POLITECNICO DI TORINO

Master degree course in Biomedical Engineering

Master Degree Thesis

Development of a graph reduction method for minimizing RNA secondary structures and speeding-up sequence-structure alignment algorithms

Supervisor
Prof. Elisa Ficarra

Correlator
Eng. Gianvito Urgese

Candidate
Rossella Resta
matricola: 222385

Academic year 2017-2018

To my grandmother.

Summary

RNA is one of the most important molecules along with DNA and proteins for the regulation of cells' life. Several types of RNA molecules participate in the process of gene expression and for their role is fundamental the structure they assume. Differently from DNA, RNA can fold into intricate structures at secondary or tertiary level. A class of RNA of great importance is the non-coding RNAs. Knowing the structure of these molecules leads to understand their function and how they influence the expression of the genes. Nowadays it is not possible to exactly determinate in vivo the secondary structures of the molecules. For this reason, several methods for their prediction have been implemented in several tools. These methods apply different physical principles (e.g. minimum free energy) using various algorithmic approaches. The goal of this thesis is the development of a tool that takes as input the secondary structures predicted by several tools and compute the consensus between the different predictions by pruning less relevant and probable interactions between molecule's nucleotides. Cosmo, the main program developed for this thesis, is written in C++ and with SeqAn code style with the addition of some python auxiliary blocks. It computes a consensus structure module for RNA secondary structures, merging the output structures of three tools of prediction, Ipknot with six configurations and RNA fold and RNAstructure that also integrate experimental data. The structure of the consensus consists in a graph whose vertex are the nucleotides and the edges represent the interactions between them. The weights of the edges, in the case of the consensus, are proportional to the number of tools that predict the interactions. Another type of structure is the one of the base pair probability matrix that consists of a graph where each vertex has edges connected with all the others and the weights are the probabilities of interaction. This type of structure is computed by RNAfold and actually is the input for the program Lara that compute the sequence structure alignment of RNA molecules. The objective of Cosmo is to give a lighter but not less accurate input to the sequence structure tools Lara and LocARNA and improve the computational time of the programs that for many sequences can take time of the order of hours. The program of Cosmo has been tested on two different sequence libraries, Rna Mapping Database (RMDB) that contains experimental data and Bralibase. The tool used to validate the improvement in the alignment is LocARNA. The obtained results show that computing the consensus structure of 50 sequences of RMDB and 476 sequences of Bralibase, the number of edges of consensus significantly decrease (98,3% for the first library and 96,9% for the second) in comparison with the base pair probability matrix. This results lead, in the second experiment, to an evident decrease in the computational time of the sequence structure alignment computed by LocARNA that

speeds-up by 69,4%. Moreover, the quality of further sequence structure alignments has been evaluated demonstrating that after Cosmo's aggregation and pruning the quality of the alignment is preserved. Consequently, it is possible to assert that giving a lighter input for the sequence structure alignment, there is an evident speed-up in the computation of sequence structure alignment without impacting the quality of the results.

Acknowledgements

I would first like to thank my supervisor Prof. Elisa Ficarra of the Dipartimento di Informatica e Automatica, for give me the possibility to realize this work. I would like also to thank Eng. Gianvito Urgese of the Dipartimento di Informatica e Automatica for his patience and for his support during the time of research and writing of this thesis. I would like to thank also the Prof. Reinert and the SeqAn team, in particular Jörg Winkler, for the collaboration during my period in Berlin.

Finally, I must express my very profound gratitude to my family for providing me support and encouragement throughout my years of study and all my friends fundamental to this university experience. This accomplishment would not have been possible without them.

Contents

List of Figures 11

1 Introduction 15

1.1 RNA molecule and gene expression . . . 17

1.2 RNA secondary structure . . . 19

1.3 RNA molecule's analysis . . . 20

2 Background 25

2.1 RNA secondary structure prediction tools . . . 26

2.1.1 RNAfold (Vienna RNA package) . . . 27

2.1.2 IPknot . . . 28

2.1.3 RNAstructure . . . 29

2.1.4 RNAex . . . 30

2.2 RNA Mapping Database . . . 30

2.3 SeqAn . . . 32

2.4 RNA sequence-structure alignment . . . 33

2.4.1 Lara . . . 34

2.4.2 LocARNA . . . 34

3 Method 37

3.1 SeqAn structures . . . 37

3.2 Prediction of secondary structure . . . 39

3.3 Cosmo application . . . 40

3.4 Bpp matrix pruning using the consensus structure . . . 43

4 Results 45

4.1 Dataset . . . 45

4.2 Implementation of secondary structure prediction . . . 45

4.3 Reduction of the edges of the base pair probability graph . . . 47

4.4 Computational time of the sequence structure alignment . . . 49

4.5 Alignment evaluation . . . 51

5 Conclusions 57

A Supplementary material 59

Bibliography 71

List of Figures

1.1 Gene expression according to the central dogma of molecular biology. The linear view of gene expression, in which DNA is transcribed by RNA polymerases to mRNA, then translated by the ribosome and tRNAs into proteins is extended by the addition of post-transcriptional regulation via miRNAs and RNA binding proteins (RBPs) [4] 16

1.2 Venn Diagram that explains the interdisciplinarity of bioinformatics: computer science for tools development, statistics for data analysis, biology as subject.[6] 16

1.3 (a) DNA is typically double stranded, whereas RNA is typically single stranded. (b) Although it is single stranded, RNA can fold upon itself, with the folds stabilized by short areas of complementary base pairing within the molecule, forming a three-dimensional [8] 17

1.4 The structural framework of RNA molecule. The nitrogenous bases are from the up uracil, cytosine, adenine, uracil, guanine. [12] 18

1.5 Illustration of the processes of gene expression [14]. 19

1.6 An RNA stem-loop secondary structure. 20

1.7 An RNA pseudoknot structure [19]. 20

1.8 Substructures typical of an RNA secondary structure without pseudoknots [21]. 21

1.9 Overview of RNA-Seq. First, RNA is extracted from the biological material of choice (e.g., cells, tissues). Second, subsets of RNA molecules are isolated using a specific protocol, such as the poly-A selection protocol to enrich for polyadenylated transcripts or a ribo-depletion protocol to remove ribosomal RNAs. Next, the RNA is converted to complementary DNA (cDNA) by reverse transcription and sequencing adaptors are ligated to the ends of the cDNA fragments. Following amplification by PCR, the RNA-Seq library is ready for sequencing [22] 22

1.10 Example of biological events in the alignment. There are two sequences, a query and a subject. The differences between the bases at a same position are mismatch if the base change, insertion if the second sequence miss the base, deletion if the second sequence has a base more [23]. 23

2.1 Main methods used for RNA secondary structure prediction. 26

2.2 Example of '.dbn' file. First row name of the sequence 'example', second row sequence, third row secondary structure and correspondent free energy. On the right the folded molecule obtained with RNAfold. 27

2.3 Example of '.ct' file for the same molecule of the Figure 2.2. 27
2.4 Example of a base pairing probability matrix (same molecule of Figure 2.2). The area of the squares is proportional to the probability of the nucleotide i on the horizonal axis to be paired with nucleotide j on the vertical axis. 28
2.5 A diagram of the iterative refinement algorithm for the base-pairing probability matrix. A constraint on secondary structure for each level is denoted by a variant of the dot-parenthesis format: a matching parenthesis '()' denotes an allowed base pair, a character 'x' indicates an unpaired base, and a dot '.' is used for an unconstrained base. 29
2.6 Example of rdat file [35]. 31
2.7 Probes used for detecting nucleotides state [36]. 32
2.8 Fig. 14 (A) RNA is selectively modified (red dots) at flexible nucleotides in an RNA. (B) Positions of adduct formation are detected by primer extension. (C) Primer extension products from the experimental, no-reagent control, and sequencing markers are resolved by capillary electrophoresis. (D) Electropherograms are computationally deconvoluted to yield normalized SHAPE reactivities [38]. 33
2.9 Example of RNA sequence-structure alignment of four sequences that share the secondary structure in the first row. [47] 34
2.10 Comparison between sequence alignment computed by the CLUSTALW program on the left and sequence-structure alignment computed by LARA on the right. 35
2.11 (a) Alignment computed by LocARNA, plot generated using RNAALIFOLD. The colour annotation shows the conservation of base pairs (b) 2D plot of consensus secondary structure with the same colour scheme [43]. . 35
3.1 Diagram scheme of Cosmo. The input structures are given by running Ipknot, RNAfold and RNAstructure; the last two integrate experimental data of the RMDB. Then the consensus structure will be the input for sequence structure alignment performed by Lara. 38
3.2 Illustration of the hierarchical organization of the RNA SeqAn structure. 38
3.3 Example of fixed graph with three interaction edges whose weights are the probability of one nucleotide to be paired with the connected one. 39
3.4 Example of graph of the base pair probability matrix. Multiple edges for the second and the fourth nucleotide. The weights are the probabilities of the base pair. 39
3.5 Input and output file formats of Ipknot, RNAfold, RNAstructure. The final '.dbn' are merged in the file that will be the input for Cosmo application. 40
3.6 For each sequence, the function readMultiStructRnaInputFile, collapse all the fixed graphs correspondent to the output structures of the different tools (Ipknot, RNAfold, RNAstructure) in a unique record. 41
3.7 Flow chart of cosmo application. The subprocess 'update the consensus graph' is shown in Figure 3.8. 42
3.8 Flow chart of the subprocess 'update the consensus graph', block of the flow chart of Figure 3.7. 43

3.9 Flow chart of the filtering of the bpp matrix in the '.ps' format. This flow is repeted for each sequence. 44

4.1 Illustration of the steps of the python script. Firstly from the .rdat file FASTA files and .dat files are extracted. FASTA files are used as input of the tools that run in different confiurations. The output files will be the input for Cosmo application. 46

4.2 Reduction of the total number of edges. In a logaritmic scale representation, the sum of the edges of the original secondary structures and the sum of the edges of the consensus structures. 49

4.3 Illustation of the base pair probability matrix for the sequence of tRNA phenylalanine (E. coli). On the left the dotplots of the entire matrix, on the right the filtered one. The smaller dots disappear and only the bigger dots that represent an high probability of nucleotide interaction, remain. . 50

4.4 Example of adjacency list computed by Cosmo. 51

4.5 Illustration of the decrease of the computational time of sequence structure alignment. The time of the algorithm has been calculated with a bash script that contain also the command line for LocARNA tool. The columns represent the time spent to compute 2500 alignments, the first with the original base pair probability matrices of the RMDB sequences as constraint, the second column refers to ones with the filtered matrices that contain only the base pairs of the consensus structure. 52

4.6 Positive differences between the alignment score of the original structures and the filtered ones (x-coordinates) and the realitive occurences (y-coordinates). 53

4.7 Negative differences between the alignment score of the original structures and the filtered ones (x-coordinates) and the realitive occurences (y-coordinates). 53

4.8 Scatter plot that shows the correlation between the score of the alignment with the original structures and the score with the filtered structures only for positive values. 54

4.9 Scatter plot that shows the correlation between the score of the alignment with the original structures and the score with the filtered structures only for negative values. 55

Chapter 1

Introduction

Ribonucleic acid (also known as RNA) is one of the three major biological macromolecules that are essential for all known forms of life along with DNA and proteins [1].
RNA usually is classified into the three major types that contribute to convert DNA code into proteins, that are messenger RNA (mRNA), transfer RNA (tRNA), ribosomal RNA (rRNA). Scientists found that RNA plays a central role in protein synthesis and that this aim is achieved by the presence of other varieties of RNAs, called noncoding RNAs (ncRNAs). Their name refers to the fact that they do not actively encode proteins but are anyway fundamental for their production. Among noncoding RNAs there are small regulatory RNAs (sRNAs) which are involved in many complex regulatory roles in cells and they are divided in subcategories according to their specific function (e.g. miRNA, siRNA) [2]. DNA is a molecule that carries the genetic instructions used in the growth, development, functioning and reproduction of all known living organisms and many viruses [3]. Through the process of gene expression from DNA, proteins and functional RNA are produced. The flow of genetic information in a biological system, from DNA to RNA to proteins, is explained in the central dogma of molecular biology, extended for the presence of noncoding RNAs as post-transcriptional regulation factors (Figure 1.1).

The action of ncRNAs influences the gene expression modifying the production of proteins [5]. Among ncRNAs, there are small nuclear RNAs (snRNAs, which are involved in splicing pre-mRNA to give rise to mature mRNA), microRNAs (miRNAs) and small interfering RNA (siRNAs) which inhibit gene expression by repressing translation. MiRNAs and siRNAs are incorporated into a complex called RISC (RNA-induced silencing complex) that inhibits the transcription of an mRNA molecule that has a sequence complementary to its RNA component. For these molecules the structure is fundamental to perform their function.
Bioinformatics starts since biological data about RNA, DNA, proteins, gene expression are produced at a phenomenal rate and computers have become indispensable to biological research. It consists in the application of computational techniques to understand and organize biological data. The aim is to allows researchers to access existing information and develop tools and resources that aid in the analysis of data to interpret the results in a biologically meaningful manner.

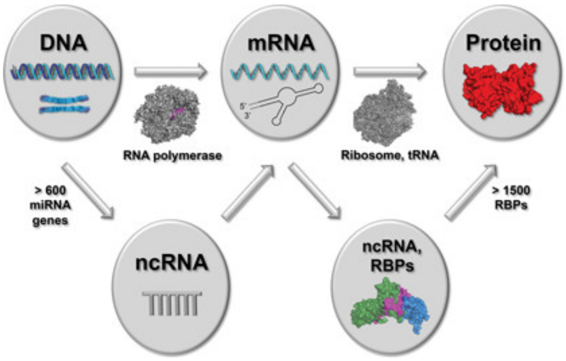


Figure 1.1: Gene expression according to the central dogma of molecular biology. The linear view of gene expression, in which DNA is transcribed by RNA polymerases to mRNA, then translated by the ribosome and tRNAs into proteins is extended by the addition of post-transcriptional regulation via miRNAs and RNA binding proteins (RBPs) [4]

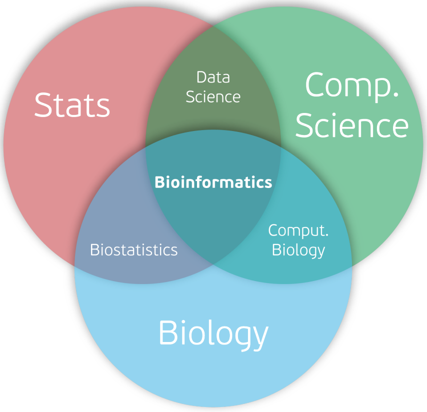


Figure 1.2: Venn Diagram that explains the interdisciplinarity of bioinformatics: computer science for tools development, statistics for data analysis, biology as subject.[6]

The data bioinformatic deals with are DNA, RNA or protein sequences, macromolecular structures and the results of functional genomics experiments [7]. The sequences are strings of an alphabet of four base letters for DNA (A, C, G, T) and RNA (A, C, G, U), the nitrogenous bases, and an alphabet of twenty letters correspondent to amino acids for proteins. Furthermore, other information included are the molecules structures closely related to their function. If DNA molecule consists of two strands coiled around each other to form a double helix, for RNA and proteins there is a big variety of structures they can assume at different levels as secondary or tertiary structures (Figure 1.3).

RNA folds into intricate structures that enable its pivotal roles in biology, ranging from regulation of gene expression to ligand sensing and enzymatic functions. Knowledge of in vivo RNA structures can reveal working mechanism of RNAs in cells, facilitates the controlled manipulation of gene expression and help in the development of molecular

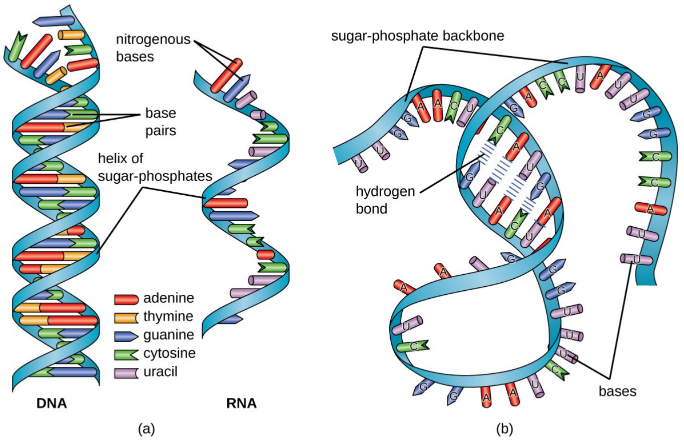


Figure 1.3: (a) DNA is typically double stranded, whereas RNA is typically single stranded. (b) Although it is single stranded, RNA can fold upon itself, with the folds stabilized by short areas of complementary base pairing within the molecule, forming a three-dimensional [8]

tools for bio and nanotechnological application [9]. Since experimental structure determination is time-consuming and expensive, resources for computational prediction of RNA structure have become indispensable.

1.1 RNA molecule and gene expression

RNA is a polymeric molecule involved in various biological roles: coding, decoding, regulation, and expression of genes. Functional RNA molecules (tRNAs, ribosomal RNAs, etc.), usually have characteristic spatial structures and therefore also characteristic secondary structures, that are prerequisites for their function. Consequently, secondary structures are highly conserved in evolution for many classes of RNA molecules [10]. At its most basic level, the structure of RNA consists of sequence of nucleotides, the basic units. Each nucleotide in RNA contains a ribose sugar, with carbons numbered from 1' to 5', a nitrogenous base that is attached to the 1' carbon position, adenine (A), cytosine (C), guanine (G), or uracil (U) and a phosphate group that have a negative charge, making RNA a charged molecule (polyanion). The nucleotides are linked to one another in a chain by chemical bonds, called ester bonds, between the sugar base of one nucleotide and the phosphate group of the adjacent nucleotide forming a sugar-phosphate backbone that defines directionality of the molecule [11] (Figure 1.4).

RNA molecules are synthesized in the process of gene expression that include the transcription, RNA splicing, translation, and post-translational modification of proteins (Figure 1.5). The transcription consists of the production of the RNA copy of the DNA. This process is performed in the nucleus by RNA polymerase, which separates the two DNA strands and adds RNA nucleotides complementary to one of the DNA strands. Each RNA base is complementary to the respective DNA one, 'A' with 'U' and 'C' with 'G'. In this way, the RNA strand will be complementary to the template DNA strand which is

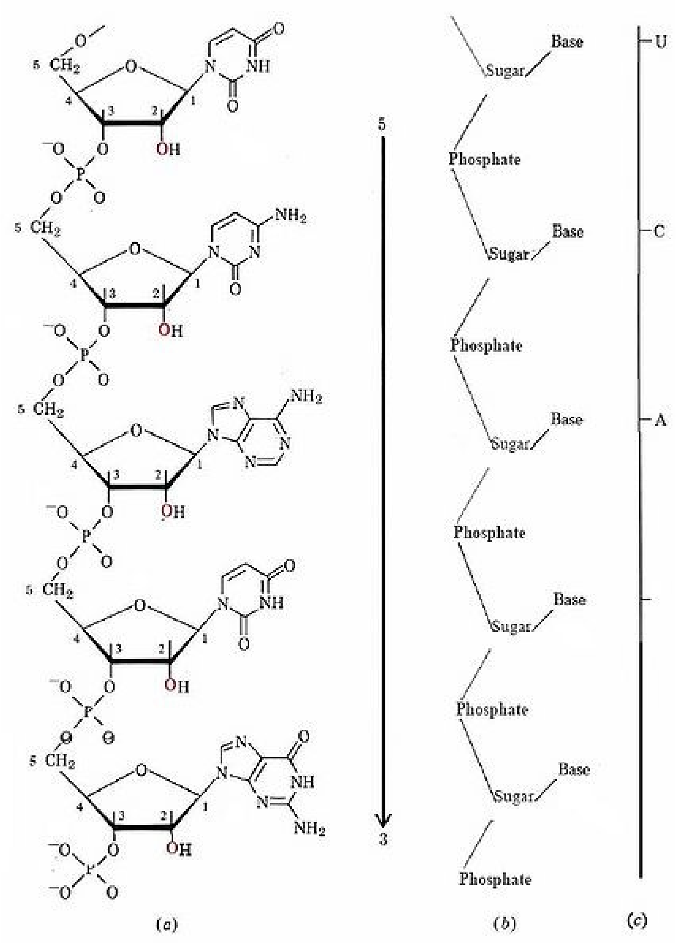


Figure 1.4: The structural framework of RNA molecule. The nitrogenous bases are from the up uracil, cytosine, adenine, uracil, guanine. [12]

complementary to the coding strand. Therefore the RNA will be equal to the DNA coding strand with the exception that the thymine is uracil in RNA. In eukaryotes there are three different type of RNA polymerase responsible for the transcription of the different RNAs, mRNA, tRNA, rRNA and non-coding RNA, these transcripts will be modified by post-processes actuated by enzymes actions. A very important modification of eukaryotic pre-mRNA is RNA splicing. The pre-mRNAs consist of alternating segments called exons and introns. During the process of splicing, an RNA-protein catalytical complex known as spliceosome catalyzes the removal of introns and then splice adjacent exons together. In certain cases, some introns or exons can be either removed or retained in mature mRNA. This so-called alternative splicing creates series of different transcripts originating from a single gene. These transcripts can be translated into different proteins, known as isoform proteins. For some RNA (e.g. non-coding RNA) the mature RNA is the final gene product [13]. In the case of mRNA the RNA is an information carrier coding for the synthesis of one or more proteins in a process called translation. The coding region of the mRNA carries information for protein synthesis. Each triplet of nucleotides of the coding region is called codon and can bind the complementary anticodon triplet in tRNA. For each anticodon, tRNAs carry the correspondent amino acid.

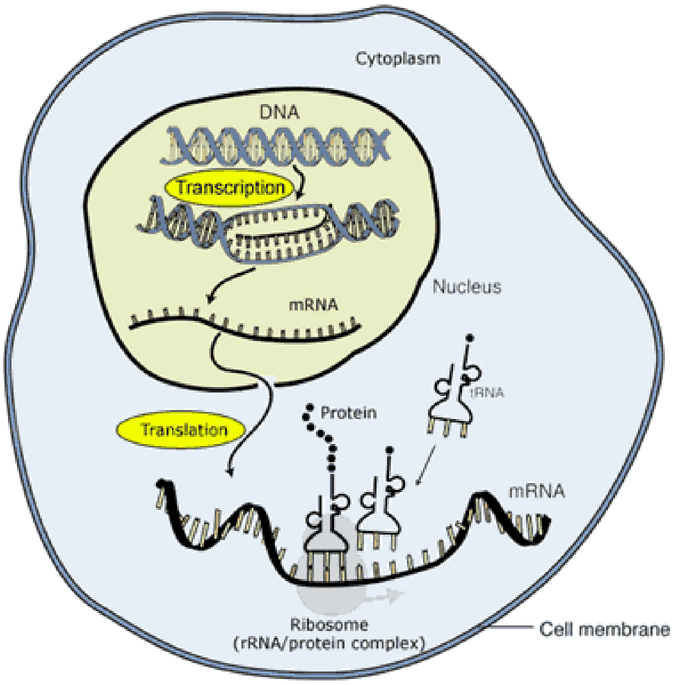


Image adapted from: National Human Genome Research Institute.

Figure 1.5: Illustration of the processes of gene expression [14].

Amino acids are then chained together by the ribosome according to the order of triplets in the coding region. The ribosome helps tRNA to bind to mRNA and takes the amino acid from each tRNA to form an amino acid chain [15][16].

Each mRNA molecule is translated into many protein molecules. Each protein exists as an unfolded polypeptide when translated from a sequence of mRNA. Amino acids interact with each other to produce a three-dimensional structure, the folded protein known as the native state. The resulting three-dimensional structure is determined by the amino acid sequence (Anfinsen's dogma)[17].

1.2 RNA secondary structure

The single-stranded nature of RNA provides the plasticity needed for it to fold into diverse secondary and tertiary structures that govern its functional roles. The first level of RNA structure is the sequence of nitrogenous bases of molecule's nucleotides. Then, base pairing between nucleotides in the same molecule create a second level of structure. The canonical bounds are A-U/U-A/C-G/G-C which correspond to energetically favorable pairings of a pyrimidine with a purine, Watson-Crick base paring. Other that occur less frequently are the wobble pairs G-U/U-G [18].

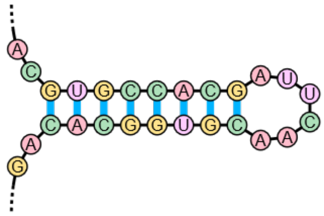


Figure 1.6: An RNA stem-loop secondary structure.

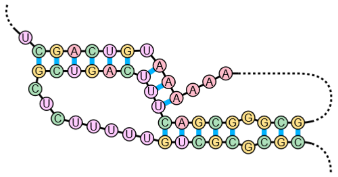


Figure 1.7: An RNA pseudoknot structure [19].

RNA molecules can fold at the second level in different ways, as shown in the pictures Figure 1.6 and Figure 1.7. The simplest one is the pseuoknot-free, for example in the Figure 1.6 there is a stem and an hairpin loop. The stem is a region where the base pairs are stacked directly to each other without any unpaired bases in between, the loop is defined by its closing base pair and all the bases between it are not paired. The RNA structure with pseudoknot contains at least two stem-loop structures in which half of one stem is intercalated between the two halves of another stem, as shown in Figure 1.7. The RNA structure pseudoknot-free can also contain other substructure (Figure 1.8) as bulge loop, loop with unpaired bases on one side (left or right) of a closing base pair, interior loop, with unpaired bases on both side, multiloop, where some base pairs bound other substructures [20].

1.3 RNA molecule's analysis

The introduction of high-throughput next-generation sequencing (NGS) technologies revolutionized transcriptomics, the study of transcriptomes. Starting from RNA-seq protocol is possible to study levels of gene expression end extract data of sequencing that are used in ulterior analysis.

A typical RNA-Seq experiment (Figure 1.9) consists of isolating RNA, converting it to complementary DNA (cDNA), preparing the sequencing library, and sequencing it on an NGS platform. To isolate different species of RNA three specific protocols can be used. Poly-A selection selects for RNA species with poly-A tail, targeted using poly-T oligos, and enriches for mRNA, ribo-depletion removes ribosomal RNA using commercially available kits and enriches for mRNA, pre-mRNA, and ncRNA and finally

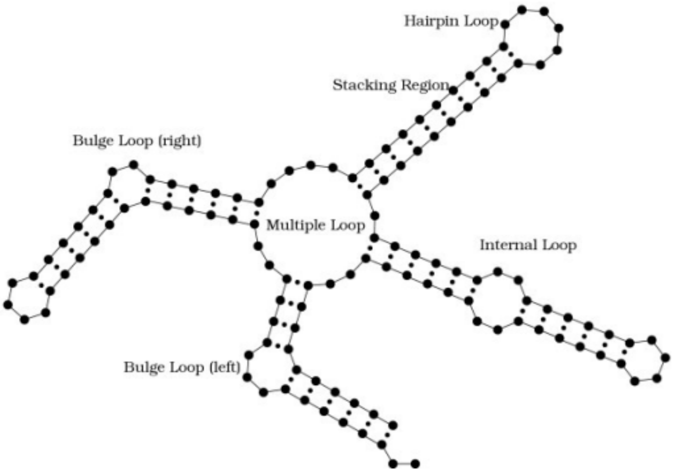


Figure 1.8: Substructures typical of an RNA secondary structure without pseudoknots [21].

size selection selects small RNA species using size fractionation by gel electrophoresis, essentially for miRNA sequencing.

An important method of RNA analysis that uses sequencing data, is the sequence alignment performed for understanding the structures, functions, and evolutionary histories of linear RNA and for finding homologs in sequence databases. Alignment is also used for DNA and proteins. The alignment is usually used to compare similar sequences and considers base per base, the similarity of the sequences classifying the events as match, mismatch, gap for insertion and gap for delection (Figure 1.10). It can take also information about the secondary structure of the molecules since sequences evolve more rapidly that the structures. For this reason, RNA secondary structure becomes indispensable for RNA analysis. Several tools have been implemented to predict it and protocols have been developed to extract information about in vivo nucleotides state.

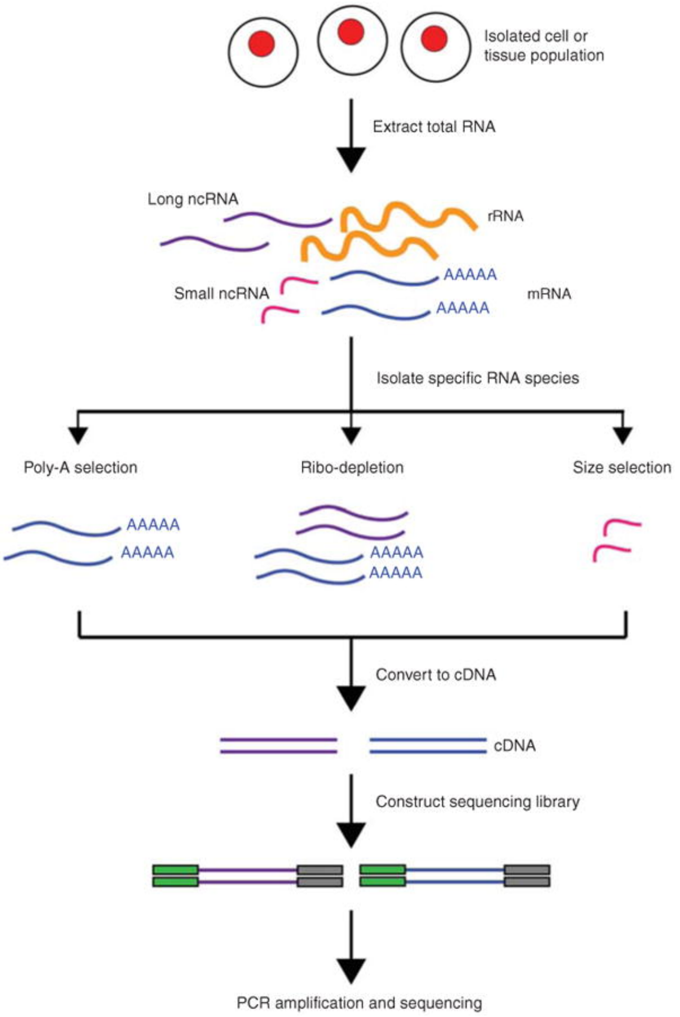
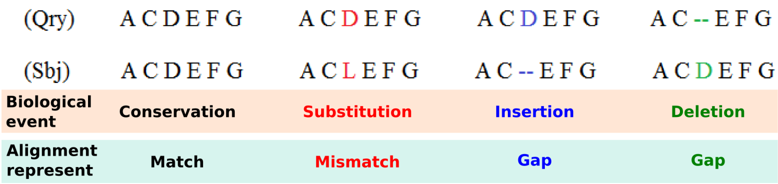


Figure 1.9: Overview of RNA-Seq. First, RNA is extracted from the biological material of choice (e.g., cells, tissues). Second, subsets of RNA molecules are isolated using a specific protocol, such as the poly-A selection protocol to enrich for polyadenylated transcripts or a ribo-depletion protocol to remove ribosomal RNAs. Next, the RNA is converted to complementary DNA (cDNA) by reverse transcription and sequencing adaptors are ligated to the ends of the cDNA fragments. Following amplification by PCR, the RNA-Seq library is ready for sequencing [22]



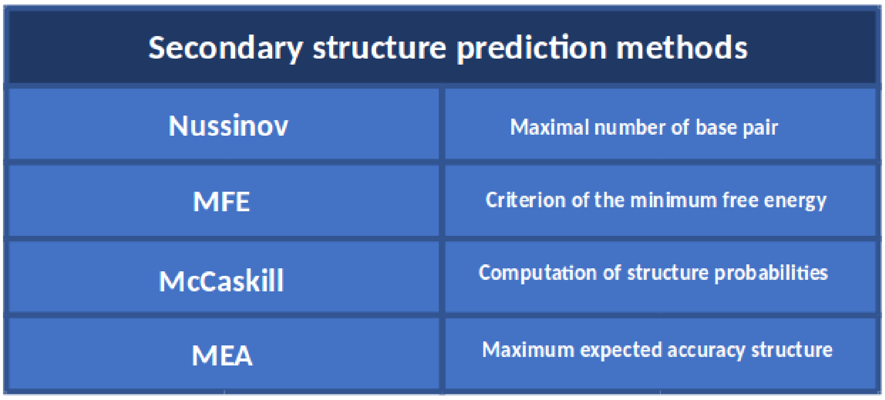
(Qry)	A C D E F G	A C D E F G	A C D E F G	A C -- E F G
(Sbj)	A C D E F G	A C L E F G	A C -- E F G	A C D E F G
Biological event	**Conservation**	**Substitution**	**Insertion**	**Deletion**
Alignment represent	**Match**	**Mismatch**	**Gap**	**Gap**

Figure 1.10: Example of biological events in the alignment. There are two sequences, a query and a subject. The differences between the bases at a same position are mismatch if the base change, insertion if the second sequence miss the base, deletion if the second sequence has a base more [23].

Chapter 2

Background

To exert their effects, RNAs assume several secondary and tertiary structures that can be studied independently, at different levels. To predict the secondary structure, numerous tools have been developed using different computational methods as well as different physical principles. The first structure prediction method was the Nussinov algorithm that enables the computation of the structure with the maximal number of base pairs for a given RNA sequence [24]. A more accurate type of structure prediction to obtain an 'optimal' structure is the criterion of the minimum free energy (MFE), since the MFE structure is not only often the most stable, but also the most probable in thermodynamic equilibrium; the energy refers to substructures energy contributions. The number of possible secondary structures a specific RNA can adopt grows exponentially with its sequence length [25] and it is difficult to enumerate all of them to assign stability scores and select the best candidate. The problem is solved efficiently by a technique called dynamic programming (DP), which recursively builds the optimal solution from solutions of smaller sub-problems. This is possible, since for pseudo-knot free structures each base pair divides the structure into two independent parts, inside and outside of the base pair [26]. The first dynamic programming algorithm to compute the MFE structure of an RNA, was published in 1981 by Zuker and Stiegler [27], about a decade after the first attempts to predict secondary structures using experimentally determined loop energy contributions. Another algorithm is the McCaskill one that enables the efficient computation of RNA structure probabilities as well as probabilities that a certain base pair is formed. Furthermore, the probabilities of unpaired bases for subsequences reflect the accessibility of RNA parts for other interactions [28]. An alternative method is the prediction of a structure with the highest sum of pairing probabilities, called the maximum expected accuracy (MEA) structure. It uses base pair probabilities and unpaired probabilities (e.g. computed via the McCaskill algorithm) to find the structure that is 'maximally accurate' in its structural elements. This approach generally maximize expected base pair accuracy as a function of base pair probabilities calculated using a partition function method [29]. Furthermore, thanks to high-throughput technologies based on enzyme cleavage or chemical modification of nucleotides, experimental data can be integrated in the prediction of RNA secondary structure. Structural states of nucleotides can be detected for the stops they cause during reverse transcription. Different protocols can be used to probe



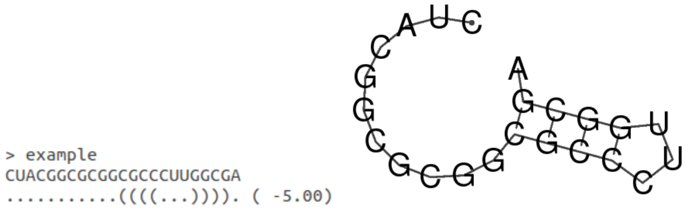
Secondary structure prediction methods	
Nussinov	Maximal number of base pair
MFE	Criterion of the minimum free energy
McCaskill	Computation of structure probabilities
MEA	Maximum expected accuracy structure

Figure 2.1: Main methods used for RNA secondary structure prediction.

RNA structures: parallel analysis of RNA structure (PARS) utilizes RNase V1 and nuclease S1 simultaneously, DMS-seq or Structure-seq uses dimethyl sulfate (DMS) to modify adenines and cytosines in single-stranded status, SHAPE protocol uses NAI-N3 to modify the backbone of all four nucleotides in single-stranded states [30]. Since these data only reveal the structural state of nucleotides (reactivity), to reveal the pairing relationships between nucleotides need to be incorporated in the proper folding algorithm. In fact, the reactivity only reveals if a nucleotide is pair on not but fails in the definition of the pairing and consequently in the definition of the secondary structure. Some prediction methods consider this data as a pseudo free energy parameter and use it as a constraint.

2.1 RNA secondary structure prediction tools

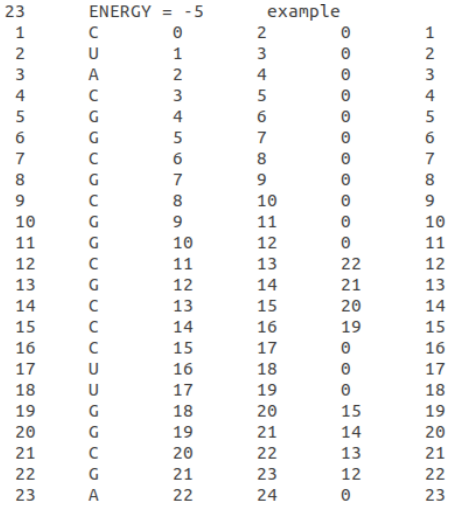
Many prediction methods for RNA secondary structure have been developed by efficient algorithms implemented in several tools. They take as input RNA sequences and give as output RNA secondary structures in different file formats. Generally the input file of a tool for RNA structure prediction is a FASTA file. It is a text-based format that can be used for nucleotide sequences or peptide sequences. The FASTA format starts with the name of the sequence preceded by the symbol '>' and at the next line the sequence of nucleotide bases. Some tools take as input a FASTA file that contains a list of sequences while others take only a sequence for file. The output file of the tools can be in different formats. Obviously, all these formats represent RNA secondary structures but in different ways. One format is the dot bracket notation (.dbn) (Figure 2.2). It contains a row for the name of the sequence preceded by '>', a row for the nucleotides sequence, a row for the secondary structure. As secondary structure, for each base there is a symbol that can be a dot, for unpaired nucleotides, a bracket '(' or ')' for the paired ones. If the bracket is open the nucleotide will be paired with one of which bracket is close. The last open bracket will be paired with the first close, the penultimate open with the second close and so on. It can include other type of bracket '[' , ']' for the pseudoknot level. Another format that has the information about RNA secondary structure is the connectivity table



```
> example
CUACGGCGCGGCGCCCUUGGCGA
...........((((...)))). ( -5.00)
```

Figure 2.2: Example of '.dbn' file. First row name of the sequence 'example', second row sequence, third row secondary structure and correspondent free energy. On the right the folded molecule obtained with RNAfold.

(.ct) format (Figure 2.3). It contains at the first row the number of nucleotides, the sequence name and additional parameters as the free energy of the structure, then a row for each nucleotide. Each row of the nucleotide is composed by the index n of the base, the base (A,C,G,U), the index $n-1$, the index $n+1$, the number of the base with which n is paired (no pair is 0). To integrate experimental data, these must be included in a .dat format. This type of format consists of one row for each nucleotide and each row consist of the index of the nucleotide and the corresponding reactivity. In the next paragraphs the description of the list of the tools for structure prediction used for the consensus structure module and RNAex that also integrates experimental data.



```
23       ENERGY = -5      example
 1       C        0       2        0       1
 2       U        1       3        0       2
 3       A        2       4        0       3
 4       C        3       5        0       4
 5       G        4       6        0       5
 6       G        5       7        0       6
 7       C        6       8        0       7
 8       G        7       9        0       8
 9       C        8       10       0       9
 10      G        9       11       0       10
 11      G        10      12       0       11
 12      C        11      13       22      12
 13      G        12      14       21      13
 14      C        13      15       20      14
 15      C        14      16       19      15
 16      C        15      17       0       16
 17      U        16      18       0       17
 18      U        17      19       0       18
 19      G        18      20       15      19
 20      G        19      21       14      20
 21      C        20      22       13      21
 22      G        21      23       12      22
 23      A        22      24       0       23
```

Figure 2.3: Example of '.ct' file for the same molecule of the Figure 2.2.

2.1.1 RNAfold (Vienna RNA package)

Vienna RNA package is a set of algorithms for analysis of RNA sequence-structure and contains in addition several scripts and utilities for plotting and input-output processing. The source code for the package is freely available and there are compiled binaries for Linux, macOS and Windows platforms. The tools provided are also available as web interface. RNAfold is a program of Vienna Package that computes RNA secondary structures. It reads RNA sequences, calculates their minimum free energy structure and prints the MFE structure and its free energy. If an option is set it also computes the partition function using McCaskill algorithm and the base pairing probability matrix, associating for each base the probability to be paired with the others [32].

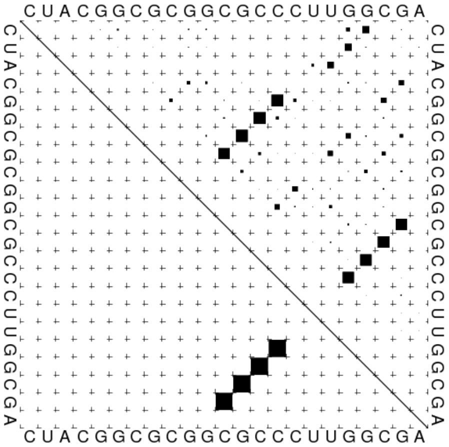


Figure 2.4: Example of a base pairing probability matrix (same molecule of Figure 2.2). The area of the squares is proportional to the probability of the nucleotide i on the horizonal axis to be paired with nucleotide j on the vertical axis.

RNA fold requires as input a FASTA file and gives as output a '.dbn' file format. Furthermore it allows to give as input structure constraints in .dat format in order to predict a structure enhanced by experimental data.

2.1.2 IPknot

IPknot predicts RNA secondary structures with pseudoknots with a method based on maximizing expected accuracy of a predicted structure, by using integer programming with threshold cut. The tool decomposes a pseudoknotted structure into a set of pseudoknot-free substructures and approximates a base-pairing probability distribution that considers pseudoknots, leading to the capability of modeling a wide class of pseudoknots. The objective is to find a secondary structure that maximizes the expectation of the gain function under a given probability distribution over the space of pseudoknotted secondary structures. The maximization of the gain function is equivalent to the maximization of the weighted sum of the base-pairing probabilities. Consequently, the base pairs whose pairing probabilities are at most a threshold are not considered, this is the threshold cut. Maximization of the gain function is solved by using the IP problem in which some constraints are set: each base can be paired with at most one base, another disallows pseudoknots within the same level and the final one ensures that each base pair of a level is pseudoknotted to at least one base pair at every lower level. Applying these constraints to a mcCaskill algorithm Ipknot predicts a secondary structure based on a refined base pair probability matrix. IPknot has several parameters that users should

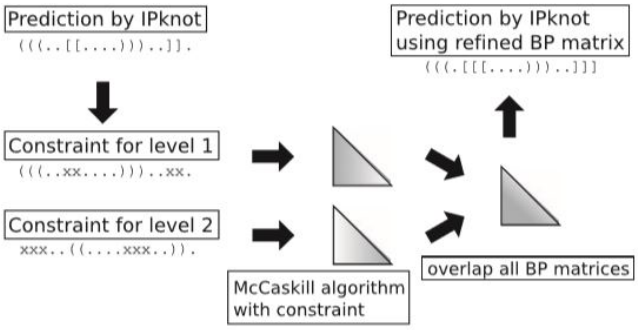


Figure 2.5: A diagram of the iterative refinement algorithm for the base-pairing probability matrix. A constraint on secondary structure for each level is denoted by a variant of the dot-parenthesis format: a matching parenthesis '()' denotes an allowed base pair, a character 'x' indicates an unpaired base, and a dot '.' is used for an unconstrained base.

select, including the weights for true positive base pairs at the levels, the number of decomposed levels of pseudoknots, and the number of iterations of the iterative refinement algorithm [31]. It takes as input a FASTA file and allow to choose three levels of prediction: pseudoknot-free, nested-pseudoknot, pseudoknotted with nested pseudoknot. Then, three possible energy models are available: McCaskill that gives the minimum energy structure, CONTRAfold that calculates the maximum expected accuracy and NUPACK based on dynamic programming.

2.1.3 RNAstructure

RNAstructure is a software package for RNA secondary structure prediction and analysis. It is publicly available with a user-friendly interface for Microsoft Windows and the package is coded in C++. This includes several algorithms for secondary structure prediction, prediction of base pair probabilities, bimolecular structure prediction, and prediction of a structure common to two sequences. The algorithms for structure prediction include free energy minimization (Fold) and maximum expected accuracy structure prediction (MaxExpect). These use nearest neighbor parameters to predict the stability of secondary structures; the parameters include both free energy and enthalpy [33]. RNAstructure requires as input a FASTA file or a SEQ file. The latter consists of at least a line for comments, a title with the sequence name and a line for the sequence. Nucleotide sequences can contain U or T interchangeably. That are interpreted according to the context of the desired operation (i.e. as U in RNA calculations or as T in DNA calculations). Spaces in the sequence are ignored. Sequences are case-sensitive and lowercase letters indicate a base that should be forced single-stranded (unpaired) in the predicted structure. "XXXX" can be used in the sequence to indicate that some bases have been left out of the prediction. It accepts file with SHAPE reactivity with any extension but with the file interface of a .dat format, previously explained. The output file will be in '.ct' format that contains secondary structure information for a sequence.

2.1.4 RNAex

RNAex is a web interface that predict RNA secondary structures enhanced by in vivo and in vitro data for both known and novel RNA transcripts in human, mouse, yeast and Arabidopsis. It provides four prediction methods, restrained MaxExpect, SeqFold, RNAstructure (Fold) and RNAfold that can be selected by the user. The first uses a posterior probabilistic model to transform various types of probing data into pairing probabilities and predict RNA secondary structure with the maximum expected accuracy (MEA) algorithm. SeqFold selects the structure centroid with minimal distance to the PARS data. RNAstructure and RNAfold allow to incorporate probing data restraints and convert them into pseudo energy contributions. Four major steps are required to process the raw probing data and predict the data-enhanced RNA secondary structure. The first is to align the reads on the genome reference, the second is to calculate the reverse transcriptase termination (RT stop) read counts because enzyme cleavage truncates the RNA transcripts and chemical modification halts reverse transcription before the modification sites. The third step consists of extracting structural reactivity from RT stops. Finally it needs to incorporate the experimental restraints into the final structure prediction. The structural reactivity derived from the third step needs to be transformed into the proper inputs (e.g. probabilities based on a statistical model) for a given structure-folding algorithm [34].

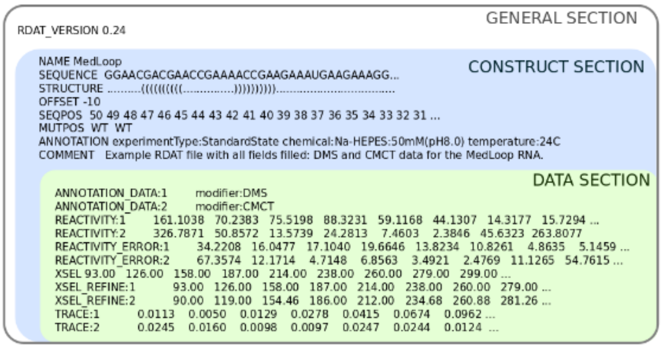
To use experimental data to enhance structure prediction, proper dataset generated by high-throughput sequencing must be chosen before submission. The structure can be predicted by providing the transcript ID or its genomic locations. When the transcript is given, RNAex extracts the sequence and structure-probing data, processes the structure-probing datasets selected by the user. The software predicts the data-enhanced RNA secondary structures using the selected folding method then it visualizes the predicted structures, the processed structure probing data and the post transcriptional regulation and mutation information. RNAex only proceeds to fold the data-enhanced structure for transcripts that have been mapped with sufficient probing data, otherwise it will proceed with the structure prediction without experimental data.

The user can set also parameters for all the four methods like percentage of nucleotides with sufficient probing reads, minimum number of reads to define sufficiency for a nucleotide and finally it is also possible choose to compare with structure prediction without data. Additionally, specific parameters for each folding method can be modified.

2.2 RNA Mapping Database

Chemical mapping is a broadly utilized technique for probing the structure and function of RNAs. RNA Mapping Database is a central location for chemical mapping data. The available database contains files in RNA Data (RDAT) text format. Each file name is given as an RMBD ID that consists of three groups of alphanumeric characters separated by underscores: the first group has 6 characters that describe the probed RNA, the second of length 3 that describe the probe used, and the third is a four digits numeric identifier. For example,TRP4P6_SHP_0003 is an ID for the Tetrahymena group

1 intron P4P6 domain, probed using 2'OH acylation (SHAPE) chemistry. The file content consists of three main sections, which are the general section, the construct section, and the data section (Figure 2.7). The general section contains information about the RDAT specification version used and serves as the root of the document. The construct section describes the specific RNA molecule that was probed in the experiment and lists information about the construct, such as nucleotide sequence, secondary structure, solution conditions as annotations and additional comments. The mapping data for each construct is then encapsulated in data sections, with two required lines as 'annotation data' (e.g., the type of probes used, the ion concentrations) and 'reactivity', for each lane capillary of an electrophoresis experiments or for each sequence position in a deep sequencing experiment.


GENERAL SECTION
CONSTRUCT SECTION
DATA SECTION

```
RDAT_VERSION 0.24

NAME MedLoop
SEQUENCE  GGAACGACGAACCGAAAACCGAAGAAAUGAAGAAAGG...
STRUCTURE ..........((((((((((..............))))))))))................................
OFFSET -10
SEQPOS  50 49 48 47 46 45 44 43 42 41 40 39 38 37 36 35 34 33 32 31 ...
MUTPOS  WT  WT
ANNOTATION experimentType:StandardState chemical:Na-HEPES:50mM(pH8.0) temperature:24C
COMMENT   Example RDAT file with all fields filled: DMS and CMCT data for the MedLoop RNA.

ANNOTATION_DATA:1      modifier:DMS
ANNOTATION_DATA:2      modifier:CMCT
REACTIVITY:1      161.1038  70.2383  75.5198  88.3231  59.1168  44.1307  14.3177  15.7294 ...
REACTIVITY:2      326.7871  50.8572  13.5739  24.2813   7.4603   2.3846  45.6323 263.8077
REACTIVITY_ERROR:1      34.2208  16.0477  17.1040  19.6646  13.8234  10.8261   4.8635   5.1459 ...
REACTIVITY_ERROR:2      67.3574  12.1714   4.7148   6.8563   3.4921   2.4769  11.1265  54.7615 ...
XSEL 93.00  126.00  158.00  187.00  214.00  238.00  260.00  279.00  299.00 ...
XSEL_REFINE:1        93.00  126.00  158.00  187.00  214.00  238.00  260.00  279.00 ...
XSEL_REFINE:2        90.00  119.00  154.46  186.00  212.00  234.68  260.88  281.26 ...
TRACE:1        0.0113   0.0050   0.0129   0.0278   0.0415   0.0674   0.0962 ...
TRACE:2        0.0245   0.0160   0.0098   0.0097   0.0247   0.0244   0.0124 ...
```

Figure 2.6: Example of rdat file [35].

The database includes experiments using base methylation by dimethyl sulfate (DMS), base adduct formation by 1-cyclohexyl-(2-morpholinoethyl) carbodiimide metho-p-toluene sulfonate (CMCT), selective 2' hydroxyl acylation with primer extension (SHAPE). The classical method to read out RNA structure probing data is gel electrophoresis (PAGE). RNA modifications induced by the different probes are detected by reverse transcription (RT) using a radio or fluorescently labeled primer; labeled cDNAs accumulate as a result of the RT stops caused by the modified molecules and are fractionated by PAGE. Capillary electrophoresis (CE) instead improves the output extending the analysis from 150 nucleotides of PAGE to approximately 500 nucleotides.

If the specificity of DMS and CMCT probes is for certain bases the SHAPE one is for all nucleotides. The reactivity derives from selective 2'-hydroxyl acylation (Figure 2.8 A), covalent SHAPE molecule modifications are detected by reverse transcriptase-mediated primer extension [37]. DNA synthesis by reverse transcriptase stops one nucleotide prior to the position of a modified one. The length of each cDNA reports the site of a SHAPE modification in the original RNA. Subtraction of the intensity of modified RNA peaks from intensities of no-reagent control peaks yields a reactivity profile [Figure 2.8 D]. Nucleotides constrained by base-pairing and tertiary interactions have low SHAPE reactivities whereas single-stranded and unconstrained nucleotides have higher reactivities. High-throughput SHAPE can be performed using fluorescently labeled primers and capillary electrophoresis using a commercial DNA sequencing instrument and analyzed using custom software.

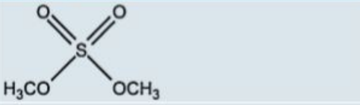
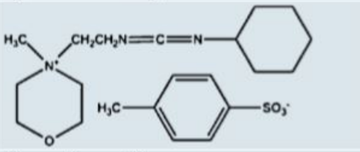
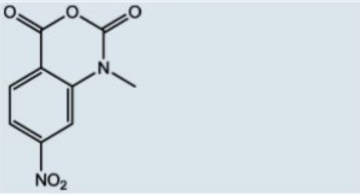
Probe name	Probe structure	Specificity
DMS		G N7 A N1 C N3
CMCT	 H_3C $CH_2CH_2N{=}C{=}N$ N^+ O H_3C SO_3^-	U N3 G N1
1M7		2′ OH

Figure 2.7: Probes used for detecting nucleotides state [36].

The reactivity data are then modified to be used as constraints for secondary structure prediction. Reactivity can be considered as a pseudo free energy term using the formula:

$$\Delta G(i) = m \ln(\text{SHAPE reactivity}(i) + 1) + b \tag{2.1}$$

This model has two free parameters, the intercept b and slope m. The intercept is negative (-0.8 kcal/mol) and represents a favorable free energy increment for pairing nucleotides at which the SHAPE reactivity is low. The slope is positive (2.6 kcal/mol) and penalizes base pairing at nucleotides with high SHAPE reactivities [39].

2.3 SeqAn

SeqAn is a C++ library of efficient data types and algorithms for sequence analysis in computational biology [40]. SeqAn has a generic programming design that guarantees an easy integration with other libraries. This design is based on four design principles. The first is the generic programming that allow to have high performance algorithms in the C++ standard library; it consists of exchangeable template types: classes and algorithms are written only once but can be applied to different data types. The second feature of SeqAn is to use global functions instead of member functions, functions members of a class, to access objects. Global functions can be added to a program at any time and without changing the existing code. Algorithms that access objects only via global functions can therefore be applied to a great variety of types, including built-in types and external classes. The third characteristic is the use of type traits. An algorithm on strings may need to know which type of characters are stored in the string, or what kind of iterator can be used for it: this is the purpose of traits. Trait classes are implemented as class templates that don't depend on a datatype. New traits and new specializations of already existing traits can be added without changing other parts of the library. Finally, the hierarchical structure called 'template argument subclassing' which means that different

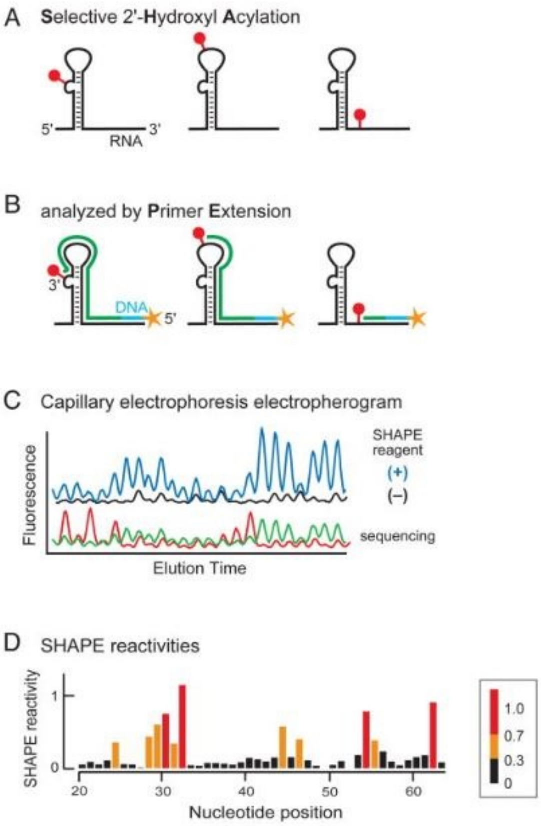


Figure 2.8: Fig. 14 (A) RNA is selectively modified (red dots) at flexible nucleotides in an RNA. (B) Positions of adduct formation are detected by primer extension. (C) Primer extension products from the experimental, no-reagent control, and sequencing markers are resolved by capillary electrophoresis. (D) Electropherograms are computationally deconvoluted to yield normalized SHAPE reactivities [38].

specializations of a given class template are specified by template arguments. As contents, SeqAn covers all areas of sequence analysis. Starting from manipulation of sequences, it contains different algorithms for pairwise and multiple sequence alignment, indexing data structure, graph type implementation, including directed graphs, undirected graphs, trees, alignment graphs. SeqAn also supports several file formats that are common in the field of bioinformatics, e.g., FASTA [41].

2.4 RNA sequence-structure alignment

RNA sequences can be aligned based on sequence similarity (i.e., primary structure) but the ability to produce good alignments in this way decreases rapidly as sequence conservation decreases. Additional criteria can be used, for example patterns of secondary structure or constraints imposed by the 3D architecture. Elements of the secondary structure that are shared by aligned molecules can serve as landmarks for alignment even in the absence of conserved sequences and can allow the alignment of more distantly related sequences, because the secondary structure evolves more slowly then the primary sequence. As a consequence alignment of distantly related RNA sequences typically require consideration of both sequence and secondary structure (Figure 2.9). Two tools of interest that perform the sequnce-structure alignment are Lara and LocARNA, described in the following sections.

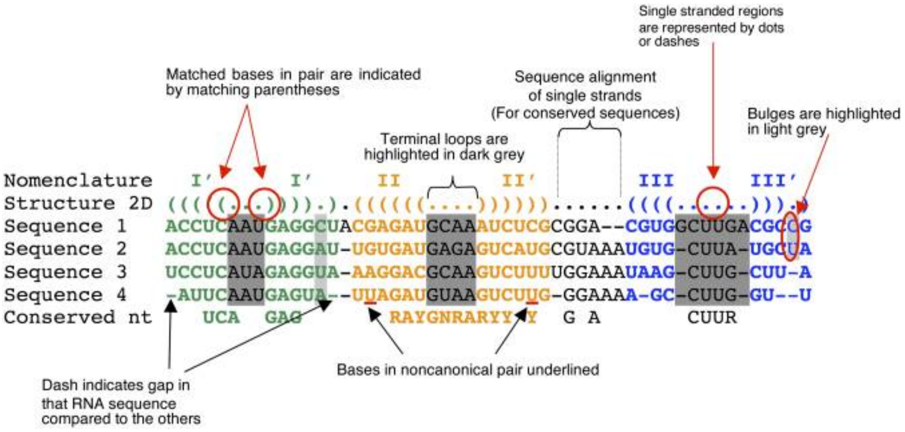

Single stranded regions are represented by dots or dashes
Matched bases in pair are indicated by matching parentheses
Sequence alignment of single strands (For conserved sequences)
Bulges are highlighted in light grey
Terminal loops are highlighted in dark grey
Nomenclature I' I' II II' III III'
Structure 2D ((((...)))).).((((((....))))))......((((.....))).)
Sequence 1 ACCUCAAUGAGGCUACGAGAUGCAAAUCUCGCGGA--CGUGGCUUGACGCCG
Sequence 2 ACCUCAAUGAGGAU-UGUGAUGAGAGUCAUGCGUAAAUGUG-CUUA-UGCUA
Sequence 3 UCCUCAUAGAGGUA-AAGGACGCAAGUCUUUUGGAAAUAAG-CUUG-CUU-A
Sequence 4 -AUUCAAUGAGUA--UUAGAUGUAAGUCUUG-GGAAAA-GC-CUUG-GU--U
Conserved nt UCA GAG RAYGNRARYYY G A CUUR
Dash indicates gap in that RNA sequence compared to the others
Bases in noncanonical pair underlined

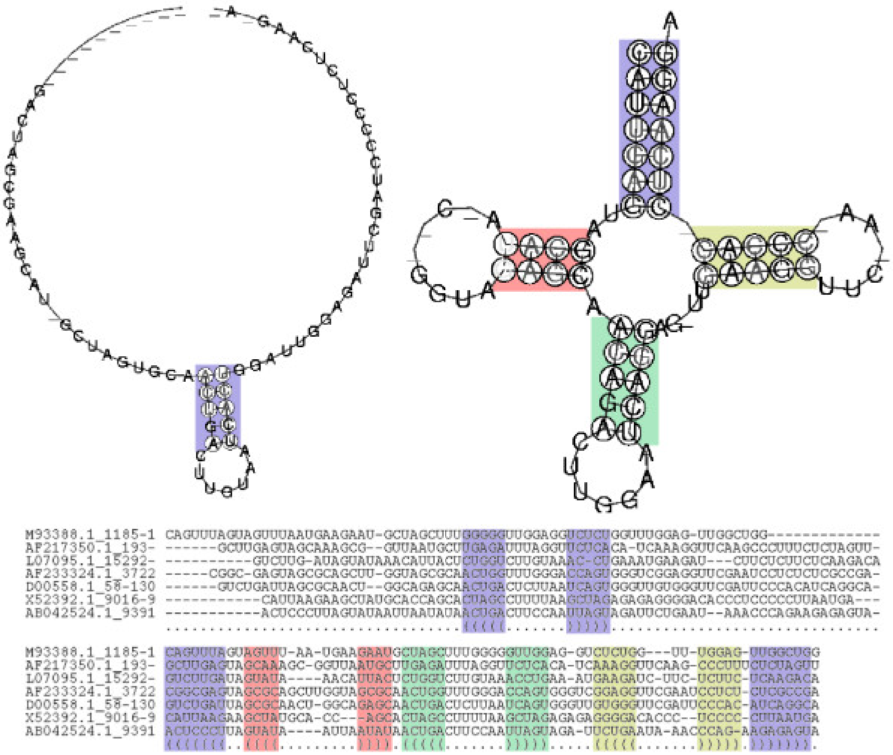
Figure 2.9: Example of RNA sequence-structure alignment of four sequences that share the secondary structure in the first row. [47]

2.4.1 Lara

Lara, lagrangian relaxation alignment, is a SeqAn program that compute RNA sequence structure alignment by using lagrangian relaxation. In Lara the nucleotides of the input sequences are converted into vertices of a graph, then weighted edges between the vertices are added and that represent either structural information or possible alignments of pairs of nucleotides. A modifyed version of Needleman-Wunsch is used as core of the program. Based on the graph model, an integer linear programming formulation is developed. The solutions are obtained by using an algorithmic approach employing methods from combinatorial optimization. The optimization consists of maximizing the edges weights by using an objective function which takes constraints into account [42].

2.4.2 LocARNA

LocARNA is a tool for multiple sequence alignment, one of the fastest and most accurate. It performs alignment and folding simultaneously. Pairwise alignments are computed using dynamic programming. Multiple alignments are built from pairwise alignments with a progressive alignment strategy. The folding algorithm, if no structure constraints are given, is based on the RNA free energy model. The input for LocARNA is a set of

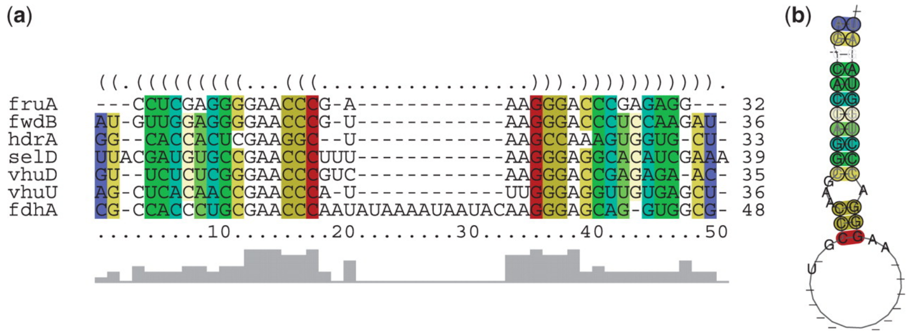


```
M93388.1_1185-1 CAGUUUAGUAGUUUAAUGAAGAAU-GCUAGCUUUGGGGGUUGGAGGUCUCUGGUUUGGAG-UUGGCUGG-------------
AF217350.1_193- ------GCUUGAGUAGCAAAGCG--GUUAAUGCUUGAGAUUUAGGUUCUCACA-UCAAAGGUUCAAGCCCUUUCUCUAGUU-
L07095.1_15292- ----------GUCUUG-AUAGUAUAAACAUUACUCUGGUCUUGUAAAC-CUGAAAUGAAGAU---CUUCUCUUCUCAAGACA
AF233324.1_3722 -----CGGC-GAGUAGCGCAGCUU-GGUAGCGCAACUGGUUUGGGACCAGUGGGUCGGAGGUUCGAAUCCUCUCUCGCCGA-
D00558.1_58-130 ------GUCUGAUUAGCGCAACU--GGCAGAGCAACUGACUCUUAAUCAGUGGGUUGUGGGUUCGAUUCCCACAUCAGGCA-
X52392.1_9016-9 -----------CAUUAAGAAGCUAUGCACCAGCACUAGCCUUUUAAGCUAGAGAGAGGGGACACCCUCCCCCUUAAUGA---
AB042524.1_9391 -----------ACUCCCUUAGUAUAAUUAAUAUAACUGACUUCCAAUUAGUAGAUUCUGAAU--AAACCCAGAAGAGAGUA-
                ..................................(((((.......)))))...............................

M93388.1_1185-1 CAGUUUAGUAGUUU-AA-UGAAGAAUGCUAGCUUUGGGGGUUGGAG-GUCUCUGG---UU-UGGAG-UUGGCUGG
AF217350.1_193- GCUUGAGUAGCAAAGC-GGUUAAUGCUUGAGAUUUAGGUUCUCACA-UCAAAGGUUCAAG-CCCUUUCUCUAGUU
L07095.1_15292- GUCUUGAUAGUAUA----AACAUUACUCUGGUCUUGUAAACCUGAA-AUGAAGAUC-UUC-UCUUC-UCAAGACA
AF233324.1_3722 CGGCGAGUAGCGCAGCUUGGUAGCGCAACUGGUUUGGGACCAGUGGGUCGGAGGUUCGAAUCCUCU-CUCGCCGA
D00558.1_58-130 GUCUGAUUAGCGCAACU-GGCAGAGCAACUGACUCUUAAUCAGUGGGUUGUGGGUUCGAUUCCCAC-AUCAGGCA
X52392.1_9016-9 CAUUAAGAAGCUAUGCA-CC---AGCACUAGCCUUUUAAGCUAGAGAGAGGGGACACCC--UCCCC-CUUAAUGA
AB042524.1_9391 ACUCCCUUAGUAUA----AUUAAUAUAACUGACUUCCAAUUAGUAGA-UUCUGAAUA-AACCCAG-AAGAGAGUA
                (((((((..((((.........)))).(((((.......))))).....(((((.......))))).))))))).
```

Figure 2.10: Comparison between sequence alignment computed by the CLUSTALW program on the left and sequence-structure alignment computed by LARA on the right.

sequences in the FASTA format that can be enriched with structure information (Figure 2.11).



```
      ((.((((((((((...(((.................))).))))))))))))).
fruA  ---CCUCGAGGGGAACCCG-A------------AAGGGACCCGAGAGG---  32
fwdB  AU-GUUGGAGGGGAACCCG-U------------AAGGGACCCUCCAAGAU-  36
hdrA  GG--CACCACUCGAAGGC--U------------AAGCCAAAGUGGUG-CU-  33
selD  UUACGAUGUGCCGAACCCUUU------------AAGGGAGGCACAUCGAAA 39
vhuD  GU--UCUCUCGGGAACCCGUC------------AAGGGACCGAGAGA-AC-  35
vhuU  AG-CUCACAACCGAACCCA-U------------UUGGGAGGUUGUGAGCU-  36
fdhA  CG-CCACCCUGCGAACCCAAUAUAAAAUAAUACAAGGGAGCAG-GUGGCG-  48
      .........10........20........30........40........50
```

Figure 2.11: (a) Alignment computed by LocARNA, plot generated using RNAALIFOLD. The colour annotation shows the conservation of base pairs (b) 2D plot of consensus secondary structure with the same colour scheme [43].

Chapter 3

Method

Cosmo, COnsensus Structue MOdule, is a consensus structure module and integrates structure prediction tools that use different prediction methods for generating a mask of the interactions selected by prediction algorithms. This mask is useful for reducing the number of nucleotide interactions evaluated during the sequence-structure alignment of RNA sequences. The Consensus Structure Module firstly computes the RNA secondary structure prediction running Ipknot, RNAfold and RNAstructure. In Cosmo there are used several configurations for each tools including the ones which integrate experimental data for RNAfold and RNAstructure. Cosmo calculates the consensus of the structures obtained looking for the common base pairs among them. If all the structures have the same base-pair this will have a major weight in the consensus module, otherwise a minor one. The code is composed by a python script that compute the input file for the main program written in C++ programming language. Cosmo is implemented using some SeqAn functions and adopting the language style. An objective of the consensus structure is to give a lighter input to tools that compute the sequence structure alignment of RNA sequences. The input of Lara and LocARNA is the base pair probability matrix computed by the RNAfold tool.

Since this method associates for each nucleotide the probability to be paired with the others in the graph structure it corresponds for n nucleotide to an nxn number of edges. Cosmo application reduces the number of edges by selecting the ones that the structure pediction tools give as output and giving them a weight proportional to the consensus.

3.1 SeqAn structures

For the consensus structure module, the SeqAn data-structures have been used because of their enanched functionality. Firstly, the sequence and all the related information are stored in RnaStructContents (Figure 3.2). This is a structure that contains a header and all records for an RNA structure file. The records consist of the string of the sequence name, an RNA alphabet string for the sequence (e.g. ACCGGCU). There is also the storage of the sequence length corresponding to the number of nucleobases and the offset that indicates where the sequence starts. There are two string sets one for reactivities and one for reactivity errors to store experimental data. The record also includes the vector

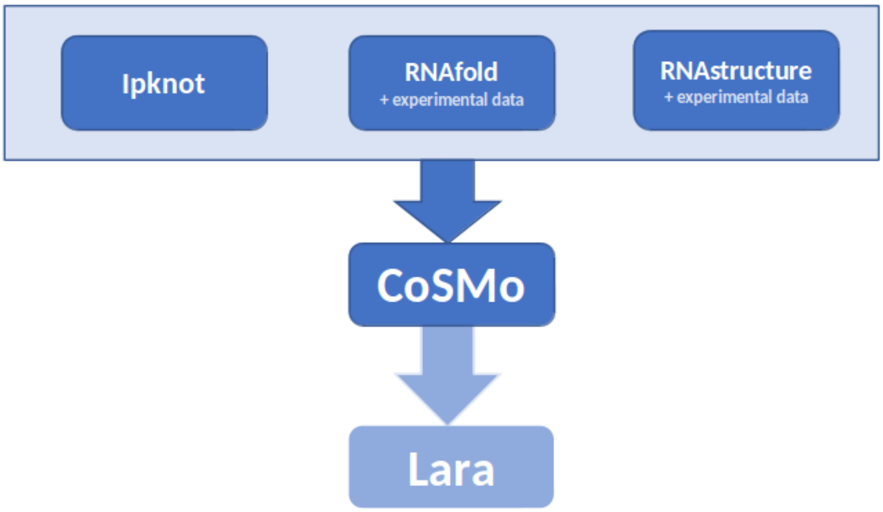


Figure 3.1: Diagram scheme of Cosmo. The input structures are given by running Ipknot, RNAfold and RNAstructure; the last two integrate experimental data of the RMDB. Then the consensus structure will be the input for sequence structure alignment performed by Lara.

of fixed graphs for the secondary structures, the vector of base pair probability graphs, a string for a related comment. A fundamental structure used for Cosmo is the graph of

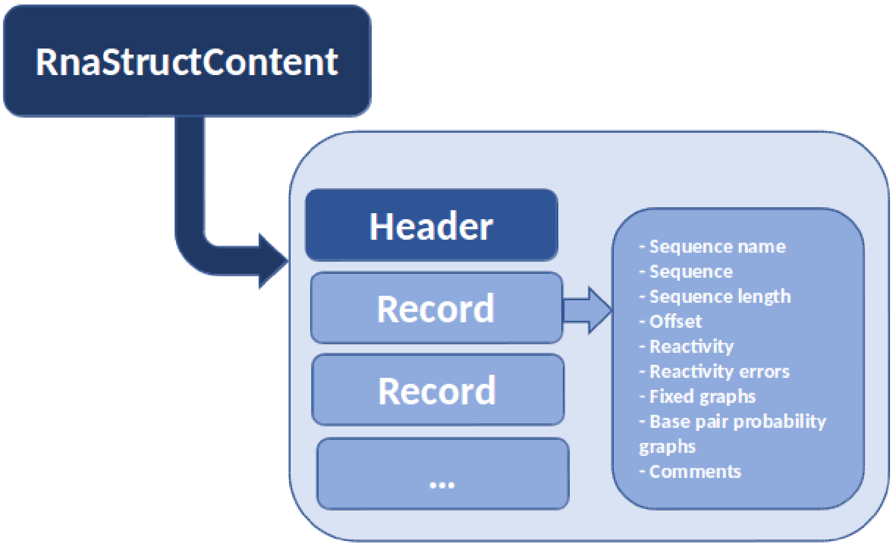


Figure 3.2: Illustration of the hierarchical organization of the RNA SeqAn structure.

the RNA record. The graph for definition is an ordered pair of a set of vertices and a set of edges. In RNA graph the vertices are nucleotides and the interaction of the secondary structure between them are represented by weighted edges where the weight can be for example the probability of a nucleotide to be paired with the other (Figure 3.3). The fixed graphs used for Cosmo are undirected, without edges directionality, and belong to the class RnaStructureGraph. The other type of graph that is used in an RNA record is the graph correspondent to the base pair probability matrix computed by RNAfold. In this case the graph is still undirected but from each nucleotide there are more then one edge and there is not the definition of a fixed structure (Figure 3.4).

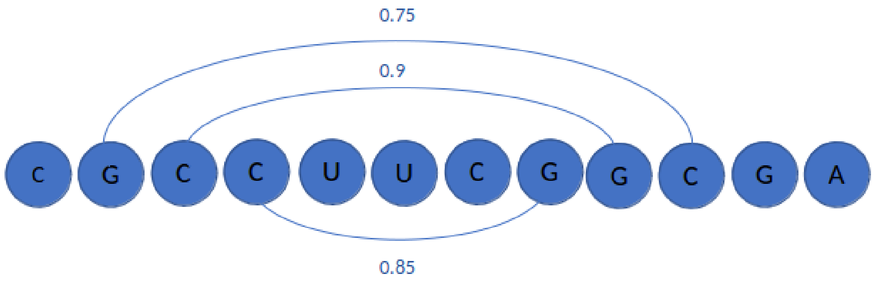


Figure 3.3: Example of fixed graph with three interaction edges whose weights are the probability of one nucleotide to be paired with the connected one.

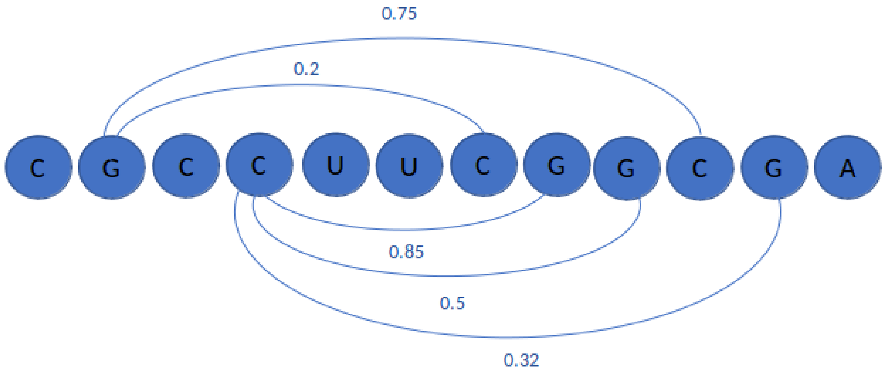


Figure 3.4: Example of graph of the base pair probability matrix. Multiple edges for the second and the fourth nucleotide. The weights are the probabilities of the base pair.

3.2 Prediction of secondary structure

To obtain the input file of Cosmo application, a python script runs Ipknot, RNAstructure and RNAfold. These tools take a FASTA file as input and compute the secondary structures of the input file sequences. For Ipknot and RNAfold a FASTA with multiple sequence is given as input instead for RNAstructure it is necessary to have a single sequence FASTA and to run the tool a number of times equal to the number of sequences. The output files for Ipknot and for RNAfold are in dot bracket notation; RNAstructure gives the output files in the connect format, one output file for sequence. Through a C++

program that combine three SeqAn functions, files in connect format are converted in dot bracket notation and resulting in a unique '.dbn' file. This program loads each input file in a structure RnaStructContents, then an output file with the 'dbn' extension is written with a SeqAn function that depending on the extension of the output file writes in the desired format. Finally all the dbn of the three tools are merged together with another python script. The resulting file will be the input for Cosmo application and will contain all the secondary structures in dot bracket notation. The fixed-structure generation using three tools and the file adaptation flow is represented in Figure 3.5.

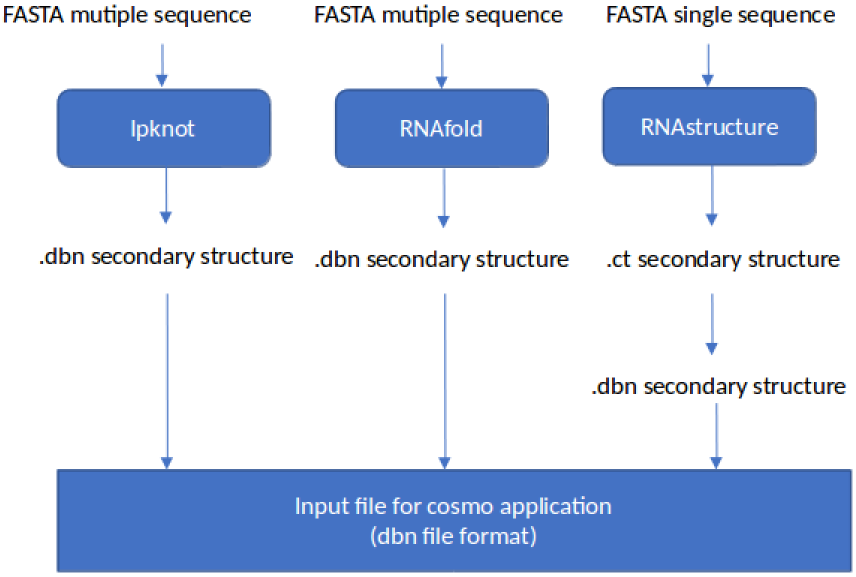


Figure 3.5: Input and output file formats of Ipknot, RNAfold, RNAstructure. The final '.dbn' are merged in the file that will be the input for Cosmo application.

3.3 Cosmo application

Cosmo application is developed to compute the consensus from several RNA secondary structures and to reduce the edges of a base pair probability graph. It takes multiple inputs from a command line that are:

- the name of the input file obtained with the python scripts,
- the name of the output file,
- the parameter used during the RNAfold execution to select the minimum energy to be considered,
- the weight of the first edge in the construction of the consensus graph,
- the weight to add to the edge in case of consensus,

- the time limit,
- the number of threads forced,
- the number of threads detected,
- the level of verbose.

In Cosmo application there are three possible levels of verbose in order to have different levels of the steps description, starting from the one that is the most general to the level three that is the most accurate. All these options are contained in the header file ‘option.h’ included in the main. The main includes also other header files that are data-types.h which contains the variable and structures definition of rna cosmo application, rna_cosmo_io.h that contains the function to read the input file, ‘vienna_rna.h’ contains functions to manage ViennaRNA objects. It consists of the declarations of the functions to compute the base pair probability matrix by using ViennaRNA and to build the corresponding graph where the edges weights are the probabilities. The first step of the program is the creation of an *RnaStructContent* where all the structures of the input file are ordered as fixed graphs of the correspondent sequence. When an RNA structure file is read there is a record for each sequence and the correspondent fixed graph. Through a function the fixed graphs of the same sequence are appended as fixed graphs of the same record and the number of records is reduced (Figure 3.6). This function is contained in the header file rna_cosmo_io.h. Then the base pair probability matrix for all the sequences is computed by using RNAfold, using a function defined in vienna-rna.h and another graph is generated. Each base pair probability graph is added to the correspondent record.

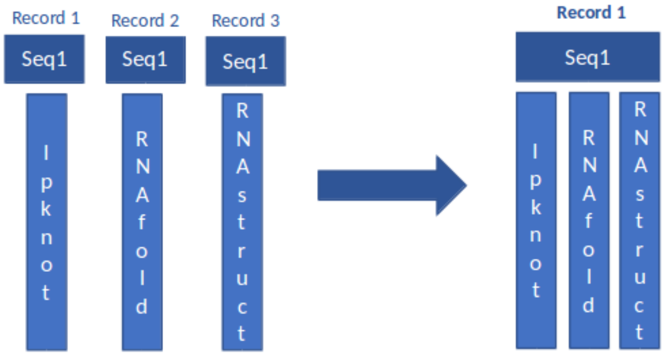


Figure 3.6: For each sequence, the function readMultiStructRnaInputFile, collapse all the fixed graphs correspondent to the output structures of the different tools (Ipknot, RNAfold, RNAstructure) in a unique record.

Then, all the graph obtained with Ipknot, RNAstructure, RNAfold contained in the records are processed in order to compute the consensus graph. The development of the next part of the code is illustrated in the flow charts, Figure 3.7. By scoring on all the graphs of a sequence, the first input graph is copied in a new one that will become the consensus graph. The initial weight of the edges of the consensus graph is equal to the parameter set as option, the default value is ‘1’; by using an edge iterator for undirected graph the weight of the edges is assigned. For the other graphs of the record the consensus one will be updated. This subprocess consists on scoring all the edges of the graphs if an edge already exists, its weight in the consensus is incremented by a number equal to the parameter set as option, default value '0.5'. If the edge doesn't exist in the consensus, it is created with the weight of initialization. In this way, the consensus graph has edges with a weight proportional to the number of graphs that have them. The consensus is obtained and then probabilities are added to the current weights.

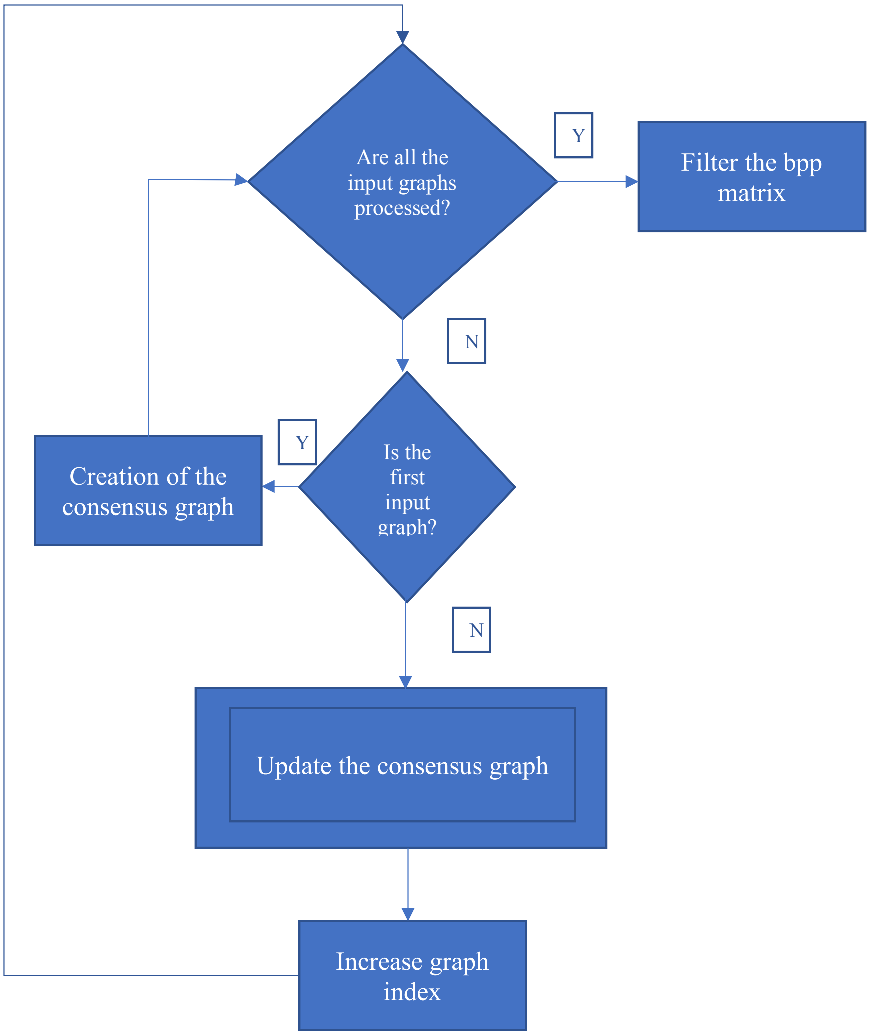

Are all the input graphs processed?
Y
Filter the bpp matrix
N
Y
Creation of the consensus graph
Is the first input graph?
N
Update the consensus graph
Increase graph index

Figure 3.7: Flow chart of cosmo application. The subprocess 'update the consensus graph' is shown in Figure 3.8.

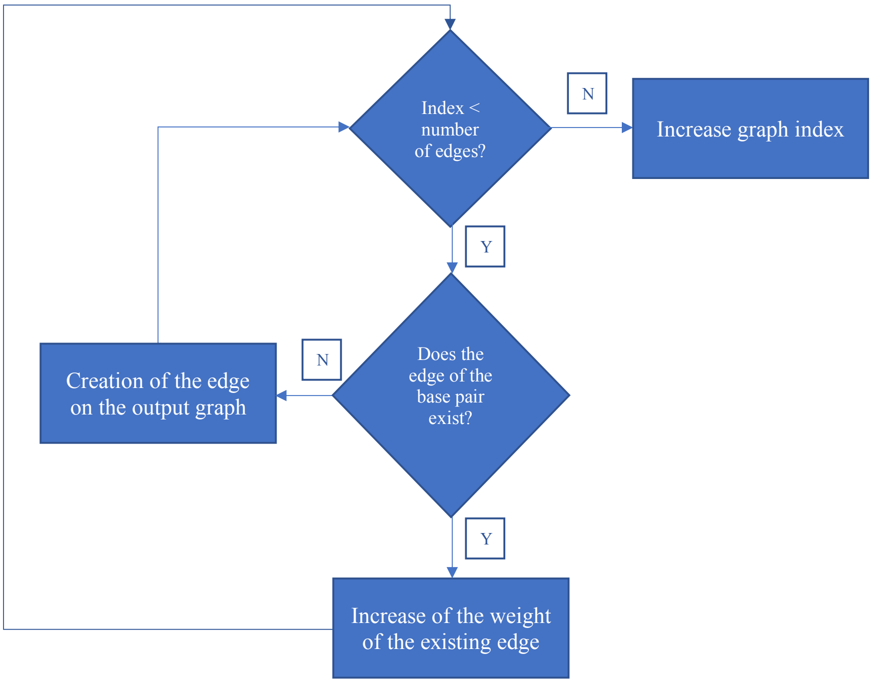


Figure 3.8: Flow chart of the subprocess ‘update the consensus graph’, block of the flow chart of Figure 3.7.

3.4 Bpp matrix pruning using the consensus structure

The consensus structure is used to filter the structure of the base pair probability matrix (bpp matrix) with a python script. Firstly, for the consensus graph, a file with all the adjacency lists of the sequences is obtained. The base pair probability matrices computed by the RNAfold tool are in ‘PostScript’ format (.ps); each file has the name of the sequence and it consists in a header and the pairs of nucleotides with the respective probabilities. Reading the sequence name from the adjacency lists, the correspondent ‘.ps’ file of the sequence is open and the header is copied in a new ‘.ps’ file. Only the pairs of the adjacency list will be maintained in the new .ps file with the respective probability. In the flow chart of Figure 3.9 the illustration of the main steps computed for each sequence.

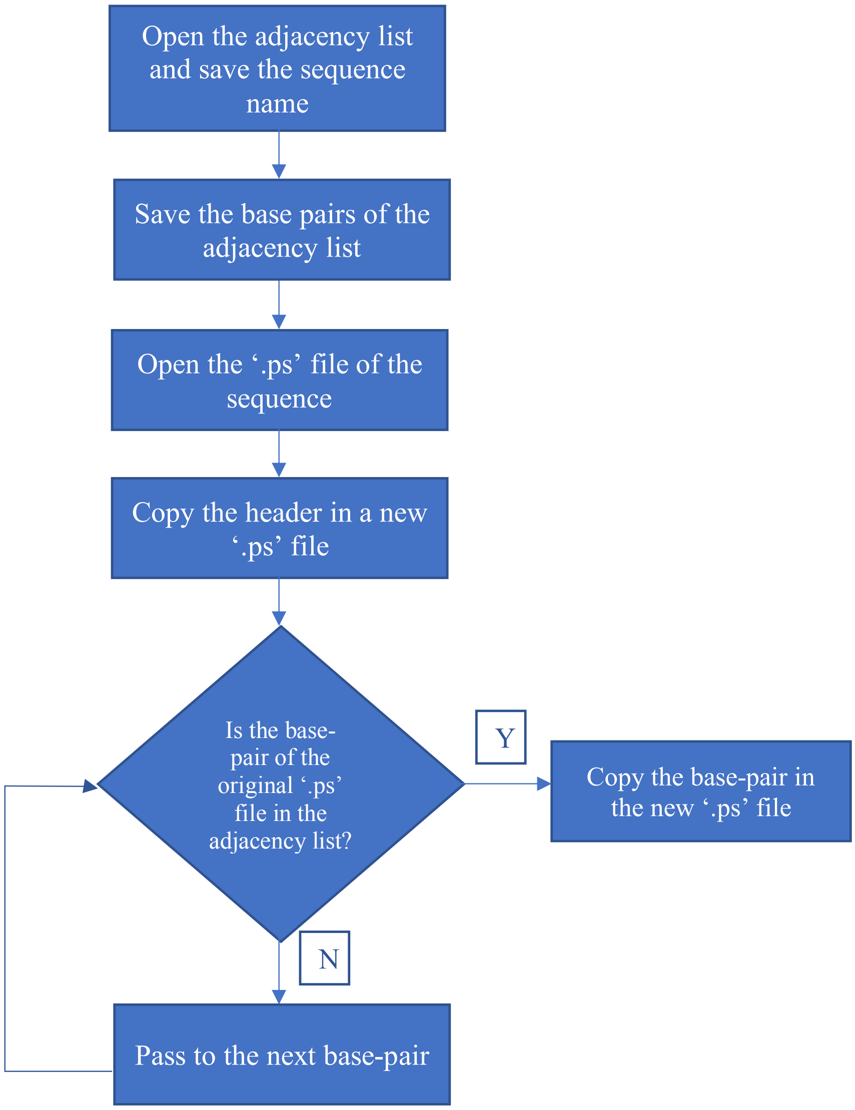

Open the adjacency list and save the sequence name
Save the base pairs of the adjacency list
Open the '.ps' file of the sequence
Copy the header in a new '.ps' file
Is the base-pair of the original '.ps' file in the adjacency list?
Y
Copy the base-pair in the new '.ps' file
N
Pass to the next base-pair

Figure 3.9: Flow chart of the filtering of the bpp matrix in the '.ps' format. This flow is repeted for each sequence.

Chapter 4

Results

The results obtained for this work take as input RNA sequences from RNA mapping database (RMDB) and from Bralibase that contains classes of RNA families. Bralibase is a benchmark alignment database [44] is used to evaluate the performance of alignment tools as well as Lara and LocARNA. The first step in order to obtain the results, is to download the sequence files from the libraries sites. Then though a python script there is the prediction of the secondary structure for each sequence of the FASTA files contained in the libraries. Ipknot runs with six different configurations, while RNAfold and RNAstructure with the default parameters. Then Cosmo computes the consensus structure of all the structures reducing significantly the edges of the base pair probability graph.

4.1 Dataset

Cosmo is validated using all the sequences of RNA mapping database that includes sequences of ribozymes, ribonucleic acid enzymes, pT181, a prototype of a family of staphylococcal plasmids that silence genes in the process of gene expression [45]. Other types of sequences are riboregulators, that silence or activate gene expression in response to different cell's signals by binding to complementary Watson-Crick base pairing [46]. There are also sequences of riboswitch that are segment of mRNA and bind small molecules resulting in changing the production of proteins of the same mRNA [48]. The second library used for validation is Bralibase and consists of 5S rRNA, tRNA, U5 spliceosomal RNA .

4.2 Implementation of secondary structure prediction

The input sequences and structures for Cosmo are obtained by using a python script. From the files of RNA mapping database are produced FASTA files, input for tools; furthermore .dat files are obtained by extracting reactivities data. The .dat format consists in two columns, one for nucleotide indeces the other for the correspondent reactivities.

Through the script the three selected tools run with different configurations. For Ipknot there are two energy models McCaskill that computes the partition function of

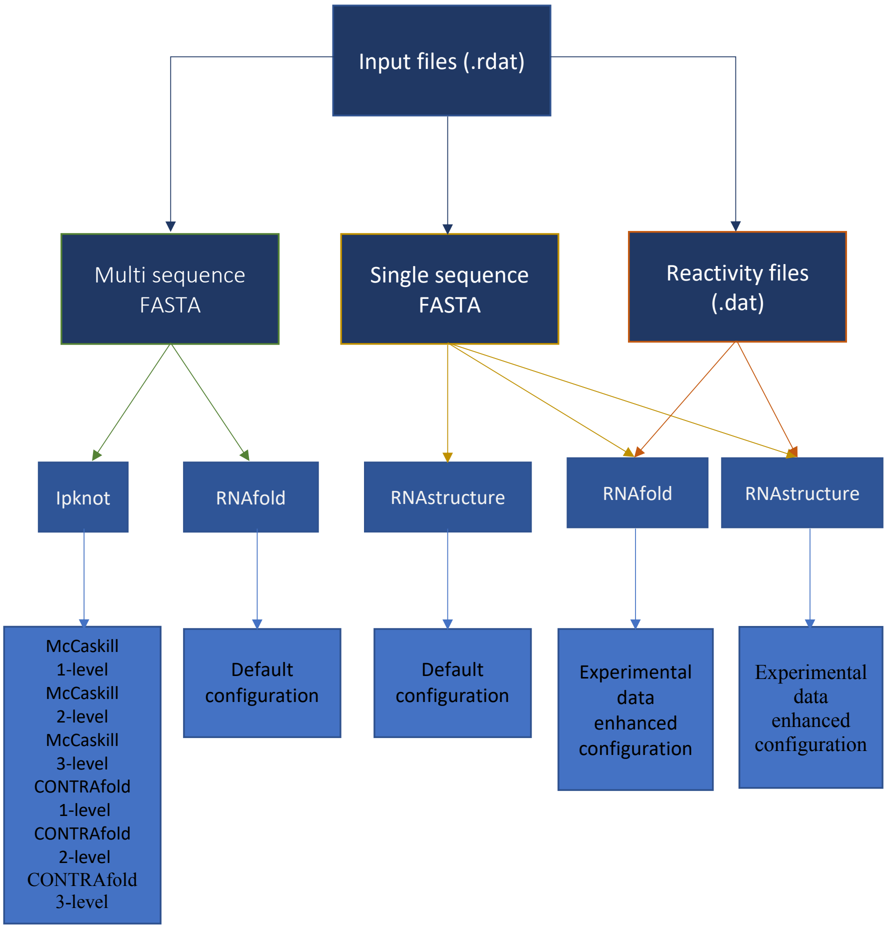


Figure 4.1: Illustration of the steps of the python script. Firstly from the .rdat file FASTA files and .dat files are extracted. FASTA files are used as input of the tools that run in different confiurations. The output files will be the input for Cosmo application.

the secondary structure and CONTRAfold that maximize the expected accuracy. Each method is used with the three available levels: pseudoknot-free, nested-pseudoknot, pseudoknotted with nested pseudoknot. RNAfold and RNAstructure are used with default configurations and each one with the configuration that integrates experimental data in .dat format. RNAfold calculates the minimum free energy structure, RNA fold uses thermodynamics and the most recent set of nearest neighbor parameters. For RNAstructure and for the two configurations with experimental data (RNAfold and RNAstructure) it was necessary for tools functionality to obtain a FASTA file per sequence. A schematic representation of the configuration of used prediction algorithms is given in Figure 4.1.

4.3 Reduction of the edges of the base pair probability graph

For the first experiment fifty sequences have been selected from the Rna Mapping Database. The graphs examined for this experiment are obtained by using RNAfold for the base pair probability graph and the consensus method for the others. The base pair probability graphs present weighted edges for each vertex, the weight is the probability of a nucleotide to be paired with the other. The other graphs obtained by Cosmo are computed with the consensus method, considering firstly all the structures obtained with the tools without shape data, then all the structures including the ones computed with shape, finally the only shape enhanced structures. As shown in the Table 4.1, the number of edges from the base pair probability graph pruned using the first consensus graph decrease of two orders of magnitude. Of course, the number of edges of the graphs is proportional to the sequence length. The number of edges in the consensus graph with shape structures is slightly lower then the graph without considering shape, this mean that the structures partially change and as consequence edges are added. In this experiment the total number of interactions decrease from 186933 to 3168, considering the structures without shape. These numbers derive from the sum of edges of all the base pair probability graphs and the sum of the edges of all the consensus graphs.

Sequence Name	length	SHAPE	A	B	C	D
Class I Ligase	187	N	4792	63	-	-
pT181 transcriptional attenuator	112	Y	2226	39	39	24
taR12 riboregulator antisense	71	Y	478	25	25	24
Ribonuclease P specificity domain, B. subtilis	156	N	4299	90	-	-
SAM I riboswitch, T. tengcongenesis	120	N	2227	48	-	-
pT181 transcriptional attenuator	120	Y	2459	72	79	29
pT181 transcriptional attenuator	112	Y	2053	48	48	24
cidGMP riboswitch, V. Cholerae	105	N	1681	51	-	-
pT181 transcriptional attenuator	108	Y	2098	46	46	25
tRNAphe, E. coli	135	N	3350	70	-	-
add adenine riboswitch	166	N	5031	72	-	-
adenine riboswitch, add	74	N	1035	27	-	-
pT181 transcriptional attenuator	118	Y	2433	52	52	30
pT181 transcriptional attenuator	106	Y	2007	46	46	24
tRNA phenylalanine (yeast)	133	N	2569	50	-	-
5S rRNA, E. coli	123	N	2728	85	-	-
taR10 riboregulator antisense	70	Y	560	24	24	22
RNA-IN S3	60	Y	411	21	21	6
Hepatitis C virus, IRES domain	338	N	21606	279	-	-
TPP riboswitch, E. coli	80	N	1159	33	-	-
P4-P6 domain, Tetrahymena ribozyme	239	N	8611	77	-	-

Hox A9 mRNA 5’ UTR	176	N	5530	102	-	-
cidGMP riboswitch, V. Cholerae	158	N	4231	75	-	-
RNA-IN S4	60	Y	409	26	29	11
add Riboswitch 13-140, V. vulnificus	128	N	3047	65	-	-
R1 translational copy number control regulator hairpin	64	Y	565	22	22	19
pT181 transcriptional attenuator	120	Y	2425	49	49	29
pT181 transcriptional attenuator	139	Y	3422	91	93	42
pT181 transcriptional attenuator	120	Y	2525	53	53	29
5S RNA, E. coli	180	N	5740	109	-	-
crR12 riboregulator UTR	70	Y	830	25	25	16
pT181 transcriptional attenuator	118	Y	2453	52	52	30
pT181 transcriptional attenuator	112	Y	2189	45	45	25
RNA-IN S4	60	Y	409	23	23	8
pT181 transcriptional attenuator	118	Y	2496	50	50	28
SAM I riboswitch, T. tengcongenesis	177	N	5174	87	-	-
16S rRNA Four-Way Junction	110	N	2062	74	-	-
Ribonuclease P specificity domain, B. subtilis	201	N	7372	104	-	-
pT181 transcriptional attenuator	139	Y	3421	66	66	32
adenine riboswitch, add	131	N	3100	47	-	-
crR10 riboregulator UTR	70	Y	850	25	25	18
P4-P6 domain, Tetrahymena ribozyme	223	N	8303	102	-	-
pT181 transcriptional attenuator	120	Y	2533	53	53	29
pMU720 translational copy number control regulator hairpin	71	Y	732	23	26	24
Tebowned	72	N	762	20	-	-
Hobartner bistable switch	89	N	1184	31	-	-
M-stableRNA	103	N	1558	27	-	-
btuB riboswitch leader sequence, E. coli	206	Y	7824	138	149	73
pT181 transcriptional attenuator	120	Y	2437	49	49	29
Hepatitis C virus, IRES domain	395	N	29537	279	-	-

Table 4.1: Decrease of the edges from the base pair probability graphs to the consensus ones. In the first column the sequences names, in the second column the length of the sequences. ‘Y’ or ‘N’ for the sequence with shape data and without respectively. In the column A the edges of the base pair probability graph, in the column B the edges of the consensus graph without considering shape enhanced structures, in the column C the edges of the consensus graph using all the structures, in the column D the consensus graph edges obtained by using only shape structures.

For the second experiment, Cosmo is validated by using 476 sequences from the Bralibase library. In this case the consensus graphs are obtained by using the tools configurations without shape data since not available for this data-set. Ipknot predicts three levels of pseudoknotted structures each one with two energy models CONTRAfold and McCaskill. RNAfold and RNAstructure predict one structure for each sequence. There is, also in this case, a considerable reduction of the number of edges from the base pair probability graph to the consensus graph computed by Cosmo, in total from 752484 to 23346 (Figure 4.2). In the barchart are shown the total number of edges of the full bpp (base pair probability) matrix computed by RNAfold and the total number of edges of the bpp matrix pruned with consensus graph computed using Cosmo. Since the number scale from the original graph to the consensus one is very different there will be a significative improvement in the computational time for sequence structure alignment. In fact, if the input is the consensus graph the alignment tool will consider a reduced number of edges during the refinement steps of the sequence-structure alignment.

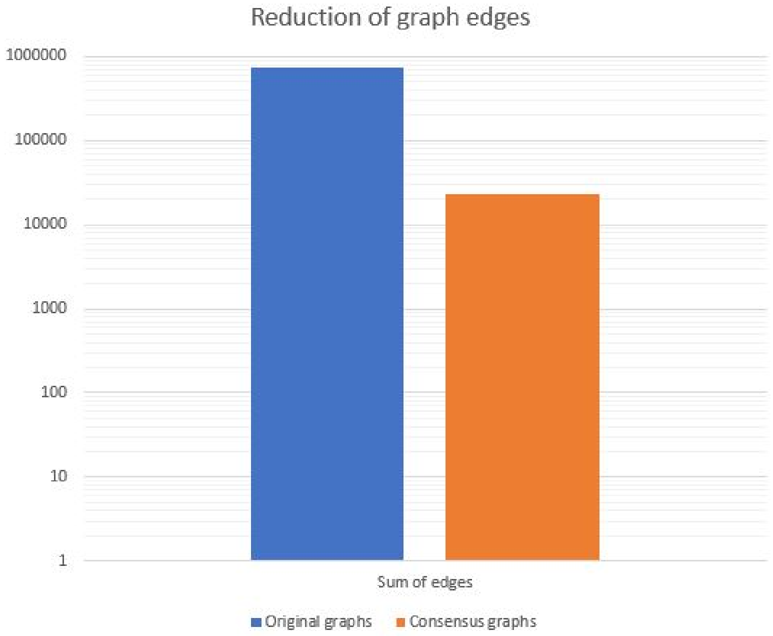


Figure 4.2: Reduction of the total number of edges. In a logaritmic scale representation, the sum of the edges of the original secondary structures and the sum of the edges of the consensus structures.

4.4 Computational time of the sequence structure alignment

The major objective of Cosmo is the improvement of the computational time of the sequence structure alignment. To test the effective performance of Cosmo the alignment is computed using the fifty sequences of RMDB (the same of the first experiment). For this experiment the alignment tool used is LocARNA that compute the alignment of RNA sequences calculating the secondary structures with RNAfold as default but accepts also structures constraint. The input for the alignment are the two sequence files that can be uploaded in several formats (Fasta, Clustal, Stockholm, LocARNA PP, ViennaRNA postscript dotplots) and can contain secondary structure information. In this experiment, the input files for LocARNA are the sequence of RMDB, each one aligned with all the others, firstly the sequences with the full bpp matrix generated by RNA fold then the same with the alignments using the bpp matrices pruned with the consensus structures. In Figure 4.3 an example of two bpp matrices of the same sequence.

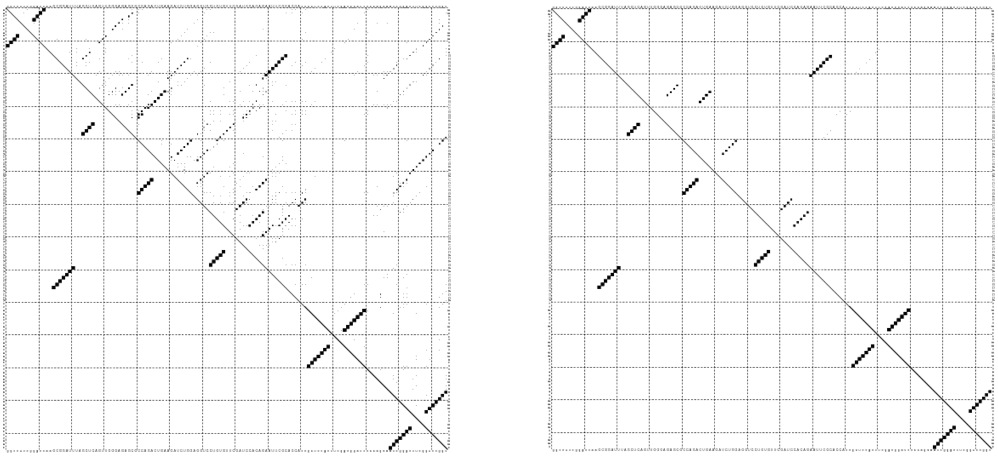
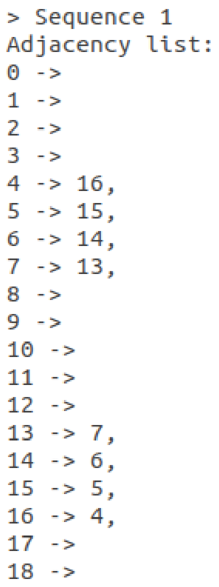


Figure 4.3: Illustation of the base pair probability matrix for the sequence of tRNA phenylalanine (E. coli). On the left the dotplots of the entire matrix, on the right the filtered one. The smaller dots disappear and only the bigger dots that represent an high probability of nucleotide interaction, remain.

The first step to obtain the original base pair probability structures is to run RNAfold with a particular configuration, giving as input the FASTA file of the fifty sequences, to compute in this way the structures in the 'ViennaRNA postscript dotplots' file format. The next step is to obtain the filtered sequence structures extracting the adjacency list of the consensus structures. It consists of the names of the sequences and the list of the nucleotides and the correspondent base pairs (Figure 4.4).

Then through a python script the ViennaRNA postscript dotplots are filtered. ViennaRNA postscript dotplots contain all the base pair and the respective probabilities computed by RNAfold, correspondent to the base pair probability matrix. The filter of these files consists of maintaining only the base pairs of the consensus structure with the respective probabilities. In this way the probability matrix will be filtered and the resulting files will be used for the alignment.

To calculate the computational time, the alignment is performed by using as input files, the base pair probability matrix that contain the original number of base pairs and the filtered base probability matrix that contain the base pairs correspondent to the consensus graph.



```
> Sequence 1
Adjacency list:
0 ->
1 ->
2 ->
3 ->
4 -> 16,
5 -> 15,
6 -> 14,
7 -> 13,
8 ->
9 ->
10 ->
11 ->
12 ->
13 -> 7,
14 -> 6,
15 -> 5,
16 -> 4,
17 ->
18 ->
```

Figure 4.4: Example of adjacency list computed by Cosmo.

To align two thousand five hundred combinations of sequences, if the structure constraint is the original ViennaRNA structure the computational time is equal to 49 minutes, whereas, using the filtered structure the time decrease to 15 minutes. There is a considerable reduction of time that can be much more effective if the number of alignments to compute is greater. It is also important that the alignment maintains the same quality in the two cases and the filter does not lead to a loss of information.

4.5 Alignment evaluation

To evaluate how the alignment quality is influenced by the filter, the output alignment files of LocARNA have been examined. These files contain the aligned sequences with match, mismatch, gap of insertion, gap of deletion (Figure 1.10).

For each file that contains the alignment of the original structures and the correspondent file that contains the alignment of the filtered structures, the CIGAR and the score are compared. Since the sequence are fifty and the alignment is computed for all the possible combinations there are two thousand five hundred files comparations. The CIGAR is a string that describes how the subject is aligned with the query. It presents the number of the events and a letter that symbolize the events. For example, ‘15M6X1D3M1I’, this CIGAR means that in the sequence alignment there are 15 match, 6 mismatches represented by ‘X’, 1 deletion, 3 match and finally 1 insertion. By comparing the CIGAR of two original structures and the CIGAR of two filtered structures it is possible to see how the consensus structure influences the alignment. Another value very significative in the alignment is the score that assign numeric values for match, mismatch, and gaps. Since the alignment score computed with LocARNA measures the sequence structure

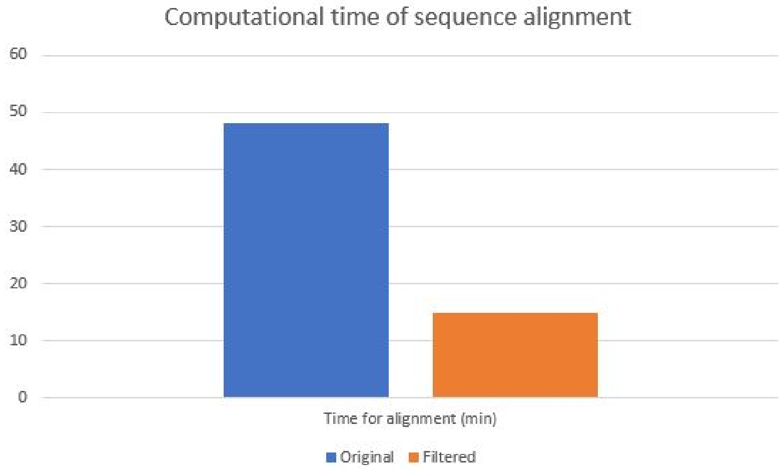


Figure 4.5: Illustration of the decrease of the computational time of sequence structure alignment. The time of the algorithm has been calculated with a bash script that contain also the command line for LocARNA tool. The columns represent the time spent to compute 2500 alignments, the first with the original base pair probability matrices of the RMDB sequences as constraint, the second column refers to ones with the filtered matrices that contain only the base pairs of the consensus structure.

alignment goodness, the score takes into account also structures similarity. In the following graphs are shown the differences between score of the aligned sequences with the full bpp matrix and the one of the aligned sequences with the filtered bpp matrix. As we can see from the Figure 4.6 and Figure 4.7 the range of values is very wide but in the first case about positive values there is a greater concentration of occurences, the negatives values are more spread. The quality of the alignment is better as higher is the score. In this context, the spread negative values can be considered as alignments with a mojority of mismatches and gaps between the examinatated sequences but this can be due to the sequences don't belong to the same family. It is also possible to see from the scatter plots the correlation between the two variants of alignment. For the negative values (Figure 4.9) there is a diagonal trend that shows a good similarity in the alignment performance. In fact, the alignment with the full bpp matrices and the alignment with the filtered ones have equal values of scores, this means that the most significative interactions of the structures are maintained in the filtering. For the other scatter-plot (Figure 4.8) the corrispondence between the two types of alignment is less, in this case the score of the alignment is slightly biased by the filtering.

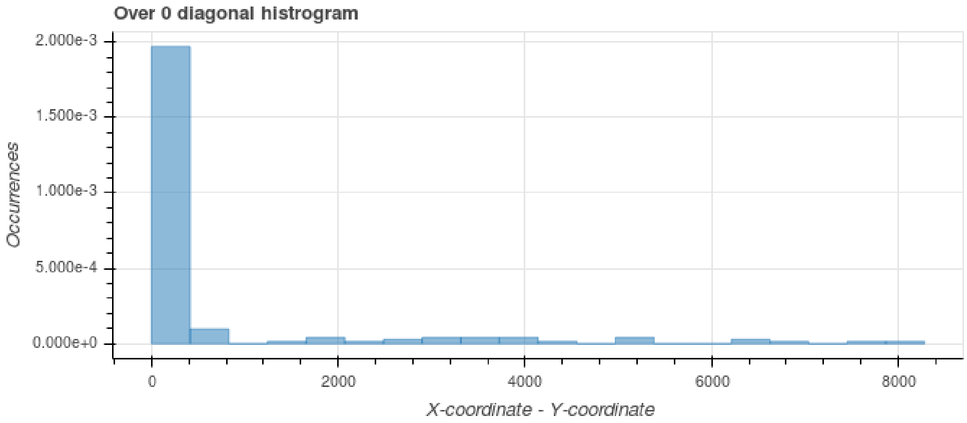


Figure 4.6: Positive differences between the alignment score of the original structures and the filtered ones (x-coordinates) and the realitive occurences (y-coordinates).

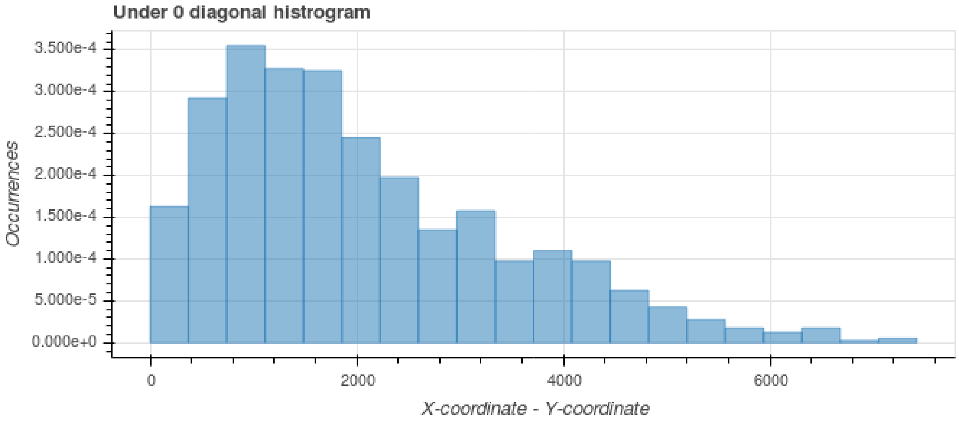


Figure 4.7: Negative differences between the alignment score of the original structures and the filtered ones (x-coordinates) and the realitive occurences (y-coordinates).

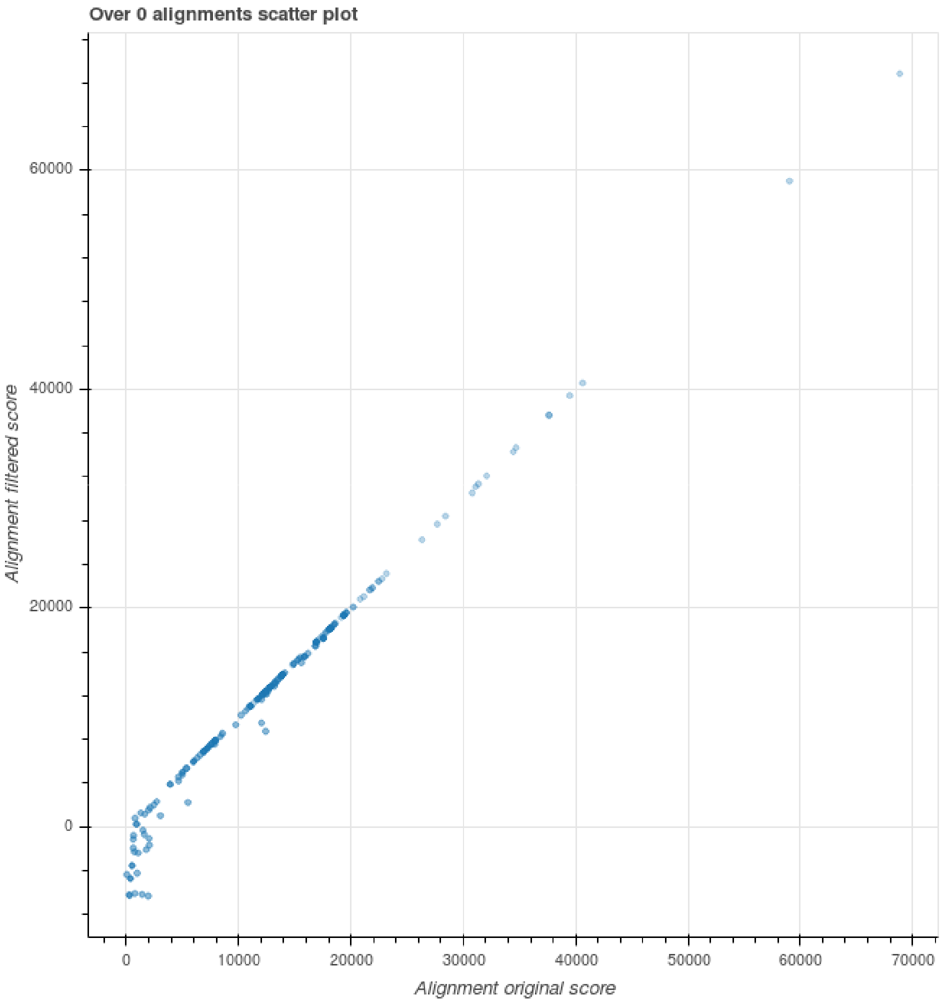


Figure 4.8: Scatter plot that shows the correlation between the score of the alignment with the original structures and the score with the filtered structures only for positive values.

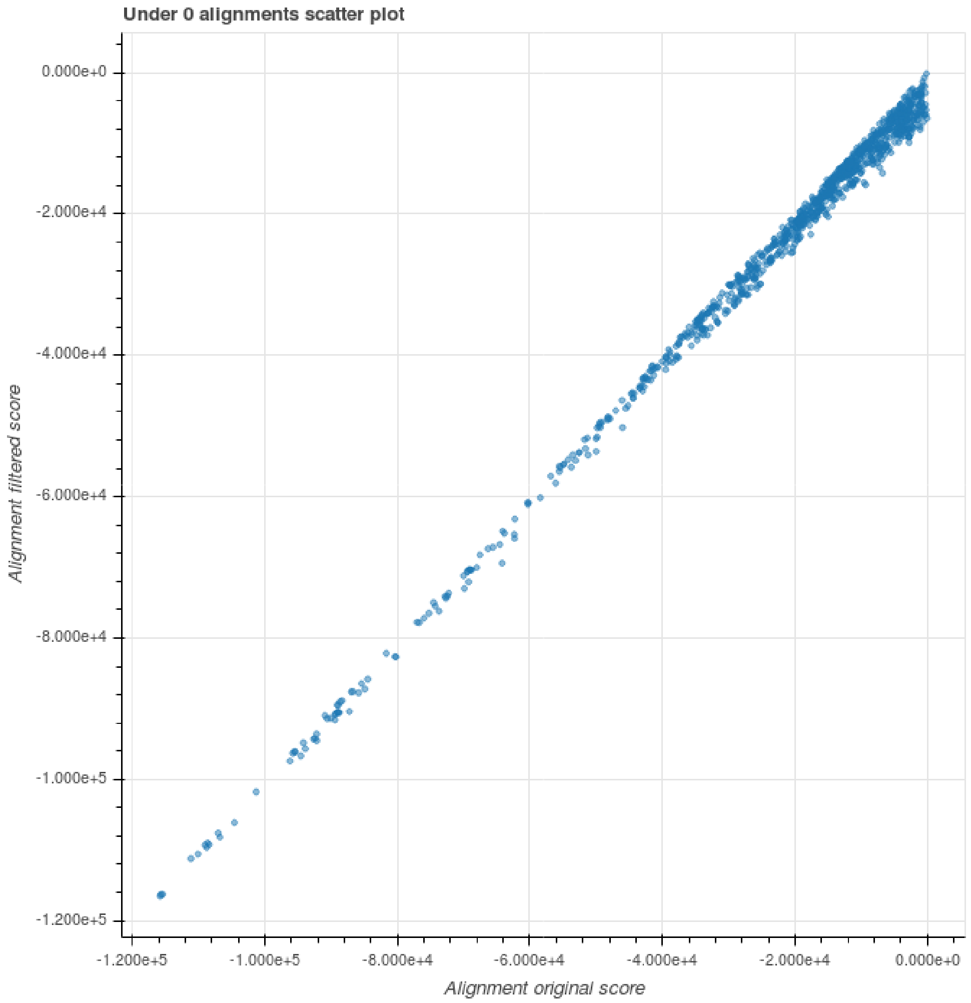


Figure 4.9: Scatter plot that shows the correlation between the score of the alignment with the original structures and the score with the filtered structures only for negative values.

Chapter 5

Conclusions

In Cosmo there is the integration of three tools that use different prediction methods for RNA secondary structure. To create a consensus structure also structures improved by experimental data are included. The main program is written in C++ language with the SeqAn code style, other supplementary blocks are implemented in python. The algorithm creates a consensus structure where each base pair of the RNA sequences is weighted according to the number of tools that present it. To store RNA secondary structures SeqAn graphs are used where the vertex are the nucleotides and the edges symbolize the pairings between two nucleotides bases.

The objective of Cosmo is to give to the sequence structure alignment computed by Lara and LocARNA a lighter input. The original input is a base pair probability matrix that consider for each nucleotide a base pair with all the others weighted on the probability the pair exists. Using Cosmo, the base pair probability matrices are filtered according to the conesnsus structures: only the edges present in the consensus structures are maintained in the matrices. Cosmo is tested on two different RNA sequences libraries, RMDB and Bralibase. For both libraries, all the sequences present a considerable decrease of the edges of the graph structures. The base pair probaility matrices pruned using Cosmo decrease as total number of edges of 98,3% for RMDB and 96,9% for Bralibase.

Cosmo output has been validated using LocARNA as alignment tool and the results clearly demonstrate that there is a relevant improvement of the computational time in performing alignment with the filtered matrices. Furthermore, the alignment quality is evaluated considering the differences of scores giving the original base pair probability matrices and the filtered ones. The differences in the score of alignment prove that the quality of alignments changes in an acceptable way. For this reason it is possible to assert that the consensus structure module can improve computationally the sequence structure alignment not losing accuracy.

Appendix A

Supplementary material

Sequence Name	bpp Matrix	Fixed all
seq184	2521	53
seq558	2407	81
seq36	2257	45
seq466	2497	51
seq304	2417	65
seq541	2403	81
seq39	2616	66
seq397	2571	79
seq555	2479	64
seq382	2453	74
seq130	2223	57
seq254	2259	101
seq460	2491	54
seq260	2300	76
seq14	2570	96
seq6	2381	76
seq11	2523	75
seq21	2932	65
seq133	2391	61
seq204	2076	63
seq468	2418	63
seq154	2398	57
seq236	2278	74
seq462	2443	58
seq29	1820	42
seq196	2068	64
seq22	3046	67
seq48	1965	48
seq375	2313	88
seq512	2396	55

seq495	2417	70
seq37	2508	47
seq193	1589	59
seq191	1717	59
seq345	2389	75
seq291	2139	56
seq59	2769	83
seq35	2524	52
seq268	2125	66
seq578	2519	61
seq377	2541	104
seq153	2460	67
seq27	1112	45
seq412	2575	79
seq42	2493	50
seq439	2454	59
seq8	2540	84
seq128	2202	97
seq23	3071	87
seq139	2162	50
seq101	2412	81
seq239	2266	77
seq305	2548	67
seq552	2414	59
seq420	2556	75
seq539	2507	63
seq18	2501	87
seq26	2970	80
seq137	2499	54
seq134	2340	50
seq25	2947	87
seq306	2410	64
seq414	2628	51
seq574	2554	85
seq542	2625	70
seq318	2351	68
seq433	2503	77
seq368	2379	55
seq387	2567	91
seq195	2040	87
seq409	2607	44
seq43	2438	59
seq12	1780	65
seq227	2180	109

seq24	3101	94
seq277	2133	61
seq450	2475	65
seq33	2563	71
seq28	1690	43
seq267	2153	63
seq410	2564	53
seq408	2489	73
seq132	2598	81
seq10	1935	60
seq504	2584	68
seq263	2416	72
seq34	2463	58
seq143	2077	99
seq573	2520	85
seq129	2237	56
seq443	2526	49
seq358	1931	48
seq167	2551	92
seq264	2159	72
seq544	2672	86
seq337	2196	59
seq31	2231	55
seq194	1625	60
seq548	2506	53
seq13	2385	82
seq183	2586	82
seq242	2112	38
seq398	2592	57
seq599	2527	71
seq201	1668	30
seq418	2387	47
seq271	2240	93
seq60	2777	75
seq428	2433	58
seq185	2445	62
seq116	2847	77
seq40	2641	78
seq406	2625	64
seq596	2470	54
seq274	2098	52
seq1041	828	36
seq893	979	38
seq457	987	31

seq1039	1040	34
seq823	977	48
seq968	1009	35
seq143	945	32
seq125	993	49
seq78	1408	45
seq220	927	36
seq681	995	30
seq928	873	39
seq1012	995	35
seq1037	1304	41
seq25	1427	47
seq1077	974	44
seq273	1049	42
seq961	898	29
seq1018	931	23
seq919	969	28
seq252	1045	58
seq884	870	42
seq340	962	26
seq156	956	38
seq434	992	31
seq189	992	44
seq86	1153	57
seq1067	1013	38
seq441	986	37
seq326	940	25
seq303	954	25
seq548	1001	42
seq274	999	31
seq906	894	31
seq375	984	42
seq904	990	21
seq913	948	31
seq370	913	33
seq186	1287	47
seq146	899	50
seq452	916	26
seq898	989	49
seq151	1006	53
seq948	1012	30
seq553	971	27
seq84	1012	44
seq1105	1355	29

seq967	979	52
seq363	993	30
seq911	981	31
seq342	1026	32
seq936	1036	27
seq950	961	30
seq485	992	32
seq446	988	41
seq155	894	31
seq276	1013	37
seq180	1213	51
seq728	928	30
seq756	902	43
seq139	1054	46
seq348	947	34
seq259	979	34
seq1066	909	33
seq104	946	36
seq758	958	41
seq789	960	37
seq366	978	47
seq292	964	23
seq749	1068	32
seq1031	939	36
seq178	935	33
seq53	982	26
seq754	923	29
seq304	925	31
seq918	1011	52
seq994	967	31
seq953	962	25
seq844	993	25
seq360	960	40
seq499	1025	34
seq325	978	47
seq1095	991	34
seq335	939	26
seq47	1042	35
seq670	1029	49
seq469	981	29
seq365	1033	54
seq1075	1010	27
seq648	942	40
seq288	958	28

seq1092	983	39
seq963	943	27
seq98	1008	54
seq101	1003	44
seq290	941	33
seq875	1019	48
seq447	984	47
seq751	1291	52
seq663	931	36
seq1098	927	27
seq1036	864	36
seq929	996	28
seq900	974	33
seq159	972	33
seq416	1026	49
seq60	959	40
seq753	966	30
seq255	929	42
seq734	961	28
seq1093	977	36
seq124	987	28
seq14	979	38
seq864	1185	51
seq462	886	43
seq316	1016	28
seq674	983	34
seq122	999	39
seq3	990	39
seq277	1015	38
seq862	962	32
seq94	914	32
seq1007	967	48
seq969	989	43
seq387	971	37
seq2	971	41
seq976	846	29
seq1058	974	34
seq294	991	59
seq389	935	26
seq791	869	35
seq250	1055	39
seq70	1023	39
seq127	910	27
seq947	952	35

seq137	912	33
seq19	1036	53
seq414	939	44
seq5	1043	52
seq11	1313	53
seq213	994	29
seq1107	901	26
seq188	982	41
seq470	939	39
seq282	1018	30
seq66	994	59
seq999	954	44
seq128	981	69
seq286	961	36
seq395	953	33
seq112	957	32
seq517	989	48
seq874	932	39
seq965	1477	44
seq436	986	54
seq678	1014	48
seq106	964	36
seq403	975	25
seq138	1012	46
seq886	1036	41
seq334	943	47
seq1070	1247	51
seq692	935	29
seq56	996	38
seq601	852	32
seq451	1035	48
seq1069	997	35
seq246	986	37
seq1046	945	38
seq944	913	30
seq153	876	41
seq428	915	25
seq672	923	37
seq305	933	26
seq406	1037	30
seq297	969	39
seq161	1078	32
seq725	824	24
seq1016	969	58

seq119	955	37
seq475	975	43
seq321	963	43
seq935	951	24
seq839	820	33
seq1050	986	52
seq54	1011	32
seq846	965	49
seq353	933	24
seq914	943	26
seq916	1266	49
seq792	998	43
seq26	971	38
seq248	1010	29
seq761	995	38
seq1010	938	35
seq1106	958	35
seq162	932	45
seq272	953	43
seq619	971	53
seq964	1025	34
seq980	965	29
seq142	1029	32
seq559	969	43
seq1074	1295	60
seq158	1276	42
seq171	992	49
seq440	910	53
seq41	977	33
seq202	2691	53
seq65	2576	68
seq119	3500	74
seq69	2438	46
seq118	2629	44
seq150	2755	59
seq88	3080	61
seq2	2005	62
seq108	1830	54
seq0	2398	47
seq5	1892	65
seq134	2573	52
seq82	2590	54
seq47	2822	60
seq129	2995	46

seq120	2933	88
seq122	2603	75
seq7	1712	59
seq154	2752	63
seq184	2591	79
seq187	2252	99
seq106	2775	59
seq44	2325	44
seq143	2719	62
seq67	2491	48
seq23	2112	47
seq125	3144	71
seq138	3343	92
seq76	3305	93
seq93	2835	49
seq41	2480	58
seq84	2186	38
seq147	2719	60
seq152	2491	47
seq144	2688	67
seq153	2712	64
seq142	3402	104
seq96	2391	69
seq107	2224	55
seq565	2469	52
seq536	2462	71
seq4	1316	43
seq484	2513	76
seq5	1301	46
seq476	2326	55
seq479	2535	78
seq510	2440	49
seq103	2371	52
seq315	2447	62
seq543	2440	45
seq584	2591	55
seq119	2496	66
seq2	1259	46
seq158	2417	88
seq473	2505	84
seq523	2516	67
seq480	2396	50
seq86	2479	92
seq365	2387	60

seq395	1935	67
seq216	2334	55
seq287	2225	75
seq198	1481	57
seq352	2393	75
seq342	2368	58
seq529	2436	94
seq0	1375	55
seq566	2546	102
seq1	1255	53
seq353	2429	49
seq142	2388	94
seq577	2413	82
seq3	2169	62
seq942	881	31
seq505	839	27
seq694	823	36
seq1000	931	57
seq708	778	33
seq534	789	33
seq179	1291	41
seq327	909	28
seq767	920	27
seq738	736	25
seq91	1224	51
seq800	940	29
seq951	1019	33
seq1040	1389	43
seq165	1398	38
seq242	1000	47
seq449	1100	35
seq1026	1475	39
seq211	1120	31
seq466	973	45
seq843	1459	53
seq747	1014	54
seq592	1099	33
seq869	928	34
seq1109	1271	28
seq683	982	29
seq105	809	23
seq668	780	46
seq411	1063	33
seq769	820	32

seq880	937	43
seq740	676	20
seq782	1064	58
seq571	828	38
seq484	1011	43
seq828	876	28
seq514	899	30
seq750	935	27
seq59	1245	31
seq927	989	24
seq573	972	30
seq801	870	47
seq50	1022	35
seq367	1000	29
seq744	703	26
seq1097	901	22
seq264	1174	44
seq307	854	30
seq200	1288	29
seq1003	1633	47
seq599	817	41
seq110	1425	54
seq798	1003	21
seq1051	954	30
seq443	1249	44
seq1002	980	36
seq433	952	44
seq1088	1288	35
seq355	920	46
seq810	888	26
seq543	812	35
seq315	935	39
seq0	1254	50
seq763	932	33
seq778	932	30
seq812	925	30
seq358	1022	40
seq834	891	25
seq97	1074	48
seq232	1653	55
seq193	2234	55
seq22	2533	40
seq124	2256	58
seq198	2506	62

seq8	1820	44
seq3	2094	59
seq163	2551	57
seq63	2514	47
seq234	2146	63
seq180	2116	61

Table A.1: Decrease of the edges from the base pair probability graphs to the consensus ones for the sequences of Bralibase.

Bibliography

[1] $https://www.rnasociety.org/about/what-is-rna/$

[2] John S. Mattick, Igor V. Makunin. *Non-coding RNA*. Human Molecular Genetics, 15:R17-R29, 2006.

[3] Alberts B, Johnson A, Lewis J, Raff M, Roberts K, Walter P. . *DNA, Chromosomes and Genomes.* Molecular Biology of the Cell. 2014

[4] $https://www.researchgate.net/figure/Gene-expression-according-to-the-central-dogma-of-molecular-biology-The-linear-view-of_{f}ig1_{2}77896017$

[5] Clancy S. *RNA Functions.* Nature Education, 1(1):102, 2008.

[6] $http://nauacmrocks.azurewebsites.net/bio_{l}ab.html$

[7] Luscombe NM, Greenbaum D., Gerstein M. . *What is bioinformatics? A proposed definition and overview of the field.* .

[8] $https://courses.lumenlearning.com/microbiology/chapter/structure-and-function-of-rna/.$

[9] Chun Kit Kwok, Yin Tang, Sarah M. Assmann, Philip C. Bevilacqua. *The RNA structurome: transcriptome-wide structure probing with next-generation sequencing.*. Trends in Biochemical Sciences 2015 RNA Secondary Structure.

[10] Ivo L. Hofacker, Peter F. Stadler. *RNA Secondary Structure.*.

[11] Barciszewski J., Frederic B., Clark C. *RNA biochemistry and biotechnology.* Springer, 72-87, 1999.

[12] $http://nchsbands.info/new/nucleic-acid-molecule.html.$

[13] Amaral PP, Dinger ME, Mercer TR, Mattick JS. *RNA Sequencing and Analysis.* Science, 319(5871):1787-9, 2008.

[14] $http://sjesci.wikispaces.com/GeneticsYr10.$
Hansen TM, Baranov PV, Ivanov IP, Gesteland RF, Atkins JF (May 2003). .

[15] Hansen TM, Baranov PV, Ivanov IP, Gesteland RF, Atkins JF. *RNA Sequencing and Analysis.* Maintenance of the correct open reading frame by the ribosome. EMBO Reports. 4(5):499-504, 2003.

[16] Berk V, Cate JH. *Insights into protein biosynthesis from structures of bacterial ribosomes.*. Current Opinion in Structural Biology. 17(3):302-9, 2007.

[17] Anfinsen CB. *The formation and stabilization of protein structure.*. The Biochemical Journal. 128(4):737-49, 1972. 2015

[18] $http://www.bioinf.man.ac.uk/resources/phase/manual/node72.html.$

[19] $https://en.wikipedia.org/wiki/Nucleic_{a}cid_{s}econdary_{s}tructure.$

[20] Rivas E, Eddy SR . *A dynamic programming algorithm for RNA structure prediction including pseudoknots.* J Mol Biol. 285: 2053-2068, 1999.

[21] $https : //openi.nlm.nih.gov/detailedresult.php?img = PMC1261154_1471 - 2105 - 6 - 224 - 1\&req = 4$.

[22] Kimberly R. Kukurba and Stephen B. Montgomery. *RNA Sequencing and Analysis.* 2015

[23] $http : //seqan.readthedocs.io/en/master/Tutorial/DataStructures/Alignment/ScoringSchemes.ht$

[24] Ruth Nussinov and Ann B. Jacobson. *Fast algorithm for predicting the secondary structure of single-stranded RNA.*

[25] M.S. Waterman, T.F. Smith. *RNA secondary structure: a complete mathematical analysis.*. Mathematical Biosciences, 42:257-266, 1978

[26] Ronny Lorenz, Michael T. Wolfinger, Andrea Tanzer, Ivo L. Hofacker. *Predicting RNA secondary structures from sequence and probing data.*.

[27] M. Zuker, P. Stiegler. *Optimal computer folding of large RNA sequences using thermodynamics and auxilary information.*. Nucleic Acids Res., 9(1):133-147, 1981

[28] J. S. McCaskill. *The equilibrium partition function and base pair binding probabilities for RNA secondary structure.* Biopolymers, 1990.

[29] Monir Hajiaghayi, Anne Condon, and Holger H. Hoos. *Analysis of energy-based algorithms for RNA secondary structure prediction.*

[30] Meiling Piao, Lei Sun, Qiangfeng Cliff Zhang. *RNA Regulations and Functions Decoded by Transcriptome-wide RNA Structure Probing.*

[31] Kengo Sato, Yuki Kato, Michiaki Hamada, Tatsuya Akutsu and Kiyoshi Asai. *IPknot: fast and accurate prediction of RNA secondary structures with pseudoknots using integer programming.* Bioinformatics. 27:i85-i93, 2011.

[32] Hofacker, I. L., Fontana, W., Stadler, P. F., Bonhoeffer, L. S., Tacker, M., Schuster, P. *Fast folding and comparison of RNA secondary structures.* Monatshefte für Chemie / Chemical Monthly. 125(2):167-188, 1994

[33] Jessica S Reuter and David H Mathews . *RNAstructure: software for RNA secondary structure prediction and analysis.* BMC Bioinformatics, 2010

[34] Yang Wu, Rihao Qu, Yiming Huang, Binbin Shi, Mengrong Liu, Yang Li and Zhi John Lu. *RNAex: an RNA secondary structure prediction server enhanced by high-throughput structure-probing data.*. Nucleic Acids Research, 2016

[35] Pablo Cordero Julius B. Lucks Rhiju Das. *An RNA Mapping DataBase for curating RNA structure mapping experiments.*. Bioinformatics, 8:3006-3008, 2012

[36] Chun Kit Kwok, Yin Tang, Sarah M. Assmann and Philip C. Bevilacqua. *The RNA structurome: transcriptome-wide structure probing with next-generation sequencing.*. Trend in Biochemical Sciences, 40:221-232, 2015

[37] Merino EJ, Wilkinson KA, Coughlan JL, Weeks KM. *RNA structure analysis at single nucleotide resolution by selective 2'-hydroxyl acylation and primer extension (SHAPE).* J Am Chem Soc. 127:4223-31, 2005

[38] Kevin M. Weeks and David M. Mauger. *Exploring RNA Structural Codes with SHAPE Chemistry.* 2012

[39] Katherine E. Deigan, Tian W. Li, David H. Mathews, and Kevin M. Weeks. *Accurate SHAPE-directed RNA structure determination.* 2008

[40] Knut Reinert, Temesgen Hailemariam Dadi, Marcel Ehrhardt, Hannes Hauswedell, Svenja Mehringer, René Rahn, Jongkyu Kimb, Christopher Pockrandt, Jörg Winkler,

Enrico Siragusa, Gianvito Urgese, David Weese. *The SeqAn C++ template library for efficient sequence analysis: A resource for programmers.* Journal of Biotechnology, 261:157-168, 2017.

[41] Andreas Döring, David Weese, Tobias Rausch and Knut Reinert. *SeqAn An efficient, generic C++ library for sequence analysis.* BMC Bioinformatics, 2008.

[42] Markus Bauer, Gunnar W Klau and Knut Reinert *Accurate multiple sequence-structure alignment of RNA sequences using combinatorial optimization.* BMC Bioinformatics, 2007.

[43] Cameron Smith Steffen Heyne Andreas S. Richter Sebastian Will Rolf Backofen. *Freiburg RNA Tools: a web server integrating IntaRNA, ExpaRNA and LocARNA.* Nucleic Acids Research, 38:W373-W377, 2010.

[44] Andreas Wilm, Indra Mainz and Gerhard Steger. *An enhanced RNA alignment benchmark for sequence alignment programs.* 2006

[45] Novick RP, Iordanescu S, Projan SJ, Kornblum J, Edelman I. *pT181 plasmid replication is regulated by a countertranscript-driven transcriptional attenuator.* 1989

[46] Volker A. Erdmann Miroslawa Z. Barciszewska Maciej Szymanski Abraham Hochberg Nathan de Groot Jan Barciszewski. *The non-coding RNAs as riboregulators.* Nucleic Acids Research. 29:189-193

[47] Nudler E, Mironov AS . *The riboswitch control of bacterial metabolism..* Trends Biochem Sci., 2004.

[48] James W. Brown, Amanda Birmingham, Paul E. Griffiths, Fabrice Jossinet, Rym Kachouri-Lafond, Rob Knight, B. Franz Lang, Neocles Leontis, Gerhard Steger, Jesse Stombaugh, and Eric Westhof. *The RNA structure alignment ontology.* RNA.15(9): 1623-1631, 2009.