Boston College

Doctoral Thesis

On the use of Coarse Grained Thermodynamic Landscapes to Efficiently Estimate Kinetic Pathways for RNA Molecules

Author:
Evan Senter

Supervisor:
Dr. Peter Clote

A thesis submitted in fulfilment of the requirements for the degree of Doctor of Philosophy

in the

Clote Lab
Department of Biology

Tuesday 30th June, 2015

Contents

Chapter 1

Ribofinder

1.1 Introduction

In this chapter, we present the `Ribofinder` program—a pipeline to facilitate the detection of putative guanine riboswitchesacross genomic data. The `Ribofinder` tool operates in three stages. First we use `Infernal` and `TransTermHP` to detect putative aptamers and expression platforms, two distinct components of riboswitchesdescribed in section 1.2. After coalescing this data into a pool of candidate riboswitches, we use `RNAfold` with constraints based on experimental data to compute the two distinct structural conformations—'gene on' and 'gene off'. In the third and final stage, we leverage `FoldAlign` to measure the similarity between our candidate pool and a canonical guanine riboswitchwell studied in the literature, the xpt G-box riboswitchfrom Bacillus subtilis.

1.1.1 Organization

This chapter is organized in the following fashion. After providing background on the structural components of a riboswitchalongside their biological significance, we outline the deficiencies in the 'state of the art' software when as it relates specifically to riboswitchdetection. We then move on to outline the three stages of `Ribofinder`: candidate selection, structural prediction, and candidate curation. Having described the approach of the software, we move on to present our findings in using `Ribofinder` to detect guanine riboswitchesacross the bacterial RefSeq database. Finally, we provide brief commentary on possible extensions of the algorithm to locate other flavors of riboswitches, of which adenine-sensitive aptamers are a straightforward extension.

1.2 Background

1.3 The Ribofinder pipeline

1.3.1 Step 1: Candidate selection

1.3.2 Step 2: Structural prediction

1.3.3 Step 3: Candidate curation

1.4 Using Ribofinder against the RefSeq database

1.5 Extending beyond guanine riboswitches

Chapter 2

FFTbor

2.1 Introduction

In this chapter, we present the `FFTbor` algorithm and accompanying software. `FFTbor` is a novel algorithm developed with the intent of efficiently computing the Boltzmann probability of those structures whom, for a given input RNA sequence $\mathtt{s}$, differ by k base pairs. By leveraging polynomial interpolation via the Fast Fourier Transform, this algorithm runs in $O(n^4)$ time and $O(n^2)$ space, a significant improvement over its predecessor. The accompanying software which implements this algorithm has been used to predict the location of expression platforms for putative riboswitchesin genomic data, and to evaluate the correlation between kinetic folding speed and landscape ruggedness.

2.1.1 Organization

This chapter is organized in the following fashion. First, we provide background on the problem which `FFTbor` aims to address, as well as a brief overview of existing approaches. We follow by a formal explanation of the problem, and proceed to describe how the energy landscape is coarsified into discrete bins. We then develop the recursions for the parameterized partition function using the Nussinov-Jacobson energy model, which allows us to exposé the novel aspects of the algorithm. After developing the recursions, we indicate how they can be reformulated as a polynomial whose coefficients $c_k = \mathbf{Z}^k_{1,n}$. We then describe how the Fast Fourier Transformcan be employed to efficiently compute the coefficients c_k, finishing our description of the underlying algorithm. Then we proceed to present two applications of `FFTbor`, an application to RNA kinetics and another to riboswitchdetection in genomic data. Finally, we give reference to the full recursions using the more accurate Turner energy model, which the underlying implementation actually uses.

2.2 Background

2.3 Formalization of the problem

FFTbor aims to compute the coefficients $p_0, \dots, p_{n-1}$ of the polynomial

$$p(x) = p_0 + p_1 x + p_2 x^2 + \cdots + p_{n-1} x^{n-1}, \tag{2.1}$$

where p_k is defined as $p_k = \frac{Z_k}{Z}$. We employ the Fast Fourier Transformto compute the inverse Discrete Fourier Transformon values $y_0, \dots, y_{n-1}$, where $y_k = p(\omega^k)$ and $\omega = e^{2\pi i/n}$ is the principal nth complex root of unity and $p(x)$ is defined in (2.1). By leveraging complex roots of unity in conjunction with the inverse Discrete Fourier Transformthe we subvert numeric instability issues observed with both Lagrange interpolation and Gaussian elimination.

Consider an RNA sequence $\mathbf{s} = s_1, \dots, s_n$, where $s_i \in \{A, C, G, U\}$, i.e. a sequence of nucleotides. We can describe a secondary structure $\mathcal{S}$ which is compatible with $\mathbf{s}$ as a collection of base pair tuples (i, j), where $1 \leq i \leq i + \theta < j \leq n$ and $\theta \geq 0$—generally taken to be 3, the minimum number of unpaired bases in a hairpin loop due to steric constraints.

To more simply develop the underlying recursions for **FFTbor**, we introduce a number of constraints on the base pairs within $\mathcal{S}$. Firstly, we require that each base pair is either a Watson-Crick or G-U wobble, i.e. base pair (i, j) for sequence $\mathbf{s}$ has corresponding nucleotides (s_i, s_j), which are restricted to the set $\mathbb{B} = \{(A, U), (U, A), (G, C), (C, G), (G, U), (U, G)\}$. With this constraint satisfied we say that $\mathcal{S}$ is *compatible* with $\mathbf{s}$, and for the remainder of this chapter will only consider those structures which are compatible with $\mathbf{s}$. Secondly, we insist that given two base pairs $(i, j), (x, y)$ from $\mathcal{S}$, $i = x \iff j = y$—bases have at most one partner. Finally, we require that $i < x < j \iff i < y < j$, no pseudoknots are allowed. While pseudoknots have been shown to be present in some biologically relevant RNAs, their inclusion greatly complicates the recursive decomposition of the structure, and thus it is common to ignore them.

Provided two secondary structures S, T, we can define a notion of distance between them. There are a number of different definitions of distance used across the literature; we will use *base pair distance* for `FFTbor`. Base pair distance is defined as the symmetric difference between the sets S, T

$$d_{BP}(S,T) = |S \cup T| - |S \cap T| \tag{2.2}$$

2.4 Coarsification of the energy landscape

2.5 Recursions for structural neighbors

2.6 Polynomial interpolation to evaluate $\mathcal{Z}_{i,j}(x)$

Chapter 3

FFTbor2D

3.1 Introduction

3.1.1 Organization

This chapter is organized in the following fashion.

Chapter 4

Hermes

4.1 Introduction

4.1.1 Organization

This chapter is organized in the following fashion.