Counting Distinguishable Secondary Structures for Multiple Sequences

Masaru Nakajima Andrew D. Smith

October 18, 2021

Abstract

Secondary structures for nucleic acid sequences, DNA and RNA, are useful abstractions when examining certain behaviors of those molecules. We examine the problem of counting secondary structures compatible with an ordered multiset of sequences. In particular, we address the issue of accounting for indistinguishable secondary structures, for which no fast algorithm has been found. We provide a cubic time algorithm for counting distinguishable secondary structures.

1 Introduction

Nucleic acids are fundamental physical components of all life. DNA serves to store and transmit genetic information, and allows for evolution. RNA is involved in a diverse cellular processes, including the well known role of messenger RNA in gene expression, with additional roles being discovered continually [**?**]. Functions and behaviors of nucleic acids depend on their physical conformation in the 3D space, which is called tertiary structure. However, tertiary structures of nucleic acid sequences are difficult to study, experimentally and computationally, as the size of the molecules (*i.e.* the number of bases in a chain) increases. An effective alternative has been to study the secondary structure, a set of paired bases, of nucleic acid sequences. Although secondary structure is a dramatically simplified representation, it still allows for accurate and useful predictions, including equilibrium thermodynamic behaviors of nucleic acid strands [**?**].

The mathematical study of nucleic acid secondary structures, usually emphasizing RNA, has seen many landmark results over the past fifty years. Algorithms to identify optimal structures for given RNA sequences, i.e. to "fold" RNA, focused on maximizing the number of paired bases [**?**] and eventually on finding minimum free energy secondary structures [**?**, **?**]. These algorithms were capable of incorporating empirically determined energy parameters and used dynamic programming to obtain globally optimal solutions [**?**]. It has long been recognized that secondary structures are dynamic, and that no single secondary structure provides an ideal model for a single stranded RNA or DNA molecule. The breakthrough algorithm of McCaskill [**?**] allowed for efficient computation of the partition function in the space of secondary structures, and enabled a full characterization of equilibrium secondary structural features in RNA sequences. Quantities derived from the partition function algorithms can be used, for example, to estimate the probability that a given base in a sequence would be paired.

Analysis of secondary structures for multiple interacting nucleic acid sequences is receiving increased attention, largely driven by applications in biotechnology [**?**]. The work of Dirks et al. (2007) remains the most comprehensive elucidation of the partition function problem for multiple interacting single-stranded oligonucleotides [**?**]. The authors frame the multi-sequence partition function problem in terms of a particular data representation consisting of a multiset of sequences under a fixed circular order. Dirks and

colleagues gave a cubic time algorithm for computing the partition function under this representation. This achievement is notable because it overcomes a challenge not seen in secondary structure analysis for individual sequences. Specifically, when multiple sequences are involved, two different secondary structures may be indistinguishable with respect to measurable quantities related to free energy. Such indistinguishable secondary structures emerge when there are symmetries in the underlying ordered sequences, which allows for symmetries in the associated secondary structures. Unfortunately, such symmetries are only apparent at a global scale, while the usual dynamic programming strategies operate on locally defined sub-problems. This leads to an issue of "over-counting," with equivalent secondary structures contributing multiple times. The solution by Dirks and colleagues circumvents this issue by jointly leveraging insights from physical and algorithmic aspects of the problem, leading to an essential algorithm in thermodynamic analysis of multi-strand nucleic acid sequences [**?**].

Dirks et al. (2007) raised the question of whether a correction for algorithmic over-counting can be separated from symmetry-dependent correction in the free energy of multi-sequence secondary structures. This motivated us to examine the following combinatoric problem: with the same data representation and assumptions, can we efficiently enumerate distinguishable secondary structures? We answer in the affirmative. In this manuscript we present an algorithm to count distinguishable secondary structures for circularly ordered multisets of sequences. The key components of our solution include a decomposition of the possible secondary structures according to the symmetries of each structure and a novel abstraction to represent only secondary structures having particular symmetries. With this foundation we present a cubic time algorithm to count distinguishable secondary structures compatible with an ordered multiset of sequences.

2 Background and notation

2.1 RNA secondary structure for one sequence

Given a natural number n, a secondary structure s is a set of unordered pairs $\{i, j\}$, with $i, j \in [0, n-1]$. We remark that we use the terminology of "pairs" to indicate that for any $\{i, j\} \in s$, we have $i \neq j$ and thus $|\{i, j\}| = 2$. We keep this assumption throughout the manuscript. To work with physically relevant secondary structures, we define *valid* secondary structures as those which satisfy the following conditions:

1. For all $\{i, j\}$ in s, $|j - i| > 1$.
2. For all $\{i, j\}$ and $\{k, l\}$ in s, if $\{i, j\} \neq \{k, l\}$, then $\{i, j\} \cap \{k, l\} = \emptyset$.
3. For all $\{i, j\}$ and $\{k, l\}$ in s, if $i < k < j$, then $i < l < j$.

The first condition reflects the steric constraint preventing pairing between sites that are adjacent. In reality, such a constraint should prevent pairs unless they are separated by several sites, for example requiring $|j - i| > 4$; we use 1 for simplicity, and note that algorithmic considerations (correctness or complexity) are not impacted by this choice. The second condition ensures that no integer $0 \leq i < n$ is involved in more than one pair. The third condition prohibits any two pairs $\{i, j\}$ and $\{k, l\}$ from having the relationship $i < k < j < l$. If this relationship is satisfied, we say the secondary structure has a pseudoknot, and we consider such structures invalid. If two pairs $\{i, j\}$ and $\{k, l\}$ cause a secondary structure to violate the third condition, we say $\{i, j\}$ and $\{k, l\}$ induce a pseudoknot. Graphical representations of some example secondary structures are shown in Figure **??**.

An RNA sequence ϕ is a string Σ^*_{RNA}, where $\Sigma_{\mathrm{RNA}} = \{\mathrm{A}, \mathrm{C}, \mathrm{G}, \mathrm{U}\}$. When considering a sequence ϕ we will use n to denote the length of ϕ. For convenience, we use zero-based indexing, so

$$\phi = \phi_0 \phi_1 \cdots \phi_{n-1}, \text{ with } \phi_i \in \Sigma_{\mathrm{RNA}}.$$

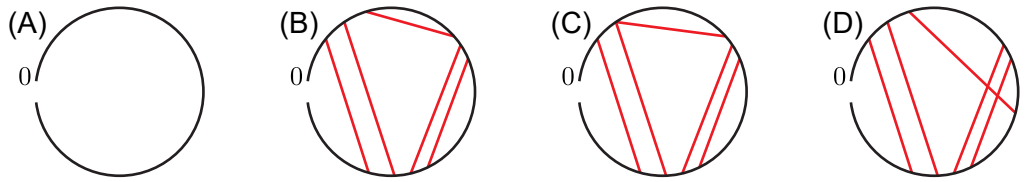

Figure 1: Example secondary structures. The sequence of length n is drawn in a circular fashion. A solid red line represents a pair. (A) The secondary structure without any pair is valid. (B) An example of a valid secondary structure with pairs. (C) An example of an invalid secondary structure that violates the second condition. (D) An example of an invalid secondary structure that violates the third condition. Note that two pairs inducing a pseudoknot result in crossing lines.

We adopt the convention of indexing ϕ in the order of the 5' end to 3' end of the RNA molecule. We will refer to each position in an RNA sequence as a site and at site i the value of ϕ_i is called a base. For a fixed n, consider a secondary structure s consisting of pairs $\{i, j\}$ with $0 \leq i, j < n$. We say that s is *compatible* with a length-n sequence ϕ if, and only if, s is valid and, for every $\{i, j\} \in s$, the bases ϕ_i and ϕ_j at those sites conform to $\{\mathrm{A}, \mathrm{U}\}$, $\{\mathrm{C}, \mathrm{G}\}$ or $\{\mathrm{G}, \mathrm{U}\}$, in which case we say that ϕ_i and ϕ_j form a *valid base pair*. The problem of counting the number of secondary structures for a given RNA sequence is defined as follows. For an RNA sequence $\phi \in \Sigma^*_{\mathrm{RNA}}$, how many secondary structures are compatible with ϕ?

2.2 Counting RNA secondary structures

The problem of counting secondary structures compatible with a given ϕ can be solved by a dynamic programming algorithm [**?**]. We first define a quantity $S(i, j)$, which corresponds to the number of secondary structures compatible with the substring $\phi[i\,..\,j]$. For $i, j \in [0\,..\,n-1]$, the quantity $S(i, j)$ is recursively defined as

$$S(i,j) = \begin{cases} 1 & \text{if } i \geq j, \\ S(i+1,j) + \sum_{i<k\leq j} p(i,k) S(i+1,k-1) S(k+1,j) & \text{otherwise.} \end{cases} \tag{1}$$

where the indicator function $p(i, j)$ is 1 if $|j - i| > 1$ and ϕ_i and ϕ_j form a valid base pair, and 0 otherwise. For $i < j$, for any secondary structure compatible with $\phi[i\,..\,j]$, site i is either not involved in any base pair, or it is involved in exactly one pair. In the first case, the number of such secondary structures is same as that of secondary structures compatible with $\phi[i+1\,..\,j]$. This value is given by $S(i+1, j)$ and corresponds to the first term of equation (**??**). In the second case, site i can potentially be paired with k with $i < k \leq j$, the validity of which is indicated by $p(i, k)$. For any secondary structure compatible with $\phi[i\,..\,j]$ and containing the pair $\{i, k\}$, any other pair $\{i', j'\}$ in the same secondary structure satisfies

$$i < i' < j' < k \;\text{ or }\; k < i' < j' < j.$$

Otherwise, the pairs $\{i, k\}$ and $\{i', j'\}$ would induce a pseudoknot, and the secondary structure would violate the third condition. This implies that the number of secondary structures compatible with $\phi[i\,..\,j]$ and containing pair $\{i, k\}$ is the product of the number of secondary structures compatible with $\phi[i+1\,..\,k-1]$ and that with $\phi[k+1\,..\,j]$. This product, $S(i+1, k-1)S(k+1, j)$, can be seen within the summation of equation (**??**). These recurrences can be evaluated by dynamic programming, as shown in Algorithm **??**.

Algorithm 1 Counting the compatible secondary structures

Input: An RNA sequence ϕ of length n

Output: The number of secondary structures compatible with ϕ

1: $S(i, j) \leftarrow 1$ for all $i, j \in [0\,..\,n-1]$
2: **for** $j \leftarrow 0$ **to** $n-1$ **do**
3: **for** $i \leftarrow j-1$ **down to** 0 **do**
4: $S(i, j) \leftarrow S(i+1, j)$
5: **for** $k \leftarrow i+1$ **to** j **do**
6: $S(i, j) \leftarrow S(i, j) + p(i, k)S(i+1, k-1)S(k+1, j)$
7: **return** $S(0, n-1)$

In that pseudocode, the most deeply nested calculations require constant time, and with three nested loops the algorithm has time complexity $\mathcal{O}(n^3)$. As values for each $S(i, j)$ must be stored for use in subsequent iterations, the space complexity of this algorithm is $\mathcal{O}(n^2)$.

3 Secondary structure for multiple sequences