Cycle decomposition of small RNA configuration space

JING QIN[1,2] AND PETER F. STADLER[1–5]

[1]Max Planck Institute for Mathematics in the Sciences, Inselstraße 22, D-04103 Leipzig, Germany
[2]Bioinformatics Group, Department of Computer Science, and Interdisciplinary Center for Bioinformatics, University of Leipzig, Härtelstrasse 16-18, D-04107 Leipzig, Germany.
[3]RNomics Group, Fraunhofer Institut for Cell Therapy and Immunology, Perlickstraße 1,D-04103 Leipzig, Germany
[4]Inst. f. Theoretical Chemistry, University of Vienna, Währingerstrasse 17, A-1090 Vienna, Austria
[5]The Santa Fe Institute, 1399 Hyde Park Rd., Santa Fe, New Mexico

Abstract It is a well known fact that the behavior of simulated annealing algorithms is tightly related to the hierarchical decomposition of their configuration spaces in cycles. We here apply the iterative routine invented by Wentzell and Freidlin [1] to construct the cycle decomposition of small RNA configuration spaces, for instance, hairpins. We furthermore explore the relationships of cycles and the barrier tree of the energy landscape.

RNA Configuration Space

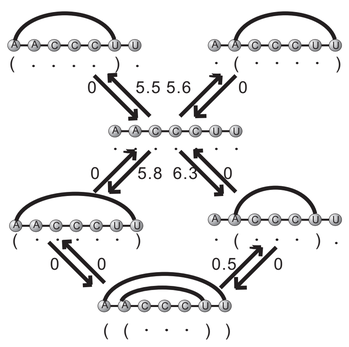
Given an RNA sequence $R = R_1R_2 \dots R_n$, where $R_i \in \{\mathbf{A}, \mathbf{U}, \mathbf{G}, \mathbf{C}\}$, a secondary structure is a graph on n vertices with three properties:

- R_i and R_{i+1} are connected for any $1 \le i \le n-1$;
- each vertex R_i can be paired to at most one other vertex R_j (exclude the arcs exist in primary structure) if $R_iR_j \in \mathsf{A} = \{\mathbf{AU}, \mathbf{UA}, \mathbf{GC}, \mathbf{CG}, \mathbf{GU}, \mathbf{UG}\}$;
- if both $R_i < R_j$ and $R_h < R_\ell$ are paired, then $i < h < j$ implies that $i < \ell < j$.

Given an RNA sequence, the configuration space (Landscape) can be viewed as a directed graph $G(V, E)$. In which, the vertex set V is formed by all the possible secondary structures with respect to the RNA sequence and we say $i \to j \in E$ if j can be obtained from i by either adding or removing an arc in i. Let $F(i)$ denote the free energy of the secondary structure i. The weight of a directed edge $i \to j$, denoted by $w(i \to j)$, is defined by

$$w(i \to j) = \begin{cases} 0 & \text{if } F(i) > F(j); \\ F(j) - F(i) & \text{otherwise.} \end{cases} \quad (1)$$

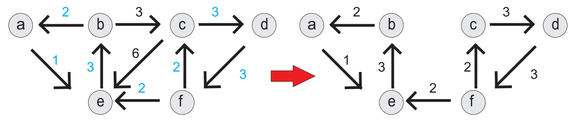
Example: A toy configuration space for AACCCUU



Exit Graph and Strongly Connected Component

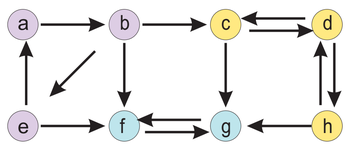
Given a directed graph $G = (V, E)$, the exit graph of G, denoted by $H = (V_H, E_H)$, is a subgraph of G, such that $V_H = V$ and $E_H = \{i \to j | C(i, j) = inf_{i \to k \in E} C(i, k)\}$.

Example: Construction of exit graph



A directed graph is called strongly connected if there is a path from each vertex in the graph to every other vertex. The strongly connected components of a directed graph G are its maximal strongly connected subgraphs with respect to set inclusion.

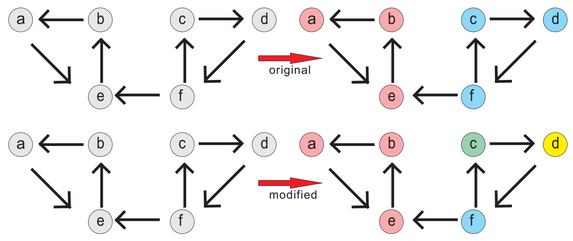
Example: Graph with strongly connected components colored in different colors



Tarjan-V: slight modification of Tarjan's Algorithm

After the exit graph $H = (V_H, E_H)$ of G is constructed, Tarjan's algorithm can be used to search for the strongly connected components in H with complexity $O(|V_H| + |E_H|)$. But here, a slight modification is we have to distinguish the "leaf" strong connected component from the others.

Example: Difference between original and modified Tarjan's algorithm



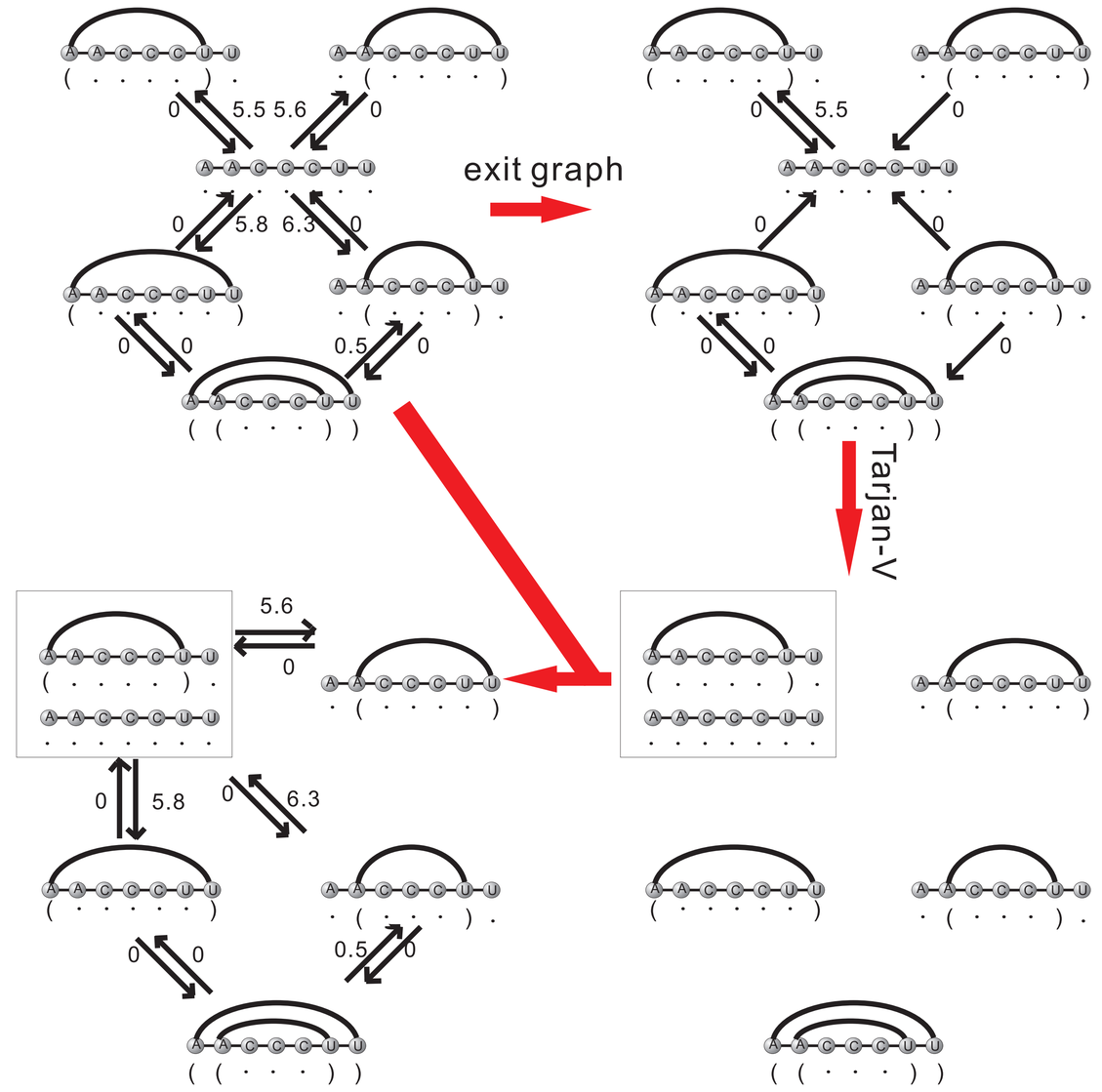
Cycle decomposition Procedure

Given the RNA sequence, the cycle decomposition in cycles is realized in an iterative way:

- Initialize $C^0(V^0, E^0)$ which is identified with the original RNA configuration space;
- Assume all the information of $C^k(V^k, E^k)$ has been derived. We construct $C^{k+1}(V^{k+1}, E^{k+1})$ as follows. Firstly, we build the exit graph of C^k and derive the components according to the Tarjan-V procedure. By contracting the vertices of C^k within the same component, we obtain the vertex-set of C^{k+1}, i.e. V^{k+1}. We say the C^k-vertices $\{v_1^k, v_2^k, \dots\}$ contracted to a single V^{k+1}-vertex m are the son-structures of m, denoted by $m \downarrow v_i^k$.
  To define E^{k+1}, given two V^{k+1}-vertices p and q, p are directed to q if and only if there at least exist one pair of V^k-vertices v_1 and v_2 such that $p \downarrow v_1$, $q \downarrow v_2$ and $v_1 \to v_2 \in E^k$. The weight of the directed edge $p \to q$, denoted by $w^{k+1}(p \to q)$ is assigned by Eqn. (2).
  $$w^{k+1}(p \to q) = \mathbf{inf}\{H_m^{k+1}(p) + w^k(v_1, v_2) - H_e^k(v_2)\}. \quad (2)$$
  In which, $H_e^s(v)$ and $H_m^s(v)$ denote the so-called escape energy and mixing energy respectively given as follows:
  $$H_e^s(v) = \mathbf{inf}\{w^s(v \to w) | v \to w \in E^s, v, w \in V^s\} \quad (3)$$
  $$H_m^s(v) = \mathbf{sup}\{H_e^{s-1}(w) | v \downarrow w, w \in V^{s-1}\}. \quad (4)$$
- The whole procedure terminates in case of $|V^{k+1}| = 1$, iterates otherwise.

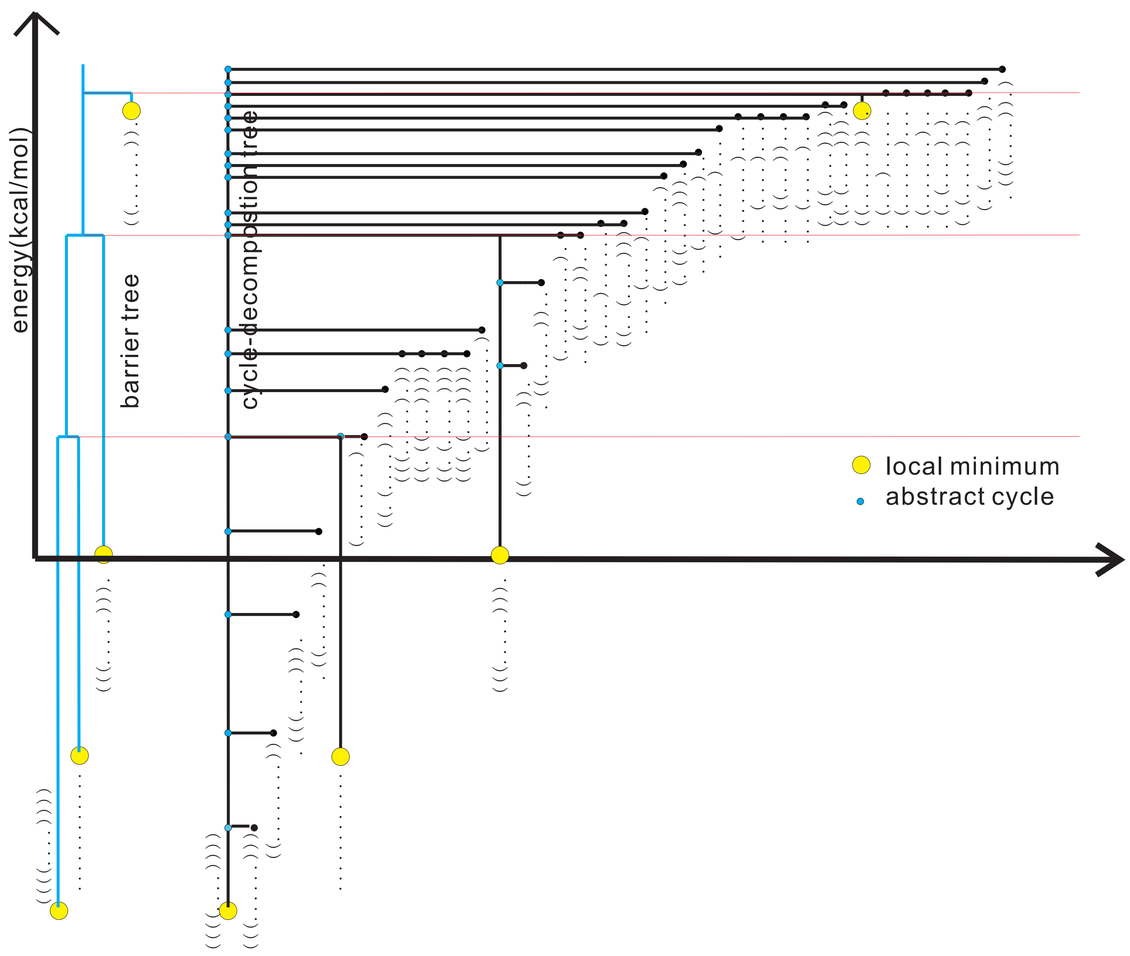
Cycle decomposition Procedure

Example: Towards Level-1 from Level-0 for AACCCUU



Computational Result

Example: Barrier tree and cycle decomposition tree for sequence GGAAUAAUUCC



Remarks:

- In the worst case, the complexity of the cycle decomposition algorithm is $O(|V|(|V| + |E|))$. In which, $O(|V| + |E|)$ is the complexity of the (modified) Tarjan's algorithm and the other factor, $|V|$, caused by (in the worst case) the number of iterations, denoted by σ, need to run before the procedure is terminated. We remark here, the integer σ is a valuable parameter that also reflects the topology of the landscape itself. For RNA configuration space, interestingly, we observe that the bigger the size of the space, the smaller the ratio $\frac{\sigma}{|V|}$ is.
- By construction, the mixing energy of the cycle (firstly) merging two local minima identified with their saddle height obtained by the `RNAbarrier`[2] included in the Vienna package. I.e. the barrier tree can be viewed as a subtree of the cycle decomposition tree. The `RNAbarrier` is based on the flooding-algorithm which need more effort in case of the degenerate RNA landscape. Comparing with `RNAbarrier`, cycle decomposition by construction avoid the problem may caused by multiple saddle points. Also, we note here the mixing energy is only one of the valuable parameters that can be read from cycle decomposition procedure.

References

[1] M.I. Freidlin and A.D. Wentzell, Perturbations of Stochastic Dynamic Systems, Springer-Verlag, 1984.

[2] C. Flamm, I.L. Hofacker, P.F. Stadler and M.T. Wolfinger, Barrier Trees of Degenerate Landscapes, Z. Phys. Chem., 2002 (216):155–173.

Gedruckt im Universitätsrechenzentrum Leipzig