Published online 16 December 2012

Nucleic Acids Research, 2013, Vol. 41, No. 3 1895–1900
doi:10.1093/nar/gks1204

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 and [2]SISSA, Scuola Internazionale Superiore di Studi Avanzati and CNR-IOM Democritos, Via Bonomea 265, I-34136 Trieste, Italy

Received April 27, 2012; Revised October 4, 2012; Accepted October 30, 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 uses a Monte Carlo algorithm with replica exchange 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 20 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 article, 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 that 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 article (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 approach, an RNA structure is viewed as a collection of stem-like structures (helices possibly comprising bulges of size 1 or internal loops of size 1×1), called 'helipoints' (7), defined in the next section. Given a certain sequence, the set of all possible helipoints is enumerated and used to build a weighted graph. The graph vertices are the helipoints and their weight is given by $-1\times$ the helipoint free energy. Two vertices are linked by an arc if and only if the corresponding helipoints are not compatible in the same secondary structure. Incompatibilities arise, for example, when two helipoints share one or more bases as this could imply the formation of base triplets, which is forbidden. Finding the MFE

*To whom correspondence should be addressed. Tel: +39 040 3787 300; Fax: +39 040 3787 528; Email: michelet@sissa.it

structure thus amounts to finding the maximum WIS 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 (MC) Metropolis scheme. As in TT2NE, there is no restriction on the pseudoknot topology 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' section, 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. We recall that a helipoint is 'an ensemble' of helices (defined as a stack of base pairs possibly comprising bulges of size 1 or internal loops of size 1×1) that are demarcated by the same extremal (initial and terminal) base pairs. Given two extremal pairs (i, j) and (k, l), the set ω_{kl}^{ij} of all helices that end with these two pairs can be generated and their individual energies calculated according to a given energy model. The free energy ΔF_{kl}^{ij} of the helipoint is then computed as

$$\exp(-\beta \Delta F_{kl}^{ij}) = \sum_{h \in \omega_{kl}^{ij}} \exp(-\beta e(h)), \quad (1)$$

where $\beta = (k_B T)^{-1}$ and $e(h)$ is the free energy of formation of helix h. In our implementation, to speed up the computation of this sum, helices of non-negative (i.e. unfavorable) energies are neglected, since their Boltzmann weight would strongly suppress their contribution. Helipoints are stem-like structural building blocks which account for all possible internal pairing possibilities that occur between their extremal pairs. 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). We stress that the set of enumerated helipoints comprises all possible helipoints, and hence is not restricted to maximal ones.

Clearly, a given RNA structure S is fully specified by a collection of compatible helipoints. It is therefore convenient to identify S with a binary vector, $\vec{\sigma}^S$, of length N and whose ith component, σ_i^S takes on the value 0 or 1 according to whether helipoint h_i belongs to S. The free energy of S can accordingly be written as:

$$F_S = \sum_{i=1}^{N} \sigma_i^S \Delta F(h_i) + \mu g(S). \quad (2)$$

The first term is the additive contribution of the free energy ΔF 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 <4. Based on previous studies (7), the default value of the genus penalty μ is set to 1.5 kcal/mol.

It is implicitly assumed that the free energy of incompatible sets of helipoints is infinite.

Advanced MC search of MFE structures

The minimization of the free energy (2) is performed by a MC exploration of structure space, which is over the set of possible $\vec{\sigma}$ vectors. Starting from a structure S where only one helipoint is present, at each MC step, one of the helipoints h_i is added ($\sigma_i = 0 \rightarrow \sigma_i = 1$) or removed ($\sigma_i = 1 \rightarrow \sigma_i = 0$). The helipoint to be modified is picked with a biased probability favouring the addition (respectively, removal) of helipoints with low (respectively, high) free energy e. The biasing is inspired by the heat-bath MC algorithm. Specifically, the a priori probability to pick helipoint h_i to be changed in structure S is given by

$$w_i^S = \frac{\sigma_i^S + (1 - \sigma_i^S) e^{-\beta \Delta F(h_i)}}{\sum'_{j=1..N} \sigma_j^S + (1 - \sigma_j^S) e^{-\beta \Delta F(h_j)}}, \quad (3)$$

where the primed sum indicates that helipoints incompatible with S are not considered. Changing the state of h_i defines a trial structure, S', which is accepted with probability

$$\min\left[1, \frac{w_i^{S'}}{w_i^S} e^{-\beta(F_{S'} - F_S)}\right]. \quad (4)$$

The above acceptance criterion is a generalization of the standard Metropolis rule and ensures that, in the long run, the generated structures are sampled with probability given by the canonical weight $\exp[-\beta F_S]$.

The stochastic generation of structures is performed within a MC algorithm with replica exchange where several simulations are run in parallel at different inverse temperatures β. The values of β are chosen so as to cover a range of thermal energies $1/\beta$, going from ~one-tenth of the smallest helipoint energy up to the largest helipoint

energy. At regular time intervals, swaps are proposed between structures at neighbouring temperatures and are accepted with the generalized Metropolis criterion described in (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 MC 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 Equation (2), 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 a set of building blocks, called primitive pseudoknots. A pseudoknot 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 (Figure 1). An arbitrary pseudoknotted structure can be decomposed in a collection of primitive pseudoknots and its total genus is the sum of the genii of its primitive constituents (11).

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 do not take into account the decomposition of the structure into primitive blocks.

RESULTS AND DISCUSSION

We have performed 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 pseudoknotted RNA molecules for which the structural data are 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 have considered 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). A visual representation of this structure can be found in the RNA STRAND database (14) under the reference TMR_00037.

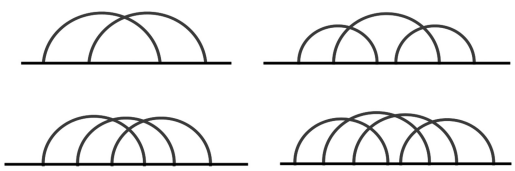

Figure 1. The only four primitive pseudoknots of genus 1 (11).

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). As McGenus and TT2NE rely on the same energy parametrization (15), 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 dataset, 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 UNAFold (4), a MFE-based algorithm restricted to secondary structures without pseudoknots. We used a customized version of UNAFold which uses the same energy parametrization as McGenus. However, it cannot yet be compared with McGenus since it outputs secondary structures in terms of base pairs rather than helipoints. To circumvent this difficulty, we generated all the lowest lying secondary structures (within 1 kcal/mol from the lowest energy structure) using the algorithm presented in (16). To match the description of the structure in terms of helipoints, we made clusters of secondary structures sharing the same extremities of their helical fragments. We then resummed them (in terms of their Boltzmann weights) and as a result the energy discrepancy between the two approaches is

negligible. In the sequel, we will refer to this process as cUNAFold.

The comparison was performed 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. To assess the efficiency of the minimization algorithm of McGenus, we ran it over our sample of 590 sequences, with the constraint $g_{max} = 0$ and compared it with the output of cUNAFold. The average MFE from McGenus with $g_{max} = 0$ is −105.1 kcal/mol while that of cUNAFold is −106.7 kcal/mol. Interestingly enough, out of the 590 sequences, 191 sequences are predicted to have identical secondary structures by both algorithms. This comparison shows the good efficiency of McGenus minimization algorithm.

In the non-zero genus case, for each of the 590 sequences, McGenus returned structures with lower free energy than cUNAFold. On the average, the free energy of the McGenus-predicted structures was −125 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 (17), HotKnots (18), ProbKnot (19) and UNAFold (20) on this set. We did not compare McGenus against PKnots (21) and gfold (22), as the original articles claim that they cannot handle sequences longer than 200 bases. We recall that UNAFold predictions are restricted to secondary structures free of pseudoknots, while ProbKnot and 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 when it is driven in the appropriate secondary structure search space.

The total number of base pairs to be predicted in the set is 56 740. The UNAFold, McQfold, ProbKnot, HotKnots and McGenus arithmetic averages of the sensitivity over all sequences are, respectively, 37, 42, 43, 39 and 43%, with a respective standard deviation of 14, 15, 14, 14 and 16%. A closer look at the secondary structures output by ProbKnot and HotKnots showed that none of them contained any pseudoknot. Therefore, the performance of McGenus is not inferior to that of the few methods that can handle sequences of comparable length. Even without resorting to advanced comparative tests (23,24), the consistent sensitivity of these five algorithms allows to conclude that their performance is very similar.

The fact that the average sensitivity of the five methods is <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, includes 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 of 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 <70 bases are present. By enforcing such knowledge-based constraint on the search space, the sensitivity of McGenus is boosted from 43 to 53% with a standard deviation of 18%. To assess the statistical significance of this improvement, we performed the Welch t-test. We find a t-value of $t = 10$, which with a total of 1168 degrees of freedom implies a P-value $<10^{-7}$, i.e. the improvement is definitely significant.

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 whereas it is −114 kcal/mol with the restriction of the pseudoknot length. Notwithstanding the reduction of the search 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 −73 kcal/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 be 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 four sequences from the PDB with g_{max} being fixed to the native genus. Their PDB ids are 1Y0Q (length = 229, $g = 1$), 3EOH (length = 412, $g = 1$), 2A64 (length = 417, $g = 1$) and 2H0W (length = 151, $g = 2$). The structures of these entries are unambiguously known 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. The generality and flexibility of the McGenus search algorithm ought to allow for incorporating any such modifications in a transparent way.

Finally, let us discuss the choice of a maximum genus. Ideally, one should perform the computation with a completely unconstrained genus. However, there are two difficulties to this approach. First, since steric constraints are only limitedly accounted for by available pseudoknot prediction algorithms (including McGenus), the predicted

structures can be sterically impossible and hence associated to an excessively high genus. Secondly, the computational time required to explore the unrestricted genus space could be impractical. To overcome these difficulties and restrict the search space, one can profitably introduce knowledge-based constraints. In particular, the statistical PDB analysis of (11) provides a quantitative indication for the dependence of the genus on the length of naturally occurring RNA sequences. The data can be clearly used to provide a phenomenological upper bound to g_{max}. Alternatively, a user could explore a few different increasing values of g_{max} and perform a supervised evaluation of the results by taking into account (i) the phenomenological constraints and (ii) the possibility that structures with excessively large genus value are returned because of the imperfect treatment of steric constraints.

To illustrate this last point, we ran McGenus on a set of 792 5S rRNA sequences of length around 150, with no pseudoknot. We set $g_{max} = 3$ which according to the study of (11) (see Figure 10 therein) is very large. The number of sequences predicted with genus 0 (i.e. without pseudoknots) is 258, with genus 1 is 500, with genus 2 is 34 and with genus 3 is 0. Consistently with the remarks made in the context of H-pseudoknots, the results indicate that performance of pseudoknot prediction algorithms could certainly benefit by improving the current handling of chain connectivity and excluded volume constraints.

CPU time

The CPU time required by McGenus to fold an RNA sequence depends on the total number of MC steps. For a tmRNA of length 400, the typical number of helipoints is 3500. For each sequence, we use 10 replicas, and overall 3000× number of helipoints steps to achieve these results. The result is typically returned in 15 min on a parallel quadcore computer (Intel Xeon CPU @2.66GHz). The current implementation of McGenus on the server is not parallelized.

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 tmRNA database (13), McGenus allows treatment of sequences of sizes up to 1000 with a typical CPU time of 15 min for a 500 long sequence on a quadcore CPU, 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 three 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 have an impact also on the parametrization of pseudoknot-free methods such as UNAFold) and (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.

FUNDING

Funding for open access charge: Italian Ministry of research, FIRB—Futuro in Ricerca N. [RBFR102PY5].

Conflict of interest statement. None declared.

REFERENCES

1. Elliot,D. and Ladomery,M. (2011) *Molecular Biology of RNA*. Oxford University Press, New York.
2. Tinoco,I. Jr and Bustamante,C. (1991) How RNA folds. *J. Mol. Biol.*, **293**, 271–281.
3. Nussinov,R., Pieczenik,G., Griggs,J.R. and Kleitman,D.J. (1978) Algorithms for loop matchings. *SIAM J. Appl. Math.*, **35**, 68–82.
4. Markham,N.R. and Zuker,M. (2008) UNAFold: software for nucleic acid folding and hybridization. In: Keith,J.M. (ed.), *Bioinformatics, Vol. II. Structure, Function and Applications*, Vol. 453, Chapter 1. Humana Press, Totowa, NJ. pp. 3–31.
5. McCaskill,J.S. (1990) The equilibrium partition function and base pair binding probabilities for RNA secondary structure. *Biopolymers*, **29**, 1105–1119.
6. Lyngso,R.B. and Pedersen,C.N.S. (2000) RNA pseudoknot prediction in energy-based models. *J. Comput. Biol.*, **7**, 409–427.
7. Bon,M. and Orland,H. (2011) TT2NE: a novel algorithm to predict RNA secondary structures with pseudoknots. *Nucleic Acids Res.*, **39**, e93–e93.
8. Orland,H. and Zee,A. (2002) RNA folding and large N matrix theory. *Nucl. Phys. B*, **620**, 456–476.
9. Vernizzi,G. and Orland,H. (2005) Large N random matrices for RNA folding. *Acta Phys. Pol. B*, **36**, 2821–2827.
10. Orlandini,E. (1998) Monte Carlo study of polymer systems by multiple Markov chain method. In: Whittington,S.G. (ed.), *Numerical Methods for Polymeric Systems (IMA Volumes in Mathematics and its Application)*Vol. 102. Springer, Berlin, pp. 33–58.
11. Bon,M., Vernizzi,G., Orland,H. and Zee,A. (2008) Topological classification of RNA structures. *J. Mol. Biol.*, **379**, 900–911.
12. Vernizzi,G. and Orland,H. (2011) *The Oxford Handbook of Random Matrix Theory*, Chapter 42. Oxford University Press, UK.
13. Zwieb,C., Gorodkin,J., Knudsen,B., Burks,J. and Wower,J. (2003) tmRDB (tmRNA database). *Nucleic Acids Res.*, **31**, 446–447.
14. Andronescu,M., Bereg,V., Hoos,H.H. and Condon,A. (2008) RNA STRAND: the RNA secondary structure and statistical analysis database. *BMC Bioinformatics*, **9**, 340.
15. Mathews,D.H., Sabina,J., Zuker,M. and Turner,D.H. (1999) Expanded sequence dependence of thermodynamic parameters improves prediction of RNA secondary structure. *J. Mol. Biol.*, **288**, 911–940.
16. Wuchty,S., Fontana,W., Hofacker,I. and Schuster,P. (1999) Complete suboptimal folding of RNA and the stability of secondary structures. *Biopolymers*, **49**, 145–165.
17. Metzler,D. and Nebel,M.E. (2008) Predicting RNA secondary structures with pseudoknots by MCMC sampling. *J. Math. Biol.*, **56**, 161–181.
18. Ren,J., Rastegari,B., Condon,A. and Hoos,H.H. (2005) HotKnots: heuristic prediction of RNA secondary structures including pseudoknots. *RNA*, **11**, 1494–1504.
19. Bellaousov,S. and Mathews,D.H. (2010) Probknot: fast prediction of RNA secondary structure including pseudoknots. *RNA*, **16**, 1870–1880.
20. Zuker,M. (2003) UNAFold web server for nucleic acid folding and hybridization prediction. *Nucleic Acids Res.*, **31**, 3406.

21. Rivas,E. and Eddy,S.R. (1999) A dynamic programming algorithm for RNA structure prediction including pseudoknots. *J. Mol. Biol.*, **285**, 2053–2068.
22. Reidys,C.M., Huang,F.W.D., Andersen,J.E., Penner,R.C., Stadler,P.F. and Nebel,M.E. (2011) Topology and prediction of RNA pseudoknots. *Bioinformatics*, **27**, 1076.
23. Xu,Z., Almudevar,A. and Mathews,D.H. (2012) Statistical evaluation of improvement in RNA secondary structure prediction. *Nucleic Acids Res.*, **40**, e26.
24. Hajiaghayi,M., Condon,A. and Hoos,H.H. (2012) Analysis of energy-based algorithms for RNA secondary structure prediction. *BMC Bioinformatics*, **13**, 22.