Modeling and design by hierarchical natural moves

Adelene Y. L. Sim[a], Michael Levitt[b], and Peter Minary[b,1]

[a]Department of Applied Physics, Stanford University, Stanford, CA 94305; and [b]Department of Structural Biology, Stanford University, Stanford, CA 94305-5126

Contributed by Michael Levitt, December 5, 2011 (sent for review November 16, 2011)

We develop a unique algorithm implemented in the program MOSAICS (Methodologies for Optimization and Sampling in Computational Studies) that is capable of nanoscale modeling without compromising the resolution of interest. This is achieved by modeling with customizable hierarchical degrees of freedom, thereby circumventing major limitations of conventional molecular modeling. With the emergence of RNA-based nanotechnology, large RNAs in all-atom representation are used here to benchmark our algorithm. Our method locates all favorable structural states of a model RNA of significant complexity while improving sampling accuracy and increasing speed many fold over existing all-atom RNA modeling methods. We also modeled the effects of sequence mutations on the structural building blocks of tRNA-based nanotechnology. With its flexibility in choosing arbitrary degrees of freedom as well as in allowing different all-atom energy functions, MOSAICS is an ideal tool to model and design biomolecules of the nanoscale.

hierarchical sampling | junctions | molecular simulation | Monte Carlo | nanostructure

Computational modeling is an important aspect of biology and nanotechnology. *In silico* design and manipulation of nanostructures often precedes experimental validation (1, 2), while effective computational structure prediction of biomolecules paves the way for biomolecular design (3). Most molecular modeling packages are either general but inefficient in modeling large molecular systems or designed to be effective for modeling only specific types of systems (e.g., coarse-grained modeling of DNA as an elastic rod) and are therefore not general purpose. These limitations are mainly a consequence of two major obstacles to computational modeling: (*i*) the high dimensionality of the systems studied and (*ii*) the complexity of the potential energy surface guiding the simulation. When attempting to model nanoscale systems at all-atom resolution, the combination of these two factors often leads to intractable complexity.

A wide range of methods has been proposed to remove obstacles presented by high dimensionality (4–6) and complex energy surfaces (7–12), but none of these studies has considered both limitations in the same context. The difficulty of such a unified approach is that limitations arising from both obstacles are related in that reducing dimensionality often results in a more complex energy surface. For example, conformational sampling using dihedral angles, which is a common solution to the high dimensionality problem of macromolecular assemblies, often results in a more complex energy surface with energy barriers that could have been easily avoided in Cartesian space. The algorithm [implemented in MOSAICS (13)] proposed here overcomes the problem of high dimensionality without increasing the complexity of the underlying energy surface. This is achieved by sampling with hierarchical variables: Degrees of freedom that introduce large conformational changes are combined with degrees of freedom that allow for local rearrangement. The former help overcome obstacles raised by large dimensionality while the latter soften the resultant energy surface. Such degrees of freedom may break the molecular chain and even spoil stereochemistry. To prevent this, we have coupled hierarchical modeling with a stochastic chain-closure algorithm (6), permitting extensive *in silico* manipulation of large molecular structures at all-atom resolution without compromising chain connectivity or stereochemistry.

Our algorithm is general and therefore can be applied to diverse types of molecular structures. However, to benchmark our technique, we turn our attention to molecular junctions, a molecular topology pervasive in biology (14) and electronics (15–17). Given the recent emergence of RNA-based nanotechnology (1), we focus on junctions found in RNA and compare our technique with other RNA computational modeling approaches. Naturally occurring RNA molecules are important in biology as they carry out a variety of roles in gene regulation while synthetically designed RNA nanostructures have increasingly been promoted as possible candidates for drug delivery (1). RNA molecules often fold in a hierarchical fashion, with the sequence determining stable base-pairings (secondary structure) that are preserved when these base-paired regions rearrange into tertiary or quaternary forms (18); RNA nanostructures are usually designed based on this premise and through the use of specific RNA structural motifs (2). Hence the hierarchical structural elements we used (see below) to define degrees of freedom for modeling RNA are natural physical descriptors of RNA (19). We emphasize that our method is general as the hierarchical organization of RNA assemblies is only one example of a more general phenomenon manifested in nanotechnology, where nanostructures are built from basic blocks that are themselves built hierarchically [e.g., nucleic acids (20, 21) or oxide nanostructures (22)].

Results and Discussion

Benchmarking with an RNA Four-Way Junction. We show that accurate sampling of a large RNA structure with all-atom representation and secondary structure constraints is impossible with currently available approaches. Specifically, they fail to generate diverse and canonical ensembles (Fig. 1). Regular Monte Carlo, molecular dynamics, and natural move Monte Carlo (4, 6) are generally confined to local structural basins close to the initial model (Fig. 1*A*). Fragment assembly, which is designed to generate a diverse set of conformations and commonly used for RNA (3, 23) and protein modeling (24), does not generate a canonically weighted ensemble of conformations because the preference for certain structures may depend on the particular fragment libraries used. As a result, thermodynamic observables cannot be calculated from such ensembles. Two state-of-the-art implementations of fragment assembly, Macromolecular Conformations by SYMbolic programming or MC-Sym (23) and Rosetta (3), produce differently biased distance distributions (Fig. 1 *B* and *C*). Both distributions feature unphysical cusps and additionally for MC-Sym, symmetry is not preserved. By contrast, our method, hierarchical natural move Monte Carlo as implemented in MOSAICS (13), has been designed to efficiently

Author contributions: A.Y.L.S., M.L., and P.M. designed research; A.Y.L.S. and P.M. performed research; A.Y.L.S. and P.M. contributed new reagents/analytic tools; A.Y.L.S., M.L., and P.M. analyzed data; and A.Y.L.S., M.L., and P.M. wrote the paper.

The authors declare no conflict of interest.

Freely available online through the PNAS open access option.

See Commentary on page 2691.

[1]To whom correspondence should be addressed. E-mail: peter.minary@stanford.edu.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1119918109/-/DCSupplemental.

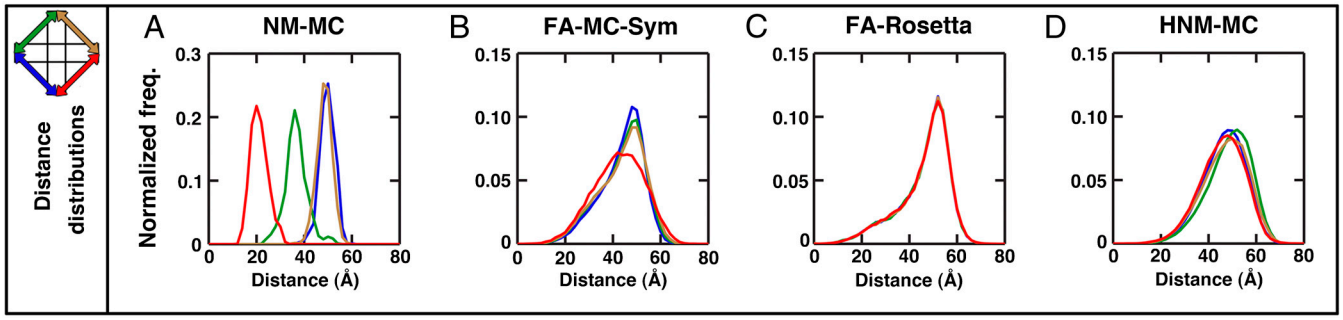


Fig. 1. Sampling a simple four-way RNA junction using four different protocols. The distributions of the four color-coded distances between the ends of helices are shown (secondary structure in Fig. S1 and additional distances in Fig. S2). If sampling is unbiased, these distributions should be smooth and identical. (*A*) Sampling with the natural move Monte Carlo (NM-MC) (4, 6) method produces smooth distributions, but they are not identical because convergence is not reached within the total number of iterations. (*B*) Fragment assembly by MC-Sym (FA-MC-Sym) (23) explores conformational space more effectively than NM-MC but cannot be used for thermodynamic sampling. MC-Sym builds RNA structures cyclically: After placement of the last helix, the acceptance criterion does not include chain connectivity but only distance constraints. Hence, the distributions for the first three distances are identical because they were derived from the same fragment libraries while the fourth distance (labeled in red) is not. (*C*) Identical distributions are obtained in the fragment assembly implementation of Rosetta (FA-Rosetta) (3), but there are nonphysical cusps in the distributions likely due to the use of native RNA fragments. (*D*) Sampling with our new hierarchical natural move Monte Carlo (HNM-MC) as implemented in MOSAICS (13) gives distance distributions that are identical and smooth. In all cases, results for 50,000 iterations are shown. Hard sphere all-atom potentials were used for both NM-MC and HNM-MC.

model large molecules and produce the canonical distribution of observables, regardless of starting structure (Fig. 1*D*).

Our hierarchical natural move Monte Carlo first generates a complex move of the molecular system by ignoring molecular connectivity, such that the system is likened to a collection of independent nucleotides, which are sequentially moved based on the user-defined degrees of freedom (a collection of independent sets of moves, L_i). This complex move may result in spoiling stereochemistry and chain breakage between any neighboring nucleotides, but these are corrected by our all-atom chain-closure algorithm (6) that adjusts a small number of dependent variables (L_d) before the actual energy evaluation and acceptance/rejection of the complex Monte Carlo move along L_i take place. (If chain closure is unsuccessful, the proposed complex move is immediately rejected.) Therefore the degrees of freedom used are the combination of both L_i and L_d.

For nucleic acids, an obvious degree of freedom is one that describes the position, orientation, and internal flexibility of each nucleotide (set L_1; Fig. 2*A* and ref. 6). Although such a set of natural moves uses a small number of essential degrees of freedom to reduce dimensionality [by an order of magnitude compared to Cartesian coordinates (6)], high dimensionality and energy surface complexity still present major challenges for large macromolecules. It may seem possible to avoid the dimensionality problem by introducing new degrees of freedom (L_2) that move larger parts of the structure. In the case of RNA, these degrees of freedom could describe the relative orientation and absolute position of all its rigid base-paired helices (Fig. 2*A*). Unfortunately, sampling helices as rigid bodies alone is inadequate: Tight packing of two helices may be prevented by the artificially high energy barriers that arise from the lack of flexibility within these rigid bodies.

To overcome this limitation we embed the smaller, local moves as described by L_1 into the more global sampling by L_2 moves; i.e., our complex independent move L_i is a combination of moves from L_1 and L_2. In this way each RNA helix is no longer completely rigid and nucleotides within each helix are allowed to move. Sampling along collective degrees of freedom (L_2) solves

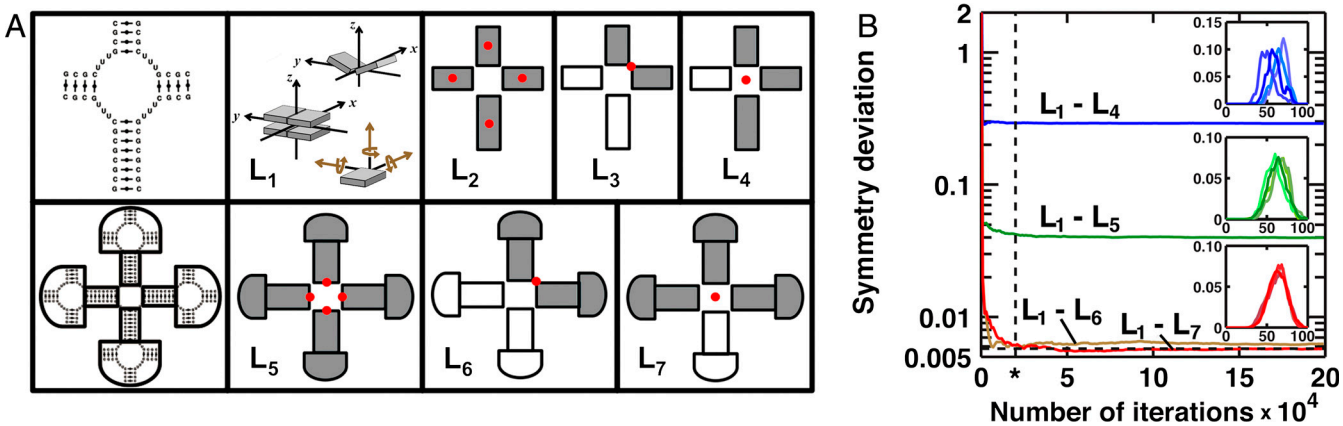


Fig. 2. Effects of adding hierarchical degrees of freedom on sampling a large symmetric RNA structure. (*A*) Hierarchical moves used. A system of this complexity has many possible collective motions. Here seven sets of independent degrees of freedom (L_1 to L_7) are defined. The base-pairs and individual nucleotides have their natural degrees of freedom (L_1). Regions of continuous base-pairing form helices (represented by rectangles), and these helices can be regarded as rigid bodies moving either independently (L_2), as pairs of helices (L_3), or as groups of three (L_4) or more helices (L_5 to L_7). Rigid body (gray) motion requires the definition of rotation centers shown as red dots. These centers are selected to preserve the symmetry of the system. (*B*) Convergence is accelerated by higher order rigid body moves. When nested hierarchical moves L_1 to L_7 were used, rapid convergence to the limiting distribution (horizontal dashed line showing the limiting "symmetry deviation" as defined in *Materials and Methods*) is reached within 2×10^4 iterations (vertical dashed line labeled *). Sampling with fewer nested hierarchical moves (L_1-L_4 and L_1-L_5) did not converge with 2×10^5 iterations, suggesting that at least a 10-fold speedup is achieved by using the additional sets L_5 to L_7. Inset in *B* shows individual distributions of end-to-end distances between pairs of full arms (see *Materials and Methods*) with 2×10^4 iterations for three different sets of hierarchical moves: L_1-L_4 (blue), L_1-L_5 (green), and L_1-L_7 (red). The *y* axis is the normalized frequency; the *x* axis is the distance in Å.

the dimensionality problem, while the increased complexity of the energy surface is circumvented by introducing internal flexibility within rigid bodies (see *Materials and Methods*).

Analogously for protein complexes, global moves of protein domains (say L_3) can be incorporated with moving secondary structure elements in each domain independently (L_2). Further flexibility can be added with residue-level degrees of freedom (L_1).

Improving Sampling Efficiency by Adding Hierarchical Degrees of Freedom. As an illustration of the effects of modeling with different embedded degrees of freedom, consider a biomolecular assembly with many levels of hierarchical complexity like the large symmetric RNA system with multiple junctions illustrated in Fig. 2*A*. We can define seven arbitrary independent sets of degrees of freedom, L_1 to L_7. We use the distance between the ends of neighboring full arms that consist of four helices each as our physical observable. By symmetry, all distance distributions are expected to converge to the same distribution; the overlap between distributions is a quantitative indicator of reaching convergence with unbiased sampling (see *Materials and Methods*). Fig. 2*B* shows that successive incorporation of degrees of freedom that introduce additional collective global rearrangements systematically reduces the number of Monte Carlo iterations needed to reach convergence. Specifically, sampling with sets L_1 to L_7 leads to convergence using at least an order of magnitude less computational effort than sampling with only sets L_1 to L_4. The effects of introducing additional degrees of freedom appear to gradually saturate, as judged by the similarity between sampling with L_1 to L_6 and L_1 to L_7. This suggests that additional sets are not required.

Application 1: Modeling RNA Nanostructure Flexibility. While we made use of synthetic four-way RNA junctions as model systems to benchmark our technique, four-way junctions are common in naturally occurring biological nucleic acids. For instance, during homologous recombination, four-way DNA Holliday junctions are formed and have preferential stacking and orientations (14). The flexibility of RNA junctions has strong influence on its dynamics and fold and affects the RNA's functionality and capability to bind to different ligands (25). Therefore understanding the behavior of nucleic acid junctions has important biological implications and our method facilitates extensive *in silico* study of the physics of such systems.

Beyond naturally occurring systems, nucleic acid junctions are also important in bionanotechnology. There has been extensive research on DNA and RNA nanostructures because of their potential in a variety of fields like drug delivery (1), material science [e.g., DNA nanotubes (26)], biomolecular computing [e.g., algorithmic self-assembly (27)], and nanoelectronics (1, 27). Nanostructures consist of basic motifs of different junction types to design appropriate bends and kinks that are assembled into large and complex structures. The flexibility of these motifs typically depends on sequences of the single stranded regions connecting the helices and/or the presence of tertiary contacts that help stabilize the RNA/DNA junctions to particular conformations. From a practical standpoint related to experimental synthesis, it is important to understand the effects of specific sequences on the flexibility of the motif as the yield and stability of large nanostructures is affected by the flexibility of its smaller building blocks. For example, it was shown that a tRNA-square (four tRNA monomers connected via kissing loops to form a square; see Fig. 3) was less stable when its individual monomers carried sequence mutations that negated the tertiary interaction likely crucial in maintaining the right-angle motif within the monomer (28). This series of mutations also resulted in a decreased yield of the tRNA-square, instead facilitating formation of smaller dimers or trimers (28).

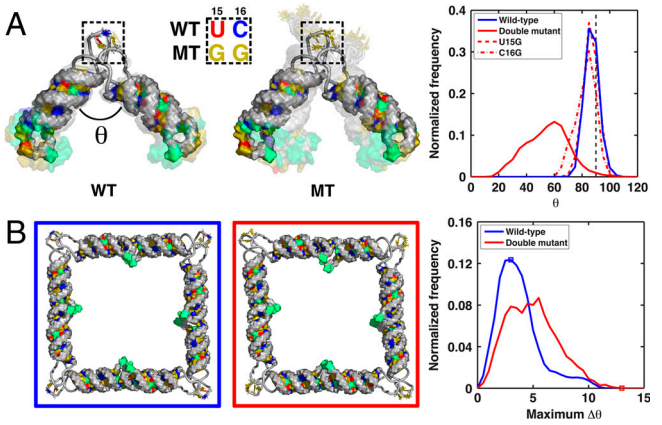


Fig. 3. Effects of sequence mutations on the flexibility of nanostructure building blocks. (*A*) The double mutation (MT) significantly increases the flexibility of the tRNA monomer compared to the wild type (WT). The initial starting structures are shown in solid while 10 randomly selected models from our simulations are shown in the background. The right-angle tRNA motif preferentially explores lower angles (θ) when both mutations U15G and C16G are done in concert. The increased flexibility of the monomer could facilitate formation of dimers or trimers, thereby decreasing the yield of the tRNA-square. (*B*) Based on our simulations, the double-mutant tRNA-square is more likely to explore distorted squares (extreme structure boxed in red; matching maximum $\Delta\theta$ shown in histogram plot) than its wild-type equivalent (most common structure boxed in blue), which could increase the strain on the kissing loop motifs that connect tRNA monomers together. This may explain the reduced stability of the double mutant tRNA-square found experimentally (28).

Using hierarchical move-sets and a simple base-pairing "potential" (see *Materials and Methods*) we show that single mutations of the tertiary contact do not strongly affect the angular distribution of the junction within each tRNA monomer (Fig. 3*A*). However, when both mutations U15G and C16G are made in concert, the tRNA monomer has significantly higher flexibility (see Fig. 3*A*), which could result in an increased ability to form complexes of different order, thereby decreasing tRNA-square yield. Further, based on thermal melting experiments, it was found that the same mutations affect the tRNA-square stability after it forms (28). Our simulations of the tRNA-square suggest that double mutations on each monomer result in more distorted squares than in the absence of any mutations (Fig. 3*B*). These distortions could lead to strains on the kissing loops that connect individual monomers, thereby decreasing the stability of the mutated tRNA-square relative to the native. Distortions in the tRNA-square could then propagate and result in lower yield and/or distortions of larger assemblies (such as a one-dimensional ladder).

Application 2: Interpreting Limited Experimental Data with Modeling Results. Another application of our algorithm is to run simulations with appropriate selection of degrees of freedom to aid the interpretation of experiments. As an example, we made use of unique move sets to identify the relative orientation of two-stacked helices within an RNA four-way junction (Fig. 4), specifically an RNA four-way junction derived from the hairpin ribozyme with its helix D stacked on A and its helix C on B. These two sets of stacked helices can orient in parallel or antiparallel, as determined by the direction of continuous RNA strands (pointing in the same or opposite directions, respectively) (14). The interhelical angle θ is 0 for the extreme parallel structure and 180° for the antiparallel equivalent. For $0 < \theta < 180°$, the junction can take either a left- or right-handed form (Fig. 4*A*). Using distance measurements between the ends of helices determined by fluorescence resonance energy transfer (FRET), experimentalists identified this junction as preferring the antiparallel conforma-

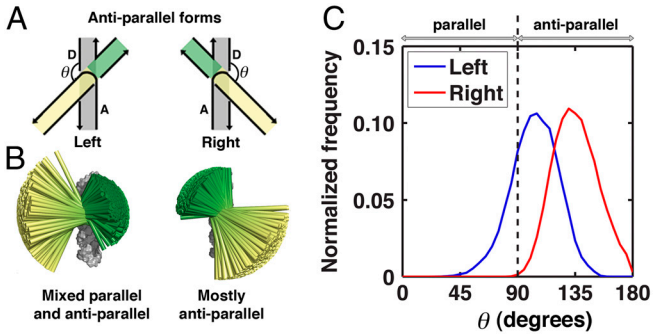


Fig. 4. Determining four-way junction handedness. The thorough sampling achieved by different types of hierarchical moves and constraints facilitates direct comparisons between modeling and experiments. (*A*) Experimental distances by fluorescence resonance electron transfer (FRET) indicated that a particular four-way junction preferentially has its helix D stacked on A and C (green) on B (yellow). The stacked helices are preferentially oriented in an antiparallel form ($\theta > 90^{\circ}$), but the distance measurements are unable to distinguish between the left- and right-handed antiparallel forms. The experimental data further showed that switching between antiparallel and parallel forms of the four-way junction does not take place in the absence of helices AD and BC unstacking. (*B*) Constrained sampling (helices kept stacked) starting from the left- and right- handed antiparallel forms ($90^{\circ} < \theta < 180^{\circ}$) indicates that the left-handed form was able to switch between parallel and antiparallel conformations (crossing $\theta = 90^{\circ}$) without the unstacking of helices. Conversely, the right-handed form was mostly confined to only antiparallel conformations. Models were superimposed relative to helix AD (gray) and each stacked helix BC is shown as a thin rod (yellow to green). For clarity, only results for 2,000 randomly chosen models are shown. (*C*) The distributions of rotation angles (θ) of helix BC relative to helix AD, as defined in *A*. When starting from a left-handed conformation the system easily switches between parallel ($\theta < 90^{\circ}$) and antiparallel ($90^{\circ} < \theta \leq 180^{\circ}$) forms. However, starting from a right-handed conformation, transitions between parallel and antiparallel forms were not observed in the absence of helix unstacking. Thus, our *in silico* results indicate that the right-handed antiparallel form of this four-way junction is most consistent with the FRET data (29).

tion in the presence of Mg^{2+}, but the handedness of the antiparallel form could not be determined from this experimental FRET data (29). In the same set of experiments, it was found that switching of the four-way junction from its antiparallel form to its corresponding parallel form does not take place in the absence of helix unstacking (i.e., rotation of stacked helix BC relative to AD is not sufficient).

In our RNA simulations (see *Materials and Methods*) that conserved the stacking of helix D on A and B on C, starting from left- and right-handed conformations, we found that the two types of handedness can be easily distinguished (Fig. 4*B*): With helices kept stacked, the left-handed form can rotate between parallel and antiparallel types, while the right-handed form is confined to mostly antiparallel conformations. It appears that steric and connectivity constraints imposed by the junction prevent the right-handed antiparallel conformation from rotating to its corresponding parallel type. Therefore, conversion from the right-handed antiparallel conformation to the right-handed parallel conformation requires more than a simple rotation of stacked helices, and helix unstacking may be necessary. Our results indicate that a right-handed antiparallel conformation is most consistent with the FRET data. Subsequent gel electrophoresis experiments showed that this four-way junction indeed prefers the right-handed antiparallel form (30).

Conclusion

We are currently applying our algorithm in the field of structural biology to model biological complexes with limited/low-resolution experimental information (such as cryoelectron microscopy and small-angle X-ray scattering density maps). A good ensemble of structures is required for such modeling, and our algorithm allows us to generate this ensemble even for large molecular systems. Moreover, due to the generality in our implementation, we can model structures with any scoring or energy function derived either from experimental input or by a physics-based or knowledge-based potential. Hence our hierarchical natural move Monte Carlo technique can be used in structure prediction of RNA and protein or else to assess the quality of molecular modeling force fields by comparing observables derived from *in silico* ensembles to experimentally determined ones. We are currently combining hierarchical natural move Monte Carlo with continuous sequence space sampling (treating the fractional presence of each type of nucleotide at each sequence position as an additional independent variable) so as to allow simultaneous exploration of sequence-structure space. We hope these approaches pave the way for large scale design of nanostructures (1, 2, 31) or molecules that mimic biological processes (3).

Materials and Methods

Descriptions about modeling the tRNA nanostructure and four-way junction derived from the hairpin ribozyme are provided in *SI Text*.

Sampling with Embedded Hierarchical Moves. The typical objective of sampling is to obtain the expectation value and distribution of system specific observables over a phase space (Ω) of interest that is described by a probability distribution function, $f: \Omega \to \Re$ (function f maps elements from phase space Ω to real numbers $\Re$). Then, for any observable, $\alpha: \Omega \to \Re$, the expectation value is defined as

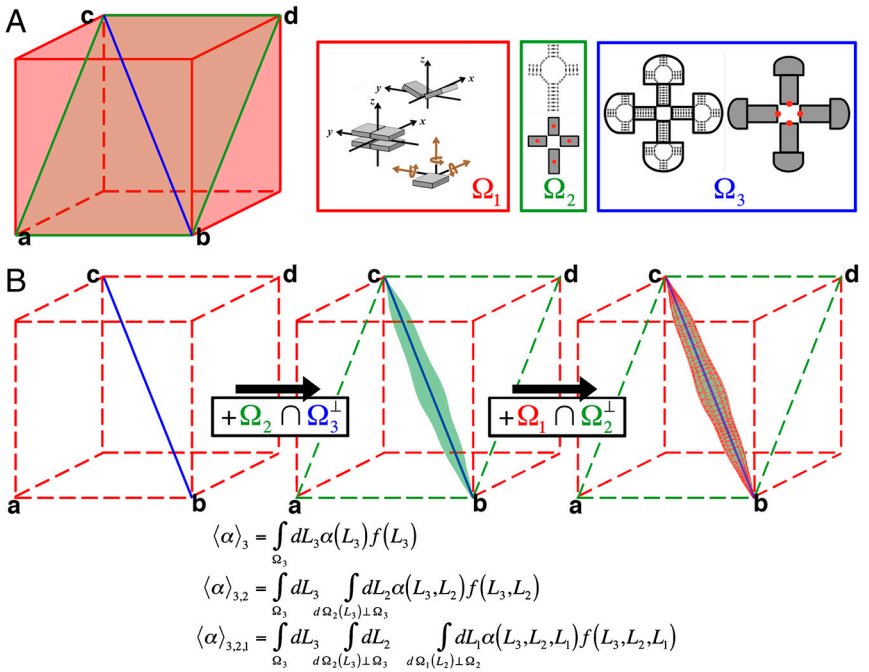
$$\langle \alpha \rangle = \int_{\Omega} dL \alpha(L) f(L) \qquad [1]$$

where L spans Ω. In the case of the canonical ensemble, the probability distribution function f is given by the Boltzmann distribution, $f(L) = \exp(-\beta E(L))/Q$, where the function $E: \Omega \to \Re$ represents the energy of the system, $\beta = 1/kT$ (k is the Boltzmann constant and T the temperature) and Q is the partition function defined as $Q = \int_{\Omega} dL \exp(-\beta E(L))$. Therefore,

$$\langle \alpha \rangle = 1/Q \int_{\Omega} dL \alpha(L) \exp(-\beta E(L)) \qquad [2]$$

A common method for obtaining $\langle \alpha \rangle$ is to average over all states visited by a Markov chain Monte Carlo procedure sampling from probability distribution f. In a brute force approach, such sampling is performed along a particular set of independent degrees of freedom, L_f, covering phase space, Ω_f, (i.e., L_f spans Ω_f) with $\Omega_f \subset \Omega$ (Ω_f is a proper subspace of Ω). In many practical applications, however, we are interested only in a proper subspace, $\Omega_s \subset \Omega_f$ spanned by a small number of generalized degrees of freedom, L_s (L_s spans Ω_s). For example, consider a long polymer chain modeled by an energy function. If the observable of interest is the end-to-end distance of the polymer, a convenient choice for Ω_s is the space spanned by the dihedral angles about single bonds (torsion angle space with constant bond lengths or angles). On the other hand, Ω_f, would be the space spanned by the Cartesian degrees of freedom, which do not constrain bond lengths or angles. Because displacements from ideal bond angles and lengths are governed by stiff harmonic forces, $\langle \alpha \rangle_f$ and $\langle \alpha \rangle_s$ (the expectation values of α inferred from Markov chain Monte Carlo in spaces Ω_f and Ω_s, respectively) should both provide sufficiently good approximations to $\langle \alpha \rangle$. However, the number of iterations required to reach convergence to give reliable expectation values and distributions of observables may vary significantly. In practical implementations, the large number of iterations required to sample the larger conformational space Ω_f makes the evaluation of Eq. **1** impractical.

While confining sampling to Ω_s circumvents most challenges posed by high dimensionality, it may make the energy surface more difficult to explore: The energy surface in phase space Ω_s might be more rugged than that in Ω_f. As a result it is likely that the probability of moving from one state to another in subspace Ω_s is much smaller than that in subspace Ω_f. For instance, in the previous polymer example, the difficulty in attaining practical Markov chain Monte Carlo acceptance rates when sampling dihedral angles is mainly due to the rough energy surface present in torsion angle space. However, augmenting Ω_s (our subspace of interest) with sufficiently small additional phase space volume elements from $\Omega_f \cap \Omega_s^{\perp}$ (the subspace of Ω_f that is orthogonal to Ω_s, where $\Omega_s^{\perp}$ refers to the orthogonal complement of Ω_s) may eliminate the sampling limitations arising from the rough energy surface in phase



$$\langle\alpha\rangle_3 = \int_{\Omega_3} dL_3 \alpha(L_3) f(L_3)$$

$$\langle\alpha\rangle_{3,2} = \int_{\Omega_3} dL_3 \int_{d\Omega_2(L_3)\perp\Omega_3} dL_2 \alpha(L_3, L_2) f(L_3, L_2)$$

$$\langle\alpha\rangle_{3,2,1} = \int_{\Omega_3} dL_3 \int_{d\Omega_2(L_3)\perp\Omega_3} dL_2 \int_{d\Omega_1(L_2)\perp\Omega_2} dL_1 \alpha(L_3, L_2, L_1) f(L_3, L_2, L_1)$$

Fig. 5. Sampling with embedded subspaces by appropriate choice of hierarchical degrees of freedom constrains phase space exploration. (*A*) An illustration of the different integration subspaces with $\Omega_3 \subset \Omega_2 \subset \Omega_1$: Ω_1 spans the volume of cube (red), Ω_2 spans the plane abcd (green) and Ω_3 spans the line cb (blue). The corresponding analogous subspaces defined by different move sets for a large symmetric RNA are also shown. Sampling full arms as rigid bodies (Ω_3 space) explores a wide range of conformations, but close proximity of these full arms (with internal helices kept rigid) could be unfavorable due to steric clashes. These clashes could be readily alleviated through independent movement of helices (in space Ω_2). Introducing nucleotide-level movement within helices (in space Ω_1) adds flexibility that could further smooth the sampling energy landscape. (*B*) Phase space Ω_1 is described by a probability distribution function, $f: \Omega_1 \rightarrow \Re$ and α: $\Omega_1 \rightarrow \Re$ is the observable of interest. The expectation value, $\langle\alpha\rangle$ is defined by an integral over Ω_1, thus $\langle\alpha\rangle$ can be referred to as $\langle\alpha\rangle_1$. If α is primarily dependent on the relative orientations of major arms, then the integrals shown are increasingly more accurate approximations to the expectation value, $\langle\alpha\rangle$. The integration volume for L_2 is $d\Omega_2(L_3)\perp\Omega_3$; it is centered on L_3 and is orthogonal to Ω_3, while its size depends on the properties of the function f (or equivalently on the topology of the corresponding energy surface). The integration volume for L_1 is $d\Omega_1(L_2)\perp\Omega_2$; it is centered on L_2 and is orthogonal to Ω_2.

space Ω_s and so allow energy barriers to be overcome more easily. An example of such a phenomenon is the softening of the torsion energy surface by adding bond angle degrees of freedom (6).

In general, additional degrees of freedom needed to smooth the rough energy surface of Ω_s can be chosen to exploit the hierarchical nature of biological/inorganic structure. Efficient sampling can be achieved with collective moves that change conformations at a higher hierarchical level while still including moves at lower levels. Thus by embedding moves of different levels, dimensionality is reduced without significantly changing the ruggedness of the energy surface. To do this, embedded subspaces of increasing sizes are defined, from Ω_s to Ω_f such that $\Omega_s = \Omega_n \subset \Omega_{n-1} \ldots\ldots \subset \Omega_1 = \Omega_f$, with each subspace representing a desired level of the hierarchy.

Fig. 5 illustrates an example with three embedded subspaces ($n = 3$): Ω_1, Ω_2, and Ω_3. The large symmetric RNA junction as discussed in the main text (Fig. 2) is a physical analog of such a scenario. In the full $\Omega_f = \Omega_1$ space (analogous to moving in the volume of the whole cube), each nucleotide is flexible and free to move independently (Fig. 5*A*). Two examples of smaller nested subspaces are moving helices as rigid bodies independently (Ω_2, analogous to moving in the titled plane *abcd*) or moving each full arm consisting of four helices simultaneously (Ω_3, analogous to moving along the diagonal line *cb*).

These integrals are illustrated geometrically and mathematically in Fig. 5*B*. The first integration is along the line *cb*, whereas the second integration extends over the plane *abcd*. As a result, the final integration is over a volume of space (red mesh) much smaller than that of the full cube. Sampling within these embedded subspaces greatly improves efficiency: Less computational time is spent exploring conformations in phase space with negligible probability of occurrence. Using a smaller but more important region for integration solves the problem of high dimensionality.

Furthermore, augmenting spaces Ω_3 with Ω_2 and Ω_1 ensures that additional physically reasonable solutions (beyond just rigid body motions) are explored and we avoid the rough energy surface problem arising due to confining sampling to $\Omega_s = \Omega_3$.

While integrating along Ω_3 (to obtain $\langle\alpha\rangle_3$) gives a good initial approximation to $\langle\alpha\rangle_1$ (that is evaluated from a full integral over Ω_1), a more accurate approximation $\langle\alpha\rangle_{3,2,1}$ is obtained by extending the domain of integration into Ω_2 and subsequently into Ω_1 around the local neighborhood of Ω_3. This integral requires a more confined phase space exploration (red mesh opposed to full cube or space Ω_1), thereby significantly reducing the number of iterations needed to give converged averages of observables. All these lead to the main objective of our algorithm: $\langle\alpha\rangle_{3,2,1}$ is not only obtained more efficiently than $\langle\alpha\rangle_3$, it is also a more accurate estimate of $\langle\alpha\rangle$.

Sampling a Small Symmetric Four-Way Junction RNA. A small symmetric four-way junction RNA (Fig. S1) was chosen to benchmark different sampling methods. Only secondary structure information was used in both fragment assembly approaches [MC-Sym (23) and Rosetta (3)]. A local version of MC-Sym was implemented, and because we were only interested in the distance distribution between ends of neighboring helices, small chain discontinuities in the RNA were not corrected either by chain closure or minimization. Rosetta 3.0 was used, with default parameters, and its built-in minimization phase using a high resolution RNA potential.

Natural move Monte Carlo and hierarchical natural move Monte Carlo sampling were both done using the MOSAICS (13) software package. The initial starting structure consisted of four perfect A-form helices (from the make-na server) stitched together using MC-Sym and a short (100 steps) all-atom "relaxation" procedure [minimization using the AMBER 99 force field (32) and the Generalized Born treatment of electrostatics (33) with an inverse Debye–Hückel length of 0.19 $Å^{-1}$ as implemented in Nucleic Acid Builder (34)]. This approach efficiently restored chain connectivity and local stereochemistry, while not perturbing base-pair interactions significantly.

In both natural move Monte Carlo and hierarchical natural move Monte Carlo sampling, we are interested in obtaining distance distributions of a symmetric RNA. A hard sphere potential was used to increase the efficiency of exploring the conformational space using the all-atom representation. To prevent unnatural distortions of the A-form helices due to the hard sphere potential, base-pairs were only allowed to move as single units except in chain closure. The same starting structure was used to run 50 independent Markov chain Monte Carlo trajectories, and the last 1,000 conformations from each of these independently equilibrated runs were used to generate statistics for Fig. 1 and S2.

Hierarchical Natural Move Monte Carlo Sampling of a Large Symmetric RNA. The conformational sampling protocol for the large symmetric RNA was similar to the one for the symmetric four-way junction RNA. MC-Sym was used to assemble the starting structure from perfect A-form helices of eight and four base-pairs, respectively, and the chain connectivity was repaired with the "relaxation" procedure described above. Data for Fig. 2 and Fig. S3 were generated using the last 4,000 conformations of 50 independently equilibrated simulations sampled using a hard sphere potential.

Measuring Interhelical Distances. To facilitate the calculation of interhelical distances, we determined the origins and ends of all helices within the RNA. A perfect A-form RNA helix of the same length was superimposed to each base-paired region of the model (using C4′, C2, C4, and C6 atoms), and its helix origin and end were determined using the program X3DNA (4).

For the large symmetric RNA, the distance distributions between full arms were based on the average positions of the distal four-way junction (averaged over the positions of three helices of four base-pairs, and one helix of eight base-pairs).

Symmetry Deviation Metric. By symmetry, the four distributions of distances between neighboring full arms of the large symmetric RNA were expected to converge to identical distributions. Therefore, a metric to measure deviation from symmetry was defined as

$$\text{Deviation} = \sum_{j=1;j\neq i}^{4} \sum_{i=1}^{4} \sum_{d=1}^{N_{\text{bins}}} (y_i(d) - y_j(d))^2$$

where $y_i(d)$ is a histogram distribution of N_{bins} distance bins for the ith set of distances. This metric quantitatively measures the degree of deviation from the expected symmetric conformational distribution. The horizontal dashed line in Fig. 2 indicates the limiting deviation achieved by sampling the large symmetric RNA with move-sets L_1 to L_7.

ACKNOWLEDGMENTS. We thank the Jaeger Lab for generously sharing their tRNA nanostructure models used in our simulations, the Levitt Lab for useful discussions, and S. Doniach and J. A. Izaguirre for careful reading. Computations were done on Stanford's Bio-X^2 computers (National Science Foundation award CNS-0619926). A.Y.L.S. is funded by the Agency of Science, Technology, and Research, Singapore. This work was supported by National Institutes of Health Grant GM041455 and by a Human Frontier Science Program grant to M.L. M.L. is the Robert W. and Vivian K. Cahill Professor of Cancer Research.

1. Guo P (2010) The emerging field of RNA nanotechnology. *Nat Nanotechnol* 5:833–842.
2. Shapiro BA, Bindewald E, Kasprzak W, Yingling Y (2008) Protocols for the in silico design of RNA nanostructures. *Methods Mol Biol* 474:93–115.
3. Das R, Karanicolas J, Baker D (2010) Atomic accuracy in predicting and designing noncanonical RNA structure. *Nat Methods* 7:291–294.
4. Lu XJ, Olson WK (2008) 3DNA: A versatile, integrated software system for the analysis, rebuilding and visualization of three-dimensional nucleic-acid structures. *Nat Protoc* 3:1213–1227.
5. Minary P, Levitt M (2008) Probing protein fold space with a simplified model. *J Mol Biol* 375:920–933.
6. Minary P, Levitt M (2010) Conformational optimization with natural degrees of freedom: A novel stochastic chain closure algorithm. *J Comput Biol* 17:993–1010.
7. Hansmann UHE (1997) Parallel tempering algorithm for conformational studies of biological molecules. *Chem Phys Lett* 281:140–150.
8. Minary P, Levitt M (2006) Discussion of "equi-energy sampler" by Kou, Zhou and Wong. *Ann Stat* 34:1636–1641.
9. Minary P, Martyna GJ, Tuckerman ME (2003) Algorithms and novel applications based on the isokinetic ensemble. I. Biophysical and path integral molecular dynamics. *J Chem Phys* 118:2510–2526.
10. Minary P, Tuckerman ME, Martyna GJ (2004) Long time molecular dynamics for enhanced conformational sampling in biomolecular systems. *Phys Rev Lett* 93:150201.
11. Wales DJ, Scheraga HA (1999) Review: Chemistry—global optimization of clusters, crystals, and biomolecules. *Science* 285:1368–1372.
12. Minary P, Tuckerman M, Martyna G (2008) Dynamical spatial warping: A novel method for the conformational sampling of biophysical structure. *SIAM J Sci Comput* 30:2055–2083.
13. Minary P (2007) Methodologies for Optimization and SAmpling In Computational Studies (MOSAICS), version 3.8., http://csb.stanford.edu/~minary/MOSAICS.html.
14. Lilley DM (2000) Structures of helical junctions in nucleic acids. *Q Rev Biophys* 33:109–159.
15. McCreery RL (2004) Molecular electronic junctions. *Chem Mater* 16:4477–4496.
16. Reddy P, Jang SY, Segalman RA, Majumdar A (2007) Thermoelectricity in molecular junctions. *Science* 315:1568–1571.
17. Terrones M, et al. (2002) Molecular junctions by joining single-walled carbon nanotubes. *Phys Rev Lett* 89:075505.
18. Tinoco I, Bustamante C (1999) How RNA folds. *J Mol Biol* 293:271–281.
19. Westhof E, Masquida B, Jossinet F (2010) Predicting and modeling RNA architecture. *Cold Spring Harb Perspect Biol* 3:a003632.
20. Chworos A, et al. (2004) Building programmable jigsaw puzzles with RNA. *Science* 306:2068–2072.
21. Mao CD, et al. (2008) Hierarchical self-assembly of DNA into symmetric supramolecular polyhedra. *Nature* 452:198–201.
22. Ren ZF, Lao JY, Huang JY, Wang DZ (2004) Hierarchical oxide nanostructures. *J Mater Chem* 14:770–773.
23. Parisien M, Major F (2008) The MC-Fold and MC-Sym pipeline infers RNA structure from sequence data. *Nature* 452:51–55.
24. Bradley P, Misura KM, Baker D (2005) Toward high-resolution de novo structure prediction for small proteins. *Science* 309:1868–1871.
25. Bailor MH, Sun X, Al-Hashimi HM (2010) Topology links RNA secondary structure with global conformation, dynamics, and adaptation. *Science* 327:202–206.
26. Yin P, et al. (2008) Programming DNA tube circumferences. *Science* 321:824–826.
27. Condon A (2006) Designed DNA molecules: principles and applications of molecular nanotechnology. *Nat Rev Genet* 7:565–575.
28. Severcan I, Geary C, Verzemnieks E, Chworos A, Jaeger L (2009) Square-shaped RNA particles from different RNA folds. *Nano Lett* 9:1270–1277.
29. Hohng S, et al. (2004) Conformational flexibility of four-way junctions in RNA. *J Mol Biol* 336:69–79.
30. Goody TA, Lilley DM, Norman DG (2004) The chirality of a four-way helical junction in RNA. *J Am Chem Soc* 126:4126–4127.
31. Shapiro BA, Kasprzak W, Bindewald E, Kim TJ, Jaeger L (2011) Use of RNA structure flexibility data in nanostructure modeling. *Methods* 54:239–250.
32. Wang JM, Cieplak P, Kollman PA (2000) How well does a restrained electrostatic potential (RESP) model perform in calculating conformational energies of organic and biological molecules? *J Comput Chem* 21:1049–1074.
33. Tsui V, Case DA (2000) Theory and applications of the generalized Born solvation model in macromolecular simulations. *Biopolymers* 56:275–291.
34. Macke TJ, Case DA (1998) Modeling unusual nucleic acid structures. *ACS Sym Ser* 682:379–393.

Supporting Information

Sim et al. 10.1073/pnas.1119918109

SI Methods

Modeling Mutations in the tRNA Monomer and tRNA-Square. Structures and sequences of the tRNA monomer and tRNA-square as studied experimentally (1) were obtained from the Jaeger lab. Mutated sequences were threaded into the RNA backbone using an in-house design software and all structures were subsequently minimized [using AMBER 99-bs0 (2) and dielectric damping model (3) in MOSAICS (4)] with only side-chain flexibility to preserve backbone topology.

To prevent steric clashes, the structures were modeled using a hard-sphere potential and base-pairs were preserved by the appropriate choice of sampling degrees of freedom. This was done by manipulating the sampling degrees of freedom such that base-pairs can only move in concert (except during chain closure) and there were no degrees of freedom that moves individual bases alone within the base-pair.

Three different sets of mutations were done on the tRNA monomer: U15G, C16G, and both U15G and C16G simultaneously. Besides moving each base-pair as one, helices were allowed to move independently, akin to the approach taken for the simple four-way junction. To determine the angle (θ) between the stacked helices of tRNA (see Fig. 3*A*), perfect A-form helices were superimposed (using backbone P only) to the stacked helices and the helical axes were determined using X3DNA (5). θ was then determined from the dot product of both axes.

Each monomer within the mutated tRNA-square carried the double mutation, and θ for each corner was determined as previously. For statistics presented in Fig. 3*B* (12,000 structures each for wild type and mutant), maximum $\Delta\theta$ is the maximum difference in θ for all four corners of each modeled structure.

Conformations of an RNA Four-Way Junction Derived from the RNA Hairpin Ribozyme. To simulate this junction, the starting structure was generated for the sequence in reference (6). Perfect A-form helices were used for the base-paired and stacked regions (i.e., a perfect A-form helix for each D on A and C on B stacked helices). The full RNA was assembled with a starting angle $\theta = 117.5°$ (see Fig. 4) and distances as determined visually between stacked helices; ensemble results were comparable, regardless of starting θ (Fig. S4). In order to stitch the RNA, minimization was done using Nucleic Acid Builder (7). In most cases, only base-pairs close to the junction were allowed to move during minimization, but additional nucleotides were also allowed to move as needed. Sampling was constrained such that all base-pairs were kept rigid and each pair of stacked helices was regarded as a continuous helix. A hard-sphere potential was used to isolate the effects of chain connectivity and sterics on junction handedness.

From our simulations, the right-handed antiparallel conformation of this four-way junction is most consistent with experimental FRET data (6) because the antiparallel to parallel structure transition can take place for the left-handed form but not the right-handed one (with stacking intact). However, there might be other biophysical forces at play, preventing the left-handed antiparallel conformation from rotating to the left-handed parallel conformation. Alternatively, the true experimental transition might be from the preferred antiparallel conformation to a differently stacked structure that is experimentally indistinguishable from the parallel conformation (6). If so, this set of experimental data and our simulations are insufficient to distinguish the handedness of the preferred antiparallel conformation. Nonetheless, the sterics and connectivity argument is simple, and the model interpretation straightforward. The experimental data together with our simulations give a simple plausible expectation that the junction prefers the right-handed antiparallel conformation. This conclusion is later supported by other experiments (8).

General Applicability of Hierarchical Natural Move Monte Carlo and Stochastic Chain Closure. While our stochastic chain closure algorithm is robust and able to restore chain connectivity in most instances, there are some cases when stochastic chain closure fails, in which case the proposed move is rejected. Because rejection of a proposed step due to unsuccessful chain closure affects the overall Monte Carlo acceptance ratio, the user-defined move step sizes can be adjusted such that the acceptance ratio is reasonable (approximately 0.3–0.4); chain closure is likely to succeed for smaller move steps. Alternatively, it is possible to increase the chain segment dedicated to solving the closure problem, as was discussed previously (9).

In our implementation, the total number of degrees of freedom (L) is partitioned into independent and dependent degrees of freedom ($L = L_i \cup L_d$). Hence our method can be considered as a Monte Carlo minimization technique in which minimization is performed on a set of dependent degrees of freedom, or equivalently we explore the energy surface:

$$\tilde{E}(L) = \tilde{E}(L_i \cup L_d) = \min_{L_d}\{E(L_i \cup L_d)\}$$

where E is the user-defined energy function. Our hierarchical natural move Monte Carlo (HNM-MC), similar to Monte Carlo minimization, is designed to sample a transformed energy surface in which some of the transition energy barriers are removed. Thus HNM-MC is not designed to preserve microscopic reversibility or to generate canonically distributed conformational ensembles along all degrees of freedom (L) over the original energy surface (E). Nevertheless, HNM-MC may be used as a canonical sampling method along the independent degrees of freedom, L_i (that are not affected by the minimization) over $\tilde{E}$. We are currently running more rigorous tests comparing simulations by HNM-MC to reference canonical distributions obtained by N dimensional numerical integrals based on quadrature rules (an approach successfully applied in ref. 10). Extensive discussion on microscopic reversibility is beyond the scope of the current paper.

Due to the versatility of our algorithm and the generality of our implementation in MOSAICS (4), choices of moves for any particular system is completely user-defined and would clearly therefore depend on the purpose of simulation/sampling. If there are any known structural constraints, one can use correlated move-sets to preserve them, as was done for applications 1 and 2 discussed in the main text. Based on our modeling experience, the effects of using different levels of moves on sampling efficiency depend heavily on the system studied. For example, sampling efficiency for an RNA with secondary structure constraints depends significantly on the length of single-stranded regions connecting helices as this changes RNA conformational flexibility.

1. Severcan I, Geary C, Verzemnieks E, Chworos A, Jaeger L (2009) Square-shaped RNA particles from different RNA folds. *Nano Lett* 9:1270–1277.
2. Perez A, et al. (2007) Refinement of the AMBER force field for nucleic acids: improving the description of alpha/gamma conformers. *Biophys J* 92:3817–3829.
3. Rohs R, Etchebest C, Lavery R (1999) Unraveling proteins: A molecular mechanics study. *Biophys J* 76:2760–2768.
4. Minary P (2007) Methodologies for Optimization and SAmpling In Computational Studies (MOSAICS), version 3.8. http://csb.stanford.edu/~minary/MOSAICS.html.

5. Lu XJ, Olson WK (2008) 3DNA: A versatile, integrated software system for the analysis, rebuilding and visualization of three-dimensional nucleic-acid structures. *Nat Protoc* 3:1213–1227.
6. Hohng S, et al. (2004) Conformational flexibility of four-way junctions in RNA. *J Mol Biol* 336:69–79.
7. Macke TJ, Case DA (1998) Modeling unusual nucleic acid structures. *ACS Sym Ser* 682:379–393.
8. Goody TA, Lilley DM, Norman DG (2004) The chirality of a four-way helical junction in RNA. *J Am Chem Soc* 126:4126–4127.
9. Minary P, Levitt M (2010) Conformational optimization with natural degrees of freedom: A novel stochastic chain closure algorithm. *J Comput Biol* 17:993–1010.
10. Minary P, Tuckerman M, Martyna G (2008) Dynamical spatial warping: A novel method for the conformational sampling of biophysical structure. *SIAM J Sci Comput* 30:2055–2083.

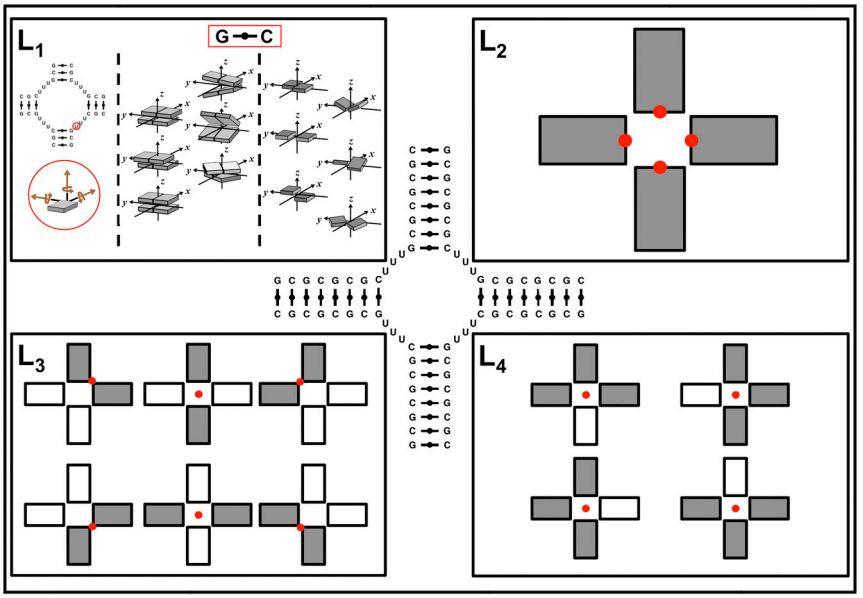


Fig. S1. Hierarchical moves for a simple four-way junction. Each nucleotide can move independently or as part of a base-pair (L_1). Regions of continuous base-pairing form helices (represented as rectangles), and these helices can be regarded as rigid bodies either independently (shaded in gray, L_2), as pairs of helices (L_3), or as groups of three (L_4). Rigid body motion requires the definition of rotation centers (shown as red dots). In the present case, the centers were defined to preserve symmetry of the system.

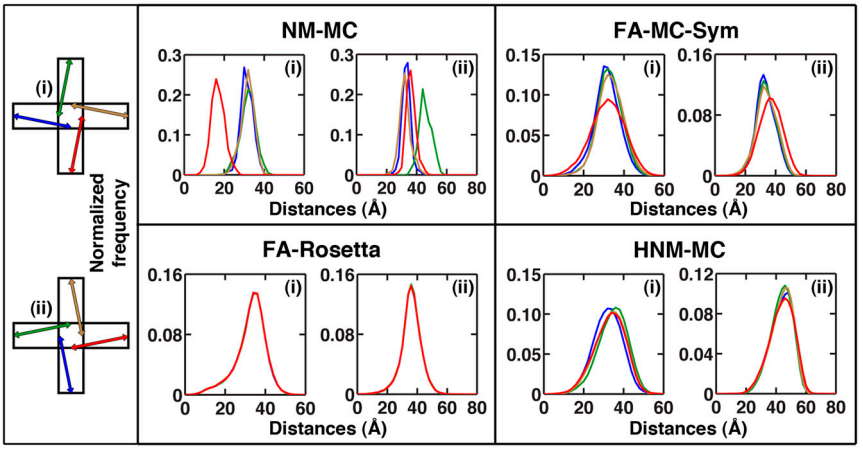


Fig. S2. Distance distributions of a simple four-way junction obtained by sampling using four different sampling protocols: natural move Monte Carlo (NM-MC), fragment assembly implemented by MC-Sym (FA-MC-Sym), fragment assembly implemented by Rosetta (FA-Rosetta) and hierarchical natural move Monte Carlo (HNM-MC). See Fig. 1 in main text for the distance distributions between ends of neighboring helices.

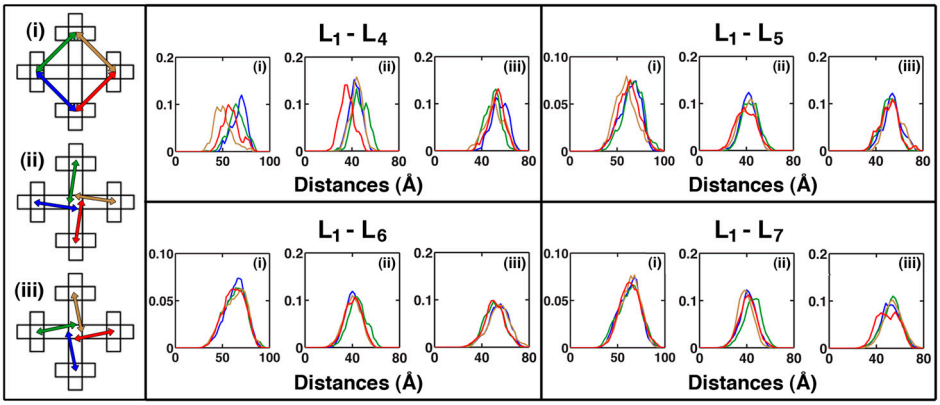


Fig. S3. Distance distributions within a large and complex symmetric RNA for four different sets of hierarchical moves (L_1-L_4, L_1-L_5, L_1-L_6, and L_1-L_7; see Fig. 2*A* for illustration of each level). Addition of higher order collective rigid body motions (i.e., adding L_5-L_7 to L_1-L_4) improved sampling efficiency significantly, and the distance distributions converged, as expected for this symmetrical system. Distance distributions were obtained from the statistics based on 2×10^4 iterations indicated by the vertical dashed line labeled * in Fig. 2*B*.

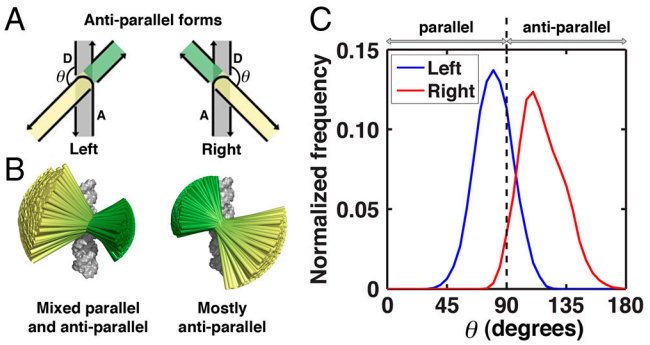


Fig. S4. Constrained sampling of a four-way RNA junction derived from the RNA hairpin ribozyme. The actual conformational distributions illustrated in Fig. 4 of the main text are dependent on the initial starting structure, due to the highly constrained nature of the junction and move-sets used. However, the conclusion that the left-handed form is more flexible (switches readily between parallel and antiparallel forms) than the right-handed one is independent of starting choice of θ. Here, a starting $\theta = 100°$ was used, compared to a starting $\theta = 117.5°$ in Fig. 4 of the main text.