COMMUNICATIONS IN INFORMATION AND SYSTEMS
Vol. 10, No. 1, pp. 53-68, 2010

004

A MOTION PLANNING APPROACH TO STUDYING MOLECULAR MOTIONS*

LYDIA TAPIA†, SHAWNA THOMAS†, AND NANCY M. AMATO†

Abstract. While structurally very different, protein and RNA molecules share an important attribute. The motions they undergo are strongly related to the function they perform. For example, many diseases such as Mad Cow disease or Alzheimer's disease are associated with protein misfolding and aggregation. Similarly, RNA folding velocity may regulate the plasmid copy number, and RNA folding kinetics can regulate gene expression at the translational level. Knowledge of the stability, folding, kinetics and detailed mechanics of the folding process may help provide insight into how proteins and RNAs fold. In this paper, we present an overview of our work with a computational method we have adapted from robotic motion planning to study molecular motions. We have validated against experimental data and have demonstrated that our method can capture biological results such as stochastic folding pathways, population kinetics of various conformations, and relative folding rates. Thus, our method provides both a detailed view (e.g., individual pathways) and a global view (e.g., population kinetics, relative folding rates, and reaction coordinates) of energy landscapes of both proteins and RNAs. We have validated these techniques by showing that we observe the same relative folding rates as shown in experiments for structurally similar protein molecules that exhibit different folding behaviors. Our analysis has also been able to predict the same relative gene expression rate for wild-type MS2 phage RNA and three of its mutants.

1. Introduction. Molecular motions play an essential role in many biochemical processes. For example, as proteins fold to their native, functional state, they sometimes undergo critical conformational changes that affect their functionality, e.g., diseases such as Mad Cow disease or Alzheimer's disease are associated with protein misfolding and aggregation [11]. Knowledge of the stability, folding, kinetics and detailed mechanics of the folding process may help provide insight into how and why the protein misfolds. In addition, it has recently been found that some RNA functions are determined not just by the sequence and the resulting native state but also by the folding process itself, e.g., RNA folding velocity may regulate the plasmid copy number [21, 33] or RNA folding kinetics can regulate gene expression at the translational level [40].

Since it is difficult to experimentally observe molecular motions, computational methods for studying such issues are essential. Traditional computational approaches for generating folding trajectories such as molecular dynamics (MD) [38, 22, 15, 18] and Monte Carlo simulation [14, 34] are so expensive that they can only be applied to relatively small structures (e.g., proteins with fewer than 130 amino acids [69]) even when they use massive computational resources, such as tens of thousands of PCs in the Folding@Home project [35, 48] or large supercomputers [69]. Statistical mechani-

*Dedicated to Michael Waterman on the occasion of his 67th birthday.

†Parasol Lab, Department of Computer Science and Engineering, Texas A&M University, College Station, TX 77845. E-mail: {ltapia, sthomas, amato}@cse.tamu.edu

cal models have been applied to compute statistics related to the energy landscape for protein [44, 1, 43, 41, 16] and RNA [9, 8, 68] molecules. While computationally more efficient than molecular dynamics or Monte Carlo simulation, these methods do not produce individual pathway trajectories and are limited to studying global averages of the energy landscape and kinetics.

In this paper, we present an overview of a novel computational method for studying molecular motions that we have developed and validated against experimental data in preliminary work [3, 2, 4, 51, 53, 56, 57, 58, 59, 60, 63, 64]. Our strategy represents a trade-off between methods such as molecular dynamics and Monte Carlo simulations that provide detailed individual folding trajectories and techniques such as statistical mechanical methods that provide global landscape statistics. Our approach, derived from *probabilistic roadmap methods (PRMs)* [30] originally developed for robotic motion planning, builds a graph corresponding to an *approximate map* of the molecule's energy landscape that encodes many (typically thousands) folding pathways. Although the individual pathways produced are not as detailed as trajectories generated from a molecular dynamics simulation, we have shown that they can be used to study landscape and pathway properties such as secondary structure formation order. We were even able to observe subtle folding differences between protein G and its mutants, NuG1 and NuG2 [64], an important 'benchmark' set developed by the Baker Lab [45]. In recent work, we developed new tools to study folding kinetics, such as the rate at which folding occurs or the time evolution of the population of particular interesting states. We validated these techniques by showing that we observe the same relative folding rates as shown in experiments for some small protein [60] and RNA [58, 59] molecules, and that our analysis predicts the same relative gene expression rate for wild-type MS2 phage RNA and three of its mutants [58, 59].

2. PRMs for Molecular Motion – Method Overview. Our approach, derived from *probabilistic roadmap methods (PRMs)* [30] originally developed for robotic motion planning, builds a graph corresponding to an *approximate map* of the molecule's energy landscape that encodes many (typically thousands of) folding pathways, see Figure 1. Our PRM-based method follows the general PRM paradigm: first conformations (graph vertices or map nodes) are sampled from the molecule's energy landscape (Figure 1(b)), and then transitions between 'nearby' conformations are encoded as graph or map edges (Figure 1(c)). As in nature, our strategy favors low energy conformations and transitions. In particular, during the sampling phase, lower energy samples have a higher retention probability, and during the node connection phase, each connection is assigned a weight to reflect its energetic feasibility. The energetic feasibility of a transition is determined by the energies of all the intermediate conformations along the transition. Thus, shortest paths in the map correspond to the most energetically feasible paths in the map, and these maps encode thousands

of feasible pathways.

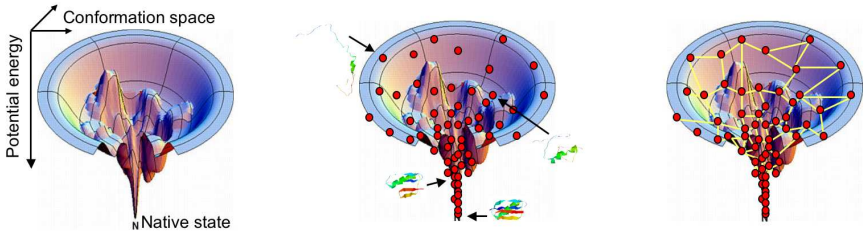

FIG. 1. *(a) The energy landscape is the set of all conformations and their associated energy. Building an approximate map of the energy landscape consists of two steps: (b) conformation sampling and (c) connecting samples together with feasible transitions.*

PRM-based approaches have been applied to several molecular domains. Singh, Latombe, and Brutlag first applied PRMs to protein/ligand binding [49]. In subsequent work, our group applied another PRM variant to this problem [7]. Our group was the first to apply PRMs to model protein folding pathways [4, 3, 53, 52, 50, 62, 61, 64, 60] and RNA folding kinetics [57, 58, 59, 55]. Subsequent to our work, a number of groups have used PRMs to study proteins. The work of Apaydin et al. [6, 5] is similarly motivated but differs from ours in several aspects. First, they model the protein at a much coarser level, considering all secondary structure elements in the native state to be already formed and rigid. Second, while our focus is on studying the transition process, their focus has been to compare the PRM approach with other computational methods such as Monte Carlo simulation. More recently, Cortes and Simeon used a PRM-based approach to model long loops in proteins [13, 12], and Chiang et al. [10] applied PRMs to calculate quantities related to protein folding kinetics such as P_{fold} and Φ-value analysis.

Map Analysis Tools for Folding Kinetics. Maps provide an approximate model of the molecule's energy landscape. With this model, we can use map-based analysis tools to study important kinetic measures such as folding rates, equilibrium distributions, population kinetics, transition states, and reaction coordinates. In recent work [57, 60, 58, 59], we developed two such techniques: Map-based Master Equation solution (MME) and Map-based Monte Carlo simulation (MMC). These tools are inspired by existing kinetics tools (namely, traditional master equation formalism and standard Monte Carlo simulation) but can be applied to much larger molecules because they work on approximate landscape models instead of the complete, detailed energy landscape.

Map-based Master Equation (MME). The traditional master equation calculation gives insight into the folding rate, the equilibrium distribution, population kinetics, and transition states [29, 66]. In master equation formalism, the stochastic folding

process is represented as a differential equation describing the probability of the folding process to be in a given state:

$$d\mathbf{p}(t)/dt = M\mathbf{p}(t) \tag{1}$$

where $\mathbf{p}$(t) is the probability of the folding to be in a given state at time t and M is a matrix of transition-rate constants. The solution to the master equation is a set of eigenvectors and eigenvalues for the matrix M. The spectrum of eigenvalues is composed of n modes where n is the number of conformations in our map. If sorted by magnitude in ascending order, the eigenvalues include the zero-valued equilibrium Boltzmann distribution and several small magnitude eigenvalues. The small, non-zero eigenvalues correspond to the eigenvectors that influence the global folding rate while large magnitude eigenvalues correspond to fast folding modes, i.e., those that fold in a burst and die away quickly.

Master equation formalism requires a detailed model of the energy landscape. This has been typically done by enumerating energy landscapes — feasible only for small molecular models or segments. Instead, we apply the master equation formalism to our maps (MME) by assigning each node in our map to a row (and column) in the matrix M. The transition rate k_{ij} is computed from the weight w_{ij} of the edge from i to j as $k_0 e^{-w_{ij}}$ where k_0 is a constant coefficient adjusted according to experimental results.

A key advantage of MME over the traditional master equation solution is that the cost of MME is proportional to the map size (i.e., the size of the landscape model) [60], whereas the traditional master equation is usually applied to a fully enumerated landscape.

Map-based Monte Carlo (MMC). Folding is a stochastic process [29]. In our early work [4, 3, 53, 50, 61, 57, 64], we simply extracted smallest weight paths from the map to study folding. However, this does not mirror the stochastic folding process. In recent work [60, 58, 59], we developed MMC to extract paths randomly based on transition probabilities. Similar to traditional Monte Carlo simulation, our method starts from a random node in the map and iteratively chooses a next node based on the transition probabilities. Just as in MME, the transition probability k_{ij} to transit from node i to node j is computed from the edge weight w_{ij} as $k_0 e^{-w_{ij}}$ where k_0 is a constant adjusted according to experimental results.

The standard Monte Carlo method [29, 46] simulates this random walk in the real (or complete) energy landscape. These simulations can be computationally intensive since at each step they must calculate the local energy landscape to choose the next step. Instead, we apply Monte Carlo simulation directly to our maps. Thus, we are able to work on larger molecules with our approximated landscape model at only a small computational cost. Previously, the size of the energy landscape limited Monte Carlo simulations to small molecules (e.g., all-atom 56 residue protein [47])

or molecules whose kinetics were restricted in some way (e.g., Higgs performed a Monte-Carlo simulation on a 135 residue RNA using only stem-based conformations [23]).

An important feature of MMC is its computational efficiency in both time and memory usage. For instance, we have shown that the cost of MMC is proportional to the map size and model complexity [59]. For 53 to 86 residue proteins, this translates into 23 to 36 minutes of computation on a 2.4 GHz desktop PC with 512 MB RAM [60]. Correspondingly, the memory usage is also reduced. For example, on a 18 nucleotide hairpin RNA, 485MB of memory is required to store 1000 traditional Monte Carlo RNA pathways produced from the program Kinfold [19]. On the other hand, 1000 MMC pathways are stored in a file of just 61MB and a map of 684KB [59].

3. Protein Motions. We have successfully applied our PRM framework for molecular motions to study protein folding and motion [4, 3, 53, 50, 52, 62, 61, 64, 60]. Here we first describe the specifics of our protein application (e.g., protein model, energy functions, map construction details) and then provide results.

3.1. Method Details. Protein Model and Energy Function. We model the protein as an articulated linkage. Using a standard modeling assumption for proteins that bond angles and bond lengths are fixed [54], the only degrees of freedom (dof) in our model are the backbone's phi and psi torsional angles which are modeled as revolute joints with values $[0, 2\pi)$.

We have used both a coarse energy function similar to [38] and an all atom energy model [36]. For the coarse model, we use a step function approximation of the van der Waals component and model all side chains as equal radii spheres with zero dof. If two spheres are too close (e.g., their centers are < 2.4Å during sampling and < 1.0Å during connection), a very high potential is returned. Otherwise, the potential is:

$$U_{tot} = \sum_{restraints} K_d\{[(d_i - d_0)^2 + d_c^2]^{1/2} - d_c\} + E_{hp} \tag{2}$$

where K_d is 100 kJ/mol and $d_0 = d_c = 2$Å as in [38]. The first term represents constraints favoring known secondary structure through main-chain hydrogen bonds and disulphide bonds, and the second term is the hydrophobic effect. The hydrophobic effect (E_{hp}) is computed as follows: if two hydrophobic residues are within 6Å of each other, then the potential is decreased by 20 kJ/mol. A detailed description of our potential can be found in [4].

Biased Sampling. As previously discussed, samples are retained based on their energy. In our protein work, a sample q, with potential energy E_q, is accepted with probability:

$$Prob(\text{accept q}) = \begin{cases} 1 & \text{if } E_q < E_{\min} \\ \frac{E_{\max} - E_q}{E_{\max} - E_{\min}} & \text{if } E_{\min} \le E_q \le E_{\max} \\ 0 & \text{if } E_q > E_{\max} \end{cases} \tag{3}$$

where $E_{\min}$ is the potential energy of the open chain and $E_{\max}$ is $2E_{\min}$.

The map produced by our technique is an approximation of the protein's energy landscape. The quality of the approximation depends on the sampling strategy. Generally, we are most interested in regions 'near' the native state and so seek to concentrate sampling there. In our original work [4, 3, 53, 50], we obtained a denser distribution of samples near the native state through an iterative sampling process where we apply small Gaussian perturbations to existing conformations, beginning with the native state. This approach works fairly well, but still requires many samples (e.g., 10,000) for relatively small proteins (e.g., 60–100 residues). In [64], we used rigidity analysis [26, 27, 28, 25, 37] to determine which portions of the protein to perturb. This approach increased the protein size we can handle.

Connection. For each node in the map, we attempt to connect it with its k nearest neighbors with a straight-line in the protein's energy landscape. The weight for the edge (q_1, q_2) is a function of the intermediate conformations along the edge $\{q_1 = c_0, c_1, \ldots, c_{n-1}, c_n = q_2\}$, where the number of intermediate conformations depends on the resolution, which is a parameter of the method. For each pair of consecutive conformations c_i and c_{i+1}, the probability P_i of transitioning from c_i to c_{i+1} depends on the difference in their potential energies $\Delta E_i = E(c_{i+1}) - E(c_i)$:

$$P_i = \begin{cases} e^{\frac{-\Delta E_i}{kT}} & \text{if } \Delta E_i > 0 \\ 1 & \text{if } \Delta E_i \leq 0 \end{cases} \tag{4}$$

This keeps the detailed balance between two adjacent states, and enables the weight of an edge to be computed by summing the negative logarithms of the probabilities for consecutive pairs of conformations in the sequence. (Negative logs are used since each $0 \leq P_i \leq 1$.) A similar weight function, with different probabilities, was used in [49].

MMC Transition Probability. We apply MMC to protein folding as described previously and set the transition probabilities as follows. We cluster the edge weights into disjoint buckets. Bucket probabilities (Q_{ij}) are assigned in a biased Gaussian fashion that favors clear discrimination of low edge weights, yet reflects the relative differences between edges of all weights. The transition probability between two states, k_{ij}, is calculated as

$$k_{ij} = \begin{cases} \frac{Q_{ij}}{1+\sum_{j=0}^{n-1} Q_{ij}} & \text{if } j \neq i \\ \frac{1}{1+\sum_{j=0}^{n-1} Q_{ij}} & \text{if } j = i \end{cases} \tag{5}$$

where n is the number of outgoing edges from node i. This ensures the sum of all probabilities (including the self-transition probability) out of node i is 1.

3.2. Results. Here we present results that have validated our technique against experimental data by comparing secondary structure formation order along folding pathways, relative folding rates, and population kinetics for several small proteins.

Secondary Structure Formation Order Validation. Proteins are composed of secondary structure elements (i.e., α-helices and β-sheets). Experimental methods, such as hydrogen exchange mass spectrometry and pulse labeling, can investigate protein folding by identifying which parts of the structure are most exposed or most protected [65]. From this data, one can infer the secondary structure formation order. In [3, 50, 53] we compared the secondary structure formation order of folding pathways extracted from our maps to experimental results [39]. We cluster paths together if they have the same secondary structure formation order. We then define the dominant formation order as the formation order of the largest path cluster. Our results are in good agreement with known experimental results for many small proteins (e.g., 60–100 amino acids) [64].

Case Study of Proteins G, L, and Two Mutants of Protein G. Proteins G, L, and mutants of protein G, NuG1 and NuG2 [45], present a good test case for our technique because they are known to fold differently despite having similar structure (see Figure 2). All proteins are composed of a central α-helix and a 4-stranded β-sheet: β strands 1 and 2 form the N-terminal hairpin (β1-2) and β strands 3 and 4 form the C-terminal hairpin (β3-4). Native state out-exchange experiments and pulse labeling/competition experiments for proteins G and L indicate that β1-2 forms first in protein L, and β3-4 forms first in protein G [39]. This is consistent with Φ-value analysis on G [42] and L [31]. In Nauli et al. [45], protein G is mutated to increase the stability of β1-2. Φ-value analysis indicates that the hairpin formation order for both NuG1 and NuG2 is switched from the wild-type. Nauli et al. also show that NuG1 and NuG2 fold 100 times faster than protein G.

Our initial iterative Gaussian sampling strategy was able to accurately capture the folding differences between protein G and L, but not between protein G and NuG1 or NuG2 [53]. Our iterative rigidity-based sampling strategy, however, was able to also capture the correct folding behavior of NuG1 and NuG2 [64], see Table 1.

In addition to detecting the correct folding behavior, our rigidity-based technique also helped explain the stability shift in NuG1 and NuG2. Figure 2 displays the rigidity maps of each protein's native state. A rigidity map is a graphical view of the rigid and flexible portions of the structure. Black regions correspond to rigid regions and green regions correspond to slightly flexible regions. In all four proteins, the central α helix remains completely rigid, and we also see increased rigidity in β1-2 from protein G to NuG1 and NuG2 as suggested in [45].

Finally, we have used MME and MMC to compute the relative folding rates between protein G, NuG1, and NuG2 from our maps [60]. Figure 3(a) shows the magnitudes of the 5 smallest eigenvalues for each protein as calculated by MME. Recall that the smallest non-zero eigenvalues represent the rate-limiting barrier in the folding process. Therefore, they have the largest impact on the global folding rate. As seen in the magnitude of the second eigenvalue in Figure 3(a), protein G folds much

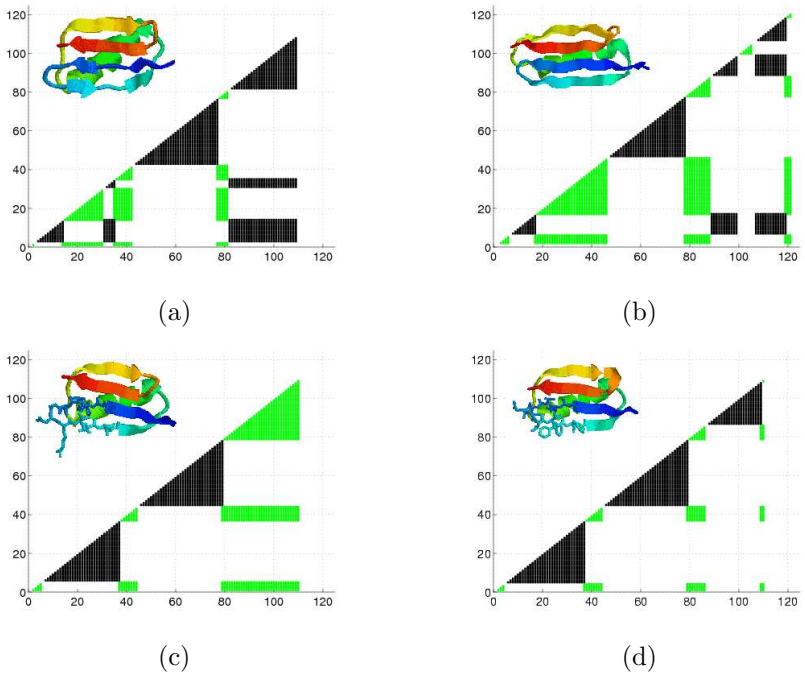

FIG. 2. *Ribbons diagrams and rigidity maps of the native state for protein G (a), protein L (b), and mutants NuG1 (c) and NuG2 (d). Mutated portions are displayed in wireframe. In the rigidity maps, rigid clusters are black and dependent hinge sets are shaded/green. Figure originally published in [64].*

slower than the two mutants, NuG1 and NuG2. Also, NuG1 and NuG2 fold at very similar rates. This trend is also seen in the curves of the MMC population kinetics for protein G (Figure 3(b)) and mutants, NuG1 (Figure 3(b)) and NuG2 (Figure 3(c)). Both of these computational results match what has been seen in lab experiments [45].

4. RNA Motion. In our previous work [57, 55, 58, 59], we developed several successful map construction techniques for RNA. In particular, the *Probabilistic Boltzmann Sampling (PBS)* method builds the smallest maps (up to 10 orders of magnitude smaller than completely enumerated maps) and enables us to study much larger RNA, up to 200 nucleotides. Similarly with proteins, we provide several map-based analysis tools including a Map-based Master Equation (MME) and Map-based Monte Carlo (MMC) simulation to extract folding kinetics.

4.1. Method Details. RNA Model and Energy Function. In the results demonstrated here, we focus on the formation of *secondary structure*. *Secondary structure* is a planar representation of an RNA conformation, which is commonly used to study RNA folding [70, 71, 24]. We adopt the definition in [24] that eliminates other types of contacts that are not physically favored. We use a common energy function

TABLE 1

Comparison of secondary structure formation orders for proteins G, L, NuG1, and NuG2 with known experimental results: [1]*hydrogen out-exchange experiments [39],* [2]*pulsed labeling/competition experiments [39], and* [3]Φ*-value analysis [45]. Brackets indicate no clear order. In all cases, our technique predicted the secondary structure formation order seen in experiment. Only formation orders greater than 1% are shown.*

Protein	Experimental Formation Order	Rigidity Formation Order	%
G	[α,β1,β3,β4], β2[1] [α,β4], [β1,β2,β3][2]	α, β3-4, β1-2	99.4
L	[α,β1,β2,β4], β3[1] [α,β1], [β2,β3,β4][2]	β1-2, α, β3-4	100.0
NuG1	β1-2, β3-4[3]	α, β1-2, β3-4	97.6
		β1-2, α, β3-4	1.6
NuG2	β1-2, β3-4[3]	α, β1-2, β3-4	96.6
		β1-2, α, β3-4	1.1
		β3-4, β1-2, α	1.1

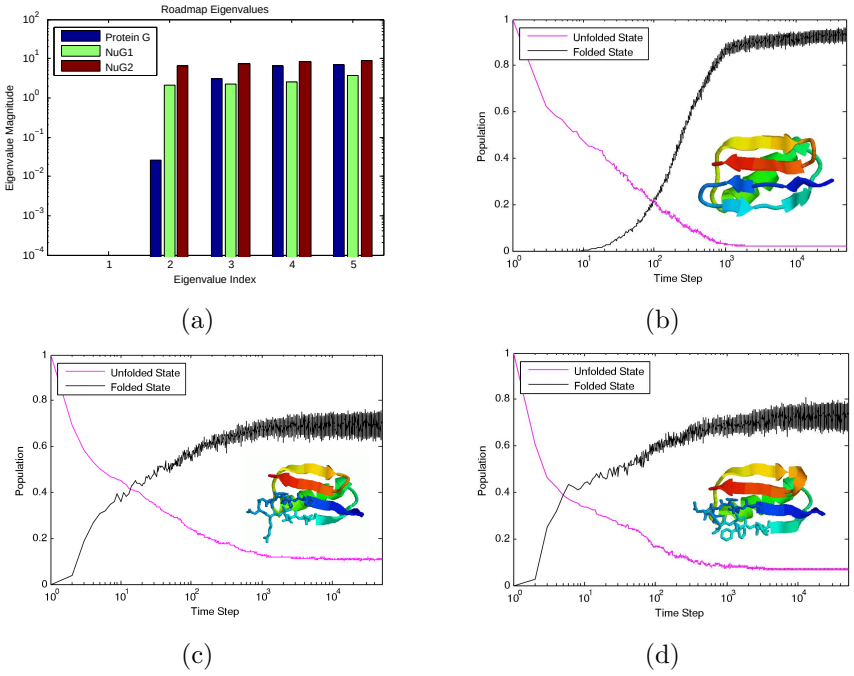

FIG. 3. *Population kinetics for protein G and mutants NuG1 and NuG2. NuG1 and NuG2 are experimentally known to fold 100 times faster than protein G [45]. (a) MME eigenvalue comparison. (b-d) MMC kinetics for protein G (b), NuG1 (c), and NuG2 (d). Figure originally published in [60].*

called the Turner or nearest neighbor rules [70].

Biased Sampling. Our sampling method, Probabilistic Boltzmann Sampling (PBS), uses Wuchty's method [67] to enumerate suboptimal (low energy) conformations within a given energy threshold. We take these suboptimal conformations as

"seeds" and include additional random conformations. Then, we use a probabilistic filter to retain a subset of the conformations based on their Boltzmann distribution factors. For a given conformation q with free energy E_q, the probability of keeping it is:

$$Prob(\text{accept q}) = \begin{cases} e^{\frac{-(E_q - E_0)}{kT}} & \text{if } (E_q - E_0) > 0 \\ 1 & \text{if } (E_q - E_0) \leq 0 \end{cases} \tag{6}$$

where E_0 is a reference energy threshold that we can use to control the number of samples kept.

Connection. Similar to Section 3.1, we calculate a weight w_{ij} for edge (q_i, q_j) that reflects the Boltzmann transition probability between q_i and q_j. First, we determine the energy barrier (the maximum energetic cost) E_b between q_i and q_j. Then, we calculate the Boltzmann transition probability k_{ij} (or transition rate) of moving from q_i to q_j using Metropolis rules [17]:

$$k_{ij} = \begin{cases} e^{\frac{-\Delta E}{kT}} & \text{if } \Delta E > 0 \\ 1 & \text{if } \Delta E \leq 0 \end{cases} \tag{7}$$

where $\Delta E = max(E_b, E_j) - E_i$, k is the Boltzmann constant, and T is the temperature. Note that the same energy barrier E_b is also used to estimate the transition probability k_{ji}, so the calculation satisfies the detailed balance. As in protein folding, the edge weight w_{ij} is the negative logarithm of the transition probability.

MMC Transition Probability. We apply MMC to RNA folding as described in Section 2. Because the edge weight w_{ij} encodes the transition probability k_{ij} between two endpoints i and j, we can calculate k_{ij} as $k_0 e^{-w_{ij}}$ where k_0 is a constant adjusted according to experimental results. Results presented here are generated using a fast variant of the standard Monte Carlo method [46]. Full technical details are available in previous work [58, 59].

4.2. Results. Here we demonstrate results for several different RNA. We have validated our technique against other computational methods and have shown that we can capture the same folding kinetics as seen in experiment.

Computational Validation: 1k2g. 1k2g (CAGACUUCGGUCGCAGAGAUGG) is a 22 nucleotide RNA with a hairpin native state [32]. Figure 4(a–e) compares the population kinetics of the native state using (a) standard Monte Carlo simulation (implemented by Kinfold [19]), (b) Map-based Monte Carlo simulation on a fully enumerated map (12,137 conformations), (c) Map-based Monte Carlo simulation on a map with our PBS sampling method (42 conformations), and (d) the master equation on a PBS map (42 conformations). While the fully enumerated map (b) is the most accurate model, it is not feasible to enumerate RNA with more than 40 nucleotides and numerical limitations in computing the eigenvalues and eigenvectors limit the master equation to small maps (e.g., up to 10,000 conformations). The population

kinetics curves all have similar features: the population first increases quickly, then gradually decreases, and eventually stabilizes at the equilibrium (final) distribution, which are all roughly 80%. Hence, these analysis methods all yield similar results and indicate that the PBS map (c,d) effectively approximates the energy landscape with less than 0.4% of all possible conformations.

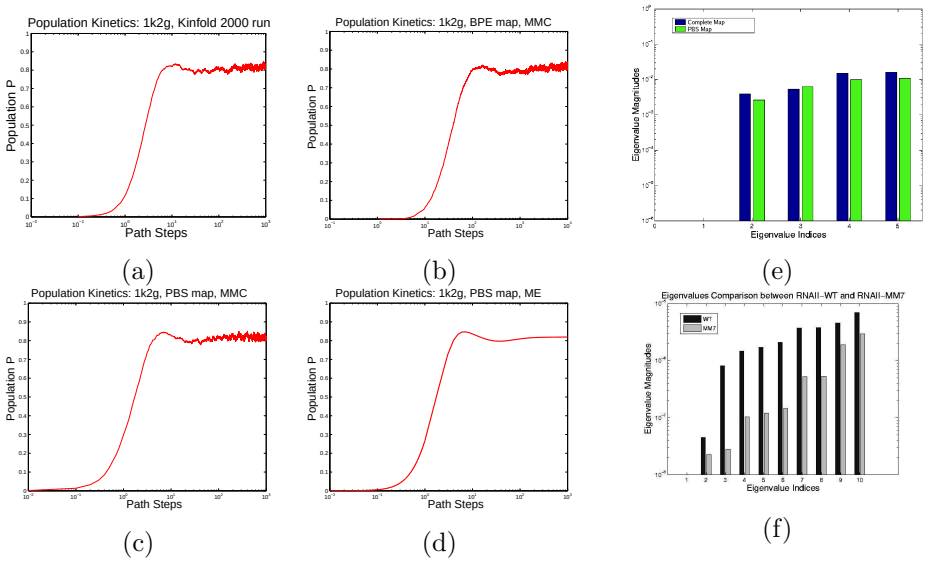

FIG. 4. *The population kinetics of the native state of 1k2g (a-e): (a) Kinfold Monte Carlo simulation, (b) our MMC simulation on a fully enumerated map (12,137 conformations), (c) our MMC simulation on a PBS map (42 conformations), and (d) master equation solution on the PBS map (42 conformations). All analysis techniques produce similar population kinetics curves and similar equilibrium distributions. (e) Comparison of the eigenvalues of 1k2g by the master equation on a fully enumerated map (12,137 conformations) and new PBS map (42 conformations). Both eigenvalues are similar between the different maps. (f) Comparison of the 10 smallest non-zero eigenvalues (i.e., the folding rates) for WT and MM7 of ColE1 RNAII as computed by the master equation. The overall folding rate of WT is faster than MM7 matching experimental data. Figure originally published in [59].*

Experimental Validation: ColE1 and Mutant MM7. ColE1 RNAII regulates the replication of E. coli ColE1 plasmids through its folding kinetics [21, 33]. The slower it folds, the higher the plasmid replication rate. A specific mutant, MM7, differs from the wild-type (WT) by a single nucleotide out of the 200 nucleotide sequence. This mutation causes it to fold slower while maintaining the same thermodynamics of the native state. Thus, the overall plasmid replication rate increases in the presence of MM7 over the WT. We studied this difference computationally by computing the folding rates of both WT and MM7 using MME and comparing their eigenvalues (the smallest non-zero eigenvalue corresponds to the folding rate). As seen in Figure 4(f), all eigenvalues of WT are larger than MM7 indicating that WT folds faster. Thus, our method correctly estimated the functional level of the new mutant.

Experimental Validation: MS2 Phage RNA Mutants. MS2 phage RNA (135 nucleotides) regulates the expression rate of phage MS2 maturation protein [20, 33] at the translational level. It works as a regulator only when a specific subsequence (the SD sequence) is open (i.e., does not form base-pair contacts). Since this SD sequence is closed in the native state, the RNA can only regulate the expression rate before the folding process finishes. Thus, its function is based on its folding *kinetics* and not the final native structure. Three mutants have been studied that have similar thermodynamic properties as the wild-type (WT) but have different kinetics and therefore different gene expression rates. Experimental results indicate that mutant CC3435AA has the highest gene expression rate, WT and mutant U32C are similar, and mutant SA has the lowest rate [20, 33].

We estimate the gene expression rate by integrating the opening probability of the SD sequence over the entire folding process. Note that the RNA regulates gene expression only when the SD opening probability is "high enough". We used thresholds ranging from 0.2 to 0.6 to estimate the gene expression rate. Thresholds higher than 0.6 will yield zero opening probability for WT and most mutants and thus cannot be correlated to experimental results. Similarly, we do not consider thresholds lower than 0.2, because otherwise mutant SA would be active even in the equilibrium condition which does not correspond to experimental results. Table 2 shows our simulation results. For most thresholds, mutant CC3435AA has the highest rate and mutant SA has the lowest rate, the same relative functional rate as seen in experiment. In addition, WT and mutant U32C have similar levels (particularly between 0.4-0.6), again correlating with experimental results. These results also suggest that the SD sequence may only be active for gene regulation when more than 40% of its nucleotides are open.

TABLE 2

Comparison of expression rates between WT and three mutants of MS2. It shows that we can predict similar relative functional rates as seen in experiment.

Mutant	Experimental Expression Rate (order of magnitude)	Our Estimation				
		$t = 0.2$	$t = 0.3$	$t = 0.4$	$t = 0.5$	$t = 0.6$
SA	0.1	0.1	0.04	0.03	0.03	0.08
WT	1	1.0	1.0	1.0	1.0	1.0
U32C	1	2.1	1.8	1.4	0.8	1.2
CC3435AA	5	7.2	8.4	3.8	3.5	9.8

5. Conclusion. We have presented an overview of a computational technique based on algorithms for robot motion planning that can study both protein and RNA motion. Our technique builds an approximate map, or model, of the molecule's energy landscape. With this model, we have extracted folding pathways and study landscape properties including relative folding rates and population kinetics. We have validated our technique by comparing it to other computational methods and to experimental

data. We have shown that our method produces pathways with the same secondary structure formation order as seen in experiment for several small proteins, including the structurally similar 'benchmark' set of proteins G, L, and mutants of G. Our technique also has reported the same relative folding rates for protein G and its mutants. For RNA, we have compared the population kinetics from our technique against another popular computational method. We also have shown that we predict the same relative folding rates for ColE1 and its mutant and the same relative gene expression rates for MS2 phage RNA and its mutants as seen in experiment.

Acknowledgments. This research supported in part by NSF Grants EIA-01037-42, ACR-0081510, ACR-0113971, CCR-0113974, ACI-0326350, CRI-0551685, CCF-0833199, CCF-0830753, by Chevron, IBM, Intel, HP, and by King Abdullah University of Science and Technology (KAUST) Award KUS-C1-016-04. Tapia supported in part by a Sloan scholarship, PEO scholarship, NIH Molecular Biophysics Training Grant (T32GM065088) and a Department of Education (GAANN) Fellowship. Thomas supported in part by an NSF Graduate Research Fellowship, a PEO scholarship, a Dept. of Education Graduate Fellowship (GAANN), and an IBM TJ Watson PhD Fellowship.

REFERENCES

[1] E. Alm and D. Baker. *Prediction of protein-folding mechanisms from free-energy landscapes derived from native structures.* Proc. Natl. Acad. Sci. USA, 96:20(1999), pp. 11305–11310.

[2] N. M. Amato, K. A. Dill, and G. Song. *Using motion planning to map protein folding landscapes and analyze folding kinetics of known native structures.* In: Proc. Int. Conf. Comput. Molecular Biology (RECOMB), pages 2–11, 2002.

[3] N. M. Amato, K. A. Dill, and G. Song. *Using motion planning to map protein folding landscapes and analyze folding kinetics of known native structures.* J. Comput. Biol., 10:3-4(2003), pp. 239–255. Special issue of Int. Conf. Comput. Molecular Biology (RECOMB) 2002.

[4] N. M. Amato and G. Song. *Using motion planning to study protein folding pathways.* J. Comput. Biol., 9:2(2002), pp. 149–168. Special issue of Int. Conf. Comput. Molecular Biology (RECOMB) 2001.

[5] M. Apaydin, D. Brutlag, C. Guestrin, D. Hsu, and J.-C. Latombe. *Stochastic roadmap simulation: An efficient representation and algorithm for analyzing molecular motion.* In: Proc. Int. Conf. Comput. Molecular Biology (RECOMB), pages 12–21, 2002.

[6] M. Apaydin, A. Singh, D. Brutlag, and J.-C. Latombe. *Capturing molecular energy landscapes with probabilistic conformational roadmaps.* In: Proc. IEEE Int. Conf. Robot. Autom. (ICRA), pages 932–939, 2001.

[7] O. B. Bayazit, G. Song, and N. M. Amato. *Ligand binding with OBPRM and haptic user input: Enhancing automatic motion planning with virtual touch.* In: Proc. IEEE Int. Conf. Robot. Autom. (ICRA), pages 954–959, 2001. This work was also presented as a poster at RECOMB 2001.

[8] S. Cao and S.-J. Chen. *Predicting RNA folding thermodynamics with a reduced chain representation model.* RNA, 11(2005), pp.1884–1897.

[9] S.-J. Chen and K. A. Dill. *RNA folding energy landscapes.* Proc. Natl. Acad. Sci. USA, 97(2000), pp. 646–651.

[10] T.-H. Chiang, D. Hsu, M. S. Apaydin, D. L. Brutlag, and J.-C. Latombe. *Predicting experimental quantities in protein folding kinetics using stochastic roadmap simulation.* In: Proc. Int. Conf. Comput. Molecular Biology (RECOMB), pages 410–424, 2006.

[11] F. Chiti and C. Dobson. *Protein misfolding, functional amyloid, and human disease.* Annu. Rev. Biochem., 75(2006), pp. 333–366.

[12] J. Cortés and T. Siméon. *Sampling-based motion planning under kinematic loop-closure constraints.* In: Algorithmic Foundations of Robotics VI, pages 75–90. Springer, Berlin/Heidelberg, 2005. book contains the proceedings of the International Workshop on the Algorithmic Foundations of Robotics (WAFR), Utrecht/Zeist, The Netherlands, 2004.

[13] J. Cortés, T. Siméon, M. Remaud-Siméon, and V. Tran. *Geometric algorithms for the conformational analysis of long protein loops.* J. Computat. Chem., 25:7(2004), pp. 956–967.

[14] D. Covell. *Folding protein α-carbon chains into compact forms by Monte Carlo methods.* Proteins: Struct. Funct. Genet., 14:4(1992), pp. 409–420.

[15] V. Daggett and M. Levitt. *Realistic simulation of naive-protein dynamics in solution and beyond.* Annu. Rev. Biophys. Biomol. Struct., 22(1993), pp. 353–380.

[16] P. Das, S. Matysiak, and C. Clementi. *Balancing energy and entropy: A minimalist model for the characterization of protein folding landscapes.* Proc. Natl. Acad. Sci. USA, 102:29(2005), pp. 10141–10146.

[17] K. A. Dill and H. S. Chan. *From Leventhal to pathways to funnels.* Nat. Struct. Biol., 4(1997), pp. 10–19.

[18] Y. Duan and P. Kollman. *Pathways to a protein folding intermediate observed in a 1-microsecond simulation in aqueous solution.* Science, 282(1998), pp. 740–744.

[19] C. Flamm. *Kinetic Folding of RNA.* PhD thesis, University of Vienna, Austria, August 1998.

[20] H. Groeneveld, K. Thimon, and J. van Duin. *Translational control of matruation-protein synthesis is phage MS2: a role of the kinetics of RNA folding?* RNA, 1(1995), pp. 79–88.

[21] A. P. Gultyaev, F. H. van Batenburg, and C. W. Pleij. *The computer simulation of RNA folding pathways using a genetic algorithm.* J. Mol. Biol., 250(1995), pp. 37–51.

[22] J. Haile. *Molecular Dynamics Simulation: elementary methods.* Wiley, New York, 1992.

[23] P. G. Higgs. *RNA secondary structure: physical and computational aspects.* Quarterly Reviews of Biophysics, 33(2000), pp. 199–253.

[24] I. L. Hofacker. *RNA secondary structures: A tractable model of biopolymer folding.* J. Theor. Biol., 212(1998), pp. 35–46.

[25] D. Jacobs. *Generic rigidity in three-dimensional bond-bending networks.* J. Phys. A: Math. Gen., 31(1998), pp. 6653–6668.

[26] D. Jacobs and M. Thorpe. *Generic rigidity percolation: The pebble game.* Phys. Rev. Lett., 75:22(1995), pp. 4051–4054.

[27] D. Jacobs and M. Thorpe. *Generic rigidity percolation in two dimensions.* Phys. Rev. E, 53:4(1996), pp. 3682–3693.

[28] D. J. Jacobs and B. Hendrickson. *An algorithm for two dimensional rigidity percolation: The pebble game.* J. Comp. Phys, 137(1997), pp. 346–368.

[29] N. G. V. Kampen. *Stochastic Processes in Physics and Chemistry.* North-Holland, New York, 1992.

[30] L. E. Kavraki, P. Švestka, J. C. Latombe, and M. H. Overmars. *Probabilistic roadmaps for path planning in high-dimensional configuration spaces.* IEEE Trans. Robot. Automat., 12:4(1996), pp. 566–580.

[31] D. E. Kim, C. Fisher, and D. Baker. *A breakdown of symmetry in the folding transition*

state of protein L. J. Mol. Biol., 298(2000), pp. 971–984.

[32] A. Kitamura, Y. Muto, S. Watanabe, I. Kim, T. Ito, Y. Nishiya, K. Sakamoto, T. Ohtsuki, G. Kawai, K. Watanabe, K. Hosono, H. Takaku, E. Katoh, T. Yamzaki, T. Inoue, and S. Yokoyama. *Solution structure of an RNA fragment with the P7/P9.0 region and the 3'-terminal guanosine of the tetrahymena group I intron.* RNA, 8(2002), pp. 440–451.

[33] P. Klaff, D. Riesner, and G. Steger. *RNA structure and the regulation of gene expression.* Plant Mol. Biol., 32(1996), pp. 89–106.

[34] A. Kolinski and J. Skolnick. *Monte Carlo simulations of protein folding.* Proteins Struct. Funct. Genet., 18:3(1994), pp. 338–352.

[35] S. M. Larson, C. Snow, and V. S. Pande. *Folding@Home and Genome@Home: Using distributed computing to tackle previously intractable problems in computational biology. Modern Methods in Computational Biology.* R. Grant, ed, Horizon Press, 2003.

[36] T. Lazaridis and M. Karplus. *Effective energy function for proteins in solution.* Proteins, 35(1999), pp. 133–152. http://mingus.sci.ccny.cuny.edu/server/.

[37] A. Lee and I. Streinu. *Pebble game algorithms and sparse graphs.* European Conference on Combinatorics, Graph Theory and Applications, 2005.

[38] M. Levitt. *Protein folding by restrained energy minimization and molecular dynamics.* J. Mol. Biol., 170(1983), pp. 723–764.

[39] R. Li and C. Woodward. *The hydrogen exchange core and protein folding.* Protein Sci., 8:8(1999), pp. 1571–1591.

[40] C. Ma, T. Kolesnikow, J. Rayner, E. Simons, H. Yim, and R. Simons. *Control of translation by mRNA secondary structure: the importance of the kinetics of structure formation.* Mol. Microbiol., 14(994), pp. 1033–1047.

[41] S. Matysiak and C. Clementi. *Optimal combination of theory and experiment for the characterization of the protein folding landscape of s6: How far can a minimalist model go?* J. Mol. Biol., 343:1(2004), pp. 235–248.

[42] E. L. McCallister, E. Alm, and D. Baker. *Critical role of β-hairpin formation in protein G folding.* Nat. Struct. Biol., 7:8(2000), pp. 669–673.

[43] V. Muñoz and W. A. Eaton. *A simple model for calculating the kinetics of protein folding from three dimensional structures.* Proc. Natl. Acad. Sci. USA, 96:20(1999), pp. 11311–11316.

[44] V. Muñoz, E. R. Henry, J. Hoferichter, and W. A. Eaton. *A statistical mechanical model for β-hairpin kinetics.* Proc. Natl. Acad. Sci. USA, 95(1998), pp. 5872–5879.

[45] S. Nauli, B. Kuhlman, and D. Baker. *Computer-based redesign of a protein folding pathway.* Nature Struct. Biol., 8:7(2001), pp. 602–605.

[46] M. E. J. Newman and G. T. Barkenma. *Monte Carlo Methods in Statistical Physics.* Clarendon Press, Oxford, 1999.

[47] J. Shimada and E. I. Shakhnovich. *The ensemble folding kinetics of protein G from an all-atom Monte Carlo simulation.* Proc. Natl. Acad. Sci. USA, 99:17(2002), pp. 11175–11180.

[48] M. Shirts and V. Pande. *Screen savers of the world unite.* Science, 290(2000), pp. 1903–1904.

[49] A. P. Singh, J.-C. Latombe, and D. L. Brutlag. *A motion planning approach to flexible ligand binding.* In: Int. Conf. on Intelligent Systems for Molecular Biology (ISMB), pages 252–261, 1999.

[50] G. Song. *A Motion Planning Approach to Protein Folding.* Ph.D. dissertation, Dept. of Computer Science, Texas A&M University, December 2004.

[51] G. Song and N. M. Amato. *Using motion planning to study protein folding pathways.* In: Proc. Int. Conf. Comput. Molecular Biology (RECOMB), pages 287–296, 2001.

[52] G. Song and N. M. Amato. *A motion planning approach to folding: From paper craft to protein folding.* IEEE Trans. Robot. Automat., 20(2004), pp. 60–71. Preliminary version

appeared in ICRA 2001, pp. 948-953.

[53] G. Song, S. Thomas, K. Dill, J. Scholtz, and N. Amato. *A path planning-based study of protein folding with a case study of hairpin formation in protein G and L.* In: Proc. Pacific Symposium of Biocomputing (PSB), pages 240–251, 2003.

[54] M. J. Sternberg. *Protein Structure Prediction.* OIRL Press at Oxford University Press, 1996.

[55] X. Tang. *Tools for Modeling and Analyzing RNA and Protein Folding Energy Landscapes.* PhD thesis, Dept. of Computer Science, Texas A&M University, December 2007.

[56] X. Tang, B. Kirkpatrick, S. Thomas, G. Song, and N. M. Amato. *Using motion planning to study RNA folding kinetics.* In: Proc. Int. Conf. Comput. Molecular Biology (RECOMB), pages 252–261, 2004.

[57] X. Tang, B. Kirkpatrick, S. Thomas, G. Song, and N. M. Amato. *Using motion planning to study RNA folding kinetics.* J. Comput. Biol., 12:6(2005), pp. 862–881. Special issue of Int. Conf. Comput. Molecular Biology (RECOMB) 2004.

[58] X. Tang, S. Thomas, L. Tapia, and N. M. Amato. *Tools for simulating and analyzing RNA folding kinetics.* In: Proc. Int. Conf. Comput. Molecular Biology (RECOMB), pages 268–282, 2007.

[59] X. Tang, S. Thomas, L. Tapia, D. P. Giedroc, and N. M. Amato. *Simulating RNA folding kinetics on approximated energy landscapes.* J. Mol. Biol., 381(2008), pp. 1055–1067.

[60] L. Tapia, X. Tang, S. Thomas, and N. M. Amato. *Kinetics analysis methods for approximate folding landscapes.* Bioinformatics, 23:13(2007), pp. 539–548. Special issue of Int. Conf. on Intelligent Systems for Molecular Biology (ISMB) & European Conf. on Computational Biology (ECCB) 2007.

[61] S. Thomas, G. Song, and N. Amato. *Protein folding by motion planning.* Physical Biology, 2(2005), pp. S148–S155.

[62] S. Thomas, G. Tanase, L. K. Dale, J. M. Moreira, L. Rauchwerger, and N. M. Amato. *Parallel protein folding with STAPL.* Concurrency and Computation: Practice and Experience, 17:14(2005), pp. 1643–1656.

[63] S. Thomas, X. Tang, L. Tapia, and N. M. Amato. *Simulating protein motions with rigidity analysis.* In: Proc. Int. Conf. Comput. Molecular Biology (RECOMB), pages 394–409, 2006.

[64] S. Thomas, X. Tang, L. Tapia, and N. M. Amato. *Simulating protein motions with rigidity analysis.* J. Comput. Biol., 14:6(2007), pp. 839–855. Special issue of Int. Conf. Comput. Molecular Biology (RECOMB) 2006.

[65] T. E. Wales and J. R. Engen. *Hydrogen exchange mass spectrometry for the analysis of protein dynamics.* Mass Spec. Rev., 25:1(2006), pp. 158–170.

[66] T. Weikl, M. Plassini, and K. Dill. *Coopertivity in two-state protein folding kinetics.* Protein Sci., 13(2004), pp. 822–829.

[67] S. Wuchty. *Suboptimal secondary structures of RNA.* Master's thesis, University of Vienna, Austria, March 1998.

[68] W. Zhang and S. Chen. *RNA hairpin-folding kinetics.* Proc. Natl. Acad. Sci. USA, 99(2002), pp. 1931–1936.

[69] R. Zhou, M. Eleftheriou, C.-C. Hon, R. S. Germain, A. K. Royyuru, and B. J. Berne. *Massively parallel molecular dynamics simulations of lysozyme unfolding.* IBM J. Res. & Dev., 52:1/2(2008), pp. 19–30.

[70] M. Zuker, D. H. Mathews, and D. H. Turner. *Algorithms and thermodynamics for RNA secondary structure prediction: A practical guide.* In: J. Barciszewski and B. F. C. Clark, editors, RNA Biochemistry and Biotechnology, NATO ASI Series. Kluwer Academic Publishers, 1999.

[71] M. Zuker and D. Sankoff. *RNA secondary structure and their prediction.* Bulletin of Mathematical Biology, 46(1984), pp. 591–621.