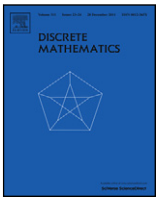
Discrete Mathematics 312 (2012) 100–112

Contents lists available at SciVerse ScienceDirect

Discrete Mathematics

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



Symmetric circular matchings and RNA folding

Ivo L. Hofacker [a], Christian M. Reidys [b,c,d], Peter F. Stadler [e,f,g,a,h]

[a] *Department of Theoretical Chemistry, University of Vienna, Währingerstraße 17, A-1090 Wien, Austria*
[b] *Center for Combinatorics, LPMC-TJKLC, Nankai University, Tianjin 300071, PR China*
[c] *College of Life Science, Nankai University, Tianjin 300071, PR China*
[d] *Department of Mathematics & Computer Science, University of Southern Denmark, Campusvej 55, DK-5230 Odense M, Denmark*
[e] *Bioinformatics Group, Department of Computer Science, and Interdisciplinary Center for Bioinformatics, University of Leipzig, Härtelstraße 16-18, D-04107 Leipzig, Germany*
[f] *Max Planck Institute for Mathematics in the Sciences, Inselstrasse 22, D-04103 Leipzig, Germany*
[g] *Fraunhofer Institut für Zelltherapie und Immunologie—IZI Perlickstraße 1, D-04103 Leipzig, Germany*
[h] *Santa Fe Institute, 1399 Hyde Park Rd., Santa Fe, NM 87501, USA*

ARTICLE INFO

Article history:
Available online 6 July 2011

Keywords:
Circular RNA folding
RNA co-folding
Symmetry correction
Matching problems

ABSTRACT

RNA secondary structures can be computed as optimal solutions of certain circular matching problems. An accurate treatment of this energy minimization problem has to account for the small — but non-negligible — entropic destabilization of secondary structures with non-trivial automorphisms. Such intrinsic symmetries are typically excluded from algorithmic approaches; however, because the effects are small, they play a role only for RNAs with symmetries at sequence level, and they appear only in particular settings that are less frequently used in practical application, such as circular folding or the co-folding of two or more identical RNAs. Here, we show that the RNA folding problem with symmetry terms can still be solved with polynomial-time algorithms. Empirically, the fraction of symmetric ground state structures decreases with chain length, so that the error introduced by neglecting the symmetry terms affects fewer and fewer predictions. We then explore the combinatorics of symmetric secondary structures in detail. Surprisingly, the singularities of the generating function coincide between symmetric and non-symmetric structures. Furthermore, generating functions and explicit asymptotic results for both the circular and the co-folding version are derived.

1. Introduction

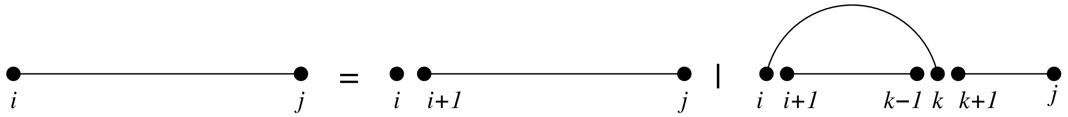
Let $G(V, E)$ be a simple finite graph. A matching M is a subset of E such that no two edges $e', e'' \in M$ are incident to the same vertex. Suppose there is a fixed natural order of the vertex set so that we can label them with integers $1 \dots n = |V|$. We say that two edges $e_1 = \{v_1', v_1''\}$ and $e_2 = \{v_2', v_2''\}$ cross if the corresponding intervals overlap, i.e., $[v_1', v_1''] \cap [v_2', v_2''] \notin \{\emptyset, [v_1', v_1''], [v_2', v_2'']\}$. A matching is *circular* if it does not contain a pair of crossing edges.

Circular matchings model the (pseudo-knot free) secondary structures of nucleic acids, i.e., RNA and DNA, in a natural way [18,22]. Here, the nucleotide sequence $(x_1, x_n, \dots, x_n)$, with $x_i \in \{A, U, G, C\}$ for RNA and $x_i \in \{A, T, G, C\}$ for DNA provides a vertex labeling. Edges are restricted to pairs of vertices that satisfy the chemical pairing rules of nucleic acids: $\{u, v\} \in E$ if and only if $\{x_u, x_v\} \in \mathfrak{B}$. The set of allowed pairs are $\mathfrak{B}_{\mathrm{RNA}} = \{\{A, U\}, \{G, C\}, \{G, U\}\}$ and $\mathfrak{B}_{\mathrm{DNA}} = \{\{A, T\}, \{G, C\}\}$, respectively.

E-mail addresses: ivo@tbi.univie.ac.at (I.L. Hofacker), duck@santafe.edu (C.M. Reidys), studla@bioinf.uni-leipzig.de, Peter.Stadler@bioinf.uni-leipzig.de (P.F. Stadler).

0012-365X/$ – see front matter © 2011 Elsevier B.V. All rights reserved.
doi:10.1016/j.disc.2011.06.004

This circular matching problem is solved by a simple recursion that is based on the observation, that every matching edge (base pair) divides the graph into two disjoint subgraphs with independent solutions:



(1)

Hence, the maximum number F of edges in a circular matching satisfies the recursion

$$F_{ij} = \max\left\{F_{i+1,j},\ \max_{\substack{k\geq i+m+1\\ \{i,k\}\in E(G)}} (F_{i+1,k-1} + F_{k+1,j} + 1)\right\} \tag{2}$$

starting from the initializations $F_{i,j} = 0$ for $j - i < m$ [18,22]. The parameter m measures the minimum number of sequence position that are located "inside" a base pair. Based on biophysical considerations, one usually sets $m = 3$ in the context of RNA. Eq. (2) immediately translates into a recursion for the number of all possible secondary structures (i.e, assuming that $G = K_n$, i.e. a complete graph):

$$s(n) = s(n-1) + \sum_{k=m}^{n-2} s(k)s(n-k-2) \tag{3}$$

with $s(n) = 0$ for $n < 0$ and $s(n) = 1$ for $0 \leq n \leq m + 1$. For $m = 0$, $s(n)$ coincides with the Catalan numbers [3]. Combinatorial problems motivated by RNA folding problems have received considerable attention over the past three decades, see e.g. [12,20,19,10,17,5,13,4]. We shall return to the combinatorial aspects in Section 3.

In contrast to the usual setting of matchings the weight (energy) associated with a particular matching M, i.e., a particular secondary structure, is not just the sum of its edges in the context of nucleic acid structures. Instead, the energy of a secondary structure is defined in terms of so-called "loops". Laying out V on a cycle in the given order and connecting consecutive vertices by additional "backbone" edges yields an outerplanar graph. The internal faces of this embedding are called "loops" in the RNA folding literature. Each face is assigned an energy contribution that depends on the number of vertices, the nucleotides (vertex labels), and the base pairs (i.e., matching edges).

Secondary structures are coarse-grained representations of the molecular structures that can be interpreted as equivalence classes of the actual spatial conformations of the molecule. The energy of the secondary structure therefore contains an entropic contribution which corresponds, according to Boltzmann's famous formula $S = R \ln \Omega$, to the diversity Ω of atomic-resolution states that are subsumed in a given secondary structure. The corresponding entropic contributions to the energy model are obtained experimentally from the melting properties of small RNA molecules [15]. These measurements are performed on homogeneous samples of linear RNA molecules. Since RNA sequences have a defined reading direction (from their 5' to their 3' ends), these molecules have no (non-trivial) symmetries.

Interactions of multiple RNA molecules as well as the structure formation of circular RNA molecules can be treated within the same model. Structures formed by two or more distinct RNA strands A, B, etc., can be dealt with by concatenating the sequences $A\$B\$\ldots Z\$$, where the sentinel character \$ is used to mark the concatenation points. For more than two strands all concatenation orders have to be considered. Formally, this leads to the same problem as folding a circular RNA sequence. The only difference is that loops that contain the \$-characters are assigned special energy contributions. In contrast to linear nucleic acids, these cyclic arrangements can have non-trivial symmetries: In fact, circular sequences have a rotational symmetry C_k if they consist of k concatenated identical copies of the same string A. Therefore, they can also form secondary structures with non-trivial symmetry. Symmetries reduce the number of *physically distinct conformations* that belong to a given secondary structure ψ. This reduction in the number of conformations is determined by the length ℓ_ψ of its orbit. Since the symmetry effect is not included in the individual energy contributions, the symmetry correction of the form

$$\varepsilon_{\text{sym}}(\psi) = RT \ln \ell_\psi \tag{4}$$

needs to be added to the standard energy model.

In practice, the effect is small and folding problems with symmetric sequences are rare. The correction (4) thus is typically neglected [9,1]. In cases where precise energies are required, one usually considers the full ensemble of Boltzmann-weighted secondary structures and computes the partition function over all secondary structures. Surprisingly, the symmetry effect is not a problem in this context since the overcounting of symmetric structures cancels exactly with an undercounting inherent in the algorithm; we refer to [2,6] for details.

From a theoretical point of view, on the other hand, there is no *a priori* relationship between the energy contributions for different structural elements and the symmetry correction. In order to properly account for the symmetries, therefore, it is necessary to account separately for secondary structures with different symmetries. At the same time, it appears natural to consider the enumerative combinatorics of secondary structures with symmetries. From a practical point of view, finally, one may ask to what extent minimum energy secondary structures of symmetric sequences are symmetric themselves, and thus how often neglecting the symmetry correction leads to incorrect results.