doi:10.1016/j.jmb.2012.06.035 J. Mol. Biol. (2012) 423, 106–122

Contents lists available at www.sciencedirect.com

Journal of Molecular Biology

journal homepage: http://ees.elsevier.com.jmb

Mapping L1 Ligase Ribozyme Conformational Switch

George M. Giambaşu[1]*, Tai-Sung Lee[1], William G. Scott[2] and Darrin M. York[1]†

[1]*BioMaPS Institute and Department of Chemistry and Chemical Biology, Rutgers University, Piscataway, NJ 08854, USA*
[2]*Center for the Molecular Biology of RNA and the Department of Chemistry and Biochemistry, University of California, Santa Cruz, Santa Cruz, CA 95064, USA*

Received 23 February 2012; received in revised form 21 May 2012; accepted 25 June 2012
Available online 3 July 2012

Edited by G. Hummer

Keywords:
ribozyme;
allostery;
conformational;
transition;
catalysis

L1 ligase (L1L) molecular switch is an *in vitro* optimized synthetic allosteric ribozyme that catalyzes the regioselective formation of a 5′-to-3′ phosphodiester bond, a reaction for which there is no known naturally occurring RNA catalyst. L1L serves as a proof of principle that RNA can catalyze a critical reaction for prebiotic RNA self-replication according to the RNA world hypothesis. L1L crystal structure captures two distinct conformations that differ by a reorientation of one of the stems by around 80 Å and are presumed to correspond to the active and inactive state, respectively. It is of great interest to understand the nature of these two states in solution and the pathway for their interconversion. In this study, we use explicit solvent molecular simulation together with a novel enhanced sampling method that utilizes concepts from network theory to map out the conformational transition between active and inactive states of L1L. We find that the overall switching mechanism can be described as a three-state/two-step process. The first step involves a large-amplitude swing that reorients stem C. The second step involves the allosteric activation of the catalytic site through distant contacts with stem C. Using a conformational space network representation of the L1L switch transition, it is shown that the connection between the three states follows different topographical patterns: the stem C swing step passes through a narrow region of the conformational space network, whereas the allosteric activation step covers a much wider region and a more diverse set of pathways through the network.

Introduction

L1 ligase (L1L) is a synthetic *in vitro* selected ribozyme that catalyzes regiospecific formation of 3′-to-5′ phosphodiester linkages, a reaction characteristic of all extant present-day RNA and DNA protein polymerases. There is no known naturally occurring ribozyme that catalyzes this phosphodiester assembly reaction, and it was not until 1993 when the first ribozyme that exhibited this activity was created through *in vitro* evolution.[1] Along with several other ligase ribozymes,[2–9] L1L adds support to the RNA world hypothesis that proposes that life originated from catalytic RNA molecules capable of replication via template-dependent assembly of RNA fragments of nucleotide monomers.[10–13]

In addition to its potential relevance to the origin of life, L1L, presumably as a fortuitous consequence of *in vitro* selection, is an allosteric ribozyme molecular switch.[14–16] It has been shown that it is possible to take advantage of this property to

*Corresponding author. E-mail address: york@biomaps.rutgers.edu.
† http://theory.rutgers.edu.

Abbreviations used: L1L, L1 ligase; MD, molecular dynamics; CSN, conformational space network; HHR, hammerhead ribozyme; CIRP, class I RNA polymerase; FSN, focused sampling on networks.

engineer new constructs whose catalytic activity can be controlled by small molecules, peptides, or even proteins.[7,17–19] Recently, there has been an increased interest in using aptamer molecules such as L1L for a large spectrum of applications including purification and biotechnology, diagnostics and biosensors, therapeutics, or combating infectious agents.[20–23]

The crystal structure of the L1L self-ligation product has been solved recently,[24] providing "a glimpse of biology's first enzyme".[25] The X-ray crystal structure of the self-ligation product shows two crystallographically independent conformations resolved in the same asymmetric unit. These conformers differ in the orientation of one of the stems (stem C) by a movement of the stem tip by 80 Å.[24] A minimal set of virtual torsions was identified to be sufficient to differentiate between the two conformers.[26] Based on the presence/absence of specific contacts in the ligation site and evolutionarily conserved regions of the bulged loop situated on stem C, it was proposed that the conformers represented catalytically active "on" and inactive "off" states.[24] We will refer to these two crystallized structures as active/docked and inactive/undocked, respectively.

L1L is a γ-shaped molecule, with three stems joined into a three-way junction (Fig. 1), which is a recurring motif in the present set of known ribozyme ligases. Relative angular orientations of helices forming RNA junctions have been shown to be decisive in deciphering the capacity of distant contacts to stabilize specific RNA conformations that can affect function.[27] The structure of the L1L stems is mainly helical, with a bulged loop on stem C and GAAA tetraloops terminating each stem. L1L helices contain canonical base pairs with the notable exception of three consecutive noncanonical base pairs between the active site and the substrate, an unusual trait in the world of synthetic ribozymes.[3,28] Several experimental sources suggest that the L1L catalytic mechanism consists of (i) binding a complementary DNA effector—an unusual fact in the context of other ribozyme ligases, (ii) binding of the substrate, (iii) allosteric activation of the catalytic site by the formation of a stem–loop contact supported by an 80 Å swing of one of the L1L stems (stem C), and (iv) catalytic self-ligation.[7,17–19,24] The crystallized construct[24] used in the present work as a

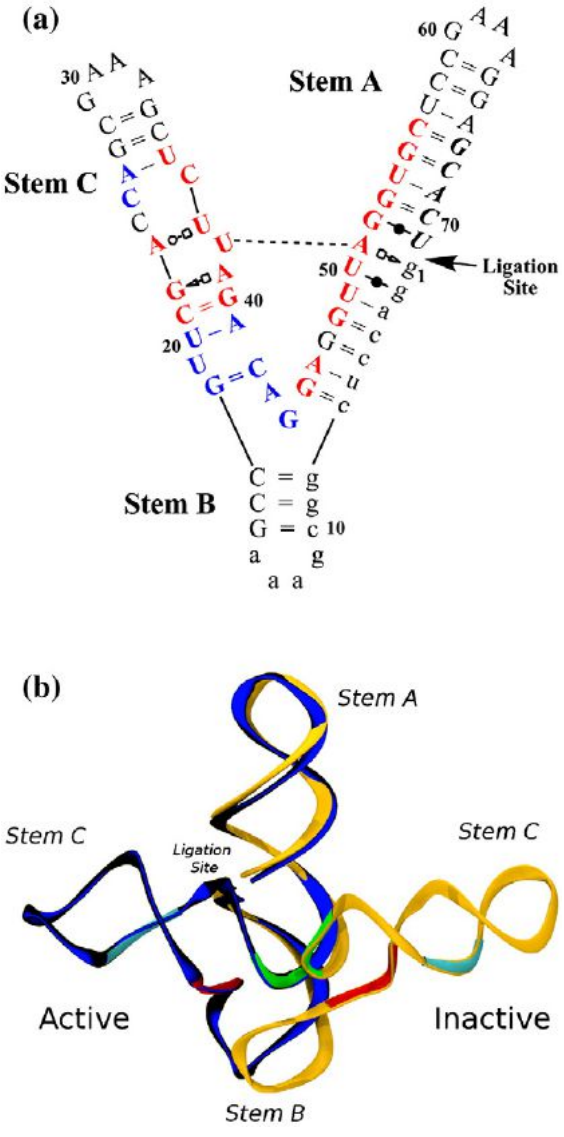

Fig. 1. L1L structure. (a) L1L is a γ-shaped molecule, with three stems (labeled A, B, and C) joined in a three-way junction. The structure of the stems is mainly helical, with the exception of a bulged loop located on stem C and with GAAA tetraloops ending each stem. L1L helical parts contain canonical base pairs with the notable exception of three consecutive noncanonical base pairs between the active site and the substrate on stem A. The docked/active conformation contains an additional tertiary contact between stem A and stem C (broken line) mediated by a U_{38}–A_{51} canonical base pair and a Mg^{2+}. Evolutionarily conserved regions are colored red for more than 95% and blue for more than 80%. Lowercase nucleotides were not varied during the *in vitro* evolution. The nucleotides that constitute the substrate are in italics.[7,17–19,24] The L1L construct studied here (l1xc6) has the substrate covalently bound. See the text for a more detailed description of the L1L allosterically controlled catalytic mechanism. (b) Overlay of the two crystallized conformations of L1L. Active/docked state is in blue and inactive/undocked state is in orange. The two co-crystallized structures are RMS fit using all the heavy atoms comprising stems A and B. This shows that for the two structures to interconvert, the tip of stem C has to travel over an 80 Å long path. We have previously shown[26] that only restricted regions of the L1L structure have to change to assist L1L conformational transition. These restricted regions can be identified using virtual torsions analysis of the crystal structures and MD simulations and are marked in green (θ_{44}), red (θ_{18}), and magenta (η_{37}, θ_{38}).

starting point for molecular dynamics (MD) simulations is missing the DNA effector binding site and has the substrate covalently bound, enabling exploration of the later two stages of the L1L catalysis: stem C swing, allosteric activation of the active site, and catalytic ligation.

In our previous analysis of the L1L structure and dynamics,[26] we identified a set of dynamical hinge points using a series of MD simulations of the precursor and product states that captured fluctuations around the active and inactive conformers. The hinge points were located in highly evolutionarily conserved regions of L1L sequence and were identified based on analysis of the distribution of virtual torsions. Specifically, two hinge points were located in the three-way junction, and another in a bulge loop on stem C (U_{38} loop). The ligation site was predicted by the simulations to visit three distinct states characterized by hydrogen bond patterns that are correlated with the formation of specific contacts implicated in catalysis.

It remains an open question how L1L can be controlled to transition between its active and inactive forms. An understanding of this transition and the factors that control it is of fundamental importance to help engineer L1L constructs that bind to a wide spectrum of analytes. This large-scale conformational change is likely a result of the artificial *in vitro* selection that targeted, in part, allosteric properties. In the case of the naturally occurring hammerhead ribozyme (HHR), adiabatic morphing shows that a fairly simple trajectory exists between the active and inactive states that can be transversed simply via a series of low energy-barrier, small-scale torsion-angle conformational changes within one of its stems and active site.[29] On the contrary, in the case of the L1L, a similar simple treatment is impeded due to steric hindrance at the three-helix junction as can be seen from the superimposition of the docked and undocked conformers in Fig. 1. L1L conformational change is so large that it defies treatment by more conventional, established means.

Here, we aim to characterize the complex conformational landscape and pathways that connect the L1L active and inactive conformations using molecular simulations. We use MD in conjunction with an enhanced sampling technique designed to increase statistical sampling along the pathway of the conformational transitions and was inspired by related methods in the literature that represent the conformational space using graph theory and apply swarms of trajectories. Based on the analysis of conformational space network (CSN) that is obtained from more than 1 μs of MD simulations, we propose that the L1L activation mechanism involves a three-state, two-step process. One of the steps corresponds to the large swing of stem C from the vicinity of the crystallized inactive conformer to the vicinity of the active conformer. This large conformational change is shown to correlate directly to changes in the three-way junction tracked by changes in a specific virtual torsion angle (θ_{44}). The second step consists in the formation of tertiary contacts between stem C and the active site, an event accompanied by a subtle change in the base-pairing dynamics between the active site and the substrate. We show that the connection between the three states follows different patterns: the stem C swing step passes through a narrow region of the CSN, whereas the allosteric activation step covers a much wider region and a more diverse set of pathways through the network.

Results and Discussion

Three substates (A, I_1, I_2) are linked during L1L switching

Here, we aim to isolate the overall large-scale characteristic motions of the dynamics of the L1L switch mechanism by constructing a CSN model. A CSN model is a graphic network representation in which each node represents a conformational state and two conformational states are linked by an edge if the transitions between these two states are observed during MD sampling.[30–36]

CSN models offer discrete representations of configurational space and of the possible transition pathways between states and have been used extensively to characterize the folding free-energy landscapes of peptides or small proteins as an alternative to projections onto limited sets of (sometimes arbitrary) order parameters.[31,32,34,36–43] CSN models also reveal the topography of the conformational space and free energy landscape. It has been shown that it is likely that nodes in the same free energy basin are well connected to each other, whereas nodes in different basins are loosely connected.[31] Additionally, finding sets of pathways through the network that connect two substates affords a detailed structural characterization of the transition. Here, our goal is to find an essential network description with a minimal number of clusters that capture the essential, structurally distinct states along the transition pathway.

Identifying the nodes representing conformational states is central for building a CSN model. Traditional clustering algorithms require a preset number of clusters for a given data set. Since the optimized overall objective similarity/dissimilarity function increases/decreases monotonically with the number of clusters,[44–46] the choice of the number of clusters is the key to establishing a meaningful data reduction. Here, we exploit features of the dynamics of the docked/active state in order to determine the

optimal number of clusters. Specifically, the docked/active state does not exhibit large conformational variation, and can, in fact, be well described by a single set of structures stably fluctuating about a well-defined average. The desired clusters should be optimally chosen so that the ensemble of structures in the docked/active state reside in a single cluster.

To achieve the desired constraint of clustering, we utilized a structure-based hybrid two-stage partitional–agglomerative clustering algorithm, which has been shown to efficiently lead to better solutions than partitional or agglomerative schemes alone.[47] (see Materials and Methods). The clustering algorithm works as follows: in the first stage, the data are partitioned in relatively large number of clusters (100 in the present case). This is a *top-down* partitional stage requiring global information about the entire data set. In the second stage, the clusters are merged until all the conformations spanned by the active state are classified into a single cluster. This is a *bottom–up* agglomerative stage that requires only local information. The utilization of the hybrid clustering approach is summarized in the tree (dendrogram) graph shown in Fig. 2a (see full details in Supplementary Material) to yield a final set of 12 coarse-grained clusters.

An initial CSN has been built using this set of 12 clusters (Fig. 2b). Interpreting the groups of nodes that strongly interconnect as forming an attractor or substate[31] leads to the identification of three major attractors (Fig. 2b): A ("Active substate", colored blue), I_1 ("Inactive substate 1", colored yellow), and I_2 ("Inactive substate 2", colored gray). Since states A and I_2 connect only with I_1, it follows that L1L switching follows a three-state/two-step mechanism: $A \rightleftharpoons I_1 \rightleftharpoons I_2$.

In order to gain more detailed insight into the structural features that support the transition between the three major states, we built a higher-resolution CSN model from the initial 100 partitional clusters (Fig. 3). The higher number of clusters allows a better refinement of the boundary of each major state and the specific interfaces where transitions occur between the major states (i.e., the nodes of the CSN that share linkages with nodes belonging to two or more of the A, I_1, or I_2 states). Along with the CSN, a set of representative

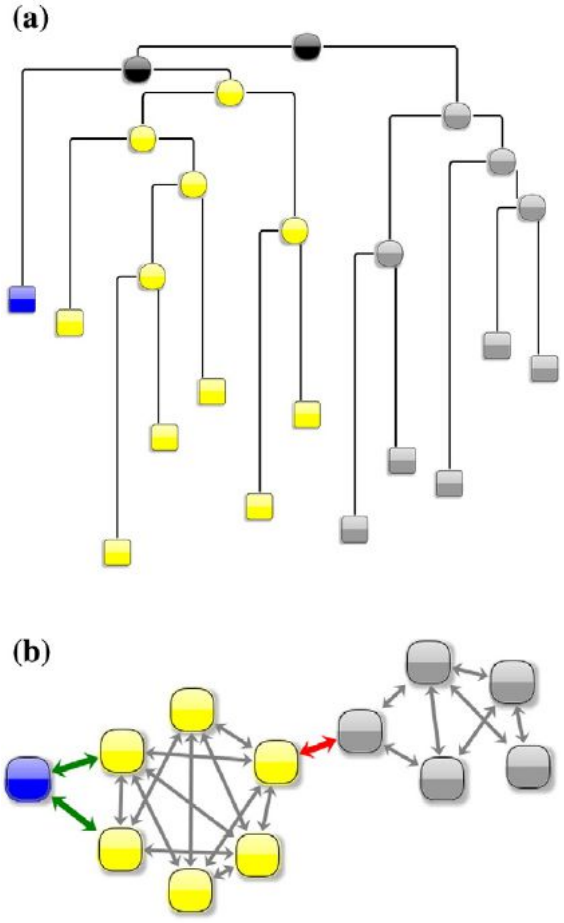

Fig. 2. Identification of the main substates spanned during L1L switching is realized in two steps. In the first step, the purpose is to form a minimum number of clusters that capture essential, structurally distinct states. For that, we used a hybrid clustering scheme that chooses the best number of clusters when the sampled active-docked conformations reside in a single cluster. The hybrid clustering consists of a partitional (*top–down*) followed by a agglomerative (*bottom–up*) stage (see Materials and Methods) and can be depicted as a dendrogram (shown in Supplementary Material). Here (a), a simplified version of the dendrogram is shown whose leaf nodes (i.e., the nodes at the bottom of the dendrogram and shown as squares) are the 12 clusters that were formed right before all the active-docked conformations were agglomerated into a single cluster. The nodes of the dendogram are placed on the vertical following their discovery order. The simplified dendrogram corresponds to the latest steps of the agglomerative (*bottom–up*) stage of the hybrid clustering procedure. In a second step (b), a CSN is obtained from the 12 leaves of the hierarchical tree. CSNs reveal the topography of the conformational space and free-energy landscape: it has been shown that it is likely that nodes in the same free energy basin (substate) are well connected among each other, whereas nodes in different basins (substates) are loosely connected.[31] Based on the connectivity pattern, three major substates (labeled A—"active substate", blue; I_1—"inactive substate 1", yellow; I_2—"inactive substate 2", gray) can be identified. The links of the graph that correspond to the transitions between these three states are colored green and red, respectively. The partitioning into the three substates is supported also by structural similarity. Indeed, merging of these three states (shown as black clusters on the dendrogram) occurs at the latest stages of the clustering scheme.

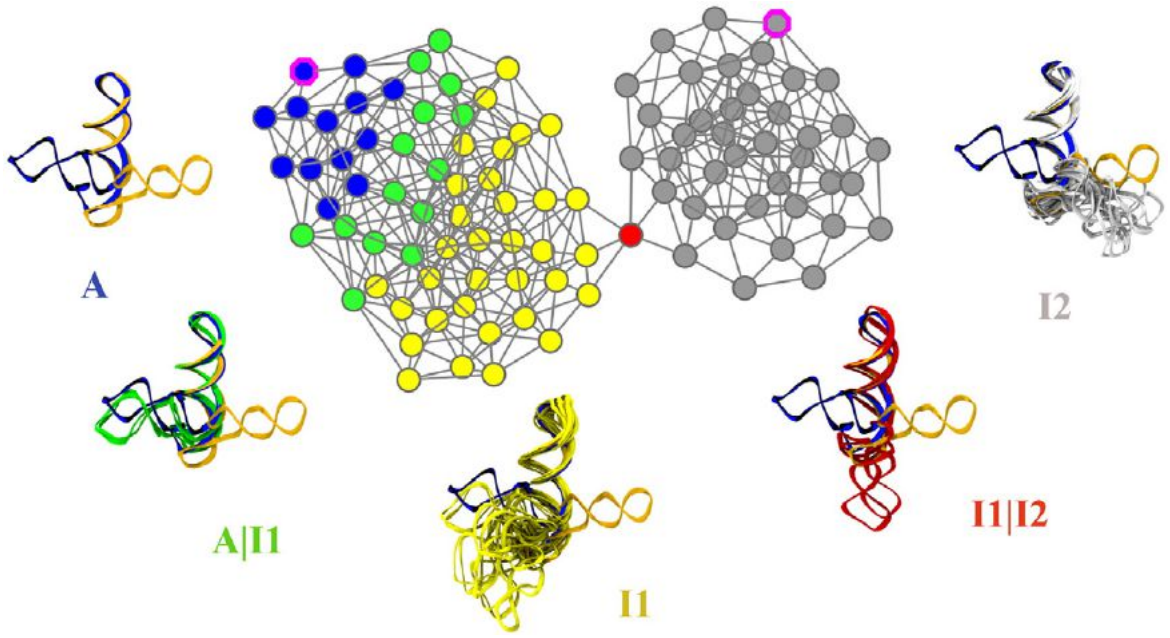

Fig. 3. Mapping the L1L switching mechanism using a CSN. A CSN is a graph whose nodes (shown here as circles) are conformations and linkages (or edges, shown here as lines connecting the circles) that represent transitions between them. These transitions are observed during MD simulations. CSNs are a discrete representation of the states as well as the paths that connect them and can reveal the topography of the conformational space. It is likely that nodes in the same free-energy basin are well connected among each other, whereas nodes in different basins are loosely connected.[31] Additionally, finding the nodes that allow two substates to connect can reveal the required structural features of that transition to occur. Here we use a CSN built using 100 conformations obtained using a partitional clustering technique described in Materials and Methods. The nodes of the network are colored based on their inclusion into the three main identified substates (A, blue; I_1, yellow; I_2, gray) or at the boundaries (interfaces) between the substates ($A \mid I_1$ interface, green; $I_1 \mid I_2$ interface, red). Clusters belonging to the A and I_2 states that contain the two crystallized constructs are marked with a magenta border. Representative sets of structures for each of the substates and their corresponding interfaces are shown. For reference, all the sets are superimposed on the crystallized conformations, shown here in blue (active) and gray (inactive). All structures are RMS fit using all heavy atoms comprising stems A and B. We find that the overall switching mechanism can be described as a three-state/two-step process. It is interesting to note that the connection between the three states follows different topographical patterns: the stem C swing ($I_1 \rightleftharpoons I_2$) step passes through a narrow region of the CSN, whereas the allosteric activation step ($A \rightleftharpoons I_1$) covers a much wider region and a more diverse set of pathways through the network.

structures for each state is shown in Fig. 3. Inspection reveals that structures associated with I_1 mainly originate from the unfolded active conformation simulations, whereas the structures included in I_2 originate from the undocked inactive-state simulations. This implies that the $A \rightleftharpoons I_1$ step corresponds to the docking/undocking of stem C from the active site, whereas the $I_1 \rightleftharpoons I_2$ step corresponds to the stem C swing.

Allosteric control mechanism: $A \rightleftharpoons I_1$

The $A \rightleftharpoons I_1$ transition corresponds to the docking/undocking of stem C, through U_{38} loop, to/from the L1L active site that is noncanonically base paired with the substrate. Both biochemical and crystallographic data[7,17–19,24] support the hypothesis that this stem–loop interaction is one of the factors that contribute to the increased L1L ligase catalytic activity and is able to promote the catalytic step. The L1L allosteric control mechanism requires a set of key interactions linked to the presumed active conformation, including noncanonical base pairing between the active site and the substrate, and tertiary interactions between the active site and a distant evolutionarily conserved part of the L1L. Here, we account for how each of these factors contributes to the initiation and control of the catalytic step.

Crystallographic data suggest that L1L can exist in (at least) two different conformations, one being presumed to be active and the other inactive.[24] The two conformations differ by a reorientation of stem C tip by ~80 Å. Based on analysis of the hydrogen-bonding patterns between the active site and substrate, we have previously shown[26] how the dynamics of the active site exhibits two completely different patterns of conformational variation in the

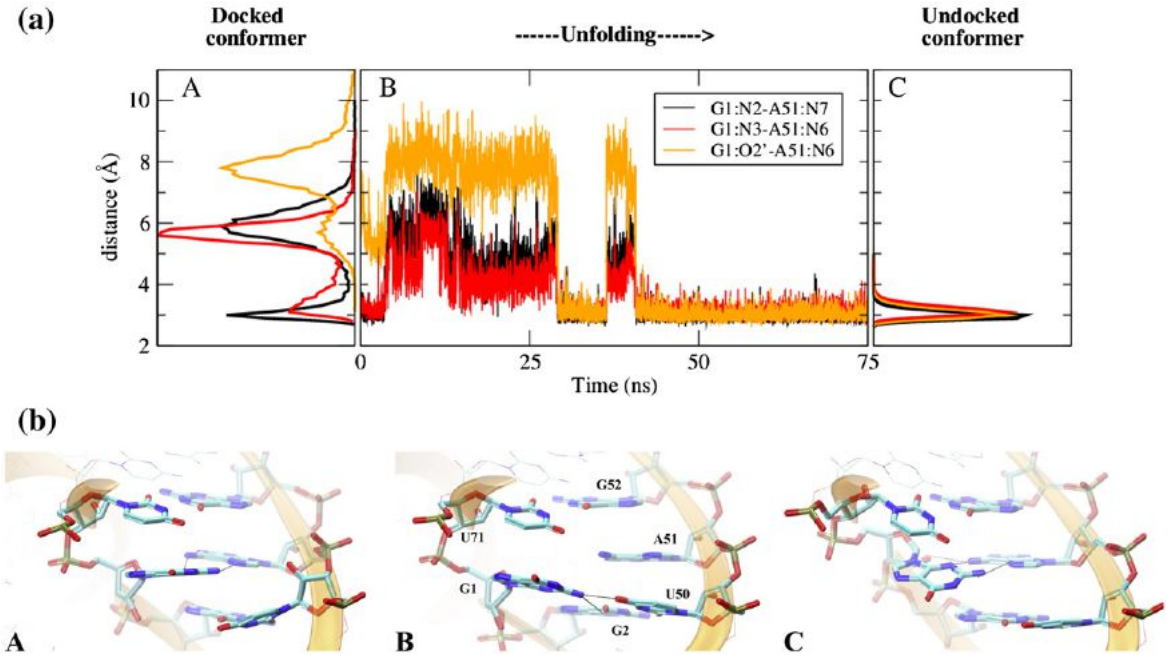

Fig. 4. The dynamical signature of the active site changes when transitioning from the active (A) to the inactive (I_1) state. (a) After removing the contacts between U_{38} and the active site, the latter recovers the dynamical signature of the inactive/undocked states. This change is followed in terms of the time series of three important hydrogen-bonding distances that are characteristic to a complete $A_{51}\square\rightarrow G_1$ (*trans* Hoogsteen/sugar edge base pair) specific to the inactive state. On the left, the multimodal distribution of these three hydrogen-bonding distances specific to the active/docked conformation simulations both in the product and in the precursor state is shown. On the right, the same distributions, which are now unimodal in nature, obtained during the MD simulations of the inactive/undocked conformer simulations are shown. In the middle, the time series evolution of the hydrogen-bonding distances, after the unfolding/undocking biasing potential has been removed, is depicted. It is important to note that the docked and undocked conformer simulation results shown here come from separate unbiased simulations. (b) Three representative snapshots along the unfolding/undocking trajectory. (A and B) Structures that are representative for the active conformation and are the two major conformations identified also in the active precursor simulation.[26] For these two structures, G_1:O2′ does not make a hydrogen bond with A_{51}:N6, typical for a complete $A_{51}\square\rightarrow G_1$. (C) Structure representative for the undocked/inactive conformation, yielding a complete $A_{51}\square\rightarrow G_1$ base pair.

active and inactive forms. More precisely, the catalytic site was shown to span three conformational states in its active form, whereas the dynamics of the inactive form was unimodal. We were able to correlate the different variability patterns with the lack/presence of a hydrogen bond that is typical of a *trans* Hoogsteen/sugar-face base pair formed by A_{51} and G_1/GTP_1 (denoted here as $A_{51}\square\rightarrow G_1$ following Ref. 48). It is important to note that A_{51} directly mediates the interaction of the catalytic site with stem C through a canonical base pair with U_{38}, along with phosphate–Mg^{2+} interactions between G_1, A_{39}, and G_{40}, respectively.

Due to the high free energy necessary to disrupt the hydrogen bonds and the phosphate–Mg^{2+} interactions that maintain the docked conformer, the conformational changes associated with the undocking of the U_{38} loop from the active site cannot be realized on the time scales of the present simulations. To overcome this impediment, we provoke the undocking by slowly forcing the disruption of the hydrogen bond network sustaining the L1L active conformation (see Materials and Methods) and monitor the changes in hydrogen-bonding patterns. Figure 4a compares the distributions of the three hydrogen bonds that define a standard $A_{51}\square\rightarrow G_1$ base pair in the case of the simulations of the docked/active and undocked/inactive conformations as well as their evolution after the removal of the tertiary contact between stem A and stem C. The distributions of these three hydrogen bonds indicate different characteristic patterns in the active and inactive states. In the case of the docked/active state, the distributions are multimodal, suggesting the existence of several conformational states and hydrogen-bonding patterns, whereas in the case of the inactive/undocked state, the distributions are unimodal. The evolution in time of these three hydrogen-bonding distances recorded after the disruption of the tertiary contacts between stem A and stem C shows that after almost 50 ns, the active site adopts the inactive/undocked

dynamical pattern, with the formation of the complete typical *trans* Hoogsteen/sugar-face base pair between A_{51} and G_1 depicted in Fig. 4b, C.

The analysis presented here suggests a link between the specific dynamical pattern of the active-site hydrogen bond network and the tertiary contacts with stem C. We have shown previously that the formation of these contacts allows the active site to adopt three main conformational states that have been shown to correlate differently with initiation of the first steps of the catalytic process.[26] We noted that only two of these three states allow the formation of a hydrogen bond between U_{71}:$H_{O3'}$ and GTP_1:$O_{2P\alpha}$ that can support the deprotonation of U_{71}:$O_{3'}$, which is the first step of the catalyzed 3′-to-5′ ligation reaction.[26] The two states that correlate positively with the U_{71}:$O_{3'}$ deprotonation require the partial disruption of the incomplete *trans* Hoogsteen/sugar-face base pair between A_{51} and G_1 that is found in the docked crystal conformation. This suggests that the role of the tertiary contacts between the active site and stem C may be to weaken the $A_{51}\square\rightarrow G_1$ base pair to be more easily disrupted, thus facilitating the first step of the catalytic process, that is, the deprotonation of U_{71}:$O_{3'}$. Experimentally, the idea of an intrinsically variable active site is supported by the fact that the optimal base pairing between L1L and its substrate has to be noncanonical.[49] For example, minimal changes of either one of the two GU wobble pairs of the active site to canonical isosteric GC pairs decrease the catalytic activity of L1L 20 times.[49]

Table 1. Structural parameters of the three substates.

(*X*,*Y*)	ang(*X*, *Y*)
$v_{cr,a}^{stemA}, v_{A}^{stemC}$	110.4 (10.7)
$v_{cr,a}^{stemA}, v_{A\,|\,I_1}^{stemC}$	130.6 (11.8)
$v_{cr,a}^{stemA}, v_{I_1}^{stemC}$	140.3 (15.5)
$v_{cr,a}^{stemA}, v_{I_1\,|\,I2}^{stemC}$	125.4 (5.0)
$v_{cr,a}^{stemA}, v_{I_2}^{stemC}$	114.8 (32.6)
$v_{cr,a}^{stemB}, v_{A}^{stemC}$	81.4 (10.5)
$v_{cr,a}^{stemB}, v_{A\,|\,I_1}^{stemC}$	61.3 (11.3)
$v_{cr,a}^{stemB}, v_{I_1}^{stemC}$	49.5 (15.9)
$v_{cr,a}^{stemB}, v_{I_1\,|\,I2}^{stemC}$	50.6 (5.4)
$v_{cr,a}^{stemB}, v_{I_2}^{stemC}$	65.5 (23.2)
$v_{cr,a}^{stemC}, v_{A}^{stemC}$	25.3 (9.0)
$v_{cr,a}^{stemC}, v_{A\,|\,I_1}^{stemC}$	48.7 (7.0)
$v_{cr,a}^{stemC}, v_{I_1}^{stemC}$	67.1 (16.9)
$v_{cr,a}^{stemC}, v_{I_1\,|\,I2}^{stemC}$	119.8 (4.7)
$v_{cr,a}^{stemC}, v_{I_2}^{stemC}$	96.0 (42.5)
$v_{cr,i}^{stemC}, v_{A}^{stemC}$	158.1 (10.7)
$v_{cr,i}^{stemC}, v_{A\,|\,I_1}^{stemC}$	146.9 (10.2)
$v_{cr,i}^{stemC}, v_{I_1}^{stemC}$	132.0 (15.9)
$v_{cr,i}^{stemC}, v_{I_1\,|\,I2}^{stemC}$	77.7 (4.7)
$v_{cr,i}^{stemC}, v_{I_2}^{stemC}$	95.0 (50.0)

Average angular distances [ang (*X*,*Y*) in degrees] and their standard deviations between the characteristic vectors (v) of stem C in the A, I_1, and I_2 states and the two crystallized conformers. Superscripts denote the stems (stems A–C) and subscripts denote the corresponding conformational states (cr,a and cr,*i* stand for the active and inactive crystallized conformers, and A, I_1, and I_2 stand for the three substates extracted from MD simulations). Prior to calculating the characteristic vectors, all the structures were RMS fit using all heavy atoms of stem A, giving rise to structural arrangements as those shown in Fig. 3. For the definition of the characteristic vectors of each stem, see Materials and Methods.

Stem C swing: $I_1 \rightleftharpoons I_2$

The $I_1 \rightleftharpoons I_2$ conformational transition corresponds to the swing of stem C from the vicinity of the inactive conformation isolated in the crystal structure (I_2 substate) to one that brings U_{38} loop in close proximity of the active site (I_1 substate). The necessity of this large conformational transition to support the L1 catalytic mechanism is suggested by structures of the two crystallized conformers.[24] The supporting hinge points were located in restricted regions of the L1L three-way junction and could be mapped to changes of only two virtual torsions (θ_{18} and θ_{44}, see Materials and Methods for details).[26]

The location of the interface between I_1 and I_2 substates is marked in red on the CSN shown in Fig. 3 along with some representative structures. It is interesting to note that the interfaces between the A, I_1, and I_2 substates differ significantly in the variety of structures and of the corresponding interconnecting paths. Whereas the $A\,|\,I_1$ interface contains a set of 14 nodes, the $I_1\,|\,I_2$ interface contains only 1. Also, the standard deviations of the angular distances with respect to crystal structures of the two interfaces (Table 1) show that the variability with respect to stems A and B is larger for the $A\,|\,I_1$ interface. This suggests that the conformational space occupied by the $A\,|\,I_1$ interface is larger than that occupied by the $I_1\,|\,I_2$ interface. In other words, the connection between the three states follows two patterns: the stem C swing step passes through a narrow region of the CSN, whereas the allosteric activation step covers a much wider region and a more diverse set of pathways through the network.

In order to obtain a greater insight into what are the structural features that support the transition between the two states, we analyze the distribution of θ_{44} as a function of the depth from the $I_1\,|\,I_2$ interface node. In Fig. 5a, the distribution of the θ_{44} virtual torsion angle is shown as a function of the depth from the $I_1\,|\,I_2$ interface node of the CSN. The *depth 1* set contains the interface node direct neighbors and the *depth 2* set contains the direct neighbors as well as the second layer of neighbors of the interface node. It can be observed that, as expected, the number of samples increases from the depth 1 to depth 2 selections; however, the region of the torsion space located approximately between −10° and 0° remains approximately with the same number of samples. This indicates that θ_{44}

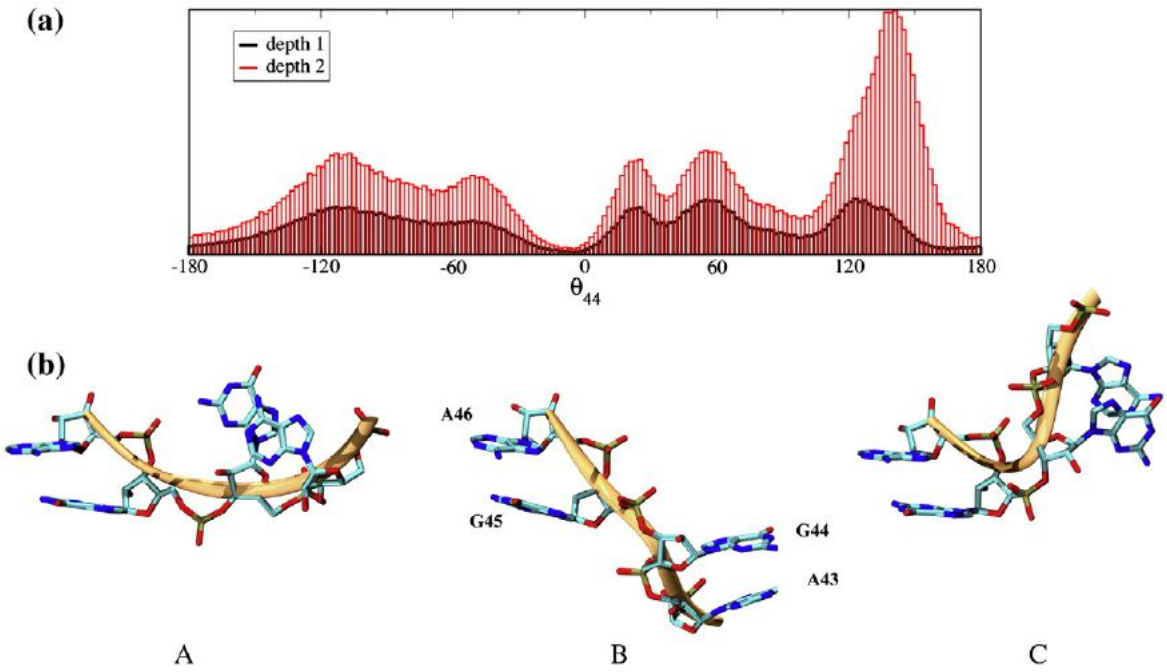

Fig. 5. The transition between I_1 and I_2 substates is supported by θ_{44} virtual torsion (A_{43}–G_{44}–G_{45}–A_{46}). (a) The distribution of θ_{44} in the vicinity of the I_1–I_2 interface node shows that θ_{44} has to traverse the −10° to 0° interval to allow the transition from I_1 to I_2. The UN-normalized distributions of θ_{44} for two sets of nodes of the CSN shown in are shown. The *depth 1 set* includes all the nodes that are at 1 edge distance from the interface node. The *depth 2 set* includes all the nodes that are a distance of 2 edges or less from the interface node. (b) Representative conformations of the region spanned by θ_{44} of the three-way junction along the conformational switch pathway. (A) Conformation specific to the vicinity of the active conformation, (B) conformation located on the −10:0 interval of θ_{44}, (C) conformation specific to the vicinity of the crystallized active conformation. The backbone is shown in green and corresponds to the region spanned by θ_{44} on L1L backbone in Fig. 1.

traverses this specific interval only when the structures are located at the interface node and its immediate neighbor nodes. In order for the transition from the I_1 to I_2 to occur, θ_{44} has to pass through this interval. This result is consistent with our previous hypothesis[26] that θ_{44} has to travel along an arc of 232° in the transition from active/docked to inactive/undocked conformational states.

In Fig. 3, we show representative conformations of the part of the three-way junction that is spanned by θ_{44} along the stem C swing. It can be observed that the G_{44} and A_{43} undergo a significant repositioning with respect to A_{46} and G_{45}, which are the constituents of the highly canonical and, consequently, relatively rigid stem A. The transition is taking place through modifications of the backbone and without a change in the base pairing or hydrogen-bonding patterns. Experimentally, G_{44} and A_{43} have been shown to be a part of a five-basis motif located on the three-way junction that is highly sensitive to mutations.[7,17–19] The present simulations also show θ_{18} to have large flexibility. However, this is not necessarily correlated with the stem C orientation (data not shown).

Comparison to conformational transitions of other ribozymes

RNA molecules have the ability to respond to their environment through changes at the secondary or tertiary level to accomplish their role in translation, transcription, posttranscriptional processing, viral replication, specific binding, assembly of the ribosome, and catalysis. RNA can be designed to respond through conformational or catalytic events to proteins, nucleic acids, metal ions, metabolites, vitamins, changes in temperature, and even RNA biosynthesis itself.[50,51] It has been shown that it is possible to engineer L1L constructs whose catalytic activity can be controlled by small molecules, peptides, or even proteins.[7,17–19]

L1L dynamics, whose complexity can be captured with the CSN representation, should contribute to the recent efforts to explore the RNA folding landscape[52–55] or the topological restraints of RNA junctions.[56,27] Three-way junctions[57,58] with tertiary contacts[59]—the L1L fold—are recurrent and versatile folds for RNA function and the characterization of L1L dynamics in solution can help us understand their important role in RNA dynamics. The use of

virtual torsions to distinguish between different folds is "a new way to see RNA".[60] We show here and elsewhere[26] that virtual torsions can be used to distinguish between L1L conformations in the crystal as well as in solution. The Mg^{2+}-mediated tertiary contact between the L1L catalytic site and an evolutionarily conserved loop on stem C (see Fig. 1) should contribute to the recent efforts to understand the role of ions for maintaining RNA fold(s) and their ability to promote catalysis.[61,62]

We can draw parallels between L1L conformational transition and presumed catalytic mechanism of other ribozymes, such as HHR and the recently crystallized class I RNA polymerase (CIRP).[9,63] Both HHR and L1L have active and inactive states, but the sequences and structures are completely different, and the conformational changes they undergo are also unique. The HHR "open", pre-catalytic state was captured in several X-ray crystal structures of the "minimal" HHR,[64–70] whereas the "closed", presumably catalytically active state was captured in the form of the "full-length" HHR.[71,72] HHR does not, strictly speaking, require divalent metal ions for catalysis,[73,74] whereas L1L has a very specific requirement for a Mg^{2+} at the active site.[7,17–19,24] The crystallographically observed active-site divalent metal ion binding mode is identical in the minimal and full-length HHR, indicating that this ion is present in both active and inactive states and may play a role in stabilizing the tertiary contacts of the former.[72] In contrast, a Mg^{2+} is bound in the active site only in the presumed active conformation.[24] In the case of the catalytic step, computational investigations suggested that when bound in the active site, Mg^{2+} can lower the pK_a of a secondary alcohol to initiate the general base step of catalysis.[75,76] Structural data obtained from MD of the reactant state of L1L[26] implied that a Mg^{2+} resolved in the crystal structure can lower the pK_a of the $H_{O3'}$ of the substrate through direct coordination.

CIRP is a synthetic ribozyme that catalyzes the same reaction as L1L although with multiple catalytic turnovers as a direct result of the *in vitro* evolutionary optimization. Both ribozymes' X-ray structures have a Mg^{2+} resolved near the catalytic site. For both ribozymes, an additional Mg^{2+} has been proposed to stabilize a triphosphate group in the active site as suggested by analysis of CIRP crystal structure[9,63] and theoretical investigations on L1L and CIRP.[26,77] These findings fit in a much larger context of DNA and RNA polymerases, nucleases, and transposases for which the proximity of two Mg^{2+} to the active site has been proposed to enhance substrate recognition and catalytic specificity.[78]

An important question is how one can translate the current simulation results into experimentally testable hypothesis. Previous studies on other RNA systems were able to extract structural and dynamical information using NMR, fluorescence resonance energy transfer, small-angle X-ray scattering, and time-resolved and single-molecule experiments[51,79–85] to explore global and local changes in the secondary and tertiary structure similar to those undergone by L1L. In this work, we (i) identify a dynamical set of hydrogen-bonding patterns in the active site that change when L1L reaches its activated state and (ii) provide a comprehensive structural characterization of the three proposed L1L substates using relative angles between the three stems (Table 1). These are observables that can be readily obtained from the aforementioned experimental techniques and can serve in testing our results.

Conclusion

L1L is a synthetic *in vitro* selected ribozyme that catalyzes regiospecific formation of 3′-to-5′ phosphodiester linkages, a reaction characteristic of all extant present-day RNA and DNA protein polymerases. The L1L allosterically controlled catalytic mechanism is a complex process consisting of large-scale conformational motions of several flexible constitutive structural motifs, such as a three-way junction and a bulged loop that is able to form tertiary distant contacts with a noncanonically base-paired, flexible active site. This is an intriguing set of properties that make L1L unique and the understanding of L1L catalytic mechanism appealing. L1L exhibits a high degree of conformational flexibility suggested by its crystal structure that reveals two largely different conformations that differ by a reorientation of one of the stems by 80 Å. It is of great interest, therefore, to gain further insight into the nature of these two states in solution and the set of pathways that allow their interconversion, as well as understand the manner in which this inherent flexibility can assist in promoting catalysis. In a first paper,[26] we identified the hinge points responsible for L1L activation that are located in restricted regions and can be characterized by a limited set of virtual torsions. In this work, we report a comprehensive characterization of the conformational landscape of L1L using an enhanced sampling method coined FSN. Using data from more than 1 μs of simulations, we employed network theory and analysis of CSNs to identify the essential three-state/two-step process whereby the L1L can transition between catalytically active and inactive states. In the first step, we show that the docking of stem C, through the U_{38} loop, disrupts the $A_{51}\square\rightarrow G_1$ base pair and elicits the adoption of the two active-site conformations with high probabilities of initiating the catalytic process. In the second step, the degree of conservation of the L1L three-way junction spanned by the θ_{44} virtual torsion is directly correlated with the stem C swing between the I_2 to I_1 states. This work provides deep insights into the

molecular mechanism of allosteric control that may be useful in guiding the design of ribozyme-based biotechnology and that have important evolutionary implications.

Materials and Methods

Molecular system setup

Initial ribozyme structures were taken from a 2.6 Å resolution crystal structure of two co-crystallized L1L ligation product conformers (Protein Data Bank ID: 2OIU)[24] and equilibrated structures from previously reported simulations.[26] The ribozyme was then immersed in a rhombododecahedral TIP3P[86] water box with an edge length of 120 Å. The ionic atmosphere consisted of Na^+ and Cl^- ions that were added at random positions at least 5.0 Å away from any RNA atom to neutralize the system and reach the physiologic concentration of 0.14 M. The total number of particles in the system was 104,000.

MD simulation protocol

Simulations were performed with the NAMD simulation package (version 2.7)[87] using the all-atom Cornell *et al.* force field (parm99),[88] generated with the AMBER 10 package[89–91] and TIP3P water model.[86] Periodic boundary conditions were used along with the isothermal–isobaric ensemble (*NPT*) at 1 atm and 300 K using Nosé–Hoover–Langevin pressure piston control[92,93] with a decay period of 100.0 fs and a damping time scale of 50 fs, and the Langevin thermostat with a damping coefficient of 0.1 ps^{-1}. The smooth particle mesh Ewald method[94,95] was employed with a B-spline interpolation order of 6 and the default κ value used in NAMD. The fast fourier transform grid points used for the lattice directions were chosen using ~1.0 Å spacing. Nonbonded interactions were treated using an atom-based cutoff of 12.0 Å with switching of nonbond potential beginning at 10.0 Å. Numerical integration was performed using the leapfrog Verlet algorithm with 1 fs time step.[96] Covalent bond lengths involving water hydrogens were constrained using the SHAKE algorithm.[97]

Equilibration procedure

The initial system first underwent 5000 steps of energy minimization and then the solute (RNA) atoms were kept fixed, letting the water and counterions and coions to be equilibrated for 15 ns. The solute atom positions were then energy minimized and then allowed to move under harmonic restraints over 5 ns. The initial harmonic force constant was 5.0 kcal mol^{-1} and exponentially released over 5 ns. The whole system was energy minimized, followed by an unconstrained dynamics simulation beginning from 30 K under constant pressure of 1 atm. The temperature then was increased to 300 K at a rate of 5 K per 10 ps. The motions and relaxation of solvent and counter- and coions are notoriously slow to converge in nucleic acid simulations,[98] and careful equilibration is critical. For each simulation, a total of 20 ns of equilibration (15 ns of water/ion relaxation and 5 ns of solvent and structure relaxation) was carried out before production of the trajectory used for analysis.[98]

Analysis

Analysis of the trajectories was carried out using VMD (version 1.8.7).[99] L1L Stem A (A) was defined as residues 1 to 7 and 45 to 71, stem B (B) was defined as residues 8 to 17, stem C (C) was defined as residues 18 to 44, the junction (J) was defined as residues 6 to 10, 15 to 22, and 39 to 47. For structural comparison, we do not include any of the three GAAA tetraloops since they were introduced in L1L structure for crystallization purposes.[24] The virtual torsions[100,101] are designated as follows: η_n is the virtual torsion angle defined by the atoms $C4'_{n-1}$, P_n, $C4'_n$, and P_{n+1}; θ_n is the virtual torsion angle defined by the atoms P_n, $C4'_n$, P_{n+1} and $C4'_{n+1}$, where n indicates the corresponding nucleotide residue number. In addition, for each stem, a characteristic vector v_{char}^{stem} is defined as follows: (1) we choose two adjacent canonical base pairs (i and j),[48] and for each, we define a vector that is determined by the positions of the C1′ atoms (v_i, v_j); (2) the characteristic vector is defined as the cross product of the two vectors: $v_{char}^{stem} = v_i \times v_j$.

The characteristic vector v_{char}^{stem} of each stem can be used to describe their relative positions and appears to be a sufficient descriptor for the L1L structure since individual stem keeps its domain structure reasonably rigid in the time scale of our simulations. The major hinge points are located in the restricted region of the junctions between stems.[26] Across this work, we make use of the characteristic vector of stem C. The two adjacent canonical base pairs that were chosen to define v_{char}^{C} were U_{20}–A_{41} and C_{21}=G_{40}. Prior to the calculation of the characteristic vectors, all structures are RMS fitted against the position of stem A in the undocked crystallized conformer.

Focused sampling on networks (FSN) method

FSN creates a network to represent the connectivity between the conformational states and identifies the most undersampled nodes that reside on pathways connecting the stable substates along the network. These nodes are used as launching points for *swarms of trajectories*[102–104] to enhance sampling and statistical analysis of these regions. The FSN method is an iterative procedure, with the following steps (see also Fig. 6):

- Build a *connected network* of conformational states from the accumulated simulation structures. Each network node represents a cluster node from a partitional incremental *k*-way clustering procedure of the entire set of simulation structures.
- Evaluate the *sampling density* for each node. The sampling density of each node is estimated through counting the number of simulation structures within a threshold distance from the centroid of the node.
- Evaluate the *traversal count* for each node, by determining how many times the node is traversed by all the shortest paths between all the nodes of the network using the Floyd–Warshall algorithm.[105,106]

- Select *launching points* for the next round of swarm of trajectories. The nodes with *lowest sampling density* and *highest traversal counts* are chosen as launching points for the next swarm of trajectories.

In this work, we use 300 nodes to build the FSN network obtained from an incremental partitional clustering of the characteristic vectors of stem C. The distance between nodes of the FSN network is the cosine similarity and the network is built by connecting each node with its closest eight neighbor nodes. For our specific case, this is a criterion to assure the building of a connected network (i.e., a network with the property that there exists at least one path connecting any two nodes).

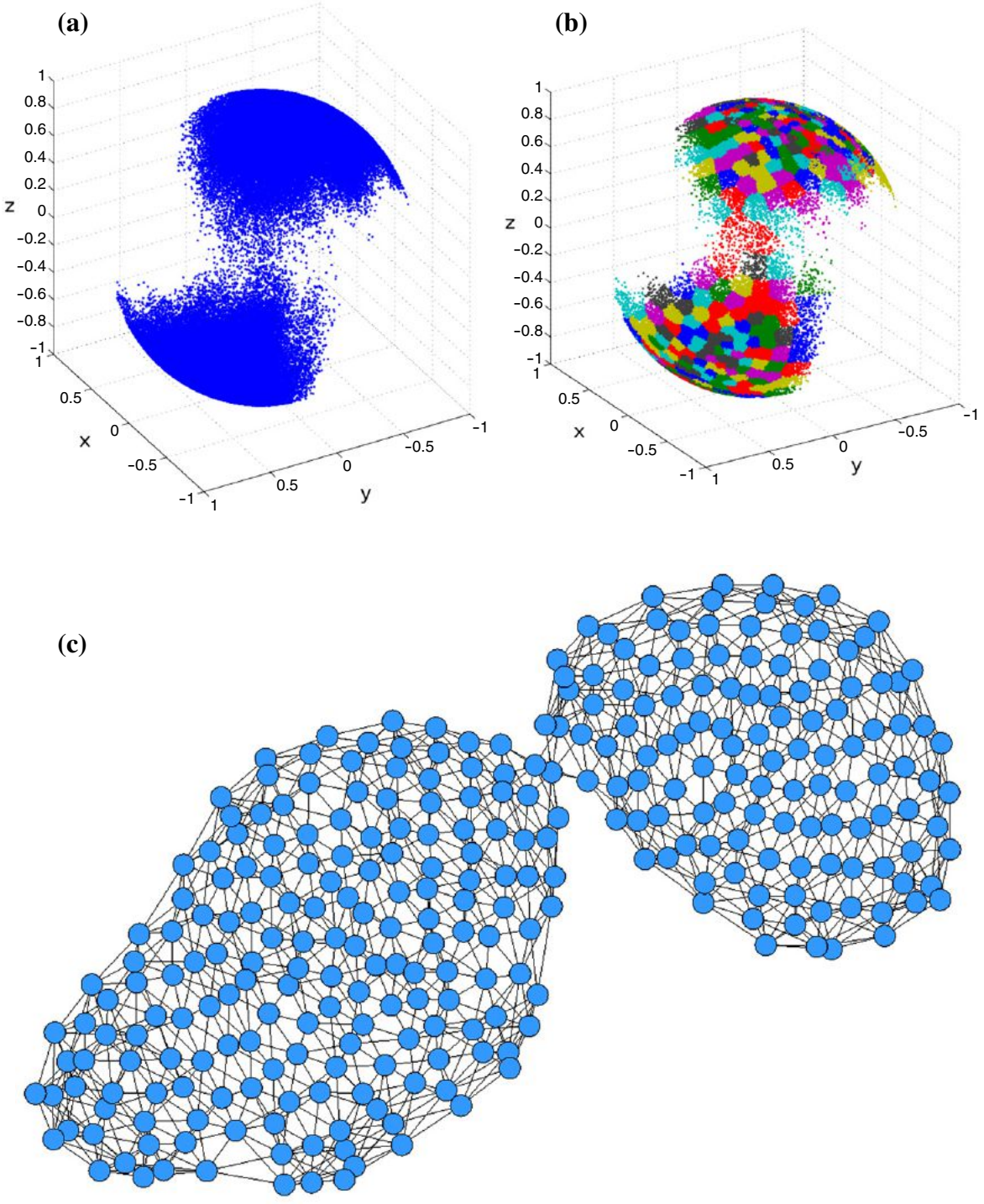

Fig. 6 (*legend on page 118*)

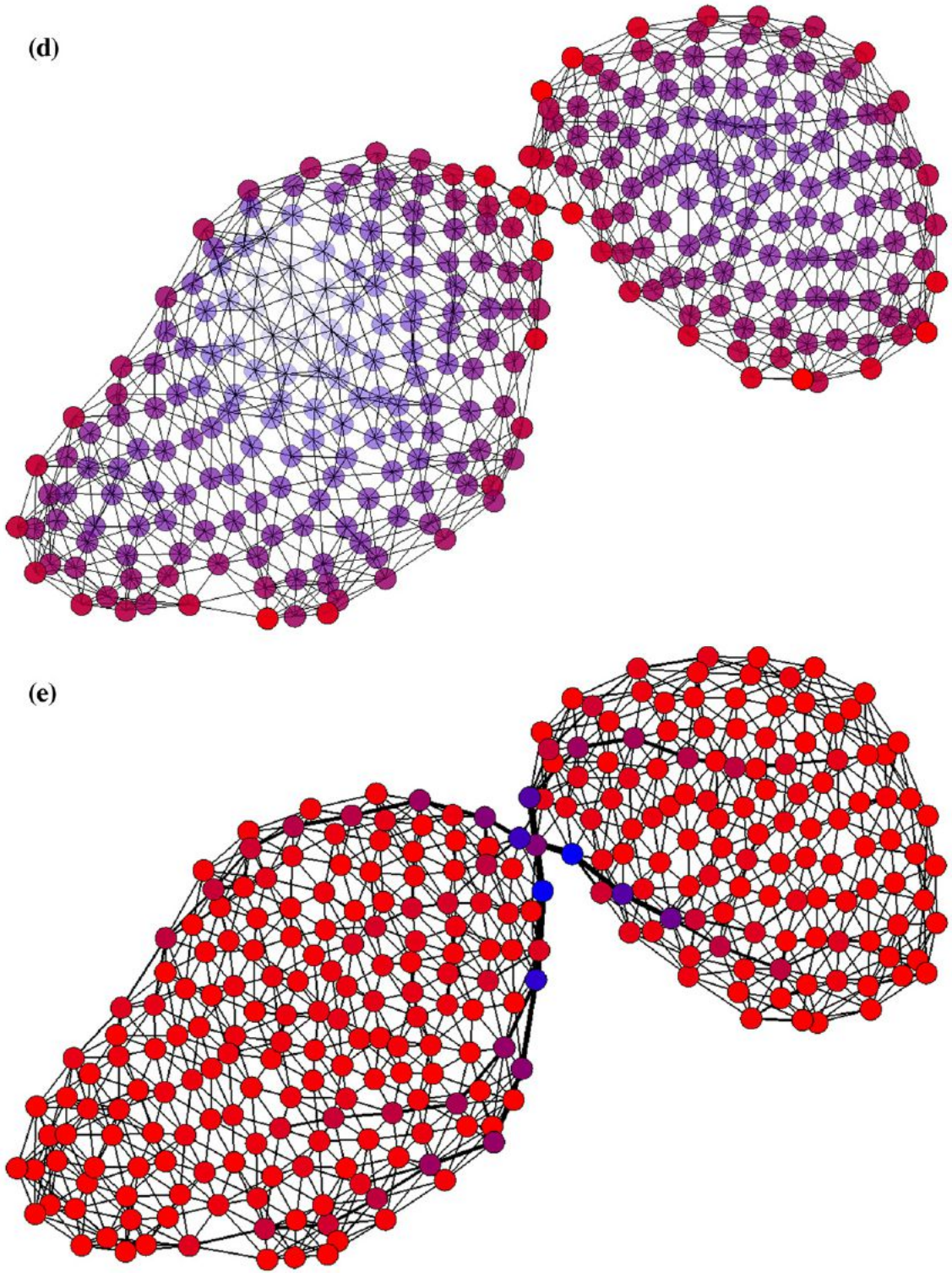

Fig. 6 (*legend on next page*)

For every iteration step of the FSN procedure, a set of five launching points are selected for a trajectory swarm. An MD simulation is started with the coordinates and velocities from each specific FSN iteration point. The number of nodes used to build the FSN network (300) was chosen, anticipating the need for better isolation on the origin of the swarm of trajectories, a decision made before data collection. As the CSN built after data collection shows, the number of clusters needed to isolate the interface nodes can be lower (100).

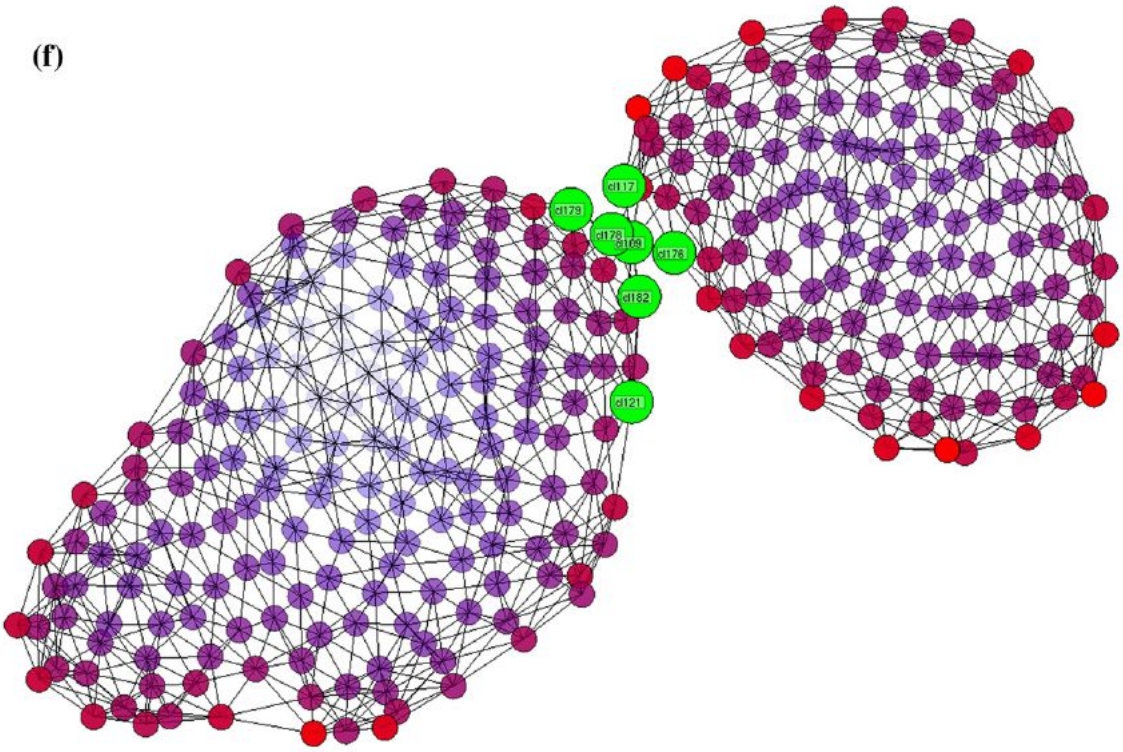

Fig. 6. To focus the sampling around relatively undersampled regions of the conformational space that might reside along the L1L switching pathway, we designed an iterative strategy (FSN). The projection of the structures on a set of order parameters (a) is clustered using an incremental partitional 300-way clustering technique (see Materials and Methods). Data points belonging to the same cluster have the same color. A network is built so that each one of the previously determined clusters is assimilated with a node (b). Each node is connected to its closest eight neighbors using a cosine similarity/distance-based function (c). For each of the clusters, a sampling density is calculated that is displayed here using a red–blue color map, with the red corresponding to low densities and the blue corresponding to the high densities (d). Each of the network nodes is associated to a traversal count, that is, the number of times it is crossed by all the shortest paths along the graph between all the nodes of the graph (e). A similar red–blue color map is used here, with red corresponding to lower values and blue corresponding to higher values of the traversal counts. The swarm of trajectories' launching points (here colored green) are chosen as the nodes that have a low sampling density and high traversal counts (f).

Swarms of trajectories have been used before in conjunction with so-called *string methods* enhanced sampling techniques.[102,104,107–109] Other methods have been used to study the thermodynamic or kinetic aspects of conformational transitions of nucleic acids using an explicit treatment of solvent and ionic atmosphere. For example, combinations of umbrella sampling,[110] targeted MD,[111–113] transition path sampling,[114,115] nudged elastic band,[116] replica exchange,[117] or long MD simulation were applied to RNA hairpins folding,[118,119] DNA A to B form transition,[120,121] or local openings of duplexes.[122–124] Other methods such as milestoning[125] or stochastic path approach[125,126] have been used to study other types of biomolecular systems.

Clustering

All clustering calculations were performed with CLUTO[127] software. The *k*-way incremental partitional clustering scheme used here is realized using $k-1$ repeated bisections, followed by a global optimization of the solution. It is a *top–down* clustering approach. The algorithm maximizes the cosine similarity function between each data point and the centroid of the cluster that is assigned to. The clustering procedure is repeated 20 times to avoid local minimums of the overall similarity function.[128] During agglomerative clustering, each of the objects are initially assigned to their own cluster and then pairs of clusters are repeatedly merged based on the same similarity function until all the initial objects are grouped in one cluster. Agglomerative clustering is a *bottom–up* procedure. Hybrid schemes that in the first stage cluster the data using a partitional method and in the second use agglomerative method are called constrained agglomerative algorithms and have been shown to efficiently lead to better solutions than partitional or agglomerative schemes alone.[47]

Network manipulation, rendering, and analysis were carried out with NAViGaTOR[129,130] and yED graph editor (yFiles software, Tubingen, Germany). Shortest path searches between all the nodes of the network as well as the determination of the traversal counts were done using a Floyd–Warshall algorithm[105,106] as implemented in NAViGaTOR.

Forced unfolding/undocking simulation

The unfolding simulation of stem C from the active site starting from the docked conformation was prepared as described previously,[26] except that a biasing potential was added to disrupt the canonical base pair between U_{38} and A_{51}. The unfolding biasing potential is a step-like potential of 5 kcal mol^{-1} turned on gradually from 3.5 Å to 5.0 Å during a 15-ns simulation period. During this period, the base pairing of the active site was restrained using harmonic potentials (using a spring constant of 5 kcal mol·$Å^{-2}$) centered at the equilibrium values specific to the hydrogen binding pattern found in crystal.

Acknowledgements

The authors are grateful for financial support provided by the National Institutes of Health (GM084149 to D.M.Y. and GM087721 to W.G.S.). Computational resources were provided by the Minnesota Supercomputing Institute and the National Science Foundation (TeraGrid grant TG-CHE100072).

Supplementary Data

Supplementary data to this article can be found online at http://dx.doi.org/10.1016/j.jmb.2012.06.035

References

1. Bartel, D. P. & Szostak, J. W. (1993). Isolation of new ribozymes from a large pool of random sequences. *Science*, **261**, 1411–1418.
2. Bagby, S. C., Bergman, N. H., Shechner, D. M., Yen, C. & Bartel, D. P. (2009). A class I ligase ribozyme with reduced Mg^{2+} dependence: selection, sequence analysis, and identification of functional tertiary interactions. *RNA*, **15**, 2129–2146.
3. Ekland, E. H., Szostak, J. W. & Bartel, D. P. (1995). Structurally complex and highly active RNA ligases derived from random RNA sequences. *Science*, **269**, 364–369.
4. Ikawa, Y., Tsuda, K., Matsumura, S. & Inoue, T. (2004). De novo synthesis and development of an RNA enzyme. *Proc. Natl Acad. Sci. USA*, **101**, 13750–13755.
5. Jaeger, L., Wright, M. C. & Joyce, G. F. (1999). A complex ligase ribozyme evolved in vitro from a group I ribozyme domain. *Proc. Natl Acad. Sci. USA*, **96**, 14712–14717.
6. McGinness, K. E. & Joyce, G. F. (2003). In search of an RNA replicase ribozyme. *Chem. Biol.* **10**, 5–14.
7. Robertson, M. P. & Ellington, A. D. (1999). In vitro selection of an allosteric ribozyme that transduces analytes to amplicons. *Nat. Biotechnol.* **17**, 62–66.
8. Rogers, J. & Joyce, G. F. (1999). A ribozyme that lacks cytidine. *Nature*, **402**, 323–325.
9. Shechner, D. M., Grant, R. A., Bagby, S. C., Koldobskaya, Y., Piccirilli, J. A. & Bartel, D. P. (2009). Crystal structure of the catalytic core of an RNA-polymerase ribozyme. *Science*, **326**, 1271–1275.
10. Crick, F. H. (1968). The origin of the genetic code. *J. Mol. Biol.* **38**, 367–379.
11. Orgel, L. E. (1968). Evolution of the genetic apparatus. *J. Mol. Biol.* **38**, 381–393.
12. Robertson, M. P. & Joyce, G. F. (2010). The origins of the RNA world. *Cold Spring Harbor Perspect. Biol.*
13. Woese, C. (1968). *Genetic Code.* Harper & Row, New York, NY.
14. Chen, X., Li, N. & Ellington, A. D. (2007). Ribozyme catalysis of metabolism in the RNA world. *Chem. Biodivers.* **4**, 633–655.
15. Ellington, A. D. & Szostak, J. W. (Aug 1990). In vitro selection of rna molecules that bind specific ligands. *Nature*, **346**, 818–822.
16. Marshall, K. A. & Ellington, A. D. (1999). Training ribozymes to switch. *Nat. Struct. Biol.* **6**, 992–994.
17. Robertson, M. P. & Ellington, A. D. (2000). Design and optimization of effector-activated ribozyme ligases. *Nucleic Acids Res.* **28**, 1751–1759.
18. Robertson, M. P. & Ellington, A. D. (2001). In vitro selection of nucleoprotein enzymes. *Nat. Biotechnol.* **19**, 650–655.
19. Robertson, M. P., Knudsen, S. M. & Ellington, A. D. (2004). In vitro selection of ribozymes dependent on peptides for activity. *RNA*, **10**, 114–127.
20. Bunka, D. H. J. & Stockley, P. G. (2006). Aptamers come of age—at last. *Nat. Rev. Microbiol.* **4**, 588–596.
21. Famulok, M. & Mayer, G. (2011). Aptamer modules as sensors and detectors. *Acc. Chem. Res.* **44**, 1349–1358.
22. Gold, L., Janjic, N., Jarvis, T., Schneider, D., Walker, J. J., Wilcox, S. K. & Zichi, D. (2012). Aptamers and the RNA world, past and present. *Cold Spring Harbor Perspect. Biol.* **4**, 1–9.
23. Keefe, A. D., Pai, S. & Ellington, A. (2010). Aptamers as therapeutics. *Nat. Rev. Drug Discov.* **9**, 537–550.
24. Robertson, M. P. & Scott, W. G. (2007). The structural basis of ribozyme-catalyzed RNA assembly. *Science*, **315**, 1549–1550.
25. Joyce, G. F. (2007). A glimpse of biology's first enzyme. *Science*, **315**, 1507–1508.
26. Giambaşu, G. M., Lee, T.-S., Sosa, C. P., Robertson, M. P., Scott, W. G. & York, D. M. (2010). Identification of dynamical hinge points of the L1 ligase molecular switch. *RNA*, **16**, 769–780.
27. Bailor, M. H., Sun, X. & Al-Hashimi, H. M. (2010). Topology links RNA secondary structure with global conformation, dynamics, and adaptation. *Science*, **327**, 202–206.
28. Landweber, L. F. & Pokrovskaya, I. D. (1999). Emergence of a dual-catalytic RNA with metal-specific cleavage and ligase activities: the spandrels of RNA evolution. *Proc. Natl Acad. Sci. USA*, **96**, 173–178.
29. Scott, W. G. (2007). Morphing the minimal and full-length hammerhead ribozymes: implications for the cleavage mechanism. *Biol. Chem.* **388**, 727–735.
30. Bowman, G. R., Voelz, V. A. & Pande, V. S. (2011). Taming the complexity of protein folding. *Curr. Opin. Struct. Biol.* **21**, 4–11.

31. Caflisch, A. (2006). Network and graph analyses of folding free energy surfaces. *Curr. Opin. Struct. Biol.* **16**, 71–78.
32. Gfeller, D., De Los Rios, P., Caflisch, A. & Rao, F. (2007). Complex network analysis of free-energy landscapes. *Proc. Natl Acad. Sci. USA*, **104**, 1817–1822.
33. Krivov, S. V. & Karplus, M. (2004). Hidden complexity of free energy surfaces for peptide (protein) folding. *Proc. Natl Acad. Sci. USA*, **101**, 14766–14770.
34. Noé, F. & Fischer, S. (2008). Transition networks for modeling the kinetics of conformational change in macromolecules. *Curr. Opin. Struct. Biol.* **18**, 154–162.
35. Noé, F., Horenko, I., Schütte, C. & Smith, J. C. (2007). Hierarchical analysis of conformational dynamics in biomolecules: transition networks of metastable states. *J. Chem. Phys.* **126**, 155102.
36. Rao, F. & Caflisch, A. (2004). The protein folding network. *J. Mol. Biol.* **342**, 299–306.
37. Böde, C., Kovács, I. A., Szalay, M. S., Palotai, R., Korcsmáros, T. & Csermely, P. (2007). Network analysis of protein dynamics. *FEBS Lett.* **581**, 2776–2782.
38. Bowman, G. R. & Pande, V. S. (2010). Protein folded states are kinetic hubs. *Proc. Natl Acad. Sci. USA*, **107**, 10890–10895.
39. Ihalainen, J. A., Bredenbeck, J., Pfister, R., Helbing, J., Chi, L., van Stokkum, I. H. M. *et al.* (2007). Folding and unfolding of a photoswitchable peptide from picoseconds to microseconds. *Proc. Natl Acad. Sci. USA*, **104**, 5383–5388.
40. Ihalainen, J. A., Paoli, B., Muff, S., Backus, E. H. G., Bredenbeck, J., Woolley, G. A. *et al.* (2008). Alpha-helix folding in the presence of structural constraints. *Proc. Natl Acad. Sci. USA*, **105**, 9588–9593.
41. Li, C.-B., Yang, H. & Komatsuzaki, T. (2008). Multiscale complex network of protein conformational fluctuations in single-molecule time series. *Proc. Natl Acad. Sci. USA*, **105**, 536–541.
42. Prada-Gracia, D., Nes, J. G. G., Echenique, P. & Falo, F. (2009). Exploring the free energy landscape: from dynamics to networks and back. *PLoS Comput. Biol.* **5**, e1000415.
43. Yang, S., Banavali, N. K. & Roux, B. (2009). Mapping the conformational transition in Src activation by cumulating the information from multiple molecular dynamics trajectories. *Proc. Natl Acad. Sci. USA*, **106**, 3776–3781.
44. Han, J., Kamber, M. & Tung, A. K. H. (2001). Spatial clustering methods in data mining: a survey. In *Geographic Data Mining and Knowledge Discovery* (Miller, H. & Han, J., eds), pp. 1–29, Taylor and Francis, London, UK.
45. Jain, A. K., Murty, M. N. & Flynn, P. J. (1999). Data clustering: a review. *ACM Comput. Surv.* **31**, 264–323.
46. Zhao, Y. & Karypis, G. (2005). Data clustering in life sciences. *Mol. Biotechnol.* **31**, 55–80.
47. Zhao, Y., Karypis, G. & Fayyad, U. (2005). Hierarchical clustering algorithms for document datasets. *Data Min. Knowl. Discovery*, **10**, 141–168.
48. Leontis, N. B., Stombaugh, J. & Westhof, E. (2002). The non-Watson–Crick base pairs and their associated isostericity matrices. *Nucleic Acids Res.* **30**, 3497–3531.
49. Robertson, M. P., Hesselberth, J. R. & Ellington, A. D. (2001). Optimization and optimality of a short ribozyme ligase that joins non-Watson–Crick base pairings. *RNA*, **7**, 513–523.
50. Dethoff, E. A., Chugh, J., Mustoe, A. M. & Al-Hashimi, H. M. (2012). Functional complexity and regulation through RNA dynamics. *Nature*, **482**, 322–330.
51. Haller, A., Soulière, M. F. & Micura, R. (2011). The dynamic nature of RNA as key to understanding riboswitch mechanisms. *Acc. Chem. Res.* **44**, 1339–1348.
52. Cruz, J. A. & Westhof, E. (2009). The dynamic landscapes of RNA architecture. *Cell*, **136**, 604–609.
53. Solomatin, S. V., Greenfeld, M., Chu, S. & Herschlag, D. (2010). Multiple native states reveal persistent ruggedness of an RNA folding landscape. *Nature*, **463**, 681–684.
54. Woodside, M. T., Anthony, P. C., Behnke-Parks, W. M., Larizadeh, K., Herschlag, D. & Block, S. M. (2006). Direct measurement of the full, sequence-dependent folding landscape of a nucleic acid. *Science*, **314**, 1001–1004.
55. Woodson, S. A. (2010). Compact intermediates in RNA folding. *Annu. Rev. Biophys.* **39**, 61–77.
56. Bailor, M. H., Mustoe, A. M., Brooks, C. L. & Al-Hashimi, H. M. (2011). Topological constraints: using RNA secondary structure to model 3D conformation, folding pathways, and dynamic adaptation. *Curr. Opin. Struct. Biol.* **21**, 296–305.
57. Lescoute, A. & Westhof, E. (2006). Topology of three-way junctions in folded RNAs. *RNA*, **12**, 83–93.
58. Lilley, D. M. (2000). Structures of helical junctions in nucleic acids. *Q. Rev. Biophys.* **33**, 109–159.
59. de la Peña, M., Dufour, D. & Gallego, J. (2009). Three-way RNA junctions with remote tertiary contacts: a recurrent and highly versatile fold. *RNA*, **15**, 1949–1964.
60. Keating, K. S., Humphris, E. L. & Pyle, A. M. (2011). A new way to see RNA. *Q. Rev. Biophys.* **44**, 433–466.
61. Draper, D. E. (2008). RNA folding: thermodynamic and molecular descriptions of the roles of ions. *Biophys. J.* **95**, 5489–5495.
62. Draper, D. E., Grilley, D. & Soto, A. M. (2005). Ions and RNA folding. *Annu. Rev. Biophys. Biomol. Struct.* **34**, 221–243.
63. Piccirilli, J. A. & Koldobskaya, Y. (2011). Crystal structure of an RNA polymerase ribozyme in complex with an antibody fragment. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* **366**, 2918–2928.
64. Dunham, C. M., Murray, J. B. & Scott, W. G. (2003). A helical twist-induced conformational switch activates cleavage in the hammerhead ribozyme. *J. Mol. Biol.* **332**, 327–336.
65. Feig, A. L., Scott, W. G. & Uhlenbeck, O. C. (1998). Inhibition of the hammerhead ribozyme cleavage reaction by site-specific binding of Tb(III). *Science*, **279**, 81–84.
66. Murray, J. B., Szöke, H., Szöke, A. & Scott, W. G. (2000). Capture and visualization of a catalytic RNA enzyme–product complex using crystal lattice trapping and x-ray holographic reconstruction. *Mol. Cell*, **5**, 279–287.
67. Murray, J. B., Terwey, D. P., Maloney, L., Karpeisky, A., Usman, N., Beigelman, L. & Scott, W. G. (1998).

The structural basis of hammerhead ribozyme self-cleavage. *Cell*, **92**, 665–673.
68. Pley, H. W., Flaherty, K. M. & McKay, D. B. (1994). Three-dimensional structure of a hammerhead ribozyme. *Nature*, **372**, 68–74.
69. Scott, W. G., Finch, J. T. & Klug, A. (1995). The crystal structure of an all-RNA hammerhead ribozyme: a proposed mechanism for RNA catalytic cleavage. *Cell*, **81**, 991–1002.
70. Scott, W. G., Murray, J. B., Arnold, J. R. P., Stoddard, B. L. & Klug, A. (1996). Capturing the structure of a catalytic RNA intermediate: the hammerhead ribozyme. *Science*, **274**, 2065–2069.
71. Martick, M., Lee, T.-S., York, D. M. & Scott, W. G. (2008). Solvent structure and hammerhead ribozyme catalysis. *Chem. Biol.* **15**, 332–342.
72. Martick, M. & Scott, W. G. (2006). Tertiary contacts distant from the active site prime a ribozyme for catalysis. *Cell*, **126**, 309–320.
73. Curtis, E. A. & Bartel, D. P. (2001). The hammerhead cleavage reaction in monovalent cations. *RNA*, **7**, 546–552.
74. Murray, J. B., Seyhan, A. A., Walter, N. G., Burke, J. M. & Scott, W. G. (1998). The hammerhead, hairpin and VS ribozymes are catalytically proficient in monovalent cations alone. *Chem. Biol.* **5**, 587–595.
75. Lee, T.-S., Silva Lopez, C., Giambaşu, G. M., Martick, M., Scott, W. G. & York, D. M. (2008). Role of Mg^{2+} in Hammerhead ribozyme catalysis from molecular simulation. *J. Am. Chem. Soc.* **130**, 3053–3064.
76. Wong, K.-Y., Lee, T.-S. & York, D. M. (2011). Active participation of the Mg^{2+} ion in the reaction coordinate of RNA self-cleavage catalyzed by the hammerhead ribozyme. *J. Chem. Theory Comput.* **7**, 1–3.
77. Sgrignani, J. & Magistrato, A. (2012). The structural role of Mg^{2+} ions in a class I RNA polymerase ribozyme: a molecular simulation study. *J. Phys. Chem. B*, **116**, 2259–2268.
78. Yang, W., Lee, J. Y. & Nowotny, M. (2006). Making and breaking nucleic acids: two-Mg^{2+}-ion catalysis and substrate specificity. *Mol. Cell*, **22**, 5–13.
79. Al-Hashimi, H. M. & Walter, N. G. (2008). RNA dynamics: it is about time. *Curr. Opin. Struct. Biol.* **18**, 321–329.
80. Bardaro, M. F. & Varani, G. (2011). Examining the relationship between RNA function and motion using nuclear magnetic resonance. *Wiley Interdiscip. Rev.: RNA*, **3**, 122–132.
81. Pollack, L. (2011). Time resolved SAXS and RNA folding. *Biopolymers*, **95**, 543–549.
82. Rinnenthal, J., Buck, J., Ferner, J., Wacker, A., Fürtig, B. & Schwalbe, H. (2011). Mapping the landscape of RNA dynamics with NMR spectroscopy. *Acc. Chem. Res.* **44**, 1292–1301.
83. Tinoco, I., Chen, G. & Qu, X. (2010). RNA reactions one molecule at a time. *Cold Spring Harbor Perspect. Biol.* **2**, a003624.
84. Zhao, L. & Xia, T. (2009). Probing RNA conformational dynamics and heterogeneity using femtosecond time-resolved fluorescence spectroscopy. *Methods*, **49**, 128–135.
85. Zhuang, X. (2005). Single-molecule RNA science. *Annu. Rev. Biophys. Biomol. Struct.* **34**, 399–414.
86. Jorgensen, W. L., Chandrasekhar, J., Madura, J. D., Impey, R. W. & Klein, M. L. (1983). Comparison of simple potential functions for simulating liquid water. *J. Chem. Phys.* **79**, 926–935.
87. Phillips, J. C., Braun, R., Wang, W., Gumbart, J., Tajkhorshid, E., Villa, E. *et al.* (2005). Scalable molecular dynamics with NAMD. *J. Comput. Chem.* **26**, 1781–1802.
88. Cornell, W. D., Cieplak, P., Bayly, C. I., Gould, I. R., Ferguson, D. M., Spellmeyer, D. C. *et al.* (1995). A second generation force field for the simulation of proteins, nucleic acids and organic molecules. *J. Am. Chem. Soc.* **117**, 5179–5197.
89. Case, D., Darden, T., Cheatham, T., III, Simmerling, C., Wang, J., Duke, R. *et al.* (2002). *AMBER 10.* University of California San Francisco, San Francisco, CA.
90. Case, D. A., Cheatham, T. E., III, Darden, T., Gohlke, H., Luo, R., Merz, K. M. *et al.* (2005). The Amber biomolecular simulation programs. *J. Comput. Chem.* **26**, 1668–1688.
91. Pearlman, D. A., Case, D. A., Caldwell, J. W., Ross, W. R., Cheatham, T., III, DeBolt, S. *et al.* (1995). AMBER, a package of computer programs for applying molecular mechanics, normal mode analysis, molecular dynamics and free energy calculations to simulate the structure and energetic properties of molecules. *Comput. Phys. Commun.* **91**, 1–41.
92. Feller, S., Zhang, Y., Pastor, R. & Brooks, B. (1995). Constant pressure molecular dynamics simulation: the Langevin piston method. *J. Chem. Phys.* **103**, 4613–4621.
93. Martyna, G. J., Tobias, D. J. & Klein, M. L. (1994). Constant pressure molecular dynamics algorithms. *J. Chem. Phys.* **101**, 4177–4189.
94. Essmann, U., Perera, L., Berkowitz, M. L., Darden, T., Hsing, L. & Pedersen, L. G. (1995). A smooth particle mesh Ewald method. *J. Chem. Phys.* **103**, 8577–8593.
95. Sagui, C. & Darden, T. A. (1999). molecular dynamics simulations of biomolecules: long-range electrostatic effects. *Annu. Rev. Biophys. Biomol. Struct.* **28**, 155–179.
96. Allen, M. & Tildesley, D. (1987). *Computer Simulation of Liquids.* Oxford University Press, Oxford, UK.
97. Ryckaert, J. P., Ciccotti, G. & Berendsen, H. J. C. (1977). Numerical Integration of the cartesian equations of motion of a system with constraints: molecular dynamics of *n*-alkanes. *J. Comput. Phys.* **23**, 327–341.
98. Ponomarev, S. Y., Thayer, K. M. & Beveridge, D. L. (2004). Ion motions in molecular dynamics simulations on DNA. *Proc. Natl Acad. Sci. USA*, **101**, 14771–14775.
99. Humphrey, W., Dalke, A. & Schulten, K. (1996). VMD: visual molecular dynamics. *J. Mol. Graphics*, **14**, 33–38.
100. Duarte, C. M. & Pyle, A. M. (1998). Stepping through an RNA structure: a novel approach to conformational analysis. *J. Mol. Biol.* **284**, 1465–1478.
101. Wadley, L. M., Keating, K. S., Duarte, C. M. & Pyle, A. M. (2007). Evaluating and learning from RNA pseudotorsional space: quantitative validation of a reduced representation for RNA structure. *J. Mol. Biol.* **372**, 942–957.
102. Gan, W., Yang, S. & Roux, B. (2009). Atomistic view of the conformational activation of Src kinase using the string method with swarms-of-trajectories. *Biophys. J.* **97**, L8–L10.

103. Pan, A. C. & Roux, B. (2008). Building Markov state models along pathways to determine free energies and rates of transitions. *J. Chem. Phys.* **129**, 064107.
104. Pan, A. C., Sezer, D. & Roux, B. (2008). Finding transition pathways using the string method with swarms of trajectories. *J. Phys. Chem. B*, **112**, 3432–3440.
105. Floyd, R. W. (1962). Algorithm 97: shortest path. *Commun. ACM*, **5**, 345.
106. Warshall, S. (1962). A theorem on Boolean matrices. *J. Assoc. Comput. Mach.* **9**, 11–12.
107. E, W., Ren, W. & Vanden-Eijnden, E. (2002). String method for the study of rare events. *Phys. Rev. B*, **66**, 052301.
108. W., E., Ren, W. & Vanden-Eijnden, E. (2005). Finite temperature string method for the study of rare events. *J. Phys. Chem. B*, **109**, 6688–6693.
109. Maragliano, L., Fischer, A., Vanden-Eijnden, E. & Ciccotti, G. (2006). String method in collective variables: minimum free energy paths and isocommittor surfaces. *J. Chem. Phys.* **125**, 24106.
110. Torrie, G. & Valleau, J. (1977). Nonphysical sampling distributions in Monte Carlo free-energy estimation: umbrella sampling. *J. Comput. Phys.* **23**, 187–199.
111. Coluzza, I., Sprik, M. & Ciccotti, G. (2003). Constrained reaction coordinate dynamics for systems with constraints. *Mol. Phys.* **101**, 2885–2894.
112. Mülders, T., Krüger, P. & Schlitter, J. (1996). Free energy as the potential of mean constraint force. *J. Chem. Phys.* **104**, 4869.
113. Schlitter, J., Engels, M. & Krüger, P. (1994). Targeted molecular dynamics: a new approach for searching pathways of conformational transitions. *J. Mol. Graphics*, **12**, 84–89.
114. Bolhuis, P. G., Chandler, D., Dellago, C. & Geissler, P. L. (2002). Transition path sampling: throwing ropes over rough mountain passes, in the dark. *Annu. Rev. Phys. Chem.* **53**, 291–318.
115. Dellago, C., Bolhuis, P. G. & Chandler, D. (1998). Efficient transition path sampling: Application to Lennard-Jones cluster rearrangements. *J. Chem. Phys.* **108**, 9236.
116. Jonsson, H. & Mills, G. (1997). Nudged elastic band method for finding minimum energy paths of transitions. In *Classical and Quantum Dynamics in Condensed Phase Simulations.* (Berne, B. J., Ciccotti, G. & Coker, D. F., eds) Ch. 6.
117. Sugita, Y. & Okamoto, Y. (1999). Replica-exchange molecular dynamics method for protein folding. *Chem. Phys. Lett.* **314**, 141–151.
118. Bowman, G. R., Huang, X., Yao, Y., Sun, J., Carlsson, G., Guibas, L. J. & Pande, V. S. (2008). Structural insight into RNA hairpin folding intermediates. *J. Am. Chem. Soc.* **130**, 9676–9678.
119. DePaul, A. J., Thompson, E. J., Patel, S. S., Haldeman, K. & Sorin, E. J. (2010). Equilibrium conformational dynamics in an RNA tetraloop from massively parallel molecular dynamics. *Nucleic Acids Res.* **38**, 4856–4867.
120. Banavali, N. K. & Roux, B. (2005). Free energy landscape of A-DNA to B-DNA conversion in aqueous solution. *J. Am. Chem. Soc.* **127**, 6866–6876.
121. Noy, A., Pérez, A., Laughton, C. A. & Orozco, M. (2007). Theoretical study of large conformational transitions in DNA: the B↔A conformational change in water and ethanol/water. *Nucleic Acids Res.* **35**, 3330–3338.
122. Bergonzo, C., Campbell, A. J., de los Santos, C., Grollman, A. P. & Simmerling, C. (2011). Energetic preference of 8-oxoG eversion pathways in a DNA glycosylase. *J. Am. Chem. Soc.* **133**, 14504–14506.
123. Hagan, M. F., Dinner, A. R., Chandler, D. & Chakraborty, A. K. (2003). Atomistic understanding of kinetic pathways for single base-pair binding and unbinding in DNA. *Proc. Natl Acad. Sci. USA*, **100**, 13922–13927.
124. Hu, J., Ma, A. & Dinner, A. R. (2008). A two-step nucleotide-flipping mechanism enables kinetic discrimination of DNA lesions by AGT. *Proc. Natl Acad. Sci. USA*, **105**, 4615–4620.
125. Faradjian, A. K. & Elber, R. (2004). Computing time scales from reaction coordinates by milestoning. *J. Chem. Phys.* **120**, 10880–10889.
126. Elber, R., Meller, J. & Olender, R. (1999). Stochastic path approach to compute atomically detailed trajectories: application to the folding of C peptide. *J. Phys. Chem. B*, **103**, 899–911.
127. Karypis, G. (2006). CLUTO: A Clustering Toolkit (release 2.1.2).
128. Zhao, Y. & Karypis, G. (2004). Criterion functions for document clustering: experiments and analysis. *Mach. Learn.* **55**, 311–331.
129. Brown, K. R., Otasek, D., Ali, M., McGuffin, M. J., Xie, W., Devani, B. *et al.* (2009). NAViGaTOR: network analysis, visualization and graphing. *Bioinformatics*, **25**, 3327–3329.
130. McGuffin, M. J. & Jurisica, I. (2009). Interaction techniques for selecting and manipulating subgraphs in network visualizations. *IEEE Trans. Vis. Comput. Graph.* **15**, 937–944.