin structures is proportional to the number of generated structures (inset below).

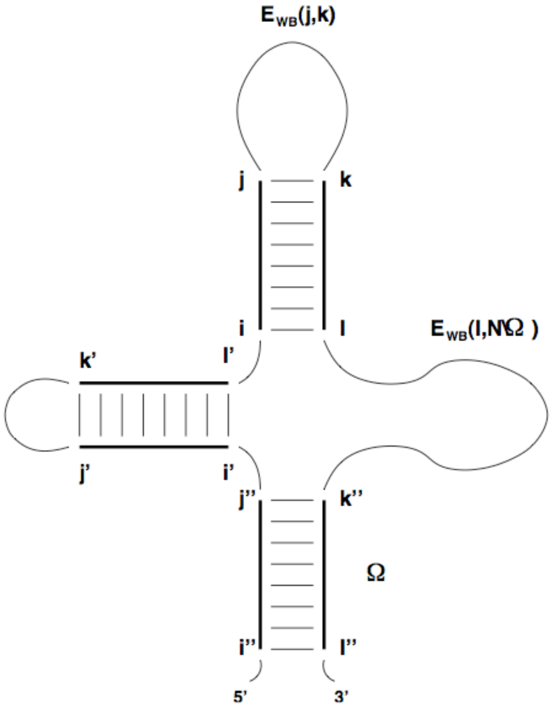


Figure S12 | Multi-branch construction. Stems are represented by *i* < *j* < *k* < *l*. E$_{WB}$(*j,k*) represents the best energy between positions j to k, as found in the dynamic programming table at entry (*j,k*). Ω represents the positions that were previously assigned in stems.

a.

```
> mcfold "GGGUGCUCAGUACGAGAGGAACCGCACCC"

Explored 1232736 structures in 00:00:23.

Top 10 solutions:

GGGUGCUCAGUACGAGAGGAACCGCACCC

(((((((((.((((..))))))))))))) -27.90
(((((((.(.((((..))))).))))))) -26.86
(((((((.((((((..))))))))))))) -26.86
(((((((((.(((....)))))))))))) -26.68
((((((((.(((((..))))))))))))) -26.68
((((((.(((((((..))))))))))))) -26.56
(((((((..(((((..))).))))))))) -26.15
((((((..((((((..)))))).)))))) -25.95
((((((.((.((((..)))))).)))))) -25.93
(((((((..(((((..)))).)))))))) -25.87
```

b.

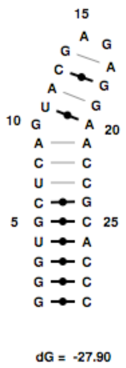


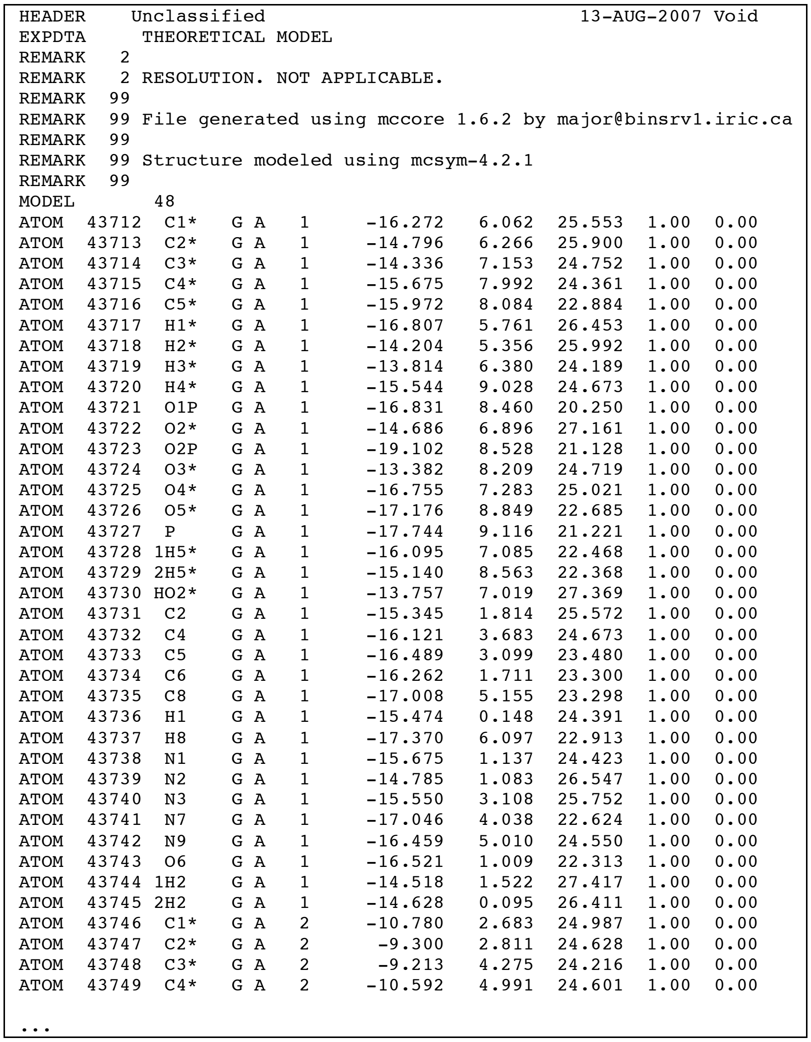
Figure S13 | MC-Fold call and output. **a**. MC-Fold is invoked in a Unix shell with the sequence of the rat 28S rRNA Loop E. The structures are generated, evaluated, and sorted by energies, indicated by the numbers on the right of each solution shown in dot-bracket notation. The number of solutions returned is an option of the program, 10 is the default value. **b**. Secondary structure of the best solution. A dot-bracket can be converted in a secondary structure representation. The dotted lines represent canonical base pairs; the lines non-canonical base pairs.

```
> mcsym IRE.mcc
```

```
//========== Sequence ==========
sequence( r A1 GGAGUGCUUCAACAGUGCUUGGACGCUCC )
//             ((((((.((((((...).)))))))))))
//========== NCMs ==========

ncm_01 = library(
        pdb( "MCSYM-DB/5/CAGUG/*.pdb.gz" ) #1:#5 <- A13:A17
        rmsd( 0.1 sidechain && !( pse || lp || hydrogen ) ) )
ncm_02 = library(
        pdb( "MCSYM-DB/2_3/ACGCU/*.pdb.gz" ) #1:#2, #3:#5 <- A12:A13, A17:A19
        rmsd( 0.1 sidechain && !( pse || lp || hydrogen ) ) )
ncm_03 = library(
        pdb( "MCSYM-DB/2_2/AAUU/*.pdb.gz" ) #1:#2, #3:#4 <- A11:A12, A19:A20
        rmsd( 0.5 sidechain && !( pse || lp || hydrogen ) ) )
ncm_04 = library(
        pdb( "MCSYM-DB/2_2/CAUG/*.pdb.gz" ) #1:#2, #3:#4 <- A10:A11, A20:A21
        rmsd( 0.5 sidechain && !( pse || lp || hydrogen ) ) )
ncm_05 = library(
        pdb( "MCSYM-DB/2_2/UCGG/*.pdb.gz" ) #1:#2, #3:#4 <- A9:A10, A21:A22
        rmsd( 0.1 sidechain && !( pse || lp || hydrogen ) ) )
ncm_06 = library(
        pdb( "MCSYM-DB/2_2/UUGA/*.pdb.gz" ) #1:#2, #3:#4 <- A8:A9, A22:A23
        rmsd( 0.1 sidechain && !( pse || lp || hydrogen ) ) )
ncm_07 = library(
        pdb( "MCSYM-DB/3_2/GCUAC/*.pdb.gz" ) #1:#3, #4:#5 <- A6:A8, A23:A24
        rmsd( 0.1 sidechain && !( pse || lp || hydrogen ) ) )
ncm_08 = library(
        pdb( "MCSYM-DB/2_2/UGCG/*.pdb.gz" ) #1:#2, #3:#4 <- A5:A6, A24:A25
        rmsd( 0.1 sidechain && !( pse || lp || hydrogen ) ) )
ncm_09 = library(
        pdb( "MCSYM-DB/2_2/GUGC/*.pdb.gz" ) #1:#2, #3:#4 <- A4:A5, A25:A26
        rmsd( 0.1 sidechain && !( pse || lp || hydrogen ) ) )
ncm_10 = library(
        pdb( "MCSYM-DB/2_2/AGCU/*.pdb.gz" ) #1:#2, #3:#4 <- A3:A4, A26:A27
        rmsd( 0.5 sidechain && !( pse || lp || hydrogen ) ) )
ncm_11 = library(
        pdb( "MCSYM-DB/2_2/GAUC/*.pdb.gz" ) #1:#2, #3:#4 <- A2:A3, A27:A28
        rmsd( 0.5 sidechain && !( pse || lp || hydrogen ) ) )
ncm_12 = library(
        pdb( "MCSYM-DB/2_2/GGCC/*.pdb.gz" ) #1:#2, #3:#4 <- A1:A2, A28:A29
        rmsd( 0.5 sidechain && !( pse || lp || hydrogen ) ) )
//=========== Backtrack ===========
stem_01 = backtrack(
        ncm_01
        merge( ncm_02 0.3 )
        merge( ncm_03 0.3 )
        merge( ncm_04 0.3 )
        merge( ncm_05 0.3 )
        merge( ncm_06 0.3 )
        merge( ncm_07 0.3 )
        merge( ncm_08 0.3 )
        merge( ncm_09 0.3 )
        merge( ncm_10 0.3 )
        merge( ncm_11 0.3 )
        merge( ncm_12 0.3 ) )
// ========= Constraints / Restraints =========
clash                 ( stem_01 1.5 !( pse || lp || hydrogen ) )
ribose_rst            ( stem_01 method = ccm, threshold = 0.2, pucker = C3p_endo )
backtrack_rst         ( stem_01 method = probabilistic )
implicit_phosphate_rst( stem_01 sampling = 90% )
// ========= Search =========
explore(
        stem_01
        option( model_limit = 5000, time_limit = 24h )
        rmsd( 1.2 sidechain && !( pse || lp || hydrogen ) )
        pdb( "Build/IRE" zipped ) )
```

Figure S14 | MC-Sym input script for the IRE consensus sequence. This script has been generated by MC-Fold. It can be submitted to MC-Sym without any editing. It produces the 3-D structure of the main manuscript Fig. 2a, shown superimposed with an NMR structure of the IRE.



```
HEADER    Unclassified                           13-AUG-2007 Void
EXPDTA     THEORETICAL MODEL
REMARK   2
REMARK   2 RESOLUTION. NOT APPLICABLE.
REMARK  99
REMARK  99 File generated using mccore 1.6.2 by major@binsrv1.iric.ca
REMARK  99
REMARK  99 Structure modeled using mcsym-4.2.1
REMARK  99
MODEL       48
ATOM  43712  C1*   G A   1     -16.272   6.062  25.553  1.00  0.00
ATOM  43713  C2*   G A   1     -14.796   6.266  25.900  1.00  0.00
ATOM  43714  C3*   G A   1     -14.336   7.153  24.752  1.00  0.00
ATOM  43715  C4*   G A   1     -15.675   7.992  24.361  1.00  0.00
ATOM  43716  C5*   G A   1     -15.972   8.084  22.884  1.00  0.00
ATOM  43717  H1*   G A   1     -16.807   5.761  26.453  1.00  0.00
ATOM  43718  H2*   G A   1     -14.204   5.356  25.992  1.00  0.00
ATOM  43719  H3*   G A   1     -13.814   6.380  24.189  1.00  0.00
ATOM  43720  H4*   G A   1     -15.544   9.028  24.673  1.00  0.00
ATOM  43721  O1P   G A   1     -16.831   8.460  20.250  1.00  0.00
ATOM  43722  O2*   G A   1     -14.686   6.896  27.161  1.00  0.00
ATOM  43723  O2P   G A   1     -19.102   8.528  21.128  1.00  0.00
ATOM  43724  O3*   G A   1     -13.382   8.209  24.719  1.00  0.00
ATOM  43725  O4*   G A   1     -16.755   7.283  25.021  1.00  0.00
ATOM  43726  O5*   G A   1     -17.176   8.849  22.685  1.00  0.00
ATOM  43727  P     G A   1     -17.744   9.116  21.221  1.00  0.00
ATOM  43728 1H5*   G A   1     -16.095   7.085  22.468  1.00  0.00
ATOM  43729 2H5*   G A   1     -15.140   8.563  22.368  1.00  0.00
ATOM  43730 HO2*   G A   1     -13.757   7.019  27.369  1.00  0.00
ATOM  43731  C2    G A   1     -15.345   1.814  25.572  1.00  0.00
ATOM  43732  C4    G A   1     -16.121   3.683  24.673  1.00  0.00
ATOM  43733  C5    G A   1     -16.489   3.099  23.480  1.00  0.00
ATOM  43734  C6    G A   1     -16.262   1.711  23.300  1.00  0.00
ATOM  43735  C8    G A   1     -17.008   5.155  23.298  1.00  0.00
ATOM  43736  H1    G A   1     -15.474   0.148  24.391  1.00  0.00
ATOM  43737  H8    G A   1     -17.370   6.097  22.913  1.00  0.00
ATOM  43738  N1    G A   1     -15.675   1.137  24.423  1.00  0.00
ATOM  43739  N2    G A   1     -14.785   1.083  26.547  1.00  0.00
ATOM  43740  N3    G A   1     -15.550   3.108  25.752  1.00  0.00
ATOM  43741  N7    G A   1     -17.046   4.038  22.624  1.00  0.00
ATOM  43742  N9    G A   1     -16.459   5.010  24.550  1.00  0.00
ATOM  43743  O6    G A   1     -16.521   1.009  22.313  1.00  0.00
ATOM  43744 1H2    G A   1     -14.518   1.522  27.417  1.00  0.00
ATOM  43745 2H2    G A   1     -14.628   0.095  26.411  1.00  0.00
ATOM  43746  C1*   G A   2     -10.780   2.683  24.987  1.00  0.00
ATOM  43747  C2*   G A   2      -9.300   2.811  24.628  1.00  0.00
ATOM  43748  C3*   G A   2      -9.213   4.275  24.216  1.00  0.00
ATOM  43749  C4*   G A   2     -10.592   4.991  24.601  1.00  0.00

...
```

Figure S15 | Header of a PDB file generated by MC-Sym.

References

1. Nussinov, R. & Jacobson, A. B. Fast algorithm for predicting the secondary structure of single-stranded RNA. *Proc Natl Acad Sci U S A* **77**, 6309-6313 (1980).
2. Waterman, M. S. & Byers, T. H. A dynamic programming algorithm to find all solutions in the neighborhood of the optimum. *Math. Biosci.* **77**, 179-188 (1985).
3. Haralick, R. & Elliott, G. Increasing tree search efficiency for constraint satisfaction problems. *Artificial Intelligence* **14**, 263-313 (1980).
4. Aalberts, D. P. & Hodas, N. O. Asymmetry in RNA pseudoknots: observation and theory. *Nucleic Acids Res.* **33**, 2210-2214 (2005).
5. Berman, H. M. et al. The Protein Data Bank. *Nucleic Acids Res.* **28**, 235-242 (2000).
6. Reeder, J. & Giegerich, R. Consensus shapes: an alternative to the Sankoff algorithm for RNA consensus structure prediction. *Bioinformatics* **21**, 3516-3523 (2005).
7. Bazaraa, M. S., Sherali, H. D., & Shetty, C. M., *Nonlinear pogramming theory and algorithms*, 3rd ed. (John Wiley & Sons, Inc., Hoboken, NJ, 2006).
8. Staple, D. W. & Butcher, S. E. Solution structure and thermodynamic investigation of the HIV-1 frameshift inducing element. *J Mol Biol* **349**, 1011-1023 (2005).
9. Nissen, P. et al. RNA tertiary interactions in the large ribosomal subunit: The A-minor motif. *Proc. Natl Acad. Sci. U S A.* **98**, 4899-4903 (2001).
10. Elgavish, T. et al. AA.AG@helix.ends: A:A and A:G base-pairs at the ends of 16 S and 23 S rRNA helices. *J Mol Biol* **310**, 735-753 (2001).
11. Gaudin, C. et al. Structure of the RNA signal essential for translational frameshifting in HIV-1. *J Mol Biol* **349**, 1024-1035 (2005).
12. Dulude, D., Baril, M., & Brakier-Gingras, L. Characterization of the frameshift stimulatory signal controlling a programmed -1 ribosomal frameshift in the human immunodeficiency virus type 1. *Nucleic Acids Res* **30**, 5094-5102 (2002).
13. Baril, M. et al. Efficiency of a programmed -1 ribosomal frameshift in the different subtypes of the human immunodeficiency virus type 1 group M. *Rna* **9**, 1246-1253 (2003).
14. Hofacker, I. L. Vienna RNA secondary structure server. *Nucleic Acids Res.* **31**, 3429-3431 (2003).
15. Girnary, R. et al. Structure-function analysis of the ribosomal frameshifting signal of two human immunodeficiency virus type 1 isolates with increased resistance to viral protease inhibitors. *J. Gen. Virol.* **88**, 226-235 (2007).
16. Do, C. B., Woods, D. A., & Batzoglou, S. CONTRAfold: RNA secondary structure prediction without physics-based models. *Bioinformatics* **22**, e90-98 (2006).