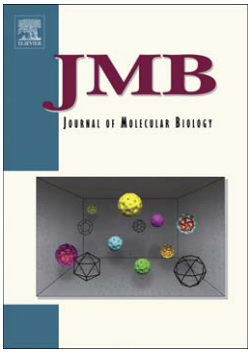


Simulating RNA Folding Kinetics on Approximated Energy Landscapes

Xinyu Tang, Shawna Thomas, Lydia Tapia, David P. Giedroc, Nancy M. Amato

PII: S0022-2836(08)00173-3
DOI: doi: 10.1016/j.jmb.2008.02.007
Reference: YJMBI 60205

To appear in: *Journal of Molecular Biology*

Received date: 8 October 2007
Revised date: 26 January 2008
Accepted date: 3 February 2008

Please cite this article as: Tang, X., Thomas, S., Tapia, L., Giedroc, D.P. & Amato, N.M., Simulating RNA Folding Kinetics on Approximated Energy Landscapes, *Journal of Molecular Biology* (2008), doi: 10.1016/j.jmb.2008.02.007

Simulating RNA Folding Kinetics on Approximated Energy Landscapes

Xinyu Tang[1], Shawna Thomas[1], Lydia Tapia[1], David P. Giedroc[2] and Nancy M. Amato[1]

[1]Department of Computer Science, Texas A&M University, College Station, TX 77843-3112 USA

[2]Department of Chemistry, Indiana University, Bloomington, IN 47405-7102 USA

To whom correspondence should be addressed: Nancy M. Amato, Department of Computer Science, Texas A&M University, College Station, TX 77843-3112, USA. Tel: 979-862-2275, Fax: 979-458-0718. Email: amato@cs.tamu.edu

Running Title: RNA folding landscapes

Abstract:

We present a general computational approach to simulate RNA folding kinetics that can be used to extract population kinetics, folding rates and the formation of particular substructures that might be intermediates in the folding process. Simulating RNA folding kinetics can provide unique insight into RNA whose functions are dictated by folding kinetics and not always by nucleotide sequence or the structure of the lowest free energy state. The method first builds an approximate map (or model) of the folding energy landscape from which the population kinetics are analyzed by solving the Master Equation on the map. We present results obtained using an analysis technique, Map-based Monte Carlo (MMC) simulation, which stochastically extracts folding pathways from the map. Our method compares favorably with other computational methods that begin with a comprehensive free energy landscape, illustrating that the smaller, approximate map captures the major features of the complete energy landscape. As a result our method scales to larger RNAs. For example, in this paper we validate kinetics of RNA of more than 200 nucleotides. Our method accurately computes the kinetics-based functional rates of wild-type and mutant ColE1 RNAII and MS2 phage RNAs showing excellent agreement with experiment.

Keywords: RNA folding, folding kinetics, motion planning, Master Equation, Monte Carlo simulation

Introduction

Ribonucleic acid (RNA) performs diverse and important functions in the cell. Messenger RNAs, transfer RNAs and ribosomal RNAs function as integral components of the protein synthesis machinery whose workings can be understood at nearly atomic detail,[1] while other RNAs are involved in the mRNA maturation as part of the spliceosome.[2] Recent work has highlighted the ability of noncoding (nc) and plant microRNAs to base pair with complementary regions of mRNAs which targets them for degradation.[3] Another class of folded RNA domains, termed riboswitches, are found in the 5' untranslated regions of mRNAs and function as molecular switches by binding a specific metabolite that changes the structure of the mRNA; this, in turn, controls transcription termination/anti-termination or translation initiation.[4] Detailed structural and physicochemical studies of metabolite sensing riboswitches has led to the hypothesis that the function of these domains may be kinetically controlled, with regulation of the Flavin mononucleotide (FMN) riboswitch, in particular, controlled by the relative rates of FMN binding and RNA transcription under the prevailing intracellular concentrations of metabolite.[5] Thus, this ligand-controlled conformational switch may operate under kinetic control, with the relative stabilities of the two, often mutually exclusive, conformational states (with and without ligand) perhaps less important.

Earlier work on the control of *E. coli* ColE1 plasmid copy number and translation initiation in bacterial and bacteriophage mRNAs hinted strongly that the kinetics of folding might be capable of dictating biological function. For example, accelerating the refolding rate of RNA II can increase the ColE1 plasmid copy number.[6; 7] In the control of translation initiation, mutations of IS10 transposase mRNA that slowed the folding kinetics of structure formation increased the rate of ribosome-binding resulting in a higher expression of IS10 transposase.[8] For the bacteriophage MS2 maturation protein, the mRNA is efficiently translated only when the

region around the Shine-Dalgarno sequence, which base pairs with the 16S ribosomal RNA, is found in an unpaired or "open" conformational state. Since it is closed (paired) in the native state, this can only happen before folding finishes. The longer the RNA remains in the open "metastable" state, the higher the gene expression rate. Recently coined RNA thermometers associated with the heat-shock response in α- and γ-proteobacteria are thought to function in the same way, by controlling translation initiation via a temperature-dependent conformational change around the Shine-Dalgarno sequence.[9; 10] These few examples highlight the importance of developing a robust computational method that can be used to study both the global properties of RNA folding and to provide more detailed insights of the kinetics of folding RNA substructures as well as the rates of interconversion between these substructures as a function of temperature.

A number of computational approaches have been developed to investigate the kinetics of RNA folding. For instance, folding pathways have been identified using both Monte Carlo-based algorithms[11; 12; 13] and genetic algorithms.[14; 15; 16] Other approaches utilize dynamic programming to calculate the partition function $Q = \sum_s exp(-\Delta G(s)/kT)$ over all secondary structures s to gain insight into RNA folding kinetics.[17] The ViennaRNA package implements McCaskill's algorithm in addition to an expanded set of nearest neighbor free energies to account for noncanonical pairings.[18] Ding and Lawrence extended this algorithm to generate statistical samplings of RNA structures based on the partition function.[19] Some methods incorporate computation of the RNA folding energy landscape. Dill and Chen used matrices to compute the partition function over all possible structures in order to approximate the complete folding landscape.[20] Wuchty modified Zuker's algorithm to generate all secondary structures within some given free energy of the native structure,[21] and Clote developed an algorithm to generate all secondary structures below a threshold with respect to the Nussinov-Jacobson energy model.[22] Wolfinger and Flamm extended this algorithm to identify local minima within some energy threshold of the native state

and connect them via energy barriers,[11; 23; 24] with the resulting energy barrier tree representing the energy landscape. In order to calculate the energy barrier, these authors used a flooding algorithm that is exponential in the size of the RNA; it is therefore impractical for large RNAs. Waldispühl and Clote created an algorithm to generate all saturated secondary structures having *k* base pairs.[25] Statistical mechanical methods have also been used to study RNA folding kinetics. For example, the Master Equation has been used to compute the population kinetics of the folding landscape, with a matrix of differential equations used to calculate the probability of a transition between different conformations.[26; 27; 28] Once solved, the dominant modes of the solution describe the general folding kinetics.[29] In addition, one can also approximate the folding kinetics by performing a statistical analysis on an ensemble of Monte Carlo folding pathways.[30; 31; 32]

In this work, we present a novel suite of computational tools that can be used to approximate the folding energy landscape and extract both global properties and detailed features of the folding process. The key advantage of our approach over the other computational techniques discussed above is that it is fast and efficient while providing both macroscopic and microscopic properties of the folding kinetics, *e.g.*, population kinetics and low-resolution folding details. Our method first builds a map, or model, of the RNA folding energy landscape. We then adapt standard energy landscape analysis tools to analyze the energy landscape represented by our map; the solution of the Master Equation on our map (Map-based Master Equation or MME) allows us to investigate properties of the ensemble during the time-course of folding. We also present another technique, termed a Map-based Monte Carlo (MMC) simulation, to extract microscopic folding pathways stochastically from our maps. These tools allow us to investigate the formation rates of transition states and therefore provide a comprehensive picture of the rates at which folding intermediates are formed and lost, as well as the final equilibrium

distribution of folded conformers. A new statistical sampling method ensures that our method scales well and is applicable to large RNAs containing hundreds of nucleotides. Finally, we validate our method against traditional Monte Carlo simulation methods that require a complete energy landscape, as well as against existing experimental data in two systems.

Results and Discussion

Summary of the Method

Our work provides computational tools to approximate the RNA folding energy landscape and extract global properties and detailed features of the folding process. The key advantage of our approach over other computational techniques is that it is fast and efficient while bridging the gap between high-level folding events and low-level folding details. Our method builds an approximate representation (called a map) of the RNA's folding energy landscape, and then uses specialized analysis techniques to extract folding kinetics from the map. We develop a new sampling strategy called *Probabilistic Boltzmann-Filtered Suboptimal Sampling* (PBS) that approximates the folding landscape with much smaller maps, enabling us to handle RNA with hundreds of nucleotides. We also present a new analysis technique, *Map-based Monte Carlo* (MMC) simulation, to stochastically extract folding pathways from the map. These tools allow us to study population kinetics, folding rates, and the folding of particular subsequences.

We validate our methods against other computational methods, *e.g.*, Monte Carlo simulation and experimental data. We demonstrate that our smaller, approximate maps efficiently capture the major features of much larger energy landscapes by comparing kinetics metrics extracted from our maps with those computed using a complete energy landscape. We also show that our method scales to large RNAs containing hundreds of nucleotides. Finally, we present two case studies to show that we compute the same relative functional rates of MS2

phage RNA and ColE1 RNA II as observed experimentally.

Table 1 compares the capabilities and limitations of the three different computational techniques used in this work: Monte Carlo simulation (MC), Map-based Monte Carlo simulation (MMC), and Map-based Master Equation (MME). MC requires significantly more time and space than MMC because it operates on the complete energy landscape at every timestep while our method, MMC, is able to operate on an approximation of the energy landscape. This landscape approximation also accounts for the memory storage requirement reduction from MC to MMC. Whereas the standard Master Equation method requires full enumeration of the conformation space, MME only considers the approximated energy landscape and hence can be applied to much larger RNA. Because MME does not produce (and store) individual folding trajectories, it has the smallest memory requirements. MME is instead limited numerically by eigenvector and eigenvalue solvers resulting in longer running times.

Computational validation

In order to determine the degree to which our sampling method captures major features of the complete landscape as well as to explicitly probe the scalability of the method, we performed folding simulations for three RNAs of increasing complexity using maps calculated with our method and compared them with results obtained using a complete energy landscape. We analyzed these maps using MME, MC and MMC (see Methods) as described below.

RNA1. RNA1 is an 18-nucleotide RNA (5'-CGCGCUACUCCUAGAGCU) that adopts a hairpin conformation in the "native" state. Fig. 1 shows the population kinetics of the four most significant conformations calculated using the BPE, SPE, and PBS maps (see Methods). In the Map-based Master Equation (MME) solution, these conformations have the largest fractional population during or after the folding process (individual conformations are designated by the

standard bracket nomenclature), so their existence is more likely to be observed in experiment. As can be seen (Fig. 1a-c), the population kinetics calculated from the BPE, SPE and PBS maps are quite similar during the folding process. Thus, for this small RNA, the reduced sampling SPE and PBS maps are good approximations of the complete energy landscape. They preserve the main characteristics of the energy landscape while using notably fewer conformations (22 vs. 19 vs. 876). In addition, the BPE, SPE and PBS maps yield similar population kinetics to those generated by Kinfold[11] (Fig. 1(d)), with minor discrepancies caused by different energy and transition rate constants.

Fig. 2(a)-(c) demonstrates the similarities of the eigenvalues and eigenvectors between the three maps. Most significant is the fact that the eigenvalues for the BPE, SPE and PBS are all approximately the same (Fig. 2(a)). In addition, the components of the eigenvectors (Fig. 2(b)-(c)) are comparable in magnitude as well, with the equilibrium distribution (Fig. 2(b)) defined by λ_0, very nearly identical for each of the three methods and matching the Boltzmann distribution for this molecule. Fig. 2(c) illustrates only small differences in magnitude of the components of the second eigenvector (λ_1) calculated from all three maps.

1k2g. 1k2g is a 22 nucleotide RNA (5'-CAGACUUCGGUCGCAGAGAUGG) that also has a hairpin native state structure. Fig. 3 compares the kinetics of the native state formation using a standard Monte Carlo (MC) simulation implemented by Kinfold[11] (Fig. 3(a)), a MMC simulation (Fig. 3(b)) on a BPE map (12,137 conformations), a MME result (Fig. 3(c)) and a MMC simulation (Fig. 3(c)) on a SPE map (70 conformations) respectively, a MME result (Fig. 3(e)) and a MMC simulation (Fig. 3(f)) on a PBS map (42 conformations) respectively. Although the fully enumerated map is the most accurate representation of the ensemble, all population kinetics curves have similar features.

In each figure, the population first increases quickly, and then it gradually decreases and

eventually stabilizes to the equilibrium distribution. Note that the equilibrium (final) distributions are very close to each other at 82%, even though the PBS map (Fig. 3(e-f)) contains less than 0.4% of all possible conformations. Also notice that the equilibrium distribution of the SPE map (Fig. 3(c-d)) is higher than the PBS map (Fig. 3(e-f)) even though it involves more conformations. This reflects that the SPE sampling method misses some conformations that represent significant fractional populations. Despite this, the SPE sampling method is still able to capture important features of the population kinetics. Thus, the SPE and PBS maps capture the main features of the energy landscape, with the PBS and BPE methods essentially interchangeable.

Fig. 4 compares the ten smallest eigenvalues of the BPE map (*blue*, 12,137 conformations) to the SPE map (*red*, 70 conformations) and to the PBS map (*green*, 42 conformations) computed by the MME. All the eigenvalues, *i.e.*, the folding rates of individual non-native structures, are similar to one another. This comparison indicates that our sparse SPE and PBS maps (70 and 42 conformations vs. 12,137 conformations) not only capture the major features of the equilibrium population distribution, but also effectively capture the most significant features of the folding kinetics for this short RNA.

Leptomonas Collosoma Spliced Leader RNA. Finally, we compare our simulations on a larger 56 nucleotide Leptomonas Collosoma Spliced Leader RNA which previous work suggests is capable of forming many metastable structures.[19] This RNA has approximately $2.0*10^{14}$ conformations, so it is not feasible to enumerate even all of the stack-pair conformations, let alone the entire conformational ensemble. Thus, we are only able to compare the folding kinetic simulations on PBS maps, analyzed using a Kinfold Monte Carlo (MC) simulation and our Map-based Monte Carlo (MMC) simulation. The MMC simulation approximates the entire conformational space of $2.0*10^{14}$ conformations with only a tiny subset ($5.0*10^{3}$) of conformers.

In each case, 1000 different folding pathways were simulated and combined to calculate the population kinetics of a particular conformation by summing its appearance in each pathway at every time increment in the usual manner.

Fig. 5 shows that although we have only 5,033 conformations in our PBS map, the MMC simulation results (Fig. 5(b)) are qualitatively similar to the Kinfold MC simulation (Fig. 5(a)). To quantitatively compare the two simulations, we fit parameters to a two-state kinetic model (unfolded⇔folded) to the Kinfold data (*black line*, Fig. 5(a)). The three parameters from the model represent the unfolded equilibrium, folded equilibrium, and rate of transition between the two states. While keeping the two equilibrium parameters constant and allowing the rate to change, we performed a fit of the model on the simulation data derived from MMC (*black line*, Fig. 5(b)). The agreement is excellent, with only the simulated rate changing from 89 for the Kinfold data to 130 for the MMC simulations. For comparison, a fit of the model on MMC data is shown with all parameters variable (*red line*), *e.g.*, without bias from the Kinfold parameters. This fit better captures the equilibrium distribution and reduced the simulated rate to 90. These results provide further evidence that our sparse map approach captures the main features of the energy landscape for even this complex RNA. An additional benefit of the MMC simulation over the conventional MC simulation is that it requires approximately one order of magnitude fewer iterations to stabilize, while using far less storage space (1G vs. 8G for Kinfold).

Experimental validation

We next show that experimentally determined folding rates can be recovered from our sparse sampling maps analyzed with the MMC method for closely similar, biologically different RNAs in two different systems. The success of these simulations speaks to the predictive power of the approach.

ColE1 RNAII. RNAII regulates the replication of *E. coli* ColE1 plasmids through its folding kinetics, with the plasmid replication rate inversely related to the folding rate.[6; 7] A specific mutant, denoted MM7, differs from the wild-type (WT) sequence by a single nucleotide in a 200 nucleotide sequence. This mutation causes RNAII to fold more slowly while maintaining the same thermodynamic stability of the native state and increases the overall plasmid replication rate in the MM7 mutant relative to the WT RNA. It is suggested that the MM7 mutant folds more slowly due to an increased half-life of a metastable structure.[6] As was done previously,[6] we investigated the folding rates of each RNA by comparing their eigenvalues, with the smallest non-zero eigenvalue corresponding to the global folding rate. In the previous study, the authors solved the Master Equation on a greatly simplified energy landscape using a specific subsequence containing only 130 of 200 nucleotides, and 9 stems hand-picked from 30 possible conformations. Here, we simulate the folding kinetics of the entire 200-nucleotide sequence using a PBS map containing approximately 4,000 conformations and analyzed using our MMC method (Fig. 6). All eigenvalues of the WT RNA are larger than for the MM7 RNA, indicating that WT sequence does indeed fold faster, a finding consistent with experiment. As in the previous study, we ignore binding effects of RNAII to RNAI and focus on kinetic differences, as it is believed that the kinetic differences are responsible for functional differences.

MS2 bacteriophage RNA. For the bacteriophage MS2 maturation protein, the mRNA (135 nucleotides) is efficiently translated only when the region around the Shine-Dalgarno sequence, which base pairs with the 16S ribosomal RNA, is found in an unpaired or "open" conformational state allowing ribosome binding and initiation of translation. Three mutants have been studied that have similar thermodynamic properties with the wild-type (WT) but different kinetics and therefore different gene expression rates. Experimental results indicate that mutant CC3435AA has the highest gene expression rate, WT and mutant U32C have similar rates, and mutant SA

has the lowest rate.[33]

Here, we first simulate the folding process by generating 1000 folding pathways for each mutant using MMC. Then, we analyze the pathways and calculate the opening probability of the Shine-Dalgarno sequence for each mutant, the latter calculated as the percentage of open nucleotides in the Shine-Dalgarno sequence. Fig. 7 presents the time evolution plot of the Shine-Dalgarno opening probability for the WT RNA and for each of the three mutants. Note that mutant CC3435AA has the longest duration at a relatively high level of opening probability while the mutant SA has the shortest duration, consistent with the functional data. On the other hand, while the opening probability of U32C decreases earlier than WT, it also opens longer than WT. Therefore, it is not clear which one has a larger total opening probability during folding. This explains why U32C has a functional expression rate that is similar to the WT RNA.

We next estimate the functional rates from the opening probability by defining a parameter called the opening threshold to determine whether the Shine-Dalgarno sequence is open enough to be functionally active (see Methods). Then we calculate the opening probability for each mutant until it is lower than the threshold. Table 2 compares our estimated functional rates of those mutant RNAs with experimental measurements. 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 thresholds 0.4-0.6), again correlating with experimental results. Our successful correlation using the opening probability of the Shine-Dalgarno sequence suggests that the opening of the Shine-Dalgarno sequence may regulate translation initiation. Our results on this RNA also suggest that the Shine-Dalgarno sequence may become available for ribosome binding only when more than 40% of its nucleotides are unpaired. Note, however, that we do not explicitly model ribosome binding to the Shine-Dalgarno sequence but instead focus on when

this sequence is available for base pairing to the 16S ribosomal RNA.

Note that in principle we can also use MME to study substructure formation. To do so, we would first need to combine the individual population kinetics solutions from MME for all conformations containing an unpaired Shine-Dalgarno sequence into a single solution. We would then repeat this for each of the mutants and compare the area under these new solution curves to determine the relative functional rates. However, this process becomes infeasible as the size of the conformation space grows exponentially.

Conclusions

We present a new model and new map-based analysis tools that can be used to study RNA folding kinetics and provide specific insights into both local and global features of the process. These new tools enable us to study larger RNAs than before, up to hundreds of nucleotides. The method is validated against known experimental data for two classic cases in detail. We anticipate that our method will be a valuable tool for discovering such relationships for other RNAs that have not yet been characterized experimentally. Finally, in the results shown we focused on RNA secondary structure and Watson-Crick base pairing. However, as new free energy models become available, our method can easily expand to include other RNA structural motifs such as non-canonical pairing, pseudoknots and other RNA tertiary structures, *e.g.*, A-minor motifs, ribose zippers, and triple base pairing that characterize complex folded RNA molecules.

Our method would also seem to be well suited to the study of ligand-binding riboswitches where ligand binding kinetics and thermodynamics could be incorporated into the map-building process with their influence of the folding energy landscape systematically investigated. Cao and Chen[34] identified the kinetic intermediates from analysis of the population kinetics and proposed

that the activity of human telomerase may be kinetically controlled by a pseudoknot-to-hairpin conformational switch. Our method provides comprehensive kinetics information to study these and other conformational switches[35] but involving much larger RNA fragments, *e.g.*, those found in the 3' untranslated regions of many plant and animal RNA viruses.[36] Studies along these lines are in progress in our laboratories.

Methods

General computational methods

Our method first constructs a map that approximates the energy landscape and then uses a number of map-based tools to analyze the approximated energy landscape. In our current implementation, we are able to generate three types of maps based on complete base-pair enumeration (BPE), stack-pair enumeration (SPE), and probabilistic Boltzmann-filtered suboptimal sampling (PBS). While a BPE map describes the complete energy landscape, it is not feasible for large RNA, *e.g.*, more than approximately 40 nucleotides. SPE maps are a subset of BPE maps and are one or two orders of magnitude smaller but are still not practical for large RNAs. PBS maps are the smallest, up to 10 orders of magnitude smaller than BPE maps, and scale well for larger RNAs consisting of hundreds of nucleotides. PBS maps are also a subset of BPE maps. Map-based analysis tools are then developed to provide insight into the folding kinetics. Map-based analysis tools used here include a Map-based Master Equation (MME) and a Map-based Monte Carlo (MMC) simulation to study folding kinetics. We show how MME can be used to extract global properties such as folding rates and transition states and demonstrate how MMC can be used to extract microscopic features of the folding process, *e.g.*, the rates and orders of formation of partially folded substructures.[31; 32] Thus, MME and MMC complement each other: MME provides a mathematical solution and MMC provides a statistical solution.

Since our methods generate and operate on approximate models of the energy landscape, their time and memory requirements are smaller than complete methods and easily run in a short amount of time on a desktop PC. Fig. 9 demonstrates the runtime (seconds) versus the number of nucleotides for a set of 9 RNA of varying lengths (9 to 200). For the 200nt RNA on a desktop PC (Intel P4 2.4GHz with 512MB RAM), just over 1 hour was required to construct an approximate landscape and less than 3 hours were required for MMC. In related work[32], we have shown that the computation time of MME scales linearly with map size. Another benefit of approximate models is the space requirements. For example, the 18nt hairpin RNA1 takes 485MB memory to store 1000 Kinfold MC pathways. On the other hand, 1000 MMC pathways are stored in a file of just 61MB and a map of 684KB. In comparison, the output from a MME calculation for this hairpin is 1.1MB.

Using maps to describe energy landscapes

The goal of map construction is to approximate the energy landscape and capture the landscape's most important features. The quality of this approximation highly depends on the quality of the sampling and connection methods. Three map node sampling and connection methods used here are discussed in turn.

Complete Base-Pair Enumeration (BPE). Let S be the set of all possible base-pair contacts. To generate a valid conformation, we first select one contact in S. Then we remove all contacts from S that would yield an invalid secondary structure[37] if combined with already selected contacts. The process of selecting a valid contact from S and then removing invalid contacts from S continues until S is empty. Each time a new contact is selected, a new secondary structure is defined. To enumerate the entire space, all possible combinations of a valid set of contacts from S are enumerated as above. As an example, Fig. 8 shows the complete

enumeration for the RNA sequence ACGUCACGU.

Stack-Pair Enumeration (SPE). This enumeration contains only those conformations containing stack-pair contacts. A *stack-pair contact* is a set of adjacent base-pair contacts, *i.e.*, no contacts are isolated from the others. More formally, if a stack-pair contact has a contact [i , j], where $i < j$, then it must also have at least one of the contacts $[i-1, j+1]$ or $[i+1, j-1]$. For example, the contacts in Fig. 8(c) form a stack, but the contacts in Fig. 8(f) do not because they are not adjacent. A conformation is a valid *stack-pair conformation* if it has only stack-pair contacts, *i.e.*, no isolated base pairs. The conformations shown in Fig. 8(a), 8(c), 8(d), 8(h), and 8(j) thus represent the enumeration of all stack-pair conformations for RNA sequence ACGUCACGU. We note that this simplification has been used previously[28] and justified on the basis of the low stability of isolated base pairs. The stack-pair enumeration is implemented similarly to the base-pair enumeration except that S contains stacks instead of base-contact pairs. Unfortunately, this method does not scale well as the number of nucleotides increases, given that RNA of approximately 40 nucleotides would have over 10^5 stack-pair configurations.

Probabilistic Boltzmann-Filtered Suboptimal Sampling (PBS). Wuchty [21] used dynamic programming to generate low energy conformations within a given energy threshold. One can then use these low energy conformations as "seeds" for map construction; making is possible to easily change the map size by simply adjusting the input energy threshold in Wuchty's algorithm. Unfortunately, as the size of the RNA or the energy threshold increases, the number of suboptimal conformations generated increases exponentially; furthermore, this method fails to generate high energy conformations by design.

In our approach, we augment this suboptimal sampling technique with additional random conformations. We use a probabilistic Boltzmann filter to retain a subset of the conformations from Wuchty's algorithm and from the additional random conformations based on their

Boltzmann distribution factors. For a given conformation i with free energy E_i, the probability P_i to keep it is:

$$P_i = \begin{cases} e^{\frac{-(E_i - E_0)}{kT}} & if\,(E_i - E_0) > 0 \\ 1 & if\,(E_i - E_0) \le 0 \end{cases} \quad (1)$$

where E_0 is a reference energy threshold used to control the number of samples, k is the Boltzmann constant, and T is the temperature of folding. Use of the Boltzmann distribution in this way allows us to generate more conformations probabilistically. We do not place requirements of the percentage of retained conformations from Wuchty's algorithm or of the percentage retained from the additional random conformations. Instead, we generate conformations, both from Wuchty's algorithm and randomly, until a minimum number of conformations are retained. In the results presented in this work, the minimum threshold is set between 5000 and 8000 conformations. As described previously, this sampling method appears to capture the important features of the energy landscape well while remaining computationally efficient for the RNA studied. In our experience with these RNA, smaller, sparser maps do not adequately model the energy landscape and larger, denser maps do not model the energy landscape significantly better to warrant the additional computational cost. We use Turner rules implemented by the ViennaRNA package[18] to compute conformation energies.

Map Node Connection

Individual members of a conformational ensemble generated by each of the three sampling methods above are then connected to form an approximate map of the energy landscape. It is impractical (and generally not necessary) to attempt all possible connections; instead, we attempt to connect a conformation with the k closest neighboring conformations

according to some distance metric, where k is some small constant[38]. Each pair of neighboring conformations is then connected using a local planner. For the results in this paper, we use k =50.

Distance Metrics. The distance metric defines which conformations are close to each other and which are far apart. Here we use base-pair distance, *i.e.*, the number of base-pair contacts that differ between two conformations. This denotes the number of base pairs that have to be opened or closed to transform one conformation into another. Our approach can also utilize other distance metrics such as string edit distance or tree edit distance[39], but we found that base-pair distances perform well on the RNA studied here.

Connecting Node Pairs. To connect a given pair of conformations, we not only wish to compute a representative transition pathway, *i.e.*, a set of intermediate conformations between them, but we also wish to assign an edge weight to approximate the Boltzmann transition probability. Note that these two goals are not always the same. If two conformations are far apart from each other, there may be many possible transition paths, while none dominates the transition probability. In our previous work,[30] we used a simple greedy algorithm to generate a single transition path and compute the transition probability/edge weight from that path. It works well when conformations are close to each other. However, as the size of RNA increases and thus the feasible sampling density decreases, this method fails. Here we present methods designed to compute transition probabilities and generate transition pathways that facilitate scaling to larger RNA.

Generating Transition Pathways. First, the stable subunits (stems) between the start and goal conformations are identified and the nucleation cost (the energy barrier to form each stem) for each of them is calculated. Then, a transition pathway connecting the start and the goal conformations is generated by probabilistically opening/closing the stems. Similar to a Monte

Carlo simulation, at every step the algorithm chooses a stem probabilistically on the basis of its nucleation cost. An analogous method will also be used as part of the analysis tools (see below).

Computing the Transition Probability. When an edge (q_i, q_j) is added to the map, it is assigned a weight W_{ij} that reflects the Boltzmann transition probability between its two end points q_i and q_j. First, we find the stable subunits (stems) that are different between q_i and q_j. We calculate the nucleation cost for each stem and calculate the maximum cost, defined as energy barrier E_b the folding process must go over to form all the stems. We use E_b to estimate the transition probability between q_i and q_j. This strategy is widely used in Monte Carlo simulations[12; 13] and in genetic algorithms for obtaining folding pathways.[14; 15] We calculate the Boltzmann transition probability K_{ij} (or transition rate) of moving from q_i to q_j using Metropolis rules:[40]

$$K_{ij} = \begin{cases} e^{\frac{-\Delta E}{kT}} & if \Delta E > 0 \\ 1 & if \Delta E \le 0 \end{cases} \qquad (2)$$

where $\Delta E = max(E_b, E_j) - E_i$, k is the Boltzmann constant, and T is the temperature of folding. Note that the same energy barrier E_b is also used to estimate the transition probability from K_{ji}, so the transition probabilities satisfy the detailed balance:

$$\frac{K_{ij}}{K_{ji}} = e^{\frac{-(E_j - E_i)}{kT}} \qquad (3)$$

The edge weight W_{ij} is therefore:

$$W_{ij} = -log(K_{ij}) = \frac{-\Delta E}{kT}. \qquad (4)$$

Note that the negative log is used since $0 \le K_{ij} \le 1$. By assigning the edge weights in this manner,

the most energetically feasible path in our map is readily extracted using simple graph search algorithms for the smallest-weighted path.

Map-based analysis tools

Map-based Monte Carlo Simulation. Transitioning from one conformation to another is probabilistically biased by the Boltzmann transition probabilities. Simulating this random walk on the real (or complete) energy landscape is called the Monte Carlo method[29], with Kinfold a well-known implementation within the ViennaRNA Package.[18] However, these simulations can be computationally intensive since at each step they must calculate the complete local energy landscape to choose the next step. We apply Monte Carlo simulations directly to our maps to mirror the stochastic folding process[29], with the map an approximation of the energy landscape with edge weights reflecting Boltzmann transition probabilities.

Similar to Monte Carlo simulation, our method starts from a random conformation in the map and iteratively chooses a next conformation probabilistically based on the transition probabilities. Because the edge weight W_{ij} encodes the transition probability between two endpoints i and j (see eq 4), we recalculate the transition probability K_{ij} as $K_0 e^{-W_{ij}}$ where K_0 is a constant adjusted according to experimental results. In the results presented here, we study relative rates, not absolute rates, so adjusting K_0 is unnecessary.

For both Monte Carlo simulation and MMC, we gradually increase the number of time steps simulated until the population reaches the equilibrium distribution. We first check that at the end of the simulation the population kinetics of states of interest have stabilized. We then verify that this population distribution matches the equilibrium distribution as determined by the their free energy. We continue increasing the simulation time until both criteria are met.

Population Kinetics and Map-based Master Equation. The Master Equation formalism

has been developed for folding kinetics in a number of earlier studies.[28] The stochastic process of folding is represented as a set of transitions among all n conformations (states). The time evolution of the population of each state, $P_i(t)$, can be described by the following differential equation:

$$dP_i(t)/dt = \sum_{i \neq j}^{n} (K_{ji} P_j(t) - K_{ij} P_i(t)) \tag{5}$$

where K_{ij} denotes the transition rate (probability) from state i to state j. Thus, the change in population $P_i(t)$ is the difference between transitions *to* state i and transitions *from* state i. We compute transition rates from the map's edge weights: $K_{ij} = K_0 e^{-W_{ij}}$ where K_0 is a constant adjusted according to experimental results.

If we use an n-dimensional column vector $\mathbf{p}(t) = (P_1(t), P_2(t), \ldots, P_n(t))'$ to denote the population of all n conformational states, then we can construct an $n \times n$ matrix M to represent the transitions, where

$$\begin{cases} M_{ij} = K_{ji} & i \neq j \\ M_{ii} = -\sum_{i \neq j} K_{ij} \end{cases} \tag{6}$$

The Master Equation can be represented in matrix form:

$$d\mathbf{p}(t)/dt = M\mathbf{p}(t). \tag{7}$$

The solution to the Master Equation is:

$$P_i(t) = \sum_k \sum_j N_{ik} e^{\lambda_k t} N_{kj}^{-1} P_j(0) \tag{8}$$

where N is the matrix of eigenvectors N_i for the matrix M in eq 6 and Λ is the diagonal matrix of its eigenvalues λ_i. $P_j(0)$ is the initial population of conformation j. From eq 8, we see that the eigenvalue spectrum is composed of n modes. If sorted by magnitude in ascending order,

the eigenvalues include $\lambda_0 = 0$ and several small magnitude eigenvalues. Since all the eigenvalues are negative, the population kinetics will stabilize over time. The population distribution $\mathbf{p}(t)$ will converge to the equilibrium Boltzmann distribution, and no mode other than the mode with the zero eigenvalue contributing to the equilibrium. Thus the eigenmode with eigenvalue $\lambda_0 = 0$ corresponds to the stable distribution, and its eigenvector corresponds to the Boltzmann distribution of all conformations at equilibrium. Large magnitude eigenvalues correspond to fast folding modes, *i.e.*, those that fold in a burst. Their contribution to the population will die away quickly. Conversely, small magnitude eigenvalues have a large influence on the global folding process; as a result, the global folding rates are determined by the slow modes.

Method to Calculate the Expression Rates for MS2 Mutants

To estimate the expression rates, we first simulate the folding process by generating 1000 folding pathways for each mutant using MMC. We found that 1000 pathways provide stable results with small variances while reasonably limiting memory requirements. Then we analyze the pathways and calculate the opening probability of the Shine-Dalgarno sequence for each mutant. We calculate the opening probability as the percentage of open nucleotides in the Shine-Dalgarno sequence. Higgs[12] performed a similar study using a stem-based Monte Carlo simulation. However, in that work, they simulated the folding process only when the RNA sequence is growing. Their results may be sensitive on the selection of growth rate. In contrast, our simulation results do not require this growth rate parameter and thus can be used to quantitatively predict the functional level of a new mutant in a more reliable way. In addition, our results have better correlations to the experimental values.

To estimate the functional rates quantitatively, we compute the integration of the opening

probability (Fig. 7) over the entire folding process until it is lower than a given threshold. 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.

Acknowledgements

We acknowledge grants from the NSF (EIA-0103742, ACR-0081510, ACR-0113971, CCR-0113974, ACI-0326350, CRI-0551685 to N. M. A.), the NIH (AI040187 to D. P. G.), the Department of Energy and the Hewlett-Packard Foundation (to N. M. A.). Ms. Thomas was supported in part by an NSF Graduate Research Fellowship, a PEO scholarship, a Dept. of Education Graduate Fellowship (GAANN), and an IBM TJ Watson Ph.D. Fellowship. Ms. Tapia supported in part by an NIH Molecular Biophysics Training Grant (T32 GM065088) and a Dept. of Education GAANN Fellowship.

References

1. Korostelev, A. & Noller, H. F. (2007). The ribosome in focus: new structures bring new insights. *Trends Biochem Sci*.
2. Valadkhan, S. (2007). The spliceosome: caught in a web of shifting interactions. *Curr Opin Struct Biol* **17**, 310-5.
3. Bartel, D. P. (2004). MicroRNAs: genomics, biogenesis, mechanism, and function. *Cell* **116**, 281-97.
4. Winkler, W. C. (2005). Riboswitches and the role of noncoding RNAs in bacterial metabolic control. *Curr Opin Chem Biol* **9**, 594-602.
5. Wickiser, J. K., Winkler, W. C., Breaker, R. R. & Crothers, D. M. (2005). The speed of RNA transcription and metabolite binding kinetics operate an FMN riboswitch. *Mol Cell* **18**, 49-60.
6. Gultyaev, A. P., van Batenburg, F. H. & Pleij, C. W. (1995). The influence of a metastable structure in plasmid primer RNA on antisense RNA binding kinetics. *Nucleic Acids Res* **23**, 3718-25.
7. Klaff, P., Riesner, D. & Steger, G. (1996). RNA structure and the regulation of gene expression.

Plant Mol Biol **32**, 89-106.
8. Ma, C. K., Kolesnikow, T., Rayner, J. C., Simons, E. L., Yim, H. & Simons, R. W. (1994). Control of translation by mRNA secondary structure: the importance of the kinetics of structure formation. *Mol Microbiol* **14**, 1033-47.
9. Narberhaus, F., Waldminghaus, T. & Chowdhury, S. (2006). RNA thermometers. *FEMS Microbiol Rev* **30**, 3-16.
10. Waldminghaus, T., Heidrich, N., Brantl, S. & Narberhaus, F. (2007). FourU: a novel type of RNA thermometer in Salmonella. *Mol Microbiol* **65**, 413-24.
11. Flamm, C., Fontana, W., Hofacker, I. L. & Schuster, P. (2000). RNA folding at elementary step resolution.
12. Higgs, P. G. (2000). RNA secondary structure: physical and computational aspects. *Q Rev Biophys* **33**, 199-253.
13. Xayaphoummine, A., Bucher, T., Thalmann, F. & Isambert, H. (2003). Prediction and statistics of pseudoknots in RNA structures using exactly clustered stochastic simulations. *Proc Natl Acad Sci U S A* **100**, 15310-5.
14. Gultyaev, A. P., van Batenburg, F. H. & Pleij, C. W. (1995). The computer simulation of RNA folding pathways using a genetic algorithm. *J Mol Biol* **250**, 37-51.
15. Shapiro, B. A., Bengali, D., Kasprzak, W. & Wu, J. C. (2001). RNA folding pathway functional intermediates: their prediction and analysis. *J Mol Biol* **312**, 27-44.
16. Shapiro, B. A., Wu, J. C., Bengali, D. & Potts, M. J. (2001). The massively parallel genetic algorithm for RNA folding: MIMD implementation and population variation. *Bioinformatics* **17**, 137-48.
17. McCaskill, J. S. (1990). The equilibrium partition function and base pair binding probabilities for RNA secondary structure. *Biopolymer* **29**, 1105-1119.
18. Hofacker, I. L. (2003). Vienna RNA secondary structure server. *Nucleic Acids Res* **31**, 3429-3431.
19. Ding, Y. & Lawrence, C. E. (2003). A statistical sampling algorithm for RNA secondary structure prediction. *Nucleic Acids Res* **31**, 7280-301.
20. Chen, S. J. & Dill, K. A. (2000). RNA folding energy landscapes. *Proc Natl Acad Sci U S A* **97**, 646-51.
21. Wuchty, S. (1998). Suboptimal Secondary Structures of RNA, University of Vienna.
22. Clote, P. (2005). An efficient algorithm to compute the landscape of locally optimal RNA seconary structures with respect to the Nussinov-Jacobson enery model. *J. Comput. Biol.* **12**, 83-101.
23. Wolfinger, M. (2001). The Energy Landscape of RNA Folding, University of Vienna.
24. Wolfinger, M., Svrcek-Seiler, W. A., Flamm, C., Hofacker, I. L. & Stadler, P. F. (2004). Efficient computation of RNA folding dynamics, Vol. 37, pp. 4731.
25. Waldispühl, J. & Clote, P. (2007). Computing the partition function and sampling for saturated secondary structures of RNA, with respect to the Turner energy model. *J. Comput. Biol.* **14**, 190-215.
26. Ozkan, S. B., Dill, K. A. & Bahar, I. (2002). Fast-folding protein kinetics, hidden intermediates, and the sequential stabilization model. *Protein Sci* **11**, 1958-70.
27. Ozkan, S. B., Dill, K. A. & Bahar, I. (2003). Computing the transition state populations in simple protein models. *Biopolymers* **68**, 35-46.
28. Zhang, W. & Chen, S. J. (2002). RNA hairpin-folding kinetics. *Proc Natl Acad Sci U S A* **99**, 1931-6.
29. Kampen, N. G. V. (1992). *Stochastic Processes in Physics and Chemistry*, North-Holland, New York.
30. Tang, X., Kirkpatrick, B., Thomas, S., Song, G. & Amato, N. M. (2005). Using motion planning to study RNA folding kinetics. *J Comput Biol* **12**, 862-81.

31. Tang, X., Thomas, S., Tapia, L. & Amato, N. M. (2007). Tools for Simulating and Analyzing RNA Folding Kinetics. *Proc1 Int1 Conf1 Comput1 Molecular Biology (RECOMB)*.
32. Tapia, L., Tang, X., Thomas, S. & Amato, N. M. (2007). Kinetics Analysis Methods For Approximate Folding Landscapes. *Bioinformatics* **23**, i539-i548.
33. Groeneveld, H., Thimon, K. & van Duin, J. (1995). Translational control of maturation-protein synthesis in phage MS2: a role for the kinetics of RNA folding? *Rna* **1**, 79-88.
34. Cao, S. & Chen, S.-J. (2007). Biphasic Folding Kinetics of RNA Pseudoknots and Telomerase RNA Activity. *Journal of Molecular Biology* **367**, 909-924.
35. Nagel, J. H. & Pleij, C. W. (2002). Self-induced structural switches in RNA. *Biochimie* **84**, 913-23.
36. Brierley, I., Pennell, S. & Gilbert, R. J. (2007). Viral RNA pseudoknots: versatile motifs in gene expression and replication. *Nat Rev Microbiol* **5**, 598-610.
37. Zuker, M. & Sankoff, D. (1984). RNA Secondary Structure and their Prediction. *Bulletin of Mathematical Biology* **46**, 591-621.
38. Kavraki, L. E., P.~Svestka, Latombe, J. C. & Overmars, M. H. (1996). Probabilistic Roadmaps for Path Planning in High-Dimensional Configuration Spaces. In *{IEEE} Trans. Robot. Automat. %L kslo-prpp-96*, Vol. 12, pp. 566-580.
39. Sankoff, D. & Kruskal, J. B. (1983). *Time Warps, String Edits and Macromolecules: the Theory and Practice of Sequence Comparison*, Addison Wesley, London.
40. Dill, K. A. & Chan, H. S. (1997). From Levinthal to pathways to funnels. *Nat Struct Biol* **4**, 10-9.

Figure Legends

Figure 1: The population kinetics of the 18 nucleotide hairpin sequence CGCGCUACUCCUAGAGCU with the native structure .((.(((....))).)). Figure (d) gives the Kinfold folding kinetics of the four most significant conformations. Figures (a) (b) and (c) give a comparison the folding kinetics of the base-pair enumeration (BPE) map (876 conformations) and the stack-pair enumeration (SPE) map (22 conformations) and the probabilistic Boltzmann-filtered suboptimal sampling (PBS) map (19 conformations).

Figure 2: The folding kinetics of the 18 nucleotide RNA (5'-CGCGCUACUCCUAGAGCU) with the native structure .((.(((....))).)). Figure (a) illustrates the differences in the eigenvalues and overall folding rates for base-pair enumeration (BPE, 876 conformations), stack-pair enumeration (SPE, 22 conformations), and probabilistic Boltzmann-filtered suboptimal sampling (PBS, 19 conformations). Figures (b) and (c) compare the 15 biggest components of eigenvector N0 and N1 respectively.

Figure 3: The population kinetics of the native state of 1k2g: (a) Kinfold Monte Carlo simulation, (b) MMC simulation on a fully enumerated map (12,137 conformations), (c) MME solution on a SPE map (70 conformations), and (d) MMC solution on the SPE map (70 conformations) (e) MME solution on a PBS map (42 conformations) and (f) MMC solution on the PBS map (42 conformations). All analysis techniques produce similar population kinetics curves and the same final equilibrium distribution for the native state.

Figure 4: Comparison of the eigenvalues of 1k2g by the Map-based Master Equation (MME) on a fully enumerated map (12,137 conformations), a SPE map (70 conformations) and a PBS map

(42 conformations). Both eigenvalues are similar between the different maps.

Figure 5: Comparison of population kinetics of a metastable state for Leptomonas Collosoma Spliced Leader RNA using (a) Kinfold Monte Carlo simulation and (b) our MMC simulation on a PBS map with 5453 conformations. Shown on both plots are kinetic fits using parameters optimized on the Kinfold plot (black lines). On the MMC plot the red line shows an optimized kinetic fit without Kinfold bias. We capture the same kinetics while only sampling a tiny fraction of the entire conformation space.

Figure 6: 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 7: Comparison of the Shine-Dalgarno opening probability during the folding process for WT and its three mutants.

Figure 8: Complete enumeration of all conformations for RNA sequence ACGUCACGU. Conformations (a), (c), (d), (h) and (j) are stack-pair conformations.

Figure 9: Running times for map generation and MMC demonstrated on a set of nine RNAs with varying numbers of nucleotides (15, 18, 21, 22, 56, 115, 116, 135 and 200 nucleotides).

Tables

Table 1: Comparison of capabilities and limitations for Monte Carlo simulation (MC), Map-based Monte Carlo simulation (MMC), and the Map-based Master Equation (MME).

Analysis Method	Running Time	Space Required	Population Kinetics	Individual Pathways	Folding Rate	Substructure Formation
MC	10x	400x	Approximate	Yes	Approximate	Yes
MMC	1x	40x	Approximate	Yes	Approximate	Yes
MME	50x	1x	Yes	No	Yes	No

Running time and space requirements are based on average performance on the RNA studied in this paper.

Table 2: Comparison of integration of Shine-Dalgarno opening probability between WT and three mutants of MS2.

Mutant	Expression rate w.r.t. WT	Our Estimation w/ Different Thresholds		
		0.4	0.5	0.6
SA	0.1	0.03	0.03	0.08
WT	1.0	1.0	1.0	1.0
U32C	1	1.4	0.8	1.2
CC3435AA	5	3.8	3.5	9.8

The second column shows the relative expression rates from experimental measurements. Columns 3-7 present the estimations from our simulations using different thresholds.

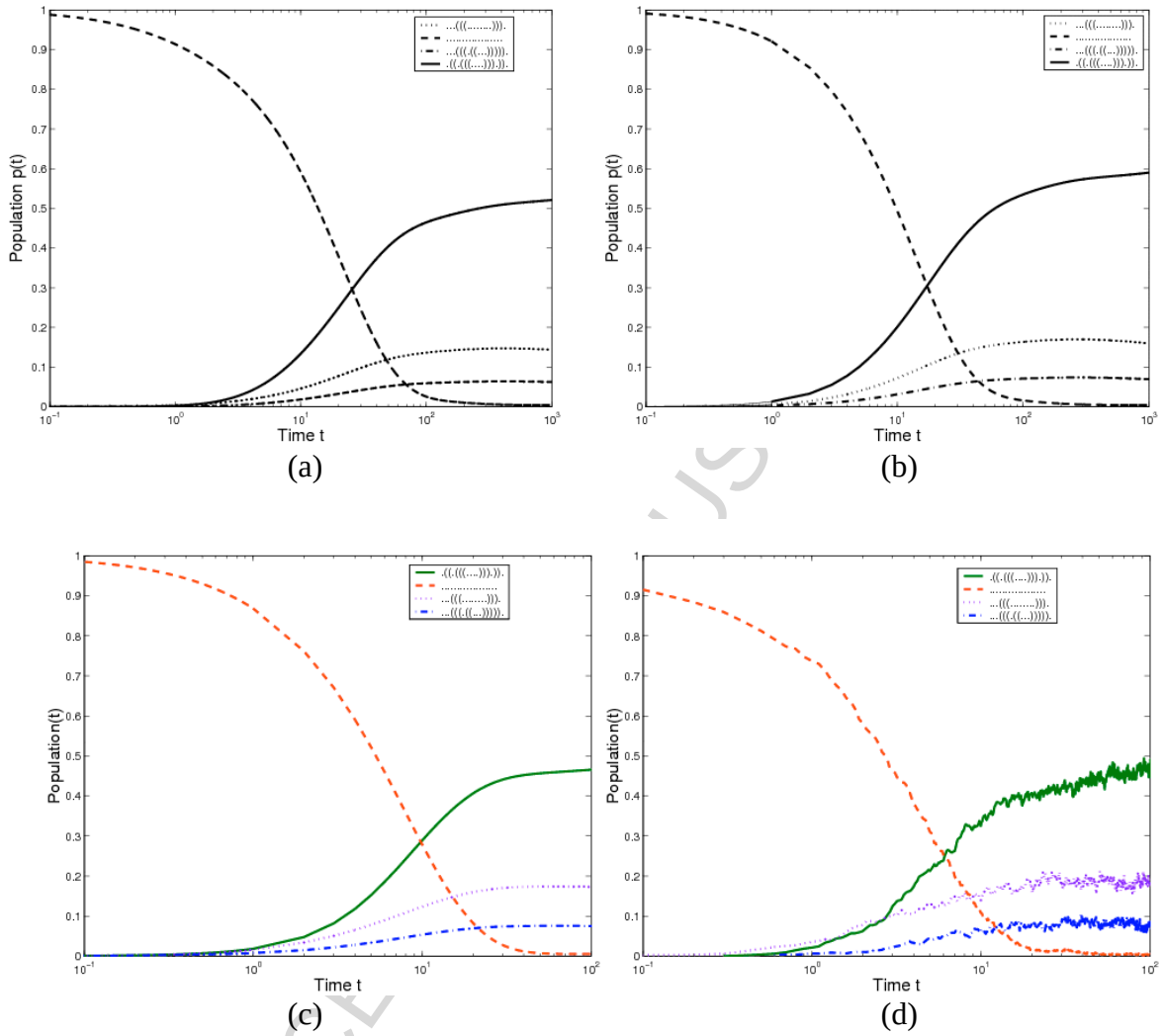
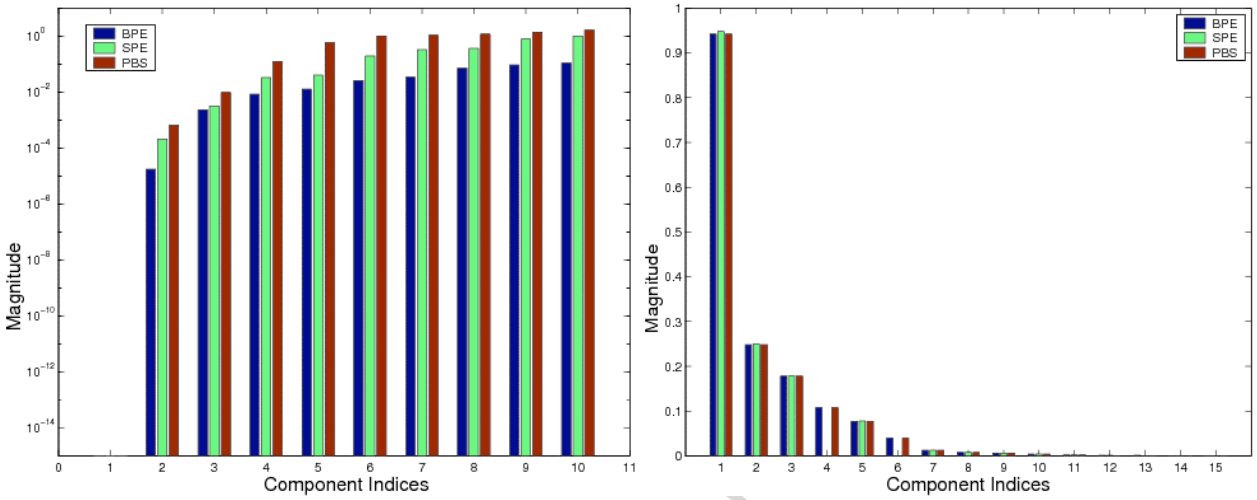
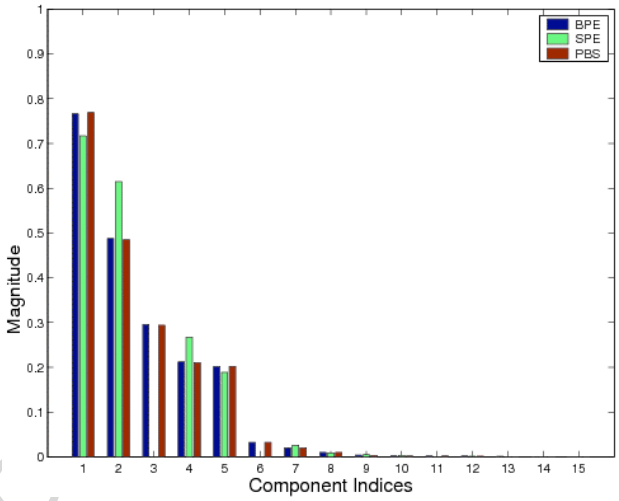


Figure 1



(a) (b)



(c)

Figure 2

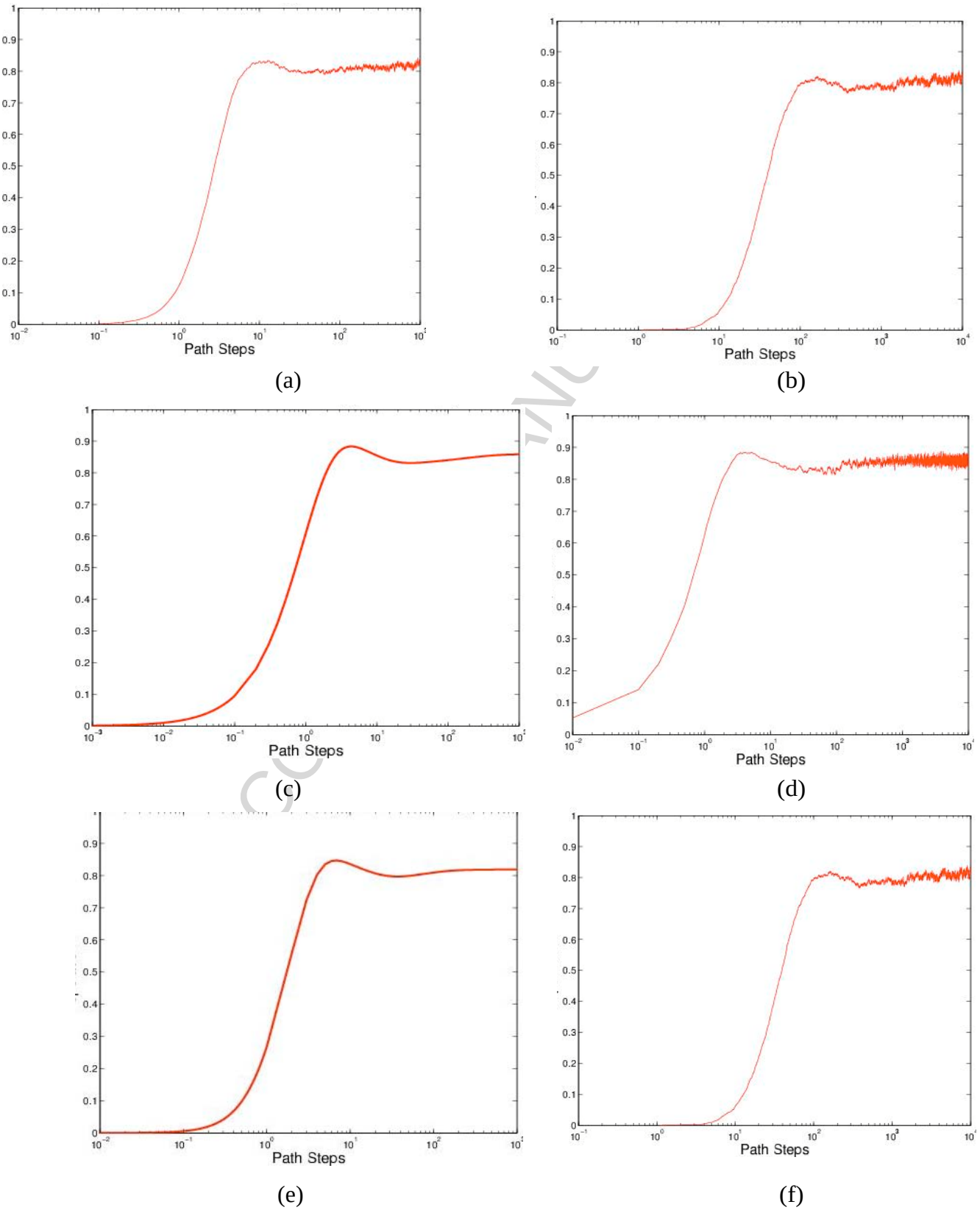


Figure 3

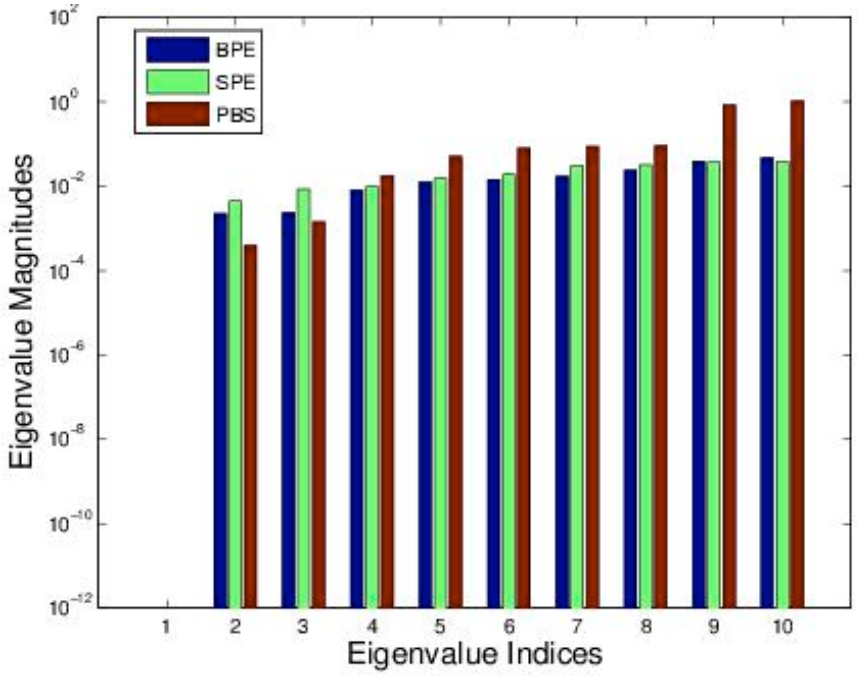
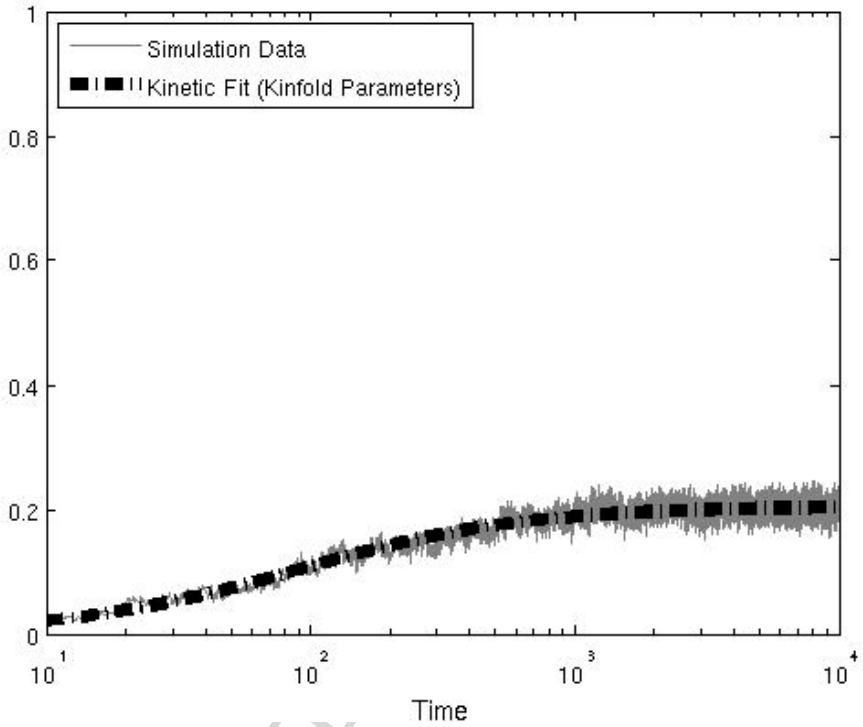
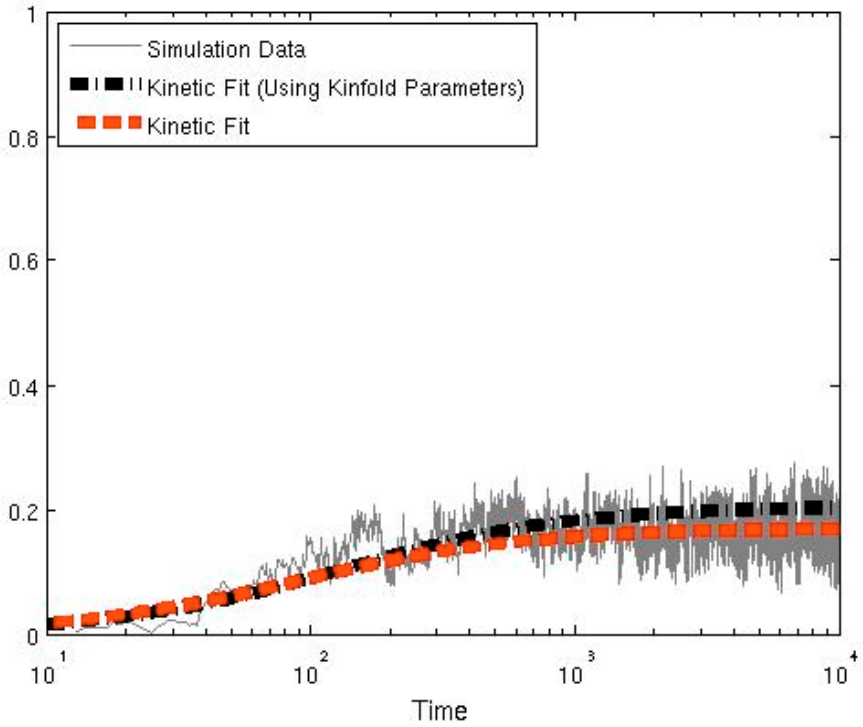


Figure 4



(a)



(b)

Figure 5

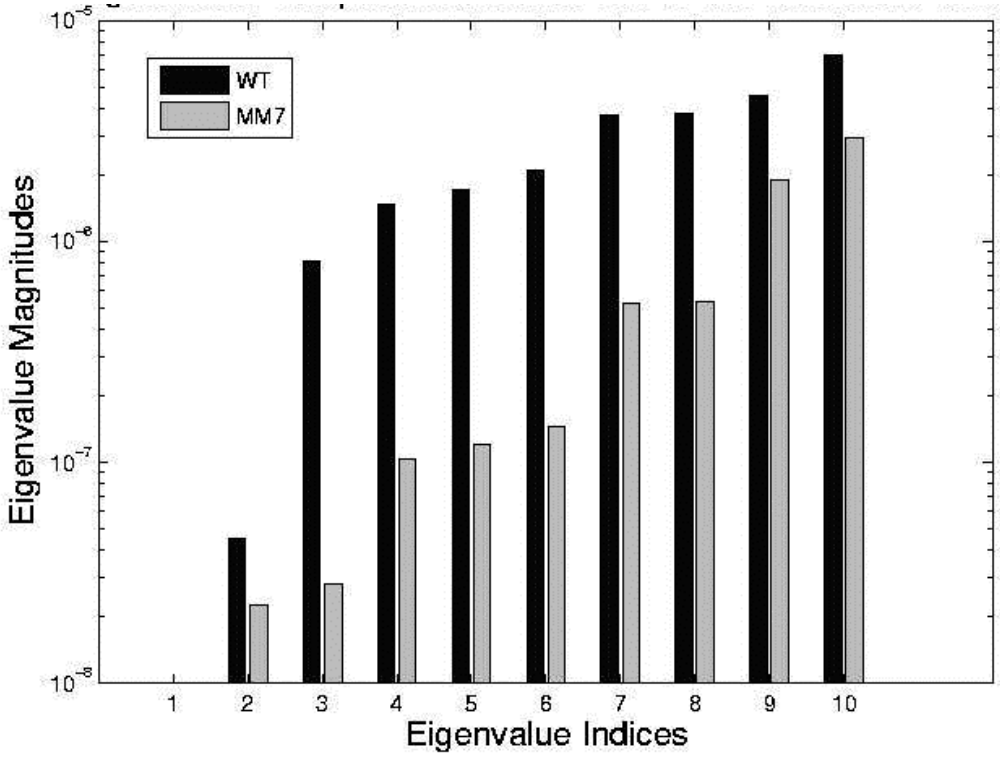


Figure 6

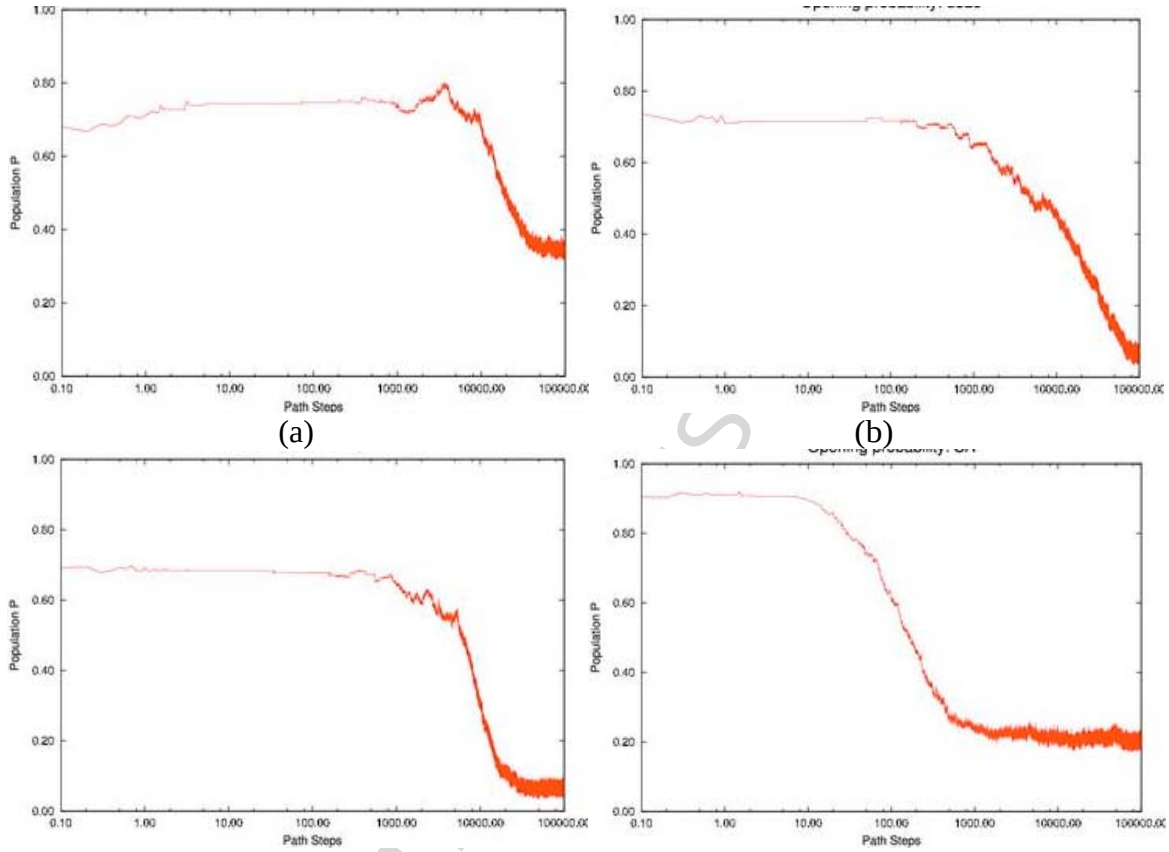
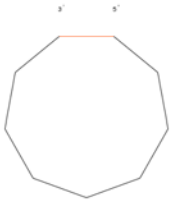
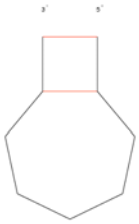


Figure 7

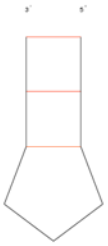
……
(a)



(…….)
(b)



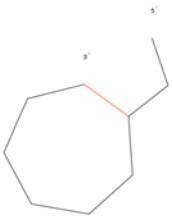
((…..))
(c)



(((…)))
(d)

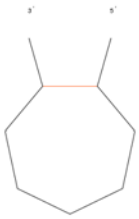
((…)..)
(e)

(.(…).)
(f)

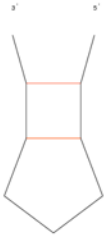


(…..)..
(g)

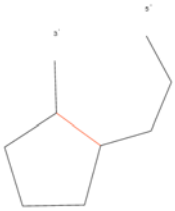
((…))..
(h)



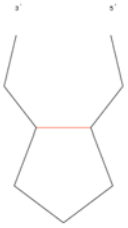
.(…..).
(i)



.((…)).
(j)



.(…)…
(k)



..(…)..
(l)

Figure 8

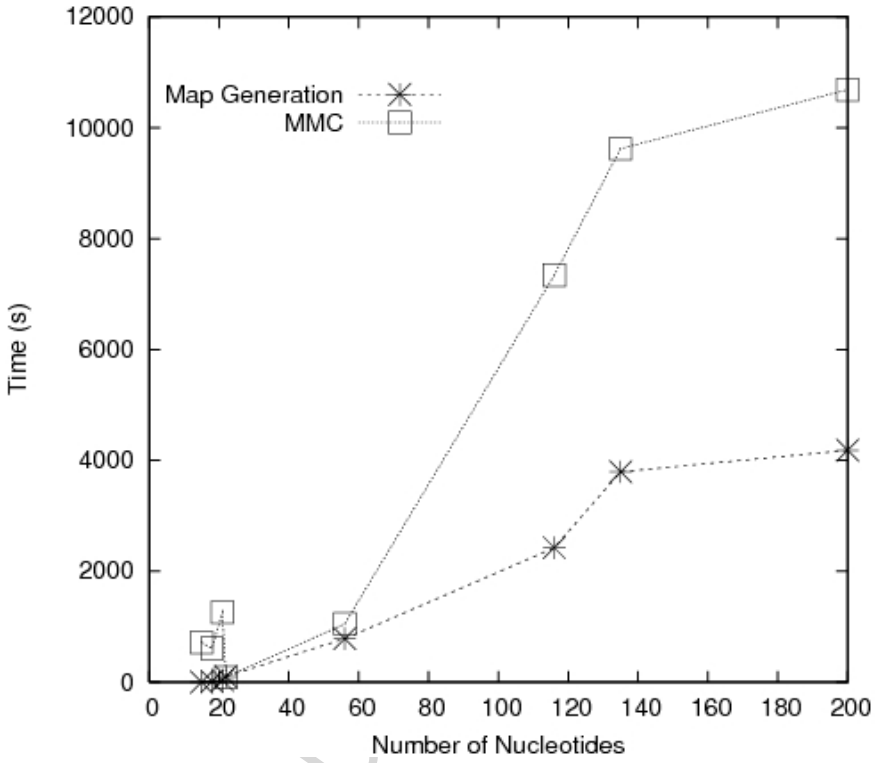


Figure 9