ALBERT LUDWIGS UNIVERISTY OF FREIBURG

MASTER THESIS

Multisite RNA-RNA Interaction Prediction

Yogapriya Ayyanarmoorthy

November 5, 2019

Contents

Chapter 1

Introduction

RNA molecules play important roles in various biological processes.Their regulation and function are mediated by interacting with other molecules. Forming base pairs between two RNAs, called RNA-RNA interactions (RRI). There are fast and reliable single interaction site (S-RRI) prediction tools like IntaRNA, that often show the additional sites within their suboptimal list, ie. are capable of modelling all sites individually but not in a joint prediction. Many RNAs interact via multiple synchronous, non-overlapping subinteractions (M-RRI), e.g. OxyS-fhlA. The simultaneous prediction of both intra- and inter-molecular base pairing allowing for multiple sites is computationally expensive. Some known approaches are IRIS, piRNA, NUPACK. Here we use a S-RRI prediction tool (namely IntaRNA) for the prediction of M-RRI.

1.1 Biological Background of RNA

In this thesis, I will focus on Ribonucleic acids (RNA).

Ribonucleic acid, or RNA is one of the three major biological macromolecules that are important for all known forms of life (along with DNA (deoxyribonucleic acid) and proteins). The central dogma of molecular biology states that the flow of genetic information in a cell is from DNA through RNA to proteins: "DNA makes RNA makes protein". Unlike double-stranded DNA, RNA is a single-stranded molecule
The RNA molecules are represented as a sequence $S \in \{A, C, G, U\}^*$, where A, C, G, U are the respective bases of the nucleotide chain, adenine (A), cytosine (C), guanine (G) and uracil (U). RNAs are classified into two major categories in accordance with their coding potential, that is, coding RNAs and noncoding RNAs. Coding RNAs mostly refers to mRNA that encodes protein to act as various components including enzymes, cell structures, and signal transductors. Noncoding RNAs act as cellular regulators without encoding proteins.
RNA sequences fold into structures which determine the function of an RNA molecule. They are created when bases form base pairs via hydrogen bonds. Due to their high binding strength, the WatsonCrick base pairs $G-C$ and $A-U$ as well as the wobble base pair $G-U$ are considered. Isolated base pairs are usually unstable. When two interacting bases that belong to the same RNA molecule they form *intra*molecular structures and if it belongs to different RNA molecules they form *inter*molecular structures.
Single stranded nucleic acid sequences contain many complementary regions that can form double helices when the molecule folds back onto itself. The resulting pattern of double helical stretches interspersed with loops is called the *Secondary* structure of an RNA.

Formally, an RNA secondary structure P of S is a set of base pairs:

$$P \subseteq \{(i,j) | 1 \leq i < j \leq n,\ Si \text{ and } Sj \text{ complementary } \},$$

where $n = |S|$ and for all $(i,j),(i',j') \in P$:

$$(i = i' \Leftrightarrow j = j') \text{ and } i \neq j'$$

To form a valid secondary structure, base pairs must satisfy several constraints. Let the bases in a sequence be numbered from 1 to N. A base pair may form between positions i and j if the bases are complementary, and if $|j - i| \geq 4$, since there must usually be at least three unpaired bases in a hairpin loop. Let bases k and l form another allowed pair. The pair $k-l$ is said to be compatible with the pair $i-j$ if the two pairs can be present in a structure simultaneously. Pairs are compatible if they are non-overlapping (e.g. $i < j < k < l$) or if one is nested within the other (e.g. $i < k < l < j$). The third case, where the pairs are interlocking (e.g. $i < k < j < l$) is known as a pseudo-knot.Such pairs are assumed to be incompatible for most dynamic programs. An allowed secondary structure is a set of base pairs that are all compatible with each other.
Nested secondary structures can be uniquely decomposed into so called loops or secondary structure elements. Depending on the number of enclosed base pairs (BP) and unpaired bases (UB), different types of secondary structure elements are distinguished.They are hairpin loop, stacking, bulge loop, internal loop, multi loop.
Let S be a fixed sequence. Further, let P be an RNA structure for S.

- a base pair $(i,j) \in P$ is a *hairpin* loop if
  $\forall i < i' \leq j' < j : (i',j') \notin P$.
- a base pair $(i,j) \in P$ is a *stacking* if
  $(i+1, j-1) \in P$
- two base pairs $(i,j) \in P$ and $(i',j') \in P$ form an *internal* loop (i,j,i',j') if
  $i < i' < j' < j$; $(i' - i) + (j - j') > 2$; no base pair (k,l) between (i,j) and (i',j')
- An internal loop is called left (right, resp.) *bulge* if
  $j = j' + 1$ or $i' = i + 1$
- A k-*multiloop* consists of multiple base pairs, $(i_1,j_1)...\ (i_k,j_k) \in P$ with a closing base pair $(j_0, i_{k+1}) \in P$ with the property that
  $\forall 0 \leq l \leq k : (j_l < i_{l'+1})$; $\forall 0 \leq l, l' \leq k$ is true that there is no base pair $(i',j') \in P$ with $i' \in [j_l ... i_{l+1}]$ and $j' \in [j'_l ... i_{l'+1}]$.
- $(i_1,j_1)...(i_k,j_k)$ are called the *helices* of the multiloop.

...It was shown by Howard DeVoe and Ignacio Tinoco Jr (1962) that vertical stackings of bases are by far the largest contribution to RNA helix stability. Thus, the directly neighbored bases are to be taken into account when estimating the stability (energy) contribution of a base or base pair, which yields the *Nearest Neighbor Model* (Tinoco et al., 1973; Borer et al., 1974).
The Nearest Neighbor Model enables the calculation of a free energy estimate for a given RNA secondary structure. The free energy can be interpreted as the amount of energy stored in a system (here an RNA structure) to perform work. A positive term provides energy (e.g. in the form of heat), a negative term can be interpreted the amount of energy necessary to destroy the

system (here to dissolve all base pairs of the RNA structure).
It is possible to define a recursive dynamic programming algorithm to compute the structure that minimizes the energy function, the so called minimum free energy (mfe) structure. This algorithm was introduced by Michael Zuker and Patrick Stiegler (1981).
It can be highly informative to compute the probability of single base pairs instead of single structures. Intuitively, highly probable base pairs frequently occur in highly probable structures. So instead of investigating the probability of complete structures, we can study the probabilities of individual structural elements. The probability of structural elements can be summarised by using base-pair probabilities.
Similarly to base pair probabilities it is possible to use the matrices computed by the McCaskill algorithm to compute the probability $Pr_u[i, j]$ that a certain subsequence i..j is unstructured, i.e. is not involved in any base pair. To this end, we have to consider all possibilities in our loop decomposition of structures that enable such an unstructured subsequence.

- 1.What is RNA
- 2.RNA representation a,c,g,u
- 3.classes of rna
- 4.base pairs of RNA
- 5.RNA secondary structure
- 6.types of rna secondary structure
- 7.nearest neighbor model
- 8.unpair probabilities

1.2 RNA-RNA Interaction

Computational prediction of RNA-RNA interactions (RRI) is a central methodology for the specific investigation of inter-molecular RNA interactions and regulatory effects of non-coding RNAs.RNA–RNA interactions are fast emerging as a major functional component in many newly discovered non-coding RNAs.

- Why RRI

1.3 RNA-RNA Interaction Prediction Approaches

There are several available methods, that can be classified according to their underlying prediction strategies, each implicating unique capabilities and restrictions often not transparent to the non-expert user.
Most computational methods for RNA structure or RNA-RNA interaction prediction are based on thermodynamic models and provide an efficient computation since Richard Bellman's principle of optimality [1] can be applied.

- 1. Approaches that predict RRI

1.3.1 Hybrid

1.3.2 General

1.3.3 Concatenation

1.3.4 Accessibility

- 1. S-RRI, M-RRI
- 2. problems with S-RRI

1.3.5 Adv. and disadv.

Hence we go for, Multi-site RRI optimization based on single-site IntaRNA predictions.

Chapter 2

Multisite Accessibility Based

Chapter 3

Results

Chapter 4

Discussion and conclusion

Bibliography

[1] Martin Raden, Mostafa Mahmoud Mohamed, Syed Mohsin Ali, and Rolf Backofen. Interactive implementations of thermodynamics-based rna structure and rna–rna interaction prediction approaches for example-driven teaching. *PLoS computational biology*, 14(8):e1006341, 2018.