Computational Biology and Chemistry 60 (2016) 43–52

Contents lists available at ScienceDirect

Computational Biology and Chemistry

journal homepage: www.elsevier.com/locate/compbiolchem

Research Article

A new heuristic method for approximating the number of local minima in partial RNA energy landscapes

Andreas A. Albrecht[a], Luke Day[b], Ouala Abdelhadi Ep Souki[b], Kathleen Steinhöfel[b,*]

[a] Middlesex University, School of Science and Technology, London, UK
[b] King's College London, Informatics Department, London, UK

ARTICLE INFO

Article history:
Received 7 March 2015
Received in revised form 17 October 2015
Accepted 10 November 2015
Available online 19 November 2015

Keywords:
RNA folding landscapes
Energy landscape analysis
Local minima
Gamma distribution
Pooling methods

ABSTRACT

The analysis of energy landscapes plays an important role in mathematical modelling, simulation and optimisation. Among the main features of interest are the number and distribution of local minima within the energy landscape. Granier and Kallel proposed in 2002 a new sampling procedure for estimating the number of local minima. In the present paper, we focus on improved heuristic implementations of the general framework devised by Granier and Kallel with regard to run-time behaviour and accuracy of predictions. The new heuristic method is demonstrated for the case of partial energy landscapes induced by RNA secondary structures. While the computation of minimum free energy RNA secondary structures has been studied for a long time, the analysis of folding landscapes has gained momentum over the past years in the context of co-transcriptional folding and deeper insights into cell processes. The new approach has been applied to ten RNA instances of length between 99 nt and 504 nt and their respective partial energy landscapes defined by secondary structures within an energy offset ΔE above the minimum free energy conformation. The number of local minima within the partial energy landscapes ranges from 1440 to 3441. Our heuristic method produces for the best approximations on average a deviation below 3.0% from the true number of local minima.

1. Introduction

The present paper aims at a fast and sufficiently accurate method for the evaluation of data obtained from partial RNA energy landscapes in the context of the approximation of the number of local minima. The partial energy landscapes can be defined by an energy offset above the minimum free energy conformation or by a bounded distance in terms of elementary transition steps from a given conformation within the energy landscape. As a continuation of previous work presented in Sahoo and Albrecht (2012), we consider energy landscapes induced by RNA secondary structures and partial energy landscapes defined by energy offsets. The method can be used, e.g., in a pre-processing step before starting a comprehensive analysis (or complete generation) of local minima in partial energy landscapes for *a priori* information about the expected number of local minima. Here, we focus on the fast and reliable evaluation of data produced by steepest decent that starts out from samples of secondary structures, whereas for pre-processing steps, such as initial sample size selection and randomised generation of sample sets, we rely on existing tools for RNA secondary structure generation and free energy calculation (tailored methods for pre-processing steps are subject of current and future research).

As in Sahoo and Albrecht (2012), we utilise the framework presented by Garnier and Kallel (2002) for approximating the number of local minima in a fitness landscape (induced by an objective function, which is called fitness value): At the initial stage, M elements of the landscape are randomly selected. Each of the elements then 'moves' towards a local/global minimum S_m based upon steepest descent (being part of the attraction basin of S_m). This way, local minima 'collect' instances originating from the M initial landscape elements, and the number of local minima having 'collected' exactly j of the M elements is denoted by β_j. The method proposed in Garnier and Kallel (2002) tries to utilise the information about the distribution of β_j for a prediction about the total number of local minima within the landscape. The authors associate with the normalised sizes of attraction basins the Gamma distribution, where an approximation of the crucial parameter γ of the density function can be obtained by using (a) a basic equation established in Garnier and Kallel (2002) for linking the Gamma distribution to expected values $\beta_{j,\gamma}$ of sampling data β_j and (b) the χ^2-test with regard to $\beta_{j,\gamma}$

* Corresponding author.
E-mail address: Kathleen.Steinhofel@kcl.ac.uk (K. Steinhöfel).

http://dx.doi.org/10.1016/j.compbiolchem.2015.11.002
1476-9271/

and β_j. The method has been applied in Sahoo and Albrecht (2012) to the approximation of the number of local minima in partial RNA folding landscapes. The application employs a minimisation procedure for the χ^2-test over a square grid for two parameters (γ, r), where γ defines the density function and r is an auxiliary parameter that eventually leads to the required approximation. Since the χ^2-minimisation is running over a square grid, finding suitable (γ, r) is relatively time-consuming. Moreover, as demonstrated in Sahoo and Albrecht (2012), the quality of approximations is affected by the values of $\beta_{j,\gamma}$ for large j (called tail values of $\beta_{j,\gamma}$, where j is close to the maximum j such that $\beta_j > 0$), together with large gaps between non-zero β_j for increasing j. The problem with tail values was observed and highlighted already in Garnier and Kallel (2002), see Section 5.3 therein. In the present paper, both problems, i.e., tail values and χ^2-minimisation, are addressed by a new heuristic that utilises a specific pooling procedure for tail values and substitutes the simultaneous χ^2-minimisation over a square grid by two linear χ^2-tests executed one after another and for γ-approximations only.

The new method is presented in the context of RNA folding landscapes, which is an important research area in Structural Biology. Research on RNA structure prediction has a long history, dating back to early work from around 1960 (Fresco et al., 1960) and further progress published in DeLisi and Crothers (1971) and Tinoco et al. (1971). Most of the computational methods related to RNA structure prediction centres around secondary structures, which are simplified models that admit a representation as outer-planar graphs, and the computation of minimum free energy conformations. The application of dynamic programming to secondary structure prediction, as presented in Nussinov et al. (1978) and Waterman and Smith (1978), eventually resulted – with various refinements and adjustments of energy functions – in the design of manageable polynomial time algorithms and powerful tools for folding simulations and associated program packages, such as `Mfold` by Zuker (1989), Zuker and Sankoff (1984) and `RNAfold` by Hofacker (2003) and Hofacker et al. (1994). In contrast, tertiary structure prediction, i.e., where RNA structures with so-called pseudoknots are included, has been shown to be NP-complete (Lyngsø and Pedersen, 2000). While, in the present paper, we also focus on secondary structures, we are not only interested in minimum free energy (MFE) conformations, but also in meta-stable structures above MFE conformations. Meta-stable structures have become the subject of recent research due to new insights into co-transcriptional folding and interactions between different types of RNA sequences.

The problem of calculating meta-stable RNA secondary structures (local minima) is considered, for example, in Flamm et al. (2002), Lorenz and Clote (2011) and Saffarian et al. (2012). The `RNAsubopt` tool by Wuchty et al. (1999) together with the `barriers` program (Flamm et al., 2002) allows the user, in principle, to identify all meta-stable conformations within an energy range ΔE above the MFE conformation. However, due to the rapid increase of conformations with increasing ΔE, the approach is only applicable to short sequences or small values of ΔE.

Lorenz and Clote (2011) describe the extension of computing the partition function over the set of locally optimal structures in RNA energy landscapes to the `RNAlocopt` tool that computes the exact, total number of locally optimal structures and the exact partition function for locally optimal structures, along with the capability of identifying the set of local minima. The underlying energy model is the Turner nearest neighbour model (Xia et al., 1998) without dangles. The algorithm is an extension of McCaskill's algorithm (McCaskill, 1990), where locally optimal structures are accounted for by additional terms in the recursion scheme. Based upon dynamic programming, `RNAlocopt` computes the total number of locally optimal structures in $O(n^3)$ time, and the associated sampling of secondary structures takes $O(n^2)$ time, which is comparable to other tools. We note that in the present paper, we target partial RNA landscapes, i.e., the information about the total number of locally optimal structures for a given RNA sequence is not directly applicable. `RNAlocopt` has been recently modified to support Turner 2004 energy parameters. A detailed and sophisticated combinatorial analysis of the asymptotic behaviour of the number of saturated RNA secondary structures and locally optimal RNA secondary structures is carried out in Clote (2006) and Fusy and Clote (2014) for a variety of energy models and structural constraints.

Saffarian et al. (2012) present an algorithm for generating all locally optimal secondary structures assembled from a set of thermodynamically stable helices. The construction of locally optimal structures is divided into two steps: First juxtaposed base pair are processed, followed by nested (within juxtaposed positions) base pairs. The main step of the construction follows a recurrent relation which reminds of dynamic programming, although an estimation of worst case or expected run-time is not provided. Each element of the intermediate set of secondary structures is then further processed in order to generate locally optimal structures. The procedure is extended to the generation of locally optimal secondary structures from a given set of thermodynamically stable helices and computational experiments for six sequences of length up to 405 nt are presented.

Kucharík et al. (2014) introduce basin hopping graphs as a new connectivity model of attraction basins within energy landscapes. Vertices represent local minima and edges connect vertices if the transition between the corresponding basins is energetically optimal in terms of the associated saddle point height. The authors present the two new tools `RNAlocmin` and `BHGbuilder` as basic implements for the approximation of basin hopping graphs. The tool `RNAlocmin` executes a modified Boltzmann sampling in order to generate sets of local minima. The modification of Boltzmann sampling tries to avoid oversampling of structures close to MFE conformation by using a parameterised thermodynamic temperature in Boltzmann weights. The `BHGbuilder` tool establishes the basin hopping graph by using a heuristic path-finding algorithm between the local minima identified by `RNAlocmin`. The authors present various comparisons to `RNAlocopt` from Lorenz and Clote (2011) regarding the coverage of local minima within a given time frame, which turn out to be in favour of `RNAlocmin`, partly with large differences in the number of detected local minima. However, one could argue that raising the temperature when running `RNAlocopt` could have increased the number of distinct locally optimal structures, which could affect the comparison to `RNAlocmin`. Kucharík et al. (2014) estimate the applicability of the overall approach to a range of RNA sequence lengths bounded by about 300 nt.

In the present paper, we aim at improved approximations of the number ν of local minima in partial RNA folding spaces. The approximation of ν can then be used for evaluating the outcome of procedures searching for local minima as presented in Kucharík et al. (2014). *A priori* knowledge about approximations of ν provides information about the distance of the current number of identified local minima to the true number of meta-stable conformations. In the present application, the number of instances as well as energy parameters of partial landscapes above minimum free energy conformations are chosen in such a way that a comparison to data generated by `RNAsubopt` and `barriers` is computationally feasible. `RNAsubopt` and `barriers` are used to calculate all secondary structures and the true set of local minima, respectively, within the partial energy landscapes. The time complexity of our heuristic method can be estimated by $O\left(n^2 E_n D \max\{M, \nu\}\right)$,