McGenus: A Monte Carlo algorithm to predict RNA secondary structures with pseudoknots

Michaël Bon [1], Cristian Micheletti [2] and Henri Orland [1]

(1) Institut de Physique Théorique, CEA Saclay,

CNRS URA 2306, 91191 Gif-sur-Yvette, France

(2) SISSA, Scuola Internazionale Superiore di Studi Avanzati and CNR-IOM Democritos,

Via Bonomea 265, I-34136 Trieste, Italy

(Dated: April 26, 2012)

Abstract

We present McGenus, an algorithm to predict RNA secondary structures with pseudoknots. The method is based on a classification of RNA structures according to their topological genus. McGenus can treat sequences of up to 1000 bases and performs an advanced stochastic search of their minimum free energy structure allowing for non trivial pseudoknot topologies. Specifically, McGenus employs a multiple Markov chain scheme for minimizing a general scoring function which includes not only free energy contributions for pair stacking, loop penalties, etc. but also a phenomenological penalty for the genus of the pairing graph. The good performance of the stochastic search strategy was successfully validated against TT2NE which uses the same free energy parametrization and performs exhaustive or partially exhaustive structure search, albeit for much shorter sequences (up to 200 bases). Next, the method was applied to other RNA sets, including an extensive tmRNA database, yielding results that are competitive with existing algorithms. Finally, it is shown that McGenus highlights possible limitations in the free energy scoring function. The algorithm is available as a web-server at http://ipht.cea.fr/rna/mcgenus.php.

INTRODUCTION

In the past twenty years, there has been a tremendous increase of interest in RNA by the biological community. This biopolymer, which was at first merely considered as a simple information carrier, was gradually proven to be a major actor in the biology of the cell [1].

Since the RNA functionality is mostly determined by its three-dimensional conformation, the accurate prediction of RNA folding from the nucleotide sequence is a central issue [2]. It is strongly believed that the biological activity of RNA (be it enzymatic or regulatory), is implemented through the binding of some unpaired bases of the RNA with their ligand. It is thus crucial to have a precise and reliable map of all the pairings taking place in RNA and to correctly identify loops. The complete list of all Watson-Crick and Wobble base pairs in RNA is called the *secondary structure* of RNA.

In this paper, we stick to the standard assumption that there is an effective free energy which governs the formation of secondary structures, so that the optimal folding of an RNA sequence is found as the minimum free energy structure (MFE for short). The problem of finding the MFE structure given a certain sequence has been conceptually solved provided the MFE is planar, *i. e.* the MFE structure contains no pair (i,j), (k,l) such that $i < k < j < l$ or $k < i < l < j$. In that case, polynomial algorithms which can treat long RNAs assuming a mostly linear free energy model have been proposed [3–5]. Otherwise, the MFE structure is said to contain pseudoknots and finding it has been shown to be an NP-complete problem with respect to the sequence length [6].

In a previous paper [7], we proposed an algorithm, TT2NE, which consists in searching for the exact MFE structure for a certain form of the energy function, where pseudoknots are penalized according to a topological index, namely their genus. TT2NE relies on the "maximum weighted independent set" (WIS) formalism. In this formalism, an RNA structure is viewed as an aggregate of stem-like structures (helices possibly comprising bulges of size 1 or internal loops of size 1×1), called "helipoints" [7]. Given a certain sequence, the set of all possible helipoints is computed and a weighted graph is built. The vertices of the graph are the helipoints, with a weight given by minus their free energy of formation. Two vertices are connected by an arc if and only if the corresponding helipoints are not compatible in the same secondary structure. Indeed, two helipoints may be mutually exclusive in a graph: this is for example the case if they share at least one base (since base triples are forbidden).

Finding the MFE structure thus amounts to finding the maximum weighted independent set of the graph, *i. e.* the set of pairwise compatible helipoints for which the overall free energy is minimum.

Both McGenus and TT2NE utilize the same energy function, defined in terms of helipoints and genus penalty as well as the same initial graph. The difference between the two lies in the search algorithm for the MFE. While in TT2NE the secondary structure is built by adding or removing helipoints in a deterministic order, in McGenus, they are added or removed one at a time according to a stochastic Monte Carlo Metropolis scheme. As in TT2NE, there is no restriction on the pseudoknots topologies that McGenus can generate. A server implementation of McGenus can be found at http://ipht.cea.fr/rna/mcgenus.php.

In the following and in the numerical implementation of McGenus, we will restrict ourselves to the energy function and genus penalty described in detail in [7]. While in TT2NE, the energy form was dictated by the requirement to allow for a branch and bound procedure, here in McGenus we insist that there is no such restriction on the form of the energy function. It can for instance include loop and pseudoknot entropies. Furthermore, the penalty for pseudoknots needs not be proportional to the genus as in TT2NE, but may depend also on the topology of each individual pseudoknot (see below).Therefore, by modifying the energy function, it is possible to improve on the results that we will present below.

As stated in the introduction, the initial graph is generated in the same way as in [7].

MATERIALS AND METHODS

In the present framework, the folded structure of a given RNA sequence is given by the set of helipoints which minimizes the free energy. For definiteness of notation, in the following we shall denote by $\{h_1, ..., h_N\}$ the set of all helipoints that can possibly arise from the pairings of nucleotides in the given sequence (their total number, N, is clearly sequence dependent). A given structure S is accordingly fully specified by the associated subset of n helipoints $\{h_{i_1}, h_{i_2}, ..., h_{i_n}\}$ and its free energy is formally given by:

$$F = \sum_{j=1}^{n} e(h_{i_j}) + \mu\, g(S)\ . \tag{1}$$

The first term is the additive contribution of the pairing and stacking energy e of individual helipoints, and is parametrized as in [7]. The second term weights the topological

complexity of the structure, measured by its genus g [8, 9]. Unlike the first term which is local, the genus, which is a non-negative integer, depends globally on all the helipoints. The parameter $\mu \geq 0$ is used to penalize structures with excessively large values of the genus, in agreement with the phenomenological observation that the genus of most naturally-occurring RNA structures of size up to 600 bases, is smaller than 4. Based on previous studies [7], the default value of the genus penalty μ is set equal to 1.5 Kcal/mol.

It is implicitly assumed that the free energy of incompatible sets of helipoints (e.g. when at least one base takes part to more than one helix) is infinite.

Advanced Monte Carlo search of MFE structures

The minimization of the free energy (1) is carried out by a Monte Carlo (MC) exploration of structure space, that is over sets of distinct helipoints $\{h_{i_1}, h_{i_2}, ..., h_{i_n}\}$, picked from the full ensemble of helipoints $\{h_1, ..., h_N\}$. To illustrate the search algorithm, it is convenient to view the structures as described by sets of integer $\{\sigma_1, ..., \sigma_N\}$, where each σ_i is equal to 1 if the helipoint i is active, i.e. present in the structure, and 0 otherwise. Starting from a structure consisting of only one active helix, at each Monte Carlo step one of the following two "moves" is tried:

- (i) helix addition, consisting of the activation of one inactive helix, $\sigma = 0 \rightarrow \sigma = 1$,
- (ii) helix removal, consisting of the inactivation of one active helix, $\sigma = 1 \rightarrow \sigma = 0$.

The helipoint whose state is changed by the MC move is picked with a biased probability favoring the activation of helipoints with low pairing/stacking energy e and the inactivation of helipoints with high values of e. The biasing is inspired by the heat-bath MC algorithm. Specifically, the helipoint $\bar{h}$ activated in case *(i)* is picked among the inactive ones with probability w defined by

$$w = \exp[-e(\bar{h})/\kappa_B T]/\mathcal{Z}_{\bar{S}} \tag{2}$$

$$\text{with} \quad \mathcal{Z}_{\bar{S}} = \sum_{h \text{ inactive in } S} \exp[-e(h)/\kappa_B T] \tag{3}$$

where κ_B is the Boltzmann constant and T is the Monte Carlo temperature. In case *(ii)*, the inactivated helix $\bar{h}$ is instead picked with probability

$$w = \exp[+e(\bar{h})/\kappa_B T]/\mathcal{Z}_S \tag{4}$$

$$\text{with } \mathcal{Z}_S = \sum_{h \text{ active in } S} \exp[+e(h)/\kappa_B T] \tag{5}$$

A generalized Metropolis criterion (which takes into account the biased choice of the activated/inactivated helices) is finally used to accept or reject the new structure and ensure that, in the long run, the generated structures are sampled with probability given by the canonical weight $\exp[-e/\kappa_B T]$.

The stochastic generation of structures is carried out within a multiple Markov chain scheme where several simulations are run in parallel at different temperatures T. The temperatures are chosen so as to cover a range of thermal energies, $\kappa_B T$, going from about one tenth of the smallest helipoint energy up to the largest helipoint energy. At regular time intervals, swaps are proposed between structures at neighboring temperatures and are accepted with the generalized Metropolis criterion described in ref. [10]. The Markov replicas at the lowest temperature progressively populate structures of low free-energy, and a record is kept of the lowest energy structures which are finally provided as output.

Finally, we point out that the Monte Carlo optimization can be performed not only within the whole space of secondary structures (unconstrained search) but is straightforwardly restricted to topologically-constrained subspaces. In particular, by introducing *ad hoc* "infinite" energy penalties in eq. 1, the search can be restricted to structures whose genus, topology or extent of pseudoknots satisfy some preassigned constraints. The web-server interface allows the user to set such thresholds, e.g. to account for knowledge based constraints.

Generalized Topological Penalties

As we have previously reported [11, 12], any RNA complex pseudoknot structure may be built from of a set of building blocks, called primitive pseudoknots. A pseudoknots is termed primitive if it is *(i)* irreducible, *i.e.* its standard diagrammatic representation cannot be disconnected by cutting one backbone line and *(ii)* contains no nested pseudoknot, that is it cannot be disconnected by cutting two backbone lines, see Fig. 1.

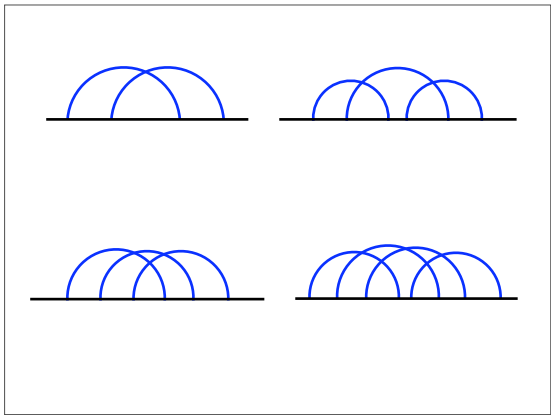

FIG. 1: *The only four primitive pseudoknots of genus 1 [11].*

As was shown in [11], most complex high genus RNA structures are built from primitive graphs of smaller genus . Therefore, it makes sense to assign different penalties to pseudoknots having same genus but with different primitive components. For example, all tmRNAs have total genus 3 or 4 and contain no primitive pseudoknots of genus larger than 1. In the present implementation, we propose only two options: i) we forbid primitive pseudoknots of genus larger than 1 (by assigning them an infinite penalty) but the overall structure can have any total genus or ii) we assign a global penalty proportional to the total genus and don't take into account the decomposition of the structure into primitive blocks.

RESULTS AND DISCUSSION

We have carried out an extensive comparison of McGenus predictions against those of other methods. For this purpose we used hundreds of RNA sequences from various sets, including: the dataset previously used for TT2NE [7], an extensive set of tmRNAs [13] and the more limited set of RNA molecules for which the structural data is available in the protein databank (PDB). Over such diverse datasets, the predictive performance is aptly conveyed by the *sensitivity* of the method, that is the fraction of pairs in the reference (native) structure that are correctly predicted by the method. Depending on the context we shall also report on the positive predicted value (PPV). The PPV corresponds to the fraction of predicted pairs that are found in the native structure, and hence measures the incidence of false positives in the predicted contacts. We shall consider this measure for the

PDB set, but not for the tmRNA set whose entries, often corresponding to putative native structures derived from homology, are known to potentially lack several native contacts, as in the paradigmatic case of Aste.yell._TRW-322098_1-426 [13].

From an overall point of view, the tests are aimed at elucidating two issues that are central to any MFE-based method. The first issue, regards the algorithmic effectiveness of the energy minimization, while the second regards the viability of the energy parametrization within the considered space of secondary structures. The former is most clearly ascertained by comparing algorithms employing the same energy parametrization. This step is crucial for the second aspect too. In fact, the appropriateness or the limitations of a given energy parametrization and/or of the considered secondary structure space, can be exposed in a non-ambiguous way only if the minimization algorithm is well-performing.

Following the above-mentioned logical order, we started by comparing the predictions of McGenus against TT2NE on a database of 47 short sequences (< 209 bases) used in [7]. Because McGenus and TT2NE rely on the same energy parametrization[14], the comparison provides a stringent test of the effectiveness of the energy-minimization procedure. In fact, we recall that TT2NE is based on an exhaustive, or nearly exhaustive search in sequence space. Despite the stochastic, non-exhaustive and much faster McGenus searches, its performance turned out to be optimal. Over the full data set, McGenus returned exactly the same MFE structures as TT2NE, as well as all the suboptimal structures.

To extend the assessment of McGenus minimization performance for longer chains, that cannot be addressed by TT2NE, we considered MFold [4], a MFE-based algorithm restricted to secondary structures without pseudoknots. We used a version of MFold which employs the same energy parametrization as McGenus.The comparison was carried out over the complete set of 590 sequences of genus 3, 4 or 5 from the tmRNA database [13] with lengths in the 200-500 range. For each of the 590 sequences, McGenus returned structures with lower free energy than MFold. On the average, the free energy of the McGenus predicted structures was -125 kCal/mol, while that predicted by MFold was -103 kCal/mol.

These two tests prove the effectiveness of the energy-minimization scheme adopted by McGenus and we accordingly turned our attention to the overall predictive performance of the method (sensitivity). For this purpose we used again the 590 sequences of genus 3, 4 or 5 from the tmRNA database [13] and compared McGenus predictions against McQfold [15], HotKnots [16], ProbKnots [17], PKnots [18], gfold [19] and Mfold [20] on this set. Besides

McGenus, only McQfold and MFold could handle all the chains in the set, which were too long to process for the other mentioned algorithms. We recall that MFold predictions are restricted to secondary structures free of pseudoknots, while McQfold can output any topology of pseudoknot. The genus of each of McGenus prediction was enforced not to exceed the genus of the native structures of the dataset. As discussed in [7], the setting of the corresponding parameter g_{max} can be decided by the user. In this report, for each test sequence, we chose to set g_{max} to the appropriate, native, value to illustrate the performance of McGenus performs when it is driven in the appropriate secondary structure search space.

The total number of base pairs to be predicted in the set is 56740. Mfold, McQfold, and McGenus respectively predicted 37%, 42% and 42% of them. Therefore the performance of McGenus is not inferior to that the few available structure prediction methods that can handle sequences of comparable length.

The fact that the average sensitivity of the three methods is below 50% poses the question of whether it can be improved by tweaking the energy parameters or by suitably further constraining the space of secondary structures over which the minimization is performed. We focus on the latter aspect as the first has been already discussed in [7]. The space of secondary structures considered by prediction schemes based on abstract, graph-theoretical representations, include structures that are unphysical, *i.e.* that cannot be realized in a three-dimensional space because of chain connectivity constraints.

The impact of this major difficulty can be lessened by excluding from further considerations those structures that present physically-unviable or atypical levels of entanglement. To illustrate this point, we note that, in the mentioned dataset of 590 molecules, only H-pseudoknots which span less than 70 bases are present. By enforcing such knowledge-based constraint on the searched space, the sensitivity of McGenus is boosted from 42% to 52%.

Introducing the constraint in structure space clearly results in higher energies for the predicted structures. In fact the average free energy was -125 kCal/mol without the constraint while it is -114 kCal/mol with the restriction of the pseudoknot length. Notwithstanding the reduction of the searched space due to the pseudoknot-length constraint, the structures returned by McGenus have an energy that is significantly lower than the reference, (putative) native structures, which is about -73kCal/mol. The free energy difference appears too large to be accounted for by the neglected contribution of loop entropy, missing chain-connectivity constraints or imperfect parametrization of the potentials, which are well established. A

more plausible source of discrepancy could the missing contacts in the homology-derived native structure of the tmRNA database.

To check this last point, we have studied the unconstrained version McGenus on a set of 4 sequences from the protein databank (PDB). Their PDB ids are: 1Y0Q (length=229), 3EOH (length=412), 2A64 (length=417) and 2H0W (length=151). The structures of these entries is known unambiguously from X-ray scattering data and contain very few long and non-hybridized RNA sequences (*i.e.* not bound to proteins, DNA or other molecules). Accordingly, the McGenus performance on this set was higher than for the tmRNA set. The sensitivity for 1Y0Q, 3EOH, 2A64 and 2H0W was equal to 87%, 39%, 50% and 72%, respectively while the PPV was equal to 90%, 38%, 35% and 84%, respectively. Again, the structures predicted by McGenus have a lower free energy than the native ones. This indicates that, besides accounting for topological effects, further improvements of secondary structure predictions would probably require a better parametrization of the free energy and of its functional form. The generality and flexibility of the McGenus search algorithm ought to allow for incorporating any such modifications in a transparent way.

CPU time

The CPU time required by McGenus to fold an RNA sequence depends on the total number of Monte Carlo steps. For a tm-RNA of length 400, the typical number of helipoints is 3500. For each sequence, we use 10 replicas, and overall 3000 $\times$ number of helipoint steps to achieve these results. The result is typically returned in 15 minutes on a parallel quadcore computer.

CONCLUSION

In this article, we presented McGenus, an efficient algorithm for RNA pseudoknot prediction, which proves that classifying pseudoknots according to their genus is a relevant and successful concept. We showed that on a set of RNA structures from the tm-RNA database [13], McGenus allows to treat sequences of sizes up to 1000 in a few minutes, with a performance that is comparable or better than the few methods that can treat sequences with comparable length.

In order to further improve the performance of McGenus, we see 3 main directions: I) improvement on the computing techniques, in particular on the parallelization of the algorithm. II) improvement of the functional form and parametrization of the energy model (likely to impact also on pseudoknot-free methods such as Mfold). III) inclusion of steric constraints.

ACKNOWLEDGEMENTS

The authors wish to thank A. Capdepon for setting up the McGenus server at http://ipht.cea.fr/rna/mcgenus.php.

[1] D. Elliot and M. Ladomery. *Molecular Biology of RNA*. Oxford University Press, 2011.

[2] I. Tinoco Jr. and C. Bustamante. How RNA folds. *Journal of Molecular Biology*, 293, 1991.

[3] R. Nussinov, G. Pieczenik, J.R. Griggs, and D.J. Kleitman. Algorithms for loop matchings. *SIAM Journal on Applied Mathematics*, 35(1):68–82, 1978.

[4] M. Zuker and P. Stiegler. Optimal computer folding of large RNA sequences using thermodynamics and auxiliary information. *Nucleic Acids Research*, 9(1):133–148, 1981.

[5] J.S. McCaskill. The equilibrium partition function and base pair binding probabilities for RNA secondary structure. *Biopolymers*, 29:1105–1119, 1990.

[6] R.B. Lyngso and C.N.S. Pedersen. RNA pseudoknot prediction in energy-based models. *Journal of Computational Biology*, 7(3-4):409–427, 2000.

[7] M. Bon and H. Orland. TT2NE: a novel algorithm to predict RNA secondary structures with pseudoknots. *Nucleic Acids Research*, 39(14):e93–e93, 2011.

[8] H. Orland and A. Zee. RNA folding and large N matrix theory. *Nuclear Physics B*, 620(3):456–476, 2002.

[9] G. Vernizzi and H. Orland. Large N Random Matrices for RNA Folding. *Acta Physica Polonica B*, 36:2821–2827, 2005.

[10] E. Orlandini. Monte carlo study of polymer systems by multiple markov chain method. *Numerical Methods for Polymeric Systems, edited by S. G. Whittington, IMA Volumes in Mathematics and Its Application*, 102:33—58, 1998.

[11] M. Bon, G. Vernizzi, H. Orland, and A. Zee. Topological classification of RNA structures. *Journal of Molecular Biology*, 379(4):900–911, 2008.

[12] G. Vernizzi and H. Orland. *The Oxford Handbook of Random Matrix Theory (chapter 42)*. Oxford University Press, 2011.

[13] C. Zwieb, J. Gorodkin, B. Knudsen, J. Burks, and J. Wower. tmRDB (tmRNA database). *Nucleic acids research*, 31(1):446–447, 2003.

[14] D.H. Mathews, J. Sabina, M. Zuker, and D.H. Turner. Expanded sequence dependence of thermodynamic parameters improves prediction of RNA secondary structure. *Journal of Molecular Biology*, 288(5):911–940, 1999.

[15] D. Metzler and M.E. Nebel. Predicting RNA secondary structures with pseudoknots by MCMC sampling. *Journal of Mathematical Biology*, 56(1):161–181, 2008.

[16] J. Ren, B. Rastegari, A. Condon, and H.H. Hoos. HotKnots: heuristic prediction of RNA secondary structures including pseudoknots. *RNA*, 11(10):1494–1504, 2005.

[17] S. Bellaousov and D.H. Mathews. Probknot: Fast prediction of RNA secondary structure including pseudoknots. *RNA*, 16(10):1870–1880, 2010.

[18] E. Rivas and S.R. Eddy. A dynamic programming algorithm for RNA structure prediction including pseudoknots. *Journal of Molecular Biology*, 285:2053–2068, 1999.

[19] C.M. Reidys, F.W.D. Huang, J.E. Andersen, R.C. Penner, P.F. Stadler, and M.E. Nebel. Topology and prediction of RNA pseudoknots. *Bioinformatics*, 27(8):1076, 2011.

[20] M. Zuker. Mfold web server for nucleic acid folding and hybridization prediction. *Nucleic Acids Research*, 31(13):3406, 2003.