RNA *in silico* Stochastic Kinetic Folding

For Dr. Dipankar Sen, MBB821

Warren Cheung

October 12, 2006

Abstract

This research proposal investigates the folding of RNA molecules *in silico* — entirely using computers. RNA is an fundamental component of the processes of cells in the human body, and investigation of the folding process will improve computer-aided rational design of RNA molecules and open up potential avenues to new therapies.

The technique proposed, `RiSKFold`, RNA *in silico* Stochastic Kinetic Folding, will initially investigate simple structures, and improve these by folding them into more complex structures. Initially, we shall look at an abstract level (nucleic acids), and then refine these structures by examining the structures at increasing levels of detail, such as that of the nucleobases and the sugars, subcomponents of nucleic acids, or even to the level of individual atoms.

To do this efficiently, the structures to be folded or refined will be chosen randomly, with preference for more stable structures. `RiSKFold` will also be designed to be easily subdivided into tasks that can be performed on multiple computers, to make this difficult problem solvable.

`RiSKFold` will look at folding pathways to the most stable structure and to other low energy structures. By giving the researcher a high degree of control, `RiSKFold` will allow for a novel way of investigating the folding of RNA molecules, and potentially allow the discovery new structures and interactions for RNA molecules.

Contents

1 Notes

I have kept as closely as possible with the format of attachments for the CIHR Research Module (`http://www.cihr-irsc.gc.ca/e/documents/resmod_e.pdf` downloaded from the webpage `http://www.cihr-irsc.gc.ca/e/797.html`). I have also included the non-technical abstract, and omitted the Summary of Progress.

As a caveat, `RiSKFold` is a hypothetical research project invented to investigate the possibilities in RNA folding simulation and related literature, and to direct this proposal towards looking at available tools for RNA structure prediction (`mfold`[Zuk89] and `Srna`[CLD05]) and RNA folding (`Kinfold`[FFHS00] and `KineFold`[XBI05]).

2 Summary of research proposal

This proposal presents `RiSKFold`, RNA *in silico* Stochastic Kinetic Folding, a computational framework for investigating RNA structure and folding. This tool will allow researchers to investigate potential RNA folding pathways efficiently, and potentially reveal novel low-energy structures.

Building on the success of previous RNA folding algorithms like `Kinfold`[FFHS00] and `KineFold`[XBI05], `RiSKFold` will be designed to compute structures by simulating the folding of RNA, modifying structures with the simpler local secondary structure interactions into more complex tertiary structures. As well, it will be able to analyse these structures at multiple levels of detail, refining the structures from the more abstract level, where nucleic acids are treated as individual subunits, to a more sophisticated model, where the nucleobases and the ribose sugars are treated independently, or even to the level of individual atoms or further to the level of quantum interactions.

Generating all possible structures is both expensive, as the number of potential structures is basically unlimited, and inefficient, as many structures are physically improbable or unstable. Therefore, `RiSKFold` will decide which structures to modify and which structures to refine by randomly choosing from the currently generated structures, thereby avoiding having to investigate all possible structures. To ensure that the results are still meaningful, it will preferentially choose more likely structures, such as those with lower predicted free energy, or with a more likely folding pathway. As well, when refining structures, the entire structure need not be refined — substructures, such as the secondary structure elements, can be refined. Again, preference will given to substructures that are more likely, such as those with lower predicted free energy, or those that are common to many of the predicted low free energy structures. Experimentally determined substructures can be obtained from existing databases, allowing `RiSKFold` to quickly obtain very detailed, low energy substructures. As well, whenever `RiSKFold` computes a substructure, it will save the result, thereby avoiding recomputing the same substructure if it occurs again.

`RiSKFold` will emphasise making the results useful and accessible. At any stage of the computation, a researcher will be able to view the structures and pathways currently computed. Moreover, the user will be able to change aspects of `RiSKFold` on-the-fly, such as increasing the randomness of the structure choice, to increase the chances of finding the higher energy transition state structures, or by specifying directly which structures to modify or refine, and to what level of detail.

3 Research Proposal

3.1 Introduction

As the power and versatility of computers increase, we can solve more complicated problems. In addition to *in vivo* and *in vitro* experiments, we now have the capability to perform experiments and analyses using computers, or *in silico*.

It has been shown that many sequences of RNA, in a given environment, will often fold to the same stable structure. The fact that the natural folding process repeatably yields the same structure is a strong indicator that the primary sequence of a given RNA encodes much of its final structure. The "holy grail" of RNA folding is to be able to determine the resulting folded RNA structure directly from the primary sequence. However, although much success has been had using energy-minimisation techniques such as `mfold`[Zuk89] for simpler structures, these are often hampered by an inability to compute structures where complex tertiary interactions occur, such as pseudoknots. As well, these techniques do not provide any intuition as to how the RNA forms the structure.

By extending on previous work in simulating RNA folding[FFHS00, XBI05], `RiSKFold`, RNA *in silico* Stochastic Kinetic Folding, presents a unique opportunity to not only improve results, by allowing increased levels of detail and types of interactions allowed, but will also be designed for a high degree of control by the researcher. `RiSKFold` can be set up to run quickly, and can early on provide "quick and dirty" solutions, but since it uses stochastic methods and variable levels of detail, `RiSKFold` can be allowed to continue to run and will give improved results over time. It will allow the researcher to jump in at any time to examine the current results, and allow for the user to change the way the program runs, or manually direct the program to investigate a particular result.

3.2 Motivation

There are several reasons motivating the development of better RNA folding tools. The number of genes and putative genes far outstrips our capacity to experimentally determine their structure. Therefore, computational tools can provide a structure from the sequence without the need for physical experimentation, at the cost of accuracy. Unlike DNA, whose purpose appears to be primarily the storage of genetic information, RNA has been shown to be be capable of binding ligands and catalysing reactions. Prediction of the folded form can therefore provide insight into binding and enzymatic activity by identifying potential active sites. This is important in the development of new therapies using RNA as a functional element, or to gain insight in the role RNA, such as viral RNA, and can also potentially determine new targets for drug therapies.

The development of RNA folding techniques also provides us with insight into the importance of various factors in the folding process, which can be applied to other problems such as inverse RNA folding, the prediction of sequences which will adopt a particular structure, paving the way towards computer-aided design of RNA molecules. It would also allow the prediction of potential interactions of newly sequenced or designed RNA molecules, providing a cost-effective alternative to *in vivo* or *in vitro* experimentation.

3.3 Energy-based Techniques

To achieve this, many existing techniques utilise the fact that structures with lower free energy are more stable. Using this physically well-founded insight, most techniques search for the structure with the lowest free energy. However, a naive search through all possible conformations is not feasible, as the number of possible conformations is extremely large. Therefore, techniques will limit the scope of the search space by only considering a subset of the possible conformations, and will also intelligently direct their search. `RiSKFold` will also use this intuition, however, it will not be limited to the lowest free energy structures, but only be guided by them. This follows our knowledge that folding will proceed between low energy structures, but can also visit higher-energy transition state intermediates.

One method of simplifying the problem sufficiently to be solved is to only solve part of the RNA folding problem, such as predicting secondary structure, which encompass most of the structures formed at physiological temperature with only the presence of Na^+ (no divalent ions present). The level of detail can range from base-pair to atom-level interactions. As well, some techniques may not solve the problem exactly, presenting only an approximate solution, or may use randomness, and therefore the quality of their output can vary each time it is used, even if given the same sequence.

As it has been experimentally shown that RNA folding tends to be hierarchical, with secondary structure forming before tertiary structure, many RNA structure prediction algorithms are focused on determining secondary structure. As well, some programs only look for the final structure, whereas others may give multiple structures, or even a set of intermediates.

To compute the free energy of RNA molecule structure requires the use of an energy model. As the complexity of the model increases, the accuracy of the prediction also increases. For the simplest models, we have algorithms that allow us to solve for the lowest possible free energy. However, more complex models are often practically impossible to solve.

3.3.1 `mfold`

To approximate the free energy of a molecule, we can note that the free energy from base interactions and base stacking are some of the biggest contributors to the free energy of an RNA molecule. Zuker's widely-used `mfold` algorithm[Zuk89] uses a set of experimentally derived "rules of thumb" as the energy model, and computes the minimum energy structures. These rules include energies for the presence of Watson-Crick canonical base pairs and adjacent base pairs for stacking energies, as well as rules for computing the energy of loops, all determined from the primary sequence of these secondary structure elements using experimentally determined values.

Certain restrictions have been imposed, however. The structures are limited to non-pseudoknotted RNA structures, and the rules are of a particular form that makes them easily expressed mathematically. By exploiting these two aspects, dynamic programming, a classic computer science technique, allows us to make a program that is guaranteed to list the lowest energy structure and all the structures within a certain energy bound.

The official `mfold` server is available online (`http://www.bioinfo.rpi.edu/applications/mfold/old/rna/form1.cgi`), however there are also several alternatives if that service is un-

available.

`RiSKFold` can exploit `mfold` and its simple yet effective energy model to find simple secondary structures quickly and efficiently. This will provide a starting point that allows us to evaluate the free energy of molecules that do not contain pseudoknots and only involve Watson-Crick base pairing, and also generate potential secondary structures. Wolfinger et al[WSSF$^+$04] describe a method which can be used to take these low energy structures and connect them into potential pathways, and demonstrate that this coarse-grained method can produce results similar to the `Kinfold` kinetic RNA folding simulator[FFHS00], which we shall examine later. As well, we can improve the free energy estimates by considering non-Watson-Crick base pairing, such as Hoogsteen and reverse Hoogsteen interactions, and other atomic level hydrogen bonding and steric interactions.

3.4 Evaluating Problem Difficulty

The time it takes a computer program to run on an example sequence is one method of evaluating the performance of a computer program. There are several measures of this, but the most common is that of CPU time, which is the amount of time spent running the program, divided by the number of computers used. For example, 30 computers that spend 2 days working on a problem have used 60 CPU days. However, these kinds of measures only of limited use — as computer technology advances, the amount of time required for a computation can be reduced by orders of magnitude. Some problems which are too expensive to be solved at present will soon be feasible, whereas others are so complex that it will be almost impossible for them to ever be solved, therefore another method must be used to separate the impossibly hard problems from those that are merely difficult, but eventually solvable.

Traditionally, the difficulty of a problem in computer science is gauged by the complexity of a problem, which is a measure of how the cost (in terms of computer time) of a problem scales with the size of the input (e.g. in our case, the length of the RNA sequence), and the complexity of a problem can often be proven to fall within several general classes. Polynomial class complexity problems, where the running time increases like a polynomial function depending on the input size (e.g. n^2 or n^3 where n is the input size), are generally considered to be the types of problems that can be solved efficiently, whereas exponential class complexity (e.g. 2^n) problems, where the running time grows exponentially depending on the input size, rapidly grow impossible to solve. Note that exponential class problems are not always impossible to solve — they can still be solved as long as we are careful to only ask for the results of small instances (e.g. in our case, for very short RNA sequences).

A special complexity class that is often mentioned is the class of NP-hard problems. These are generally considered the hardest class of problem that we still have a chance at solving. It is actually possible that NP-hard problems are polynomial problems, however current methods require exponential-class solutions. Unfortunately, many simplifications of the RNA folding/interaction problems fall into the NP-hard complexity class, and therefore no method currently exists to solve them exactly, for large RNA sequences. For example, using the Zuker energy model but allowing pseudoknotted RNA structures makes finding the lowest energy structure an NP-hard problem.

3.5 Stochastic Solutions

Even if exact solutions cannot be computed, we can alternatively look for approximations of the true result. If we allow for the possibility of error our solutions, we can employ techniques that can achieve results in more reasonable time. Note, however, that the "best solution" already has a degree of error, as we already have the error incurred by the simplifications of our energy model. Therefore, it can be argued that approximating the "best solution" is sufficient, as our true goal is not the best solution to the energy model, but to get a solution that is close to the true physical reality.

The use of randomness, or stochastic techniques, is an effective way to generate good quality solutions in reasonable time. Instead of exploring all possibilities, controlled randomness is introduced to limit the choices. As all possibilities are no longer considered, there is a chance that the best solution might be missed. However, we can intelligently constrain the random choice so that we are likely to make a good choice. The random element introduced means that no two runs of the program are the same. Therefore, the quality of the solution is variable. However, this allows the program to generate better results given more time, as the program can simply be run again, with the possibility of an different or potentially improved result.

3.5.1 Stochastic Local Search

Stochastic local search is a technique that takes a solution and attempts to improve it by modifying it. If, using some measure, the new solution is better than the previous one, it is accepted. If the solution is worse, there is a low random chance that it still will be accepted. Stochastic local search has been successfully applied to get solutions to otherwise very hard problems. Hoos and Stützle[HS04] provide a more detailed introduction to the theory behind this technique and its many applications.

`RiSKFold` will use stochastic local search to look for improved structures. We shall use free energy as our measure, and our modifications will be potential changes in the structure — the addition or removal of an interaction. From our pool of potential structures, we shall take one of the structures, and modify it. If the free energy of the new structure is lower than the original one, we are potentially getting a better structure, and this new structure is added to our pool of potential structures. If the new structure is worse, we shall keep it with some low random chance — it may be a transition state intermediate, which could be along the path to the native state.

However, there are times when the researcher may wish to direct the search manually. `RiSKFold` will be designed to be interactive, and will make it straightforward for a researcher to direct this tool, emphasising that although computers are very capable at straightforward, repetitious tasks, there are times when a researcher will have insight not available to the computer program, and will wish to investigate a particular structure or pathway in more detail.

3.6 Kinetic Folding Algorithms

3.6.1 `Kinfold`

Flamm et al.[FFHS00] proposed simulating the folding of RNA molecules using a stochastic approach, similar to the one proposed by `RiSKFold`. They limit their resolution to only base-pairing contacts, although they mention that their approach can be extended, as we propose to do, to handle more complex interactions such as tertiary interactions and non-Watson-Crick base pairing. They also do not account for more complex structures such as pseudoknots or tertiary structure. However, with these simplifications, they are able to simulate folding pathways stochastically — they allow the RNA to change shape randomly in a limited way, by making and breaking bonds, and the chains sliding, favouring changes that lower free energy. By simulating the folding of many molecules, they can gather statistics on which folding pathways occur most frequently, and the secondary structures along these pathways. The source code for `Kinfold` is available at `http://www.tbi.univie.ac.at/~xtof/RNA/Kinfold/`. The program has to be downloaded and compiled to be used.

3.6.2 `KineFold`

Xayaphoummine et al[XBI05] have developed a stochastic folding simulator which models the folding of a single RNA (or DNA) molecule over short time scales, on the order of the first few minutes of the folding process. They also use a stochastic folding technique similar to `Kinfold` and only model one randomly determined pathway, but they allow the formation of pseudoknots and associated effects on the RNA helix. This is modelled as the pseudoknots of n helical turns "twist" in one region requiring a compensating $-n$ helical turns "counter-twist" elsewhere to keep the structure stable, modelled as entropy. As well, long helices that are not part of a pseudoknot, but near the pseudoknot can be prevented from unpairing, as they are unable to uncoil without the pseudoknot being unformed first. They have successfully used this technique on molecules as long as 300–400 nucleotides, such as the 394 nt *Tetrahymena* group I intron. Sequences can be submitted to `KineFold` online (`http://kinefold.curie.fr/`), and the server can generate movies of the folding of the sequence. As the folding is done stochastically, users are recommended do several trials with their sequence to see several independent folding simulations.

`RiSKFold` is inspired by both of these programs, and aims to improve on both of their strengths. It will incorporate both the sampling of many folding pathways from `Kinfold` as well as handling the more complex structures from `KineFold`. It also aims to improve the model by investigating the local structures at higher levels of detail, to obtain improved estimates of the free energy.

3.7 Importance of Other Low Energy Structures — `Srna`

Minimising the free energy and simulating the folding pathway are both models for predicting the structure of an RNA molecule. Ding et al.[CLD05, DCL05] propose another method for predicting the structure of an RNA molecule that uses a simplified free energy model like mfold, but

improves upon it by using statistical sampling of secondary structures, favouring low energy secondary structures. This is motivated by the existence of RNAs that may exist in a population of possible structures and to compensate for inaccuracies of the simplified energy model. They then find an "average" representative from this pool of candidate structures, and show that this is often an improvement over the structure with the lowest free energy. A set of RNA-related tools are available at `http://sfold.wadsworth.org/index.pl`, with the `Srna` structure prediction tool at `http://sfold.wadsworth.org/srna.pl`

Similarly, `RiSKFold` does not look at only the lowest energy structures, but also generates a set of low energy structures. As well, by constructing potential folding pathways to the structures, we can model of the likelihood of the RNA to fold to a given structure after a period of time. This would also allow us to potentially differentiate kinetically favoured conformations from thermodynamically favoured conformations.

3.8 Computational Experimentation

`RiSKFold` is also a foray into the new field of using computer simulations as an alternative platform on which to perform experiments. Allowing the possibility of modelling extremely high levels of detail, combined with the ever increasing capabilities of commonly available computers, gives us the ability to gather data which would be otherwise impossible using normal experimental techniques. A computer simulation allows us to know every facet of the system at every time point, which can be saved and analysed at a future date, as every aspect of the system is a known, deliberately modelled quantity. Of course, compromises will always need to be made so that results can be obtained in reasonable time, and these introduce error. `RiSKFold` achieves this by initially simulating the folding process initially at a coarse level of detail, but mitigates this by improving these to high levels of detail. The ability to simultaneously capture fine details, which may be very difficult to obtain using traditional experimental techniques, may ultimately give us a new source of experimental data.

The idea of using analysing an "*in silico* reconstitution" of a natural biological process has previously been done to investigate the movement of *Listeria* via actin polymerisation[AO04]. In this case, they were able to correlate pauses in the movement of the bacterium with binding and breaking of attachments between the bacterium and the actin filaments. Modelling the folding of RNA is undoubtedly a significantly more complex task, however the potential applications are likewise of broader relevance.

3.8.1 Simulating the GCAA RNA tetraloop

Nivón et al.[NS04] detail an atomic level analysis of the folding of the GCAA RNA tetraloop motif. In this case, they fully exploit the known structure of the GCAA tetraloop and instead focus on the features revealed by pathways predicted using a simple model. In the G_0 Monte Carlo model they used, interactions that occur in the known, native structure are attractive, whereas all other possible interactions are repulsive. Therefore, although covalent interactions are being very coarsely modelled, all atoms are being modelled, emphasising steric and entropic effects. Using this model, they simulate the folding of a linear RNA sequence into its final folded form by

randomly rotating the backbone and bases. After each set of random rotations, the RNA structure is evaluated based on the simplified atom model. If it has a lower energy than before, it is accepted. If it has a higher energy, it is accepted with some probability. Even given this simplified model, they found a potential intermediate state and the possibility of misfolds.

`RiSKFold` as a platform will emphasise versatility and accessibility of results. Free energy is a commonly used measure that is well-founded in physical theory, however this can easily be augmented or replaced by different measures if we wish to emphasise or study certain interactions. For example, if the true native structure of an RNA molecule is already known, we can use the energy model of the G_0 Monte Carlo method to investigate what pathways are favoured when steric and entropic effects are emphasised. This could be applied globally to all the structures — several energy measures could be computed for all the structures while the program is running. It could also be used to only analyse the most likely pathways, if these computations take a long time to run.

Other energy functions that would also be of interest would be the energy of an RNA molecule docked with a particular ligand. In this case, we are not only interested in the folding of the RNA molecule, but also how it binds to a particular ligand. A low free energy involves not only the conformation of the RNA, but also its interactions with the ligand, allowing us to identify potential structures that could favourably interact with the ligand. This would allow us to identify binding sites as well as potential active sites, if the RNA molecule is a ribozyme.

3.9 Conclusion

We present here `RiSKFold`, RNA *in silico* Stochastic Kinetic Folding, a stochastic RNA folding algorithm designed to explore the folding pathways of an RNA molecule. To do this efficiently, it will look at multiple levels of detail, and explore the most likely folding pathways by evaluating the free energy of structures along the potential folding paths.

As yet, all prediction programs are restricted to predicting the structure in the presence of Na^+, and are unable to predict the more complex interactions which occur when divalent ions, such as Mg^{2+}, are present in solution. For example, as seen by Zarrinkar and Williamson[ZW94] in the case of the *Tetrahymena* Group I ribozyme, many of the more interesting interactions, including those critical for the activity of the ribozyme, only form in the presence of Mg^{2+}. Although prediction algorithms continue to improve, modelling this remains a lofty goal we have yet to achieve.

Ultimately, the question is not whether a system like `RiSKFold` is an effective tool using the present-day technology, but rather whether such a system will be effective in the near future when it is complete. Computers have yet to slow their phenomenal rate of progress — what is needed is to be ready to exploit this power when it becomes available.

References

[AO04] Jonathan B. Alberts and Garrett M. Odell. In silico reconstitution of *Listeria* propulsion exhibits nano-saltation. *Public Library of Science Biology*, 2(12):e412, 2004.

[CLD05] Chi Yu Chan, Charles E. Lawrence, and Ye Ding. Structure clustering features on the Sfold web server. *Bioinformatics*, 21:3926–3928, 2005.

[DCL05] Ye Ding, Chi Yu Chan, and Charles E. Lawrence. RNA secondary structure prediction by centroids in a Boltzmann weighted ensemble. *RNA*, 11:1157–1166, 2005.

[FFHS00] Christoph Flamm, Walter Fontana, Ivo L. Hofacker, and Peter Schuster. RNA folding at elementary step resolution. *RNA*, 6:325–338, 2000.

[HS04] Holger H. Hoos and Thomas Stützle. *Stochastic Local Search: Foundations and Applications*. The Morgan Kaufmann Series in Artificial Intelligence. Morgan Kaufmann, 2004.

[NS04] Lucas G. Nivón and Eugene I. Shakhnovich. All-atom Monte Carlo simulation of GCAA RNA folding. *Journal of Molecular Biology*, 334:29–45, 2004.

[WSSF+04] Michael T. Wolfinger, W. Andreas Svrcek-Seiler, Christoph Flamm, Ivo L Hofacker, and Peter F Stadler. Efficient computation of RNA folding dynamics. *Journal of Physics A: Mathematical and General*, 37:4731–4741, 2004.

[XBI05] A. Xayaphoummine, T. Bucher, and H. Isambert. Kinefold web server for RNA/DNA folding path and structure prediction including pseudoknots and knots. *Nucleic Acids Research*, 33:W605–W610, 2005.

[Zuk89] M. Zuker. The use of dynamic programming algorithms in RNA secondary structure prediction. *Mathematical Methods for DNA Sequences*, pages 159–184, 1989.

[ZW94] Patrick P. Zarrinkar and James R. Williamson. Kinetic intermediates in RNA folding. *Science*, 265(5174):918–924, 1994.