ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ
ΣΧΟΛΗ ΗΛΕΚΤΡΟΛΟΓΩΝ ΜΗΧΑΝΙΚΩΝ ΚΑΙ ΜΗΧΑΝΙΚΩΝ ΥΠΟΛΟΓΙΣΤΩΝ

Τομέας Τεχνολογίας Πληροφορικής και Υπολογιστών

Πρόβλεψη της δευτεροταγούς δομής ακολουθιών RNA με έμφαση στο μοτίβο ψευδοκόμβων, με τη χρήση τεχνικών συντακτικής αναγνώρισης προτύπων

ΔΙΔΑΚΤΟΡΙΚΗ ΔΙΑΤΡΙΒΗ

του

Ευάγγελου Ι. Μακρή

Αθήνα, Ιούλιος 2024

Εθνικο Μετσοβιο Πολυτεχνειο
ΣΧΟΛΗ ΗΛΕΚΤΡΟΛΟΓΩΝ ΜΗΧΑΝΙΚΩΝ ΚΑΙ ΜΗΧΑΝΙΚΩΝ ΥΠΟΛΟΓΙΣΤΩΝ
Τομέας Τεχνολογίας Πληροφορικής και Υπολογιστών

Πρόβλεψη της δευτεροταγούς δομής ακολουθιών RNA με έμφαση στο μοτίβο ψευδοκόμβων, με τη χρήση τεχνικών συντακτικής αναγνώρισης προτύπων

ΔΙΔΑΚΤΟΡΙΚΗ ΔΙΑΤΡΙΒΗ

του

Ευάγγελου Ι. Μακρή

Συμβουλευτική Επιτροπή: Παναγιώτης Τσανάκας
Ηλίας Μαγκλογιάννης
Χρήστος Παυλάτος

Εγκρίθηκε από την επταμελή εξεταστική επιτροπή την 3η Ιουλίου 2024.

.....	.....	.....
Παναγιώτης Τσανάκας	Ηλίας Μαγκλογιάννης	Χρήστος Παυλάτος
Καθηγητής Ε.Μ.Π.	Καθηγητής Παν.Πειραιώς	Επ. Καθηγητής Σχολή Ικάρων
.....	.....	.....
Ανδρέας Σταφυλοπάτης	Γεώργιος Ματσόπουλος	Δημήτριος Σούντρης
Καθηγητής Ε.Μ.Π.	Καθηγητής Ε.Μ.Π.	Καθηγητής Ε.Μ.Π.
	.....	
	Σωτήριος Ξύδης	
	Επ. Καθηγητής Ε.Μ.Π.	

Αθήνα, Ιούλιος 2024

Δημοσιεύσεις Υποψήφιου Διδάκτορα

[1] Makris, E., Kolaitis, A., Andrikos, C., Moulos, V., Tsanakas, P., Pavlatos, C. An intelligent grammar-based platform for RNA H-type pseudoknot prediction. *Artificial Intelligence Applications and Innovations.* AIAI 2022 IFIP WG 12.5 International Workshops. IFIP Advances in Information and Communication Technology **2022**, *vol 652*, Springer.

[2] Makris, E., Kolaitis, A., Andrikos, C., Moulos, V., Tsanakas, P., Pavlatos, C. Knotify+: Toward the Prediction of RNA H-Type Pseudoknots, Including Bulges and Internal Loops. *Biomolecules.* **2023**, *13(2),* 308.

[3] Andrikos, C., Makris, E., Kolaitis, A., Rassias, G., Pavlatos, C., Tsanakas, P. Knotify: An Efficient Parallel Platform for RNA Pseudoknot Prediction Using Syntactic Pattern Recognition. *Methods Protoc.* **2022**, *5*, 14

[4] Koroulis, C., Makris, E., Kolaitis, A., Tsanakas, P., Pavlatos, C. Syntactic Pattern Recognition for the prediction of L-type pseudoknots in RNA. *Appl. Sci.* **2023**, *13,* 5168.

[5] Pavlatos, C., Makris, E., Fotis, G., Vita, V., Mladenov, V. Utilization of Artificial Neural Networks for Precise Electrical Load Prediction. *Technologies.* **2023**, *11*, 70.

[6] Pavlatos, C., Makris, E., Fotis, G., Vita, V., Mladenov, V. Enhancing Electrical Load Prediction Using a Bidirectional LSTM Neural Network. *Electronics.* **2023**, *12*, 4652.

[7] Kolaitis, A., Makris, E., Karagiannis, A.A., Tsanakas, P., Pavlatos, C. Knotify_V2.0: Deciphering RNA Secondary Structures with H-type Pseudoknots and Hairpin Loops. *Genes.* **2024**, *15(6),* 670.

Ευάγγελος Ι. Μακρής
Διδάκτωρ Ηλεκτρολόγος Μηχανικός και Μηχανικός Υπολογιστών Ε.Μ.Π.
.....

Ευχαριστίες

Για την ολοκλήρωση της διατριβής αυτής αφιερώθηκε ένα σημαντικό χρονικό διάστημα και ενέργεια τόσο του συγγραφέα όσο και όλης της ερευνητικής ομάδας του Εθνικού Μετσόβιου Πολυτεχνείου. Προέκυψαν πολλές δυσκολίες, όπως αναμένεται σε μία διδακτορική διατριβή, αλλά με σωστή καθοδήγηση και ενθάρρυνση προέκυψε το επιθυμητό αποτέλεσμα που παρουσιάζεται στη συνέχεια. Αρχικά, θα ήθελα να ευχαριστήσω την οικογένειά μου, που από την αρχή αυτής της προσπάθειας παρείχε κάθε δυνατή βοήθεια για να εξασφαλιστεί ο χρόνος και οι κατάλληλες συνθήκες για την επιστημονική αυτή έρευνα. Περιορισμοί χωρικοί και χρονικοί καθώς και πανδημίες και εγκλεισμοί, απλά αναδιάρθρωσαν τον τρόπο διεξαγωγής της μελέτης δίνοντας κάθε φορά διαφορετικά κίνητρα για εναλλακτικές μεθόδους έρευνας και συνεργασίας. Θα ήθελα να εκφράσω την ευγνωμοσύνη μου στον επιβλέποντα μου καθηγητή της σχολής και κοσμήτορα του τμήματος κ. Παναγιώτη Τσανάκα για τη συνεχή καθοδήγηση και την υπομονή του σε κάθε δύσκολη συνθήκη που προέκυψε σε αυτή τη μακρά περίοδο. Επίσης, μεγάλη ήταν η συνεισφορά του επίκουρου καθηγητή κ. Χρήστου Παυλάτου και του καθηγητή κ. Ηλία Μαγκλογιάννη μελών της τριμελούς επιτροπής. Ο κ. Παυλάτος από την πρώτη μέρα της εκπόνησης της διατριβής ήταν ο κύριος πυλώνας υποστήριξης όλων των ζητημάτων, τόσο ερευνητικών όσο και τεχνικών που προέκυπταν και λειτούργησε ως μέντορας όλης της ερευνητικής ομάδας. Ο κ. Μαγκλογιάννης με την μεγάλη εμπειρία του, συνεισέφερε με τις συμβουλές του όποτε η ομάδα το χρειαζόταν και έδινε ουσιαστικές λύσεις και προσεγγίσεις. Όσον αφορά τη συγκεκριμένη εργασία, στόχος έχει την καλλιέργεια μίας κουλτούρας ευρύτερης διεπιστημονικής συνεργασίας μεταξύ επιστημόνων της πληροφορικής και της βιολογίας. Στο πλαίσιο αυτό θα ήθελα να ευχαριστήσω όλους τους συνεργάτες, ερευνητές και μηχανικούς με τους οποίους είτε συζητήσαμε είτε συνεργαστήκαμε, τόσο για την θετική τους προσέγγιση, τις ιδέες τους αλλά και για τη συνεισφορά τους στην υλοποίηση των προτεινόμενων συστημάτων.

Περιεχόμενα

Κατάλογος Σχημάτων

Κατάλογος Πινάκων

Abstract

Accurate "base pairing" in RNA molecules, which is an essential task in order to predict RNA secondary structures, plays a crucial role in explaining unknown biological processes. The COVID-19 pandemic, a widespread disease, has had a devastating impact on humanity in recent years. This extreme condition enhances the significance of analyzing RNA molecules, and their structures have been highlighted by SARS-CoV-2, a single-stranded RNA virus that causes COVID-19 disease. This thesis aims to develop a pioneering set of frameworks that leverage syntactic pattern recognition to predict specific RNA structures. Specifically, it focuses on introducing novel systems for predicting RNA secondary structure patterns including pseudoknots, utilizing syntactic pattern-recognition strategies, and concepts such as maximum base pairing and minimum free energy. By primarily treating pseudoknot predictions as parsing and optimization problems, the proposed methodologies formalize the prediction of RNA secondary structures. The first developed framework, called Knotify, addresses the prediction of first-order pseudoknots (H-type). It introduces a context-free grammar (CFG) capable of recognizing potential pseudoknot patterns. Knotify exhibits prediction capabilities similar to state-of-the-art frameworks, but significantly outperforms them in terms of execution time efficiency. An optimized version of Knotify is also presented, utilizing pruning techniques for further efficiency gains. To predict more complex patterns of H-type pseudoknots, Knotify+ introduced, which extends the power of CFG by searching for intricate patterns in the pseudoknot's loops such as bulges and internal loops, while incorporating the advantages of maximum base pairing and minimum free energy. Knotify+ demonstrates superior accuracy in predicting core stems compared to existing frameworks. It performs exceptionally well in small sequences and maintains a comparable accuracy rate in larger ones, all while requiring shorter execution times than well-known platforms. Finally, a new grammar-based system is presented to address the prediction of a rare but complex type of pseudoknot, the L-type pseudoknot, that proves effective in prediction and execution time when compared to other novel systems. These innovative systems and architectures aim to improve the ability of biologists to predict RNA motifs, including pseudoknots. They hold the potential for application in various biological domains, such as gene therapy, drug design, and understanding RNA functionality. Moreover, these approaches can be combined with other methodologies to improve the precision of RNA structure prediction. To support this initiative, the Knotify, Knotify+, L-type prediction framework source codes, and implementation details are available online as public repositories on GitHub.
Keywords: H-type pseudoknot structure, L-type pseudoknot structure, RNA, parsers, CFG, syntactic pattern recognition.

Chapter 1

Extended Introduction

RNA is an important molecule with significant contributions to many biological processes. Genetic information encoded in DNA is transcribed into mRNA, which in turn passes this information to the cytoplasm for protein production through the translation procedure. This fundamental process of molecular biology is also known as "central dogma" [1]. Beyond this key contribution, RNA is involved in critical biological processes including the regulation of gene expression, recognition of specific sites, and catalysis [2, 3]. All RNA types that have been examined by scientists in recent years, aside from mRNA, are referred to as non-coding RNAs, as their functions are related to other aspects than protein synthesis, highlighting the importance of a deep understanding of these molecules. A crucial task in the unveiling of RNA functionality is predicting its structure at the three-dimensional (3D) base, called tertiary structure. Techniques such as X-ray crystallography [4] and nuclear magnetic resonance [5] are employed to determine the complex structure of the molecule. To tackle the impediments of these methods, emphasis has been given to predicting RNA's secondary structure in two dimensions, i.e. predicting its secondary structure. This simpler representation consists of base pairs (A-U, C-G, and G-U) that form structural motifs including loops, bulges, and hairpins. Identifying the precise locations of these base pairs and the respective motifs is crucial in unveiling the 3D structure and the undiscovered functional mechanisms of RNA.

Recent approaches in the prediction of RNA secondary structure have focused on the solution of the problem by using scoring functions based on thermodynamics, probabilistic models, or artificial intelligence (AI). Facilitated by a variety of methods, the minimum free energy algorithm, initially introduced by Zuker, is a key methodology. This algorithm benefits from dynamic programming, optimized with experimental parameters [7]. Additionally, the Nussinov algorithm, which aims at the identification of structure with the maximal number of base pairs through dynamic programming [8], has shown improved performance when integrated with other, more complex algorithms [9]. In literature, methods utilizing stochastic methods, syntactic pattern recognition, machine learning, statistical

techniques, integer programming, and various heuristic algorithms for structure prediction have been proposed. For a comprehensive review of these methodologies, see Section 3.2.

Predicting pseudoknots, a complex structural motif, within RNA secondary structures is a very demanding task. While algorithms can accurately predict common motifs such as stems, hairpins, internal loops, and multibranch loops, when it comes to pseudoknots, the complexity arises because dynamic programming and minimum-free energy algorithms are not inherently designed to deal with their interconnected structure. Furthermore, as RNA sequences get larger, these algorithms require exponentially more time to execute. To address the challenge of accurately and efficiently predicting pseudoknots, this thesis aims to develop platforms for this purpose. These platforms, Knotify, Knotify+, and Knotify for L-type, focus on predicting H-type and L-type pseudoknots, along with bulges and internal loops. They aim to match the accuracy of established methods while significantly improving execution time efficiency.

The H-type pseudoknot, as described by [17], is characterized by its formation from two stems and two loops of varying lengths, emerging from the intersection of two base pairs (referred to as core stems here). Although pseudoknots are a common structural pattern, they serve as the basis for a variety of unique and stable RNA structures. These structures are known for their diversity in length and configuration of loops and stems (the hydrogen-bonded base pairs), playing crucial roles in numerous biological processes. This includes acting as catalysts in various ribozymes [10, 11], facilitating self-splicing of introns [12], and functioning within telomerase [13]. Furthermore, pseudoknots are significant in modulating gene expression in many viruses, sometimes even defining the mechanism [14, 15, 16].

Moreover, the bulge loop, or bulge, is another crucial motif found in RNA secondary structures. This feature emerges when a sequence of unpaired nucleotides disrupts one strand of a helix, a common occurrence across RNA secondary structures [18, 19]. Bulges are universally present in all types of structured functional RNA, indicating their widespread significance [20]. Similarly, base-base mismatches that form internal loops are prevalent and influence the stability of the molecule [18].

The study of bulges is of particular interest due to the frequent presence of bulged adenosine residues at RNA's protein-binding sites [18, 21]. These residues also play a pivotal role in the tertiary folding of RNA, serving as critical contact points [19, 23]. Bulges contribute to the formation of unique recognition sites within RNA's tertiary structure by acting as molecular handles within helical regions and, indirectly, by altering the RNA backbone. This alteration gives access to base pairs within an expanded deep groove [20]. Moreover, helical segments divided by bulges often transition between unstacked and coaxially stacked arrangements, playing a vital role in the folding and functionality of noncoding RNAs [20]. These references underscore the vital role of pseudoknots and motifs such as bulges and internal loops. They are fundamental structural components of RNAs and high-

light the critical importance of these elements in the architecture of RNAs and molecular recognition.

The existing methods for RNA structure prediction have utilized diverse algorithms, including thermodynamic, probabilistic, and AI-based approaches, to varying degrees of success. However, pseudoknot prediction remains a challenging task due to the complex interconnections and exponential execution time requirements. To address this gap, this thesis introduces a groundbreaking suite of frameworks: Knotify, Knotify+, and Knotify for L-type pseudoknots. These platforms employ novel syntactic pattern recognition strategies, combining maximum base pairing and minimum free energy concepts to accurately predict pseudoknots, bulges, and internal loops. By treating pseudoknot predictions as parsing and optimization problems, the proposed methodologies offer enhanced efficiency and performance, surpassing existing frameworks in execution time while maintaining comparable accuracy rates. Furthermore, a dedicated grammar-based system is presented for the prediction of L-type pseudoknots, showcasing its effectiveness and computational efficiency. These innovative approaches aim to empower biologists in predicting RNA motifs, facilitating advancements in gene therapy, drug design, and comprehensive understanding of RNA functionality. By making the source codes and implementation details publicly available, this research fosters collaboration and encourages the integration of these methodologies with other prediction techniques, ultimately advancing the precision of RNA structure prediction and its diverse applications in biological domains.

In this thesis, the theoretical foundation for the prediction of the RNA structure is explored in Chapter 2, providing the necessary background to understand the subsequent developments. Chapter 5 introduces Knotify, the first developed framework that addresses the prediction of first-order pseudoknots (H-type) using a context-free grammar (CFG) approach. The optimization of execution time through pruning techniques is presented in Chapter 6. Building upon the achievements of Knotify, Chapter 7 presents Knotify+, an extension that incorporates bulges and internal loops to predict more complex patterns of H-type pseudoknots. Additionally, an example demonstration showcases the effectiveness of an extension for predicting pseudoknots of type L in Chapter 8. Finally, Chapter 9 summarizes the findings of this thesis and outlines avenues for future research in the field of RNA structure prediction. The accurate prediction of RNA secondary structures, including pseudoknots, is a challenging task with significant implications in the field of genomics and healthcare. In this work, three innovative methodologies, namely Knotify, Knotify+, and Knotify for L-type, were developed to detect and predict different types of pseudoknot structures in RNA. These systems were evaluated based on several metrics, including the prediction of pseudoknot core stems, accuracy of base pair prediction, execution time, precision, recall, F1-score, and Matthews correlation coefficient (MCC). For the prediction of pseudoknot core stems, Knotify showed promising results, successfully detecting the core stems in 143 out of 262 sequences. It outperformed other widely used methodologies, such

as Knotty, HotKnots, IPknot, and IHFold. Knotify exhibited better accuracy in predicting core stems in three of four groups of RNA sequences, with performance comparable to Knotty in the remaining group. This indicates that Knotify has the potential to accurately identify the core stems of pseudoknots in various RNA sequences. Regarding base pair prediction, Knotify demonstrated excellent performance in terms of precision, achieving an average precision metric of 0.784. Although Knotty had a higher F1 score and MCC, Knotify achieved an accuracy very close to Knotty. Notably, Knotify showcased better precision in all ranges of RNA sequence lengths, while Knotty excelled primarily in larger RNA sequences. This suggests that Knotify is a reliable method for base pair prediction, particularly for smaller RNA sequences. In terms of execution time, Knotify proved to be efficient, outperforming Knotty with a speed ratio of 7.76. It also demonstrated comparable or better execution times than IPknot and HotKnots. This indicates that Knotify is a time-efficient solution for pseudoknot prediction in RNA sequences.

Moving on to Knotify+, the enhanced version of Knotify, it further improved the prediction of pseudoknot core stems. Knotify+ successfully predicted both core stems in 142 of 260 sequences, surpassing IPknot and Knotty in terms of core stem detection. In addition, Knotify+ excelled in predicting at least one core stem, outperforming Knotify, IPknot, and Knotty. This highlights the improved ability of Knotify+ to predict pseudoknot core stems, even in cases where exact prediction is challenging. When considering the evaluation metrics for Knotify+, it demonstrated superior performance compared to Knotify. Knotify+ achieved higher recall, F1-score, and MCC, reducing the gap with Knotty, which still exhibited better performance in these metrics. Knotify+ maintained better precision than Knotty, similar to Knotify. This indicates that Knotify+ improves overall prediction accuracy, particularly in terms of recall, F1-score, and MCC. However, Knotty outperformed all methods in larger RNA sequences, likely due to its ability to detect additional complex patterns, such as hairpins, that Knotify+ does not capture in this version. Future work aims to enhance Knotify+ to handle more intricate patterns within pseudoknot loops.

Finally, for the prediction of pseudoknots of type L, Knotify for type L showed promising results, surpassing other methods in terms of precision with a score of 0.844. The F1 score and MCC were higher for Knotify than for the other methods, while Knotty had a remarkable recall score. This suggests that Knotify for L-type pseudoknots is effective and accurate in predicting this specific type of pseudoknot. Overall, the results demonstrate the effectiveness of the developed systems (Knotify, Knotify+, and Knotify for L-type pseudoknots) in predicting RNA pseudoknots. These systems provide reliable predictions for pseudoknot core stems and base pairs while maintaining good execution times. Further improvements can be made to enhance the prediction accuracy for larger RNA sequences and more intricate pseudoknot patterns. In conclusion, the introduction of Knotify, Knotify+, and Knotify for L-type represents significant advancements in the field of RNA secondary structure prediction, particularly in pseudoknot detection. These

methodologies demonstrate improved accuracy, performance, and capabilities compared to existing methods. The successful implementation of Knotify, Knotify+, and Knotify for L-type opens up new avenues for future research, including the exploration of more complex patterns (i.e., hairpins, multiloops, K- and M-type pseudoknots), the development of advanced searching algorithms, and the creation of an accessible web platform, equipped with a modern graphical user interface, to enhance its accessibility and usability for researchers. These advances have the potential to contribute to the field of bioinformatics and genomics and foster unified collaboration between interdisciplinary teams of healthcare professionals, biologists, and IT professionals.

Περίληψη

Η ακριβής 'αντιστοίχιση ζευγών βάσεων' σε μόρια RNA, που οδηγεί στην πρόβλεψη των δευτεροταγών δομών του RNA, είναι ζωτικής σημασίας για την εξήγηση άγνωστων βιολογικών λειτουργιών. Πρόσφατα, ο COVID-19, μια ευρέως διαδεδομένη ασθένεια, προκάλεσε πολλούς θανάτους, επηρεάζοντας την ανθρωπότητα με ασύλληπτο τρόπο. Ο SARS-CoV-2, ένας ιός μονόκλωνου RNA, απέδειξε τη σπουδαιότητα της ανάλυσης αυτών των μορίων και των δομών τους. Αυτή η διατριβή στοχεύει στο να δημιουργήσει ένα σύνολο σύγχρονων εργαλείων στην κατεύθυνση της πρόβλεψης συγκεκριμένων δομών RNA, με χρήση τεχνικών συντακτικής αναγνώρισης προτύπων. Πιο συγκεκριμένα, στοχεύει να συμβάλει σε αυτό το πεδίο εισάγοντας ένα σύνολο καινοτόμων συστημάτων που αποσκοπούν στην πρόβλεψη μοτίβων δευτεροταγούς δομής RNA γνωστών ως ψευδοκόμβων RNA, με τη χρήση τεχνικών συντακτικής αναγνώρισης προτύπων και των έννοιων του μέγιστου πλήθους ζευγών βάσεων και της ελάχιστης ελεύθερης ενέργειας. Έχοντας επικεντρωθεί στις προβλέψεις ψευδοκόμβων, οι προτεινόμενες μεθοδολογίες διαμορφώνουν την πρόβλεψη της δευτεροταγούς δομής του RNA ως ένα πρόβλημα ανάλυσης και, δευτερευόντως, ως ένα πρόβλημα βελτιστοποίησης. Το πρώτο σύστημα που αναπτύχθηκε, το Knotify, αντιμετωπίζει το πρόβλημα της πρόβλεψης ψευδοκόμβων πρώτης τάξης (τύπου H). Εισάγει μια γραμματική χωρίς συμφραζόμενα (CFG) που επιτρέπει την αναγνώριση πιθανών προτύπων ψευδοκόμβων. Το Knotify παρουσιάζει μια παρόμοια ικανότητα πρόβλεψης με τα πιο σύγχρονα συστήματα αναγνώρισης της δευτεροταγούς δομής του RNA, αλλά είναι πολύ πιο αποδοτικό όσον αφορά τον χρόνο εκτέλεσης. Στη συνέχεια, παρουσιάζεται μια βελτιστοποιημένη έκδοση αυτού, κατά την οποία χρησιμοποιείται μια τεχνική κλαδέματος συντακτικών δένδρων, για περαιτέρω περιορισμό του χρόνου εκτέλεσης. Για την πρόβλεψη πιο περίπλοκων προτύπων ψευδοκόμβων τύπου H, αναπτύχθηκε το σύστημα Knotify+, το οποίο αντιμετωπίζει το πρόβλημα της πρόβλεψης ψευδοκόμβων τύπου H, συμπεριλαμβάνοντας ασύμμετρους βρόχους/εξογκώματα (bulges) και εσωτερικούς βρόχους (internal loops), εκμεταλλευόμενο τη δύναμη της γραμματικής χωρίς συμφραζόμενα (CFG). Βασίζεται στο Knotify, αλλά ενισχύει την εκφραστικότητά του αναζητώντας πιο πολύπλοκα πρότυπα στους βρόχους του ψευδοκόμβου. Τέλος, με στόχο την πρόβλεψη ενός σπάνιου, αλλά περίπλοκου τύπου ψευδοκόμβου, του ψευδοκόμβου τύπου L, παρουσιάζεται επίσης ένα νέο σύστημα βασισμένο σε γραμματική χωρίς συμφραζόμενα που αποδεικνύεται αποτελεσματικό στην πρόβλεψη και τον χρόνο εκτέλεσης σε σύγκριση με άλλα σύγχρονα συστήματα. Αυτά τα καινοτόμα συστήματα και οι ανάλογες αρχιτεκτονικές στοχεύουν στην ενίσχυση των δυνατοτήτων των βιολόγων στην πρόβλεψη μοτίβων RNA, με έμφαση στις διαφορετικές κατηγορίες ψευδοκόμβων. Έχουν τη δυνατότητα εφαρμογής σε διάφορους βιολογικούς τομείς, όπως η γονιδιακή θεραπεία, η σχεδίαση φαρμάκων και η κατανόηση της λειτουργίας του RNA. Επιπλέον, αυτές οι προσεγγίσεις μπορούν να συνδυαστούν με άλλες μεθοδολογίες για να βελτιώσουν την ακρίβεια της πρόβλεψης της δομής του RNA. Οι πηγαίοι

κώδικες του Knotify, Knotify+, του συστήματος πρόβλεψης τύπου L και οι λεπτομέρειες υλοποίησης είναι διαθέσιμοι στο διαδίκτυο σε αποθετήρια του GitHub για την αξιοποίησή τους από την κοινότητα, με σκοπό την υποστήριξη και την ενθάρρυνση αυτής της πρωτοβουλίας.

Λέξεις κλειδιά: H-τύπου ψευδοκόμβοι, L-τύπου ψευδοκόμβοι, RNA, συντακτικοί αναλυτές, γραμματικές χωρίς συμφραζόμενα, συντακτική αναγνώριση προτύπων.

Κεφάλαιο 2

Εισαγωγή

Το RNA και οι λειτουργίες του διαδραματίζουν έναν πολύ σημαντικό ρόλο σε διάφορες βιολογικές λειτουργίες. Χαρακτηριστικό παράδειγμα είναι η διαδικασία όπου τα μόρια του DNA, όπου αποθηκεύεται η γενετική πληροφορία, μεταγράφονται σε mRNA, το οποίο μεταφέρει την πληροφορία στο κυτταρόπλασμα, όπου λαμβάνει χώρα η μετάφραση και οδηγεί στην παραγωγή μιας πρωτεΐνης. Λόγω της σπουδαιότητας της διαδικασίας αυτής, ονομάζεται επίσης και κεντρικό δόγμα της μοριακής βιολογίας [1]. Εκτός όμως από αυτήν τη βασική λειτουργία, έχει αποδειχθεί ότι το RNA συμμετέχει σε ένα μεγάλο πλήθος κεντρικών βιολογικών φαινομένων, όπως ο έλεγχος της έκφρασης των γονιδίων, η αναγνώριση τοποθεσιών και η καταλυτική δράση [2, 3]. Όλα τα RNA που εκτελούν τις παραπάνω λειτουργίες, εκτός του mRNA, ονομάζονται μη κωδικοποιητικά, διότι εκτελούν λειτουργίες πέρα από την κωδικοποίηση πρωτεϊνών, κάτι που επιβάλλει τη λεπτομερή ανάλυσή τους. Σε αυτό το πλαίσιο, γίνεται φανερή η ανάγκη πρόβλεψης και ανάλυσης της δομής του RNA και, ειδικότερα, της τρισδιάστατης δομής του, για να κατανοήσουμε όλες αυτές τις λειτουργίες του. Αυτή η τρισδιάστατη δομή μπορεί να καθοριστεί χρησιμοποιώντας τεχνικές όπως η αναλυτική ακτινογραφία ακτινών X [4] και η πυρηνική μαγνητική κυρτότητα [5]. Ωστόσο, οι ερευνητές έχουν επικεντρωθεί στην ανάπτυξη μιας μεθοδολογίας για την πρόβλεψη μιας πιο απλής αναπαράστασης της δομής του RNA σε ένα δισδιάστατο χώρο, που ονομάζεται δευτεροταγής δομή, η οποία αποτελείται από τις βάσεις A (Αδενίνη), U (Ουρακίλη), G (Γουανίνη) και C (Κυτοσίνη) που σχηματίζουν διπλές περιοχές και ανοικτές περιοχές, δημιουργώντας σημαντικά μοτίβα γύρω από αυτές, όπως βρόχους, εξογκώματα και ψευδοκόμβους. Συνεπώς, η δευτεροταγής δομή αποτελείται από τα ζεύγη (A-U, C-G) και σπανιότερα από το ζευγος (G-U) που σχηματίζουν διάφορα μοτίβα. Η ακριβής θέση των ζευγών βάσεων και των μοτίβων, αντίστοιχα, αποτελεί ένα χρήσιμο ορόσημο και σημείο εκκίνησης για την ανακάλυψη της τρισδιάστατης δομής και, ως εκ τούτου, την κατανόηση των λειτουργιών του RNA.

Οι πρόσφατες μέθοδοι πρόβλεψης δευτεροταγών δομών του RNA βασίζονται κυρίως σε μια συνάρτηση βαθμολόγησης που μπορεί να βασίζεται σε θερμοδυναμικούς, πιθανοτικούς ή αλγορίθμους τεχνητής νοημοσύνης (AI). Το μεγαλύτερο μέρος των μεθόδων ενσωματώνει και προσαρμόζει έναν αλγόριθμο ελάχιστης ελεύθερης ενέργειας

που εισήγαγε ο Zuker, ο οποίος χρησιμοποιεί δυναμικό προγραμματισμό ενισχυμένο με παραμέτρους από πειράματα [7]. Ο αλγόριθμος Nussinov, επίσης ένας από τους πιο γνωστούς αλγορίθμους, στοχεύει στην πρόβλεψη της δευτεροταγούς δομής μέσω της πρόβλεψης του μεγαλύτερου αριθμού ζευγών βάσεων, χρησιμοποιώντας δυναμικό προγραμματισμό [8]. Η μεθοδολογία αυτή παρουσιάζει τη μεγαλύτερη αποδοτικότητά της, όταν συνδυάζεται ή ενσωματώνεται ως εσωτερικό συστατικό σε άλλους πιο εξελιγμένους αλγορίθμους, όπως στο [9]. Άλλες σύγχρονες προσεγγίσεις έχουν χρησιμοποιήσει στοχαστικές μεθόδους, συντακτική αναγνώριση προτύπων, μηχανική μάθηση, στατιστικές τεχνικές, ακέραιο προγραμματισμό ή άλλους ευρηματικούς αλγόριθμους για την αντιμετώπιση της διαδικασίας πρόβλεψης. Η ενότητα 3 περιλαμβάνει μια λεπτομερή ανάλυση της σχετικής βιβλιογραφίας με παραδείγματα μεθόδων από όλες τις παραπάνω κατηγορίες.

Σε μια δευτεροταγή δομή του RNA, η κατηγορία των ψευδοκόμβων είναι η πιο απαιτητική εργασία αναφορικά με την πρόβλεψη της. Άλλα κοινά μοτίβα είναι οι βρόχοι, τα εξογκώματα, οι ανακάμψεις, οι εσωτερικοί βρόχοι και οι πολλαπλοί βρόχοι, τα οποία πολλοί αλγόριθμοι είναι σε θέση να προβλέπουν με υψηλή ακρίβεια. Αντίθετα, η πρόβλεψη ενός ψευδόκομβου είναι πολύ δύσκολη, επειδή οι αλγόριθμοι δυναμικού προγραμματισμού και ελάχιστης ελεύθερης ενέργειας δεν είναι δομημένοι με τρόπο, ώστε να μπορούν να εφαρμοστούν σε ψευδοκόμβους λόγω της πολυπλοκότητας της δομής τους. Έναν άλλο σημαντικό λόγο αποτελεί το γεγονός ότι με την αύξηση του μήκους του RNA, αυτοί οι αλγόριθμοι απαιτούν εκθετικό χρόνο εκτέλεσης. Επομένως, η ανάγκη για μια ακριβή πρόβλεψη των ψευδοκόμβων οδήγησε αυτή την έρευνα στη δημιουργία μιας σειράς αλγορίθμων και αντίστοιχων πλατφορμών που προβλέπουν ψευδόκομβους τύπου H, στη συνέχεια συνδυασμένους με εξογκώματα και εσωτερικούς βρόχους και ψευδόκομβους τύπου L, με ακρίβεια παρόμοια με γνωστές μεθόδους και, ταυτόχρονα, πιο αποτελεσματικούς αναφορικά με τον χρόνο εκτέλεσης, που ονομάζονται Knotify, Knotify+ και Knotify για τον τύπο L αντίστοιχα.

Ο ψευδόκομβος τύπου H [17] αποτελείται από δύο περιοχές με συζευγμένες βάσεις και δύο βρόχους. Οι βρόχοι του ψευδοκόμβου διασταυρώνονται με αποτέλεσμα η μία πλευρά του ενός να κάνει δεσμούς υδρογόνου με τη μία πλευρά του άλλου. Παρόλο που ο ψευδόκομβος είναι ένα τυπικό μοτίβο, είναι ο εκκινητής για εντυπωσιακές αλλά ανθεκτικές δομές του RNA που συνδέονται με πληθώρα βιολογικών λειτουργιών, όπως η καταλυτική λειτουργία διαφόρων ριβοζύμων [10, 11], η αυτοκοπή των ιντρόνων [12] και η τελομεράση [13]. Οι ψευδόκομβοι συχνά συμβάλλουν, μερικές φορές ακόμη και στο βαθμό του καθορισμού, στην αλλαγή της έκφρασης γονιδίων πολλών ιών [14, 15, 16].

Επιπλέον, αναλύοντας περισσότερο τα μοτίβα τα οποία εξετάστηκαν, τα εξογκώματα και οι ασύμμετροι βρόχοι δημιουργούνται όταν μια αλυσίδα διακόπτεται σε μία ή περισσότερες βάσεις της αλυσίδας και παρατηρούνται συχνά στις δευτεροταγείς δομές του RNA [18, 19], καθώς εμφανίζονται σε όλα τα είδη δομημένων λειτουργικών RNA [20]. Οι αντιστοιχίες βάσεων, που δημιουργούν εσωτερικούς βρόχους, επηρεάζουν επίσης τη σταθερότητα του μορίου [18]. Ειδικά, οι ερευνητές έχουν επικεντρωθεί στη μελέτη των ασύμμετρων βρόχων, λόγω της συχνότητας με την οποία οι περιοχές

με εξογκώματα αδενίνων εμφανίζονται στα σημεία σύνδεσης με πρωτεΐνες στο RNA [18, 21], ενώ λειτουργούν επίσης ως σημεία επαφής στη δημιουργία της τριτοταγούς δομής του RNA [19, 23]. Οι ασύμμετροι αυτοί βρόχοι ή εξογκώματα δημιουργούν μοναδικές τοποθεσίες αναγνώρισης στις τριτοταγείς δομές του RNA με δύο τρόπους, πρώτον δρώντας ως μοριακές λαβές εντός των αλυσίδων και δεύτερον, με έμμεσο τρόπο, παραμορφώντας την ¨ραχοκοκαλιά' (backbone) του RNA, επιτρέποντας την πρόσβαση στα ζεύγη βάσεων με ένα διευρυμένο τρόπο [20]. Επιπλέον, είναι συχνό φαινόμενο τα στοιχεία της αλυσίδας που χωρίζονται από εξογκώματα να περνούν από μεταβάσεις μεταξύ μη στοιβαγμένων και συμμετρικά στοιβαγμένων σχηματισμών κατά τη διάρκεια της αναδίπλωσης και της λειτουργίας των μη-κωδικοποιητικών RNA [20]. Όλες οι ανωτέρω αναφορές δείχνουν τη σημασία της αναγνώρισης των ασύμμετρων βρόχων ή εξογκωμάτων και των εσωτερικών βρόχων ως κεντρικών δομικών στοιχείων σε ένα μεγάλο ποσοστό δομών ψευδοκόμβων του RNA.

Συνοψίζοντας, η ακριβής πρόβλεψη των δευτεροταγών δομών πολύπλοκων μοτίβων όπως οι ψευδόκομβοι, οι ασύμμετροι βρόχοι και οι εσωτερικοί βρόχοι, είναι εξέχουσας σημασίας στην αποκωδικοποίηση των λειτουργικών πτυχών των μορίων του RNA. Οι υπάρχουσες μεθόδοι για την πρόβλεψη της δομής του RNA έχουν χρησιμοποιήσει διάφορους αλγορίθμους, συμπεριλαμβανομένων θερμοδυναμικών, πιθανοτικών και βασισμένων σε τεχνητή νοημοσύνη προσεγγίσεων, με διαφορετικά ποσοστά επιτυχίας. Ωστόσο, η πρόβλεψη των ψευδοκόμβων παραμένει μια πρόκληση λόγω των σύνθετων αλληλεπιδράσεων και των απαιτήσεων για εκθετικό χρόνο εκτέλεσης. Για να αντιμετωπιστεί αυτό το κενό, αυτή η διατριβή παρουσιάζει μια καινοτόμο σειρά πλατφορμών: Knotify, Knotify+ και Knotify για τους ψευδοκόμβους τύπου L. Αυτές οι πλατφόρμες χρησιμοποιούν καινοτόμους αλγορίθμους συντακτικής αναγνώρισης, συνδυάζοντας έννοιες μέγιστης αντιστοίχισης βάσεων και ελάχιστης ελεύθερης ενέργειας για την ακριβή πρόβλεψη των ψευδοκόμβων τύπου H, ασύμμετρων βρόχων και εσωτερικών βρόχων σε αντίστοιχους ψευδοκόμβους. Με τον τρόπο αυτό, η διαδικασία πρόβλεψης των ψευδοκόμβων αντιμετωπίζεται ως πρόβλημα ανάλυσης και βελτιστοποίησης, προσφέροντας βελτιωμένη αποτελεσματικότητα και απόδοση, υπερβαίνοντας τις υπάρχουσες πλατφόρμες σε χρόνο εκτέλεσης, διατηρώντας παράλληλα συγκρίσιμες τιμές ακρίβειας πρόβλεψης. Επιπλέον, παρουσιάζεται ένα εξειδικευμένο σύστημα βασισμένο σε γραμματικές για την πρόβλεψη των ψευδοκόμβων τύπου L, επιδεικνύοντας την αποτελεσματικότητα και την υπολογιστική αποδοτικότητα του. Αυτές οι καινοτόμες προσεγγίσεις στοχεύουν στο να εξοπλίσουν τους βιολόγους για την πρόβλεψη των μοτίβων του RNA, διευκολύνοντας τις εξελίξεις στη γονιδιακή θεραπεία, τον σχεδιασμό φαρμάκων και την κατανόηση της λειτουργικότητας του RNA. Διαθέτοντας τον πηγαίο κώδικα και τις λεπτομέρειες υλοποίησης για δημόσια χρήση, αυτή η έρευνα προωθεί την αξιοποίηση αυτών των μεθόδων στο πλαίσιο άλλων τεχνικών, με στόχο τη βελτίωση της ακρίβειας της πρόβλεψης της δομής του RNA στις πολλαπλές εφαρμογές που εντοπίζεται στον τομέα της βιολογίας.

2.1 Περίγραμμα της Διατριβής

Στην ενότητα αυτή παρουσιάζεται ένα περίγραμμα της διατριβής, με τη δομή των Κεφαλαίων και των Παραρτημάτων. Συγκεκριμένα:

- Στο Κεφάλαιο 3 αναλύεται η θεωρητική βάση της πρόβλεψης της δομής του RNA, με ενδεικτικές μεθόδους και προσεγγίσεις που έχουν αναπτυχθεί στη βιβλιογραφία.
- Στο Κεφάλαιο 4 παρέχονται πληροφορίες για το απαραίτητο υπόβαθρο που απαιτείται για την κατανόηση των επόμενων Κεφαλαίων.
- Στο Κεφάλαιο 5 παρουσιάζεται η παλτφόρμα Knotify και ο πρώτος αλγόριθμος που αναπτύχθηκε, ο οποίος αντιμετωπίζει την πρόβλεψη ψευδοκόμβων πρώτης τάξης (τύπου H) με την χρήση μιας προτεινόμενης γραμματικής (CFG).
- Η βελτιστοποίηση του χρόνου εκτέλεσης μέσω τεχνικών κλαδέματος παρουσιάζεται στο Κεφάλαιο 6.
- Με βάση τα επιτεύγματα του Knotify, στο Κεφάλαιο 7 παρουσιάζεται η πλατφόρμα Knotify+, μια επέκταση που συμπεριλαμβάνει ασύμμετρους βρόχους ή εξογκώματα και εσωτερικούς βρόχους για την πρόβλεψη πιο πολύπλοκων προτύπων ψευδοκόμβων τύπου H.
- Μία επέκταση του Knotify για ψευδοκόμβους τύπου L παρουσιάζεται στο Κεφάλαιο 8 με μία νέα τροποποιημένη γραμματική.
- Τέλος, στο Κεφάλαιο 9 συνοψίζονται τα ευρήματα αυτής της διατριβής και παρουσιάζονται οι προοπτικές για μελλοντική έρευνα στον τομέα της πρόβλεψης της δευτεροταγούς δομής του RNA.

Αναφορικά με τα παραρτήματα που παρατίθενται στο τέλος της διατριβής:

- Στο Παράρτημα Α΄ παρουσιάζονται οι συντομογραφίες που χρησιμοποιήθηκαν, με στόχο την κατανόηση των όρων που αναπτύσσονται.
- Στο Παράρτημα Β΄ περιλαμβάνεται ένα μέρος του κώδικα που αναπτύχθηκε για την υλοποίηση των συστημάτων για την πρόβλεψη των ψευδοκόμβων.

2.2 Συνεισφορά της Διατριβής

Η συνεισφορά της παρούσας διατριβής συνοψίζεται στα ακόλουθα σημεία:

- Προτείνει εξειδικευμένα συστήματα πρόβλεψης ψευδοκόμβων με τη χρήση υβριδικών μοντέλων, τα οποία αξιοποιούν τη συντακτική αναγνώριση προτύπων, καθώς και βιολογικά και ενεργειακά κριτήρια. Η μεγάλη ακρίβεια και ταχύτητα εκτέλεσής τους στοχεύει στην αξιοποίησή τους από τους βιολόγους για την ταχύτερη απόκριση και τη βελτίωση των αποτελεσμάτων τους, όπου αυτό απαιτείται.

- Εντοπίζει σύνθετα μοτίβα ψευδοκόμβων, τα οποία πολλές φορές αγνοούνται λόγω των ελάχιστων παραδειγμάτων που έχουν εντοπιστεί. Ωστόσο, το σύστημα είναι έτοιμο να συμπεριλάβει τέτοιες δομές στην πρόβλεψή του, όταν αυτά θα παρουσιαστούν, σε πραγματικές συνθήκες, στη φύση.

- Αποτελεί τη δομή, πάνω στην οποία μπορούν να αναπτυχθούν νέα συστήματα ή να βελτιωθούν υφιστάμενα εφόσον ο κώδικας των υλοποιήσεων είναι διαθέσιμος σε δημόσια αποθετήρια.

- Δημιουργεί το πρώτο σύνολο συστημάτων, μέσα σε μία ευρύτερη εργαλειοθήκη που προετοιμάζεται και στοχεύει σε μία ολιστική προσέγγιση για την ανάλυση και την πρόβλεψη δομών RNA.

Κεφάλαιο 3

Σχετική Βιβλιογραφία

Στο Κεφάλαιο αυτό περογράφονται όλες οι σύγχρονες μέθοδοι πρόβλεψης της δευτεροταγούς δομής του RNA. Αρχικά, περιγράφονται οι πειραματικές μέθοδοι που χρησιμοποιούνται για αυτό το σκοπό, ενώ στη συνέχεια δίνεται έμφαση στις υπολογιστικές μεθόδους και στις διαφορετικές προσεγγίσεις που αυτές ακολουθούν.

3.1 Πειραματική επαλήθευση

Οι μέθοδοι πειραματικής επαλήθευσης, όπως η κρυσταλλογραφία με ακτίνες X (X-ray crystallography), η φασματοσκοπία NMP (Nuclear Magnetic Resonance spectroscopy), η χημική ανίχνευση (Chemical Probing) και η κρυο-ηλεκτρονική μικροσκοπία Cryo-electron Microscopy) συγκαταλέγονται στις κύριες εργαστηριακές μεθόδους που εφαρμόζονται για τον καθορισμό της δομής του RNA. Κάθε μία από αυτές τις μεθόδους προσφέρει μία διαφορετική διάσταση για την ανάλυση της δομής του RNA. Η κρυσταλλογραφία με ακτίνες X επιτρέπει την ολοκληρωμένη ανάλυση μοριακών δομών μέσω της κρυσταλλοποίησης, ενώ η φασματοσκοπία NMR παρέχει πληροφορίες για τις δομικές και δυναμικές ιδιότητες των μορίων σε διάλυμα. Από την άλλη, η χημική ανίχνευση επιτρέπει τον προσδιορισμό περιοχών του RNA που υφίστανται χημικές αλλαγές, προσφέροντας πληροφορίες για τις αλληλεπιδράσεις του με άλλα μόρια. Τα πειραματικά αποτελέσματα αποτελούν την πιο έγκυρη και αξιόπιστη μορφή επιβεβαίωσης της δομής του RNA, καθιστώντας αυτές τις μεθόδους ουσιαστικά αναπόσπαστο μέρος της δομικής βιολογίας. Στις ακόλουθες ενότητες παρουσιάζονται οι μέθοδοι αυτές με τις αντίστοιχες διαδικασίες που ακολουθούνται και περιγράφονται τα πλεονεκτήματα και τα μειονεκτήματά τους.

3.1.1 Κρυσταλλογραφία με χρήση ακτίνων X

Η τεχνική της κρυσταλλογραφίας με ακτίνες X (X-ray crystallography) εξετάζει λεπτομερώς τη δομή των πολύπλοκων βιολογικών μορίων, όπως οι πρωτεΐνες και τα νουκλεϊνικά οξέα. Αυτή η μέθοδος χρησιμοποιείται από τις αρχές του 20ου αιώνα

και στηρίζεται στην επεξεργασία των μορίων μέσα σε διάφορες συνθήκες και θερμοκρασίες διαλύματος («χώρος συνθηκών») για τη δημιουργία κρυστάλλου. Έπειτα, ο κρύσταλλος αυτός διαθλάται με συγκεκριμένο τρόπο με ακτίνες X, οι οποίες παράγουν μια τρισδιάστατη εικόνα της πυκνότητας ηλεκτρονίων μέσα στον κρύσταλλο, συνυπολογίζοντας τις μετρήσεις των γωνιών και των εντάσεων κατά τη διαδικασία. Αυτή η πυκνότητα παρέχει πληροφορίες για τις θέσεις των ατόμων, τους χημικούς δεσμούς, τις διαταραχές και άλλα στοιχεία του μορίου. Το RNA, λόγω των ειδικών του χαρακτηριστικών, όπως η μορφή, το μέγεθος, το μικρό αλφάβητο στοιχείων με τέσσερις βάσεις, της κυριαρχίας του ζεύγους βάσεων και της έλικας μορφής A, προσφέρει πλεονεκτήματα στην κρυσταλλογραφία. Ειδικότερα τα RNAs που είναι μικρότερα των 400 νουκλεοτιδίων, μπορούν να παρασκευαστούν σε μεγάλες ποσότητες, ενώ ταυτόχρονα η αλληλουχία είναι διαλυτή σε υψηλές συγκεντρώσεις, στοιχείο απαραίτητο για τη δημιουργία κρυστάλλου. Παρά τα οφέλη που αναφέρονται, ορισμένα μειονεκτήματα είναι το απαιτούμενο κόστος, ο χρόνος υλοποίησης και η αδυναμία πολλές φορές επαναληπτικότητας των διαδικασιών κρυσταλλοποίησης, ενώ η πλαστικότητα του RNA περιορίζει τη δυνατότητα εφαρμογής της μεθόδου αυτής.

3.1.2 Φασματοσκοπία NMR

Μια άλλη μέθοδος για τον προσδιορισμό της δομής του RNA είναι η φασματοσκοπία πυρηνικού μαγνητικού συντονισμού (NMR). Αυτή η τεχνική χρησιμοποιείται σχετικά με τη δομή και τη δυναμική των μορίων νουκλεϊκού οξέος, όπως το DNA και το RNA. Η φασματοσκοπία NMR μπορεί να εκτελέσει αναλύσεις σε διάλυμα, παρέχοντας σημαντικές πληροφορίες για τα RNA, με σχεδόν το 50% των γνωστών δομών RNA να έχουν προσδιοριστεί με τη χρησιμοποίηση αυτής της τεχνικής. Η φασματοσκοπία NMR είναι κυρίως χρήσιμη για την ανίχνευση των δομών μικρών μορίων RNA, έως και 100 νουκλεοτιδίων. Ένα από τα πλεονεκτήματά της έναντι της κρυσταλλογραφίας ακτίνων X είναι ότι τα μόρια παρατηρούνται στη φυσική τους κατάσταση διαλύματος, αντί της κρυσταλλικής μορφής, το οποίο μπορεί να επηρεάσει τις δομικές ιδιότητες του μορίου. Επιπλέον, η φασματοσκοπία NMR είναι οικονομικότερη από την κρυσταλλογραφία ακτίνων X, αλλά οι δομές που ανακαλύπτονται είναι λιγότερο ακριβείς.

3.1.3 Χημική ανίχνευση (Chemical Probing)

Η χημική ανίχνευση (Chemical Probing) αποτελεί μια άλλη πειραματική μέθοδο που ευρέως χρησιμοποιείται ως γρήγορη προσέγγιση για τον προσδιορισμό της δομής του RNA. Τα μόρια RNA υποβάλλονται σε χημική τροποποίηση σε συγκεκριμένες θέσεις βάσεων, και τα αποτελέσματα ερμηνεύονται με τη χρήση διάφορων αλγορίθμων. Σε ένα πείραμα χημικής ανίχνευσης, ένα RNA που μας ενδιαφέρει τροποποιείται με ένα χημικό ανιχνευτή όπως για παράδειγμα είναι ο θειικός διμεθυλεστέρας, που αντιδρά με νουκλεοτίδια. Αυτοί οι ανιχνευτές επιλέγονται ώστε να έχουν αντιδράσεις που εξαρτώνται από το τοπικό περιβάλλον ενός νουκλεοτιδίου. Σε σύγκριση με την

κρυσταλλογραφία ακτίνων X και τη φασματοσκοπία NMR, η χημική ανίχνευση είναι γρήγορη και εύκολη, προσφέροντας έναν απλό και ακριβή τρόπο ανάλυσης της δομής του RNA. Επιπλέον, η χημική ανίχνευση συμπλόκων RNA και RNA/protein μπορεί να παρέχει μεγάλα σύνολα δεδομένων, προσφέροντας λεπτομερείς πληροφορίες σχετικά με τη δομή του RNA. Όπως και σε άλλες πειραματικές προσεγγίσεις, η διαδικασία περιέχει διάφορες μορφές θορύβου και τα πειραματικά αποτελέσματα εξαρτώνται από τις συνθήκες, όπως το pH, κάτω από τις οποίες λαμβάνει χώρα το πείραμα. Έτσι, η δομή που προκύπτει μπορεί να παρουσιάζει αποκλίσεις από την πραγματική δομή του μορίου.

3.1.4 Κρυο-ηλεκτρονική μικροσκοπία (Cryo-electron Microscopy)

Η κρυο-ηλεκτρονική μικροσκοπία (Cryo-electron Microscopy) θεωρείται ως μία από τις πλέον ισχυρές μεθόδους στη δομική βιολογία τα τελευταία δέκα χρόνια. Στη μονοσωματική κρυο-ηλεκτρονική μικροσκοπία, βιολογικά μακρομόρια τοποθετούνται σε έναν υαλώδη πάγο με τυχαίους προσανατολισμούς. Το παγωμένο πλέγμα εξετάζεται σε ένα ηλεκτρονικό μικροσκόπιο και συλλέγονται εικόνες των σωματιδίων. Μέσω της εξισορρόπησης και του συνδυασμού προβολών σωματιδίων σε διαφορετικούς προσανατολισμούς, μπορεί να ανακατασκευαστεί ένας τρισδιάστατος χάρτης του σωματιδίου. Πρόσφατες εξελίξεις στο υλισμικό και το λογισμικό επέτρεψαν στους ερευνητές να προσδιορίσουν δομές υψηλής ανάλυσης και πολλών στόχων που θεωρούνταν παραδοσιακές προκλήσεις, όπως μεγάλα πρωτεϊνικά σύμπλοκα και πρωτεΐνες που είναι δύσκολο να κρυσταλλωθούν.

3.2 Υπολογιστικές Μέθοδοι

Το μεγαλύτερο μέρος των αλγορίθμων που έχουν αναπτυχθεί για την πρόβλεψη της δευτεροταγούς δομής του RNA έχουν ενσωματώσει τεχνικές δυναμικού προγραμματισμού στις διαδικασίες τους. Αυτές οι μεθόδοι στοχεύουν στον προσδιορισμό της πιο πιθανής δευτεροταγούς δομής με ταυτόχρονη ελαχιστοποίηση της ελεύθερης ενέργειας που συσχετίζεται με τη διαδικασία πτύσης [25, 26]. Ορισμένες προσεγγίσεις, όπως αυτή που προτάθηκε από τον Cao [27], έχουν επικεντρωθεί ειδικά στην πρόβλεψη των ψευδοκόμβων, λαμβάνοντας υπόψη παράγοντες όπως η εντροπία, η σταθερότητα και η ελάχιστη ελεύθερη ενέργεια. Η απόδειξη της NP(nondeterministic polynomial time)-completeness για το πρόβλημα πρόβλεψης της δευτεροταγούς δομής του RNA [28] έχει οδηγήσει την ερευνητική κοινότητα στην ανάπτυξη στοχαστικών και ευριστικών μεθόδων [29, 30, 31]. Ένα παράδειγμα είναι ο αλγόριθμος Knotty [32], που χρησιμοποιεί τον αλγόριθμο CCJ (Chen–Condon–Jabbari) με αραιοποίηση για την πρόβλεψη των ψευδοκόμβων. Το ProbKnot [34], από την άλλη πλευρά, υπολογίζει τις πιθανότητες σχηματισμού βάσεων των μη-ψευδοκομβικών υποδομών και κατασκευάζει τη δευτεροταγή δομή βασισμένη στο μέγιστο αναμενόμενο επίπεδο ακρίβειας. Ο IP-

knot [35] εκμεταλλεύεται τα πλεονεκτήματα του ακέραιου προγραμματισμού και των πιθανοτήτων σχηματισμού βάσεων, υπερτερώντας σε ακρίβεια των προηγούμενων μεθόδων σε συγκεκριμένα σύνολα δεδομένων. Η επέκτασή του, το IPknot2 [36], χρησιμοποιεί το μοντέλο LinearPartition και τη ψευδοαναμενόμενη ακρίβεια για τον υπολογισμό δευτεροταγών δομών με ψευδοκόμβους σε γραμμικό χρόνο. Αυτή η βελτιωμένη έκδοση μπορεί να χειριστεί μεγάλες ακολουθίες μέσα σε λογικό χρόνο εκτέλεσης, αν και υπάρχει ακόμη χώρος για βελτίωση στον τομέα της ακρίβειας. Ένας άλλος τύπος προσεγγίσεων για την πρόβλεψη της δευτεροταγούς δομής του RNA βασίζεται στη συντακτική αναγνώριση προτύπων και τις γραμματικές χωρίς συμφραζόμενα (SCFG). Μέθοδοι όπως το Pfold [37, 38], το PPfold [39] και το RNA-Decoder [40] εφαρμόζουν SCFG για την πρόβλεψη της δευτεροταγούς δομής. Αυτές οι προσεγγίσεις ειδικεύονται στην αναγνώριση μοτίβων, επιτρέποντάς τους να ανακαλύψουν ομοιότητες στις δομές του RNA. Αν αναθέσουν κατάλληλα βάρη στους κανόνες, αυτές οι μέθοδοι μπορούν να βελτιώσουν την ακρίβεια της πρόβλεψης. Διάφορα πλαίσια που βασίζονται σε SCFG, συμπεριλαμβανομένων του Evfold [42], του Infernal [43] και του Oxfold [44], έχουν αναπτυχθεί για την αντιμετώπιση της πρόβλεψης της δευτεροταγούς δομής του RNA, υπογραμμίζοντας τη σημασία της αποτελεσματικής συνεργασίας μεταξύ γραμματικών και υπολογιστικών μεθόδων, ευριστικών και πιθανοτικών αλγορίθμων και την ένταξη εννοιών όπως ο υπολογισμός της ελάχιστης ελεύθερης ενέργειας, ο μέγιστος αριθμός ζευγών βάσεων και οι πιθανότητες σχηματισμού αυτών.

Τα τελευταία χρόνια, οι αλγόριθμοι μηχανικής μάθησης έχουν κερδίσει επίσης την προσοχή στον τομέα της πρόβλεψης της δευτεροταγούς δομής του RNA. Αυτές οι μέθοδοι στοχεύουν στο να ανακαλύψουν κρυφά μοτίβα στις ακολουθίες του RNA χρησιμοποιώντας τεχνικές επιβλεπόμενης και μη-επιβλεπόμενης μάθησης. Ωστόσο, πολλές μεθόδοι μηχανικής μάθησης, ιδιαίτερα αυτές που χρησιμοποιούν βαθιά μάθηση, συχνά απαιτούν μεγάλα σύνολα δεδομένων για να αποτρέψουν την υπερμοντελοποίηση (overfitting) και να επιτύχουν καλή γενίκευση. Για παράδειγμα, ο Singh κ.ά. [45] συνδύασαν τη βαθιά μάθηση με κάποιους τριτογενείς περιορισμούς για τη βελτίωση της ακρίβειας της πρόβλεψης της δομής. Επίσης, στην εργασία [9], χρησιμοποιήθηκε ένα δίκτυο LSTM με διπλή κατεύθυνση και η αρχή IBPMP για να επιλέξει τις σωστές βάσεις και να προβλέψει τις βέλτιστες δομές. Το ContextFold [46] είναι ένας άλλος αλγόριθμος που συνδυάζει πληροφορίες ακολουθίας και πλαισίου για την πρόβλεψη της δευτεροταγούς δομής του RNA. Χρησιμοποιεί τεχνικές βαθιάς μάθησης, ειδικότερα συνελικτικά νευρωνικά δίκτυα (CNNs), για να εντοπίσει τις συμφραζόμενες εξαρτήσεις μεταξύ νουκλεοτιδίων. Για την αντιμετώπιση της πρόβλεψης των δευτεροταγών δομών με ψευδοκόμβους, το ATTfold [47] συνδυάζει μοντέλα βαθιάς μάθησης με μηχανισμό προσοχής. Με την κωδικοποίηση ενός πίνακα σκορ αντιστοίχισης βάσης και τη χρήση ενός Συνελικτικού Νευρωνικού Δικτύου (CNN) για την αποκωδικοποίηση, το ATTfold στοχεύει στην αιχμαλώτιση πολύπλοκων εξαρτήσεων στις ακολουθίες του RNA και στη βελτίωση της ακρίβειας της πρόβλεψης. Μια άλλη προσέγγιση, γνωστή ως 2dRNA [48], χρησιμοποιεί μια συζευγμένη διαδικασία δικτύου νευρώνων βάθους δύο σταδίων. Σε αυτήν τη μέθοδο, ένα δίκτυο LSTM με διπλή κατεύθυνση κωδικοποιεί τα δεδομένα σε υψηλότερη διάσταση, και ένα δίκτυο πλήρους σύνδεσης

αποκωδικοποιεί τα δεδομένα για την παραγωγή της δομής του dot-bracket, δηλαδή μία ακολουθία αποτελούμενη από τελείες, παρενθέσεις και αγκύλες που υποδεικνύουν την ύπαρξη ή όχι ζευγών βάσεων, δημιουργώντας έτσι τα παρατηρούμενα μοτίβα. Επιπλέον, υπάρχουν αρκετοί καλά εδραιωμένοι αλγόριθμοι και λογισμικά πακέτα που χρησιμοποιούνται ευρέως για την πρόβλεψη της δευτεροταγούς δομής του RNA. Το CONTRAfold [41] συνδυάζει θερμοδυναμικές και εξελικτικές πληροφορίες για την εξαγωγή της δομής. Το RNAstructure [49] προσφέρει μια σειρά αλγορίθμων για την πρόβλεψη της δευτεροταγούς δομής του RNA, συμπεριλαμβανομένων αλγορίθμων βασισμένων στη θερμοδυναμική και μεθόδων συγκριτικής ανάλυσης ακολουθιών. Το RNAalifold [50] χρησιμοποιεί μια προσέγγιση συγκριτικής ανάλυσης ακολουθιών και συμπεριλαμβάνει την εξελικτική διατήρηση, ενώ το CentroidFold [51] προβλέπει δομές χρησιμοποιώντας μια προσέγγιση βασισμένη σε κεντροειδείς. Αυτοί οι αλγόριθμοι και προσεγγίσεις υπογραμμίζουν την ποικιλία των μεθόδων που χρησιμοποιούνται στην πρόβλεψη της δευτεροταγούς δομής του RNA, συμπεριλαμβανομένων των τεχνικών δυναμικού προγραμματισμού, πιθανοτικών μοντέλων, μηχανικής μάθησης και τεχνικών βαθιάς μάθησης. Μέσω της διερεύνησης και σύγκρισης αυτών των διαφορετικών προσεγγίσεων, είναι δυνατόν να αποκτηθεί μια σφαιρική κατανόηση των τρεχουσών προηγμένων μεθόδων και να εντοπιστούν δυνητικά μονοπάτια για περαιτέρω έρευνα.

Κεφάλαιο 4

Θεωρητικό Υπόβαθρο

Στο Κεφάλαιο αυτό παρέχονται όλες οι απαραίτητες γενικές γνώσεις βιολογίας και πληροφορικής για την κατανόηση της διατριβής. Γίνεται περιγραφή της βιοπληροφορικής και εννοιών όπως το RNA, καθώς και της δευτεροταγούς δομής του με τα μοτίβα της, με έμφαση σε αυτό του ψευδοκόμβου. Στη συνέχεια, γίνεται μία εισαγωγή στη συντακτική αναγνώριση προτύπων, στις γραμματικές χωρίς συμφραζόμενα και στον αλγόριθμο του Early. Στο τέλος του Κεφαλαίου, γίνεται μία παρουσίαση του αναλυτή Yaep, που χρησιμοποιήθηκε στην υλοποίηση του συστήματος, καθώς και κάποιων εννοιών της θερμοδυναμικής, όπως η ελεύθερη ενέργεια Gibbs.

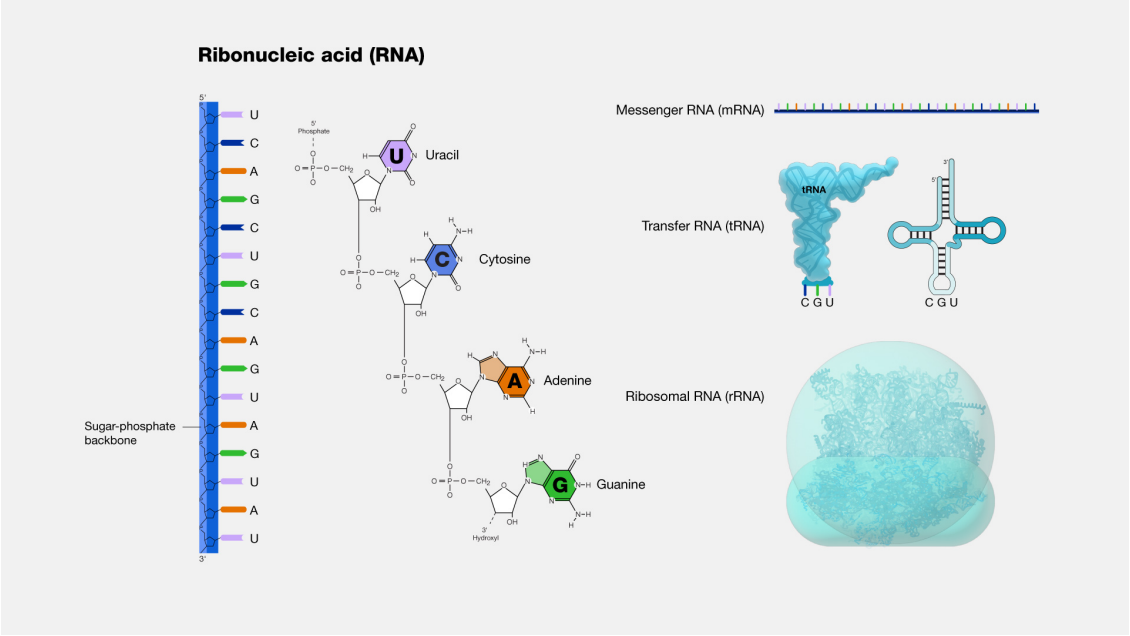
4.1 Βιοπληροφορική

Η βιοπληροφορική είναι ένας τομέας που χρησιμοποιεί μεθόδους και εργαλεία λογισμικού για την κατανόηση μεγάλων και πολύπλοκων βιολογικών δεδομένων. Είναι ένα διεπιστημονικό πεδίο που ενσωματώνει τη βιολογία, τη χημεία, τη φυσική, την επιστήμη των υπολογιστών, τη μηχανική πληροφοριών, τα μαθηματικά και τη στατιστική για την ανάλυση και την ερμηνεία βιολογικών δεδομένων. Η βιοπληροφορική χρησιμοποιείται συχνά για υπολογιστικές και στατιστικές αναλύσεις βιολογικών ερωτημάτων. Περιλαμβάνει τη χρήση προγραμματισμού υπολογιστών ως μέρος της μεθοδολογίας των βιολογικών μελετών και την ανάπτυξη ειδικών «αγωγών» ανάλυσης, ιδιαίτερα στη γονιδιωματική. Οι κοινές εφαρμογές του περιλαμβάνουν την αναγνώριση υποψήφιων γονιδίων και πολυμορφισμών μεμονωμένων νουκλεοτιδίων για την καλύτερη κατανόηση της γενετικής βάσης της νόσου, τις μοναδικές προσαρμογές, τις επιθυμητές ιδιότητες ή τις διαφορές μεταξύ των πληθυσμών. Η βιοπληροφορική βοηθά, επίσης, στην κατανόηση των οργανωτικών αρχών εντός των αλληλουχιών νουκλεϊκών οξέων και πρωτεϊνών, γνωστές ως πρωτεϊνική. Στον τομέα της γενετικής, η επεξεργασία εικόνας και σήματος διαδραματίζει κρίσιμο ρόλο στην αλληλούχιση και τον σχολιασμό των γονιδιωμάτων και των παρατηρούμενων μεταλλάξεών τους. Βοηθά στην εξόρυξη κειμένου της βιολογικής βιβλιογραφίας και στην ανάπτυξη βιολογικών και γονιδιακών οντολογιών για την οργάνωση και αναζήτηση βιολογικών δεδομένων. Τα εργαλεία

βιοπληροφορικής βοηθούν στη σύγκριση, ανάλυση και ερμηνεία γενετικών και γονιδιωματικών δεδομένων, καθώς και στην κατανόηση των εξελικτικών πτυχών της μοριακής βιολογίας. Τα εργαλεία αυτά είναι χρήσιμα για την ανάλυση και την καταλογογράφηση βιολογικών μονοπατιών και δικτύων που είναι απαραίτητα στη βιολογία συστημάτων. Η βιοπληροφορική βοηθά επίσης στην προσομοίωση και τη μοντελοποίηση του DNA, του RNA, των πρωτεϊνών και των βιομοριακών αλληλεπιδράσεων στη δομική βιολογία.

4.1.1 Στόχοι της Βιοπληροφορικής

Προκειμένου να κατανοήσουμε πώς επηρεάζονται οι φυσιολογικές κυτταρικές δραστηριότητες από διάφορες ασθένειες, είναι απαραίτητο να συνδυαστούν βιολογικά δεδομένα ώστε να επιτευχθεί μια ολοκληρωμένη κατανόηση αυτών των δραστηριοτήτων. Ο τομέας της βιοπληροφορικής αναπτύχθηκε ως απάντηση σε αυτή την ανάγκη, με πρωταρχικό στόχο την ανάλυση και την ερμηνεία διαφόρων τύπων δεδομένων, συμπεριλαμβανομένων των αλληλουχιών νουκλεοτιδίων και αμινοξέων, των πρωτεϊνικών περιοχών και των πρωτεϊνικών δομών. Αυτή η διαδικασία αναφέρεται ως υπολογιστική βιολογία και περιλαμβάνει την ανάπτυξη και εφαρμογή προγραμμάτων υπολογιστών που επιτρέπουν την αποτελεσματική πρόσβαση, διαχείριση και χρήση βιολογικών πληροφοριών. Επιπλέον, η βιοπληροφορική περιλαμβάνει την ανάπτυξη νέων μαθηματικών αλγορίθμων και στατιστικών μέτρων που αξιολογούν τις σχέσεις μεταξύ των μελών μεγάλων συνόλων δεδομένων. Ο πρωταρχικός στόχος της βιοπληροφορικής είναι να βελτιώσει την κατανόησή μας για τις βιολογικές διεργασίες, με έμφαση στην εφαρμογή υπολογιστικά εντατικών τεχνικών. Αυτό περιλαμβάνει αναγνώριση προτύπων, εξόρυξη δεδομένων, αλγόριθμους μηχανικής μάθησης και οπτικοποίηση. Το πεδίο περιλαμβάνει μια σειρά ερευνητικών τομέων, όπως ευθυγράμμιση αλληλουχίας, εύρεση γονιδίων, συγκρότηση γονιδιώματος, σχεδιασμό φαρμάκων, ευθυγράμμιση δομής πρωτεΐνης, πρόβλεψη γονιδιακής έκφρασης και αλληλεπιδράσεις πρωτεΐνης-πρωτεΐνης, πρόβλεψης των δομών των μορίων, μελέτες συσχέτισης σε όλο το γονιδίωμα και μοντελοποίηση εξέλιξης και διαίρεση/μίτωση των κυττάρων. Η βιοπληροφορική περιλαμβάνει τη δημιουργία βάσεων δεδομένων, αλγορίθμων, υπολογιστικών και στατιστικών τεχνικών και την ανάπτυξη της θεωρίας για την επίλυση τόσο τυπικών όσο και πρακτικών προβλημάτων που σχετίζονται με τη διαχείριση και την ανάλυση βιολογικών δεδομένων. Οι ραγδαίες εξελίξεις στις τεχνολογίες γονιδιωματικής και μοριακής έρευνας, σε συνδυασμό με τις τεχνολογίες της πληροφορίας, έχουν οδηγήσει σε πληθώρα πληροφοριών που σχετίζονται με τη μοριακή βιολογία, οι οποίες μπορούν να αναλυθούν και να ερμηνευτούν χρησιμοποιώντας προσεγγίσεις βιοπληροφορικής. Οι κοινές δραστηριότητες στη βιοπληροφορική περιλαμβάνουν τη χαρτογράφηση και την ανάλυση αλληλουχιών DNA και πρωτεϊνών, τη σύγκριση αλληλουχιών DNA και πρωτεϊνών και τη δημιουργία τρισδιάστατων μοντέλων πρωτεϊνικών δομών. Η βιοπληροφορική είναι ένας επιστημονικός τομέας που σχετίζεται στενά, αλλά διαφέρει από τον βιολογικό υπολογισμό. Αν και χρησιμοποιείται συχνά εναλλακτικά με την υπολογιστική βιολογία, υπάρχουν κάποιες διαφορές μεταξύ των δύο. Ο βιολογικός υπολογισμός



Σχήμα 4.1: Το RNA και ενδεικτικοί τύποι του ([52])

περιλαμβάνει τη χρήση της βιομηχανικής και της βιολογίας για τη δημιουργία βιολογικών υπολογιστών, ενώ η βιοπληροφορική χρησιμοποιεί υπολογιστικές μεθόδους για την καλύτερη κατανόηση της βιολογίας. Η βιοπληροφορική και η υπολογιστική βιολογία επικεντρώνονται στην ανάλυση βιολογικών δεδομένων, με ιδιαίτερη έμφαση στις αλληλουχίες DNA, RNA και πρωτεϊνών. Ο τομέας της βιοπληροφορικής γνώρισε σημαντική επέκταση από τα μέσα της δεκαετίας του 1990, κυρίως λόγω της προόδου στην τεχνολογία προσδιορισμού αλληλουχίας DNA και του Έργου Ανθρώπινου Γονιδιώματος. Για την εξαγωγή σημαντικών πληροφοριών από βιολογικά δεδομένα, τα προγράμματα λογισμικού σχεδιάζονται και υλοποιούνται χρησιμοποιώντας αλγόριθμους από ποικίλα πεδία, συμπεριλαμβανομένης της θεωρίας γραφημάτων, της τεχνητής νοημοσύνης, της εξόρυξης δεδομένων, της επεξεργασίας εικόνας και της προσομοίωσης. Αυτοί οι αλγόριθμοι βασίζονται σε θεωρητικά θεμέλια όπως τα διακριτά μαθηματικά, η θεωρία ελέγχου, η θεωρία συστημάτων, η θεωρία πληροφοριών και η στατιστική.

4.1.2 RNA

Το RNA διαθέτει μια μοναδική δομή που συμβάλλει σε διάφορες βιολογικές λειτουργίες. Αποτελείται από νουκλεοτίδια, τα οποία αποτελούνται από τρία βασικά συστατικά: ένα μόριο σακχάρου ριβόζης, μια φωσφορική ομάδα και μια αζωτούχο βάση και έχει διάφορους τύπους όπως φαίνεται στο Σχήμα 4.1. Το σάκχαρο ριβόζης χρησιμεύει ως το θεμέλιο του μορίου RNA, συνδέοντας τα νουκλεοτίδια. Οι φωσφορικές

ομάδες συνδέουν τα μόρια σακχάρου μεταξύ τους μέσω φωσφοδιεστερικών δεσμών, σχηματίζοντας μια συνεχή αλυσίδα. Οι βάσεις που συναντώνται στο RNA είναι η αδενίνη (A), η κυτοσίνη (C), η γουανίνη (G) και η ουρακίλη (U) και συνδέονται μεταξύ τους με 3'-5' φωσφοδιεστερικούς δεσμούς, δημιουργώντας έτσι μια μονόκλωνη αλυσίδα.
Οι λειτουργίες του RNA ποικίλουν ανάλογα με τον τύπο του και τη δομή του. Στην επόμενη Ενότητα παρουσιάζονται οι διαφορετικοί του τύποι, οι οποίοι έχουν διακριτές λειτουργίες εντός του κυττάρου, όπως τη γονιδιακή έκφραση και τη σύνθεση των πρωτεϊνών αλλά και πολλές άλλες σημαντικές λειτουργίες για τους οργανισμούς. Ακολούθως, περιγράφονται οι σπουδαιότερες δομές του και τα μοτίβα που παρουσιάζονται συχνότερα σε αυτό.

4.1.3 Τύποι RNA

Το αγγελιαφόρο RNA (mRNA) είναι το RNA που μεταφέρει πληροφορίες από το DNA στο ριβόσωμα, τις θέσεις της πρωτεϊνικής σύνθεσης (μετάφρασης) στο κύτταρο. Το mRNA είναι αντίγραφο του DNA. Η κωδικοποιητική αλληλουχία του mRNA καθορίζει την αλληλουχία αμινοξέων στην πρωτεΐνη στην οποία παράγεται. Ωστόσο, πολλά RNA δεν κωδικοποιούν την πρωτεΐνη (περίπου το 97% της μεταγραφικής παραγωγής δεν κωδικοποιεί πρωτεΐνες στους ευκαρυώτες). Αυτά τα λεγόμενα μη κωδικοποιητικά RNA (ncRNA) μπορούν να κωδικοποιηθούν από τα δικά τους γονίδια (γονίδια RNA), αλλά μπορούν επίσης να προέρχονται από ιντρόνια mRNA. Τα πιο σημαντικά παραδείγματα μη κωδικοποιητικών RNA είναι το RNA μεταφοράς (tRNA) και το ριβοσωμικό RNA (rRNA), τα οποία εμπλέκονται και τα δύο στη διαδικασία της μετάφρασης. Υπάρχουν επίσης μη κωδικοποιητικά RNA που εμπλέκονται στη ρύθμιση των γονιδίων, την επεξεργασία του RNA και άλλους ρόλους. Ορισμένα RNA είναι ικανά να καταλύουν χημικές αντιδράσεις όπως η κοπή και η απολίνωση άλλων μορίων RNA και η κατάλυση του σχηματισμού πεπτιδικού δεσμού στο ριβόσωμα, τα οποία είναι γνωστά και ως ριβόζυμα. Σύμφωνα με το μήκος της αλυσίδας RNA, το RNA κατατάσσεται σε μικρό RNA και μακρύ RNA. Συνήθως, τα μικρά RNA έχουν μήκος μικρότερο από 200 νουκλεοτίδια και τα μακρά RNA έχουν μήκος μεγαλύτερο από 200 νουκλεοτίδια. Τα μακρά RNA, που ονομάζονται επίσης μεγάλα RNA, περιλαμβάνουν κυρίως μακρύ μη κωδικοποιητικό RNA (lncRNA) και mRNA. Τα μικρά RNA περιλαμβάνουν κυρίως 5.8S ριβοσωμικό RNA (rRNA), 5S rRNA, RNA μεταφοράς (tRNA), microRNA (miRNA), μικρό παρεμβαλλόμενο RNA (siRNA), μικρό πυρηνικό RNA (snoRNAs), αλληλοεπιδρώντα με Piwi RNA (piRNA), tRNA- προερχόμενο μικρό RNA (tsRNA) και με το μικρό προερχόμενο από rDNA RNA (srRNA). Υπάρχουν ορισμένες εξαιρέσεις όπως στην περίπτωση του 5S rRNA των μελών του γένους Halococcus (Archaea), που έχουν παρεμβολή, αυξάνοντας έτσι το μέγεθός τους.
Το αγγελιαφόρο RNA (mRNA) μεταφέρει πληροφορίες σχετικά με μια αλληλουχία πρωτεΐνης στα ριβοσώματα, τα εργοστάσια πρωτεϊνοσύνθεσης στο κύτταρο. Κωδικοποιείται έτσι ώστε κάθε τρία νουκλεοτίδια (ένα κωδικόνιο) να αντιστοιχεί σε ένα αμινοξύ. Στα ευκαρυωτικά κύτταρα, μόλις το πρόδρομο mRNA (προ-mRNA) μετα-

γραφεί από το DNA, υποβάλλεται σε επεξεργασία για να ωριμάσει το mRNA. Αυτό αφαιρεί τα εσώνια του—μη κωδικοποιητικά τμήματα του pre-mRNA. Στη συνέχεια, το mRNA εξάγεται από τον πυρήνα στο κυτταρόπλασμα, όπου συνδέεται με τα ριβοσώματα και μεταφράζεται στην αντίστοιχη πρωτεϊνική του μορφή με τη βοήθεια του tRNA. Σε προκαρυωτικά κύτταρα, τα οποία δεν έχουν διαμερίσματα πυρήνα και κυτταροπλάσματος, το mRNA μπορεί να συνδεθεί με τα ριβοσώματα ενώ μεταγράφεται από το DNA. Μετά από ένα ορισμένο χρονικό διάστημα, το μήνυμα αποικοδομείται στα συστατικά του νουκλεοτίδια με τη βοήθεια ριβονουκλεασών.
Το RNA μεταφοράς (tRNA) είναι μια μικρή αλυσίδα RNA περίπου 80 νουκλεοτιδίων που μεταφέρει ένα συγκεκριμένο αμινοξύ σε μια αναπτυσσόμενη πολυπεπτιδική αλυσίδα στη ριβοσωμική θέση της πρωτεϊνικής σύνθεσης κατά τη μετάφραση. Έχει θέσεις για σύνδεση αμινοξέων και περιοχή αντικωδικονίου για αναγνώριση κωδικονίων, η οποία συνδέεται με μια συγκεκριμένη αλληλουχία στην αλυσίδα αγγελιαφόρου RNA μέσω δεσμού υδρογόνου. Το ριβοσωμικό RNA (rRNA) είναι το καταλυτικό συστατικό των ριβοσωμάτων. Το rRNA είναι το συστατικό του ριβοσώματος που φιλοξενεί τη μετάφραση. Τα ευκαρυωτικά ριβοσώματα περιέχουν τέσσερα διαφορετικά μόρια rRNA: 18S, 5.8S, 28S και 5S rRNA. Τρία από τα μόρια rRNA συντίθενται στον πυρήνα και ένα συντίθεται αλλού. Στο κυτταρόπλασμα, το ριβοσωμικό RNA και η πρωτεΐνη συνδυάζονται για να σχηματίσουν μια νουκλεοπρωτεΐνη που ονομάζεται ριβόσωμα. Το ριβόσωμα δεσμεύει το mRNA και πραγματοποιεί πρωτεϊνική σύνθεση. Πολλά ριβοσώματα μπορούν να προσκολληθούν σε ένα μόνο mRNA ανά πάσα στιγμή. Σχεδόν όλο το RNA που βρίσκεται σε ένα τυπικό ευκαρυωτικό κύτταρο είναι rRNA.

4.1.4 Σύνθεση RNA - Μεταγραφή

Η σύνθεση του RNA καταλύεται συνήθως από ένα ένζυμο, το οποίο ονομάζεται - RNA πολυμεράση - χρησιμοποιώντας το DNA ως πρότυπο, μια διαδικασία γνωστή ως μεταγραφή. Η μεταγραφή είναι η διαδικασία αντιγραφής ενός τμήματος DNA σε RNA. Τα τμήματα του DNA που μεταγράφονται σε μόρια RNA που μπορούν να κωδικοποιήσουν πρωτεΐνες λέγεται ότι παράγουν αγγελιαφόρο RNA (mRNA). Άλλα τμήματα του DNA αντιγράφονται σε μόρια RNA που ονομάζονται μη κωδικοποιητικά RNA (ncRNAs). Το mRNA αποτελεί μόνο το 1-3% των συνολικών δειγμάτων RNA. Λιγότερο από το 2% του ανθρώπινου γονιδιώματος μπορεί να μεταγραφεί σε mRNA, ενώ τουλάχιστον το 80% του γονιδιωματικού DNA θηλαστικών μπορεί να μεταγραφεί ενεργά (σε έναν ή περισσότερους τύπους κυττάρων), με την πλειοψηφία από αυτό το 80% να θεωρείται ότι είναι ncRNA.
Τόσο το DNA όσο και το RNA είναι νουκλεϊκά οξέα, τα οποία χρησιμοποιούν τα ζεύγη βάσεων νουκλεοτιδίων ως συμπληρωματική γλώσσα. Κατά τη μεταγραφή, μια αλληλουχία DNA διαβάζεται από μια RNA πολυμεράση, η οποία παράγει έναν συμπληρωματικό, αντιπαράλληλο κλώνο RNA που ονομάζεται πρωτεύουσα μεταγραφή. Η μεταγραφή προχωρά στα ακόλουθα γενικά βήματα:

- Η RNA πολυμεράση, μαζί με έναν ή περισσότερους γενικούς μεταγραφικούς παράγοντες, συνδέεται με το DNA του προαγωγέα.

- Η RNA πολυμεράση δημιουργεί μια φυσαλίδα μεταγραφής, η οποία διαχωρίζει τους δύο κλώνους της έλικας του DNA. Αυτό γίνεται με το σπάσιμο των δεσμών υδρογόνου μεταξύ συμπληρωματικών νουκλεοτιδίων DNA.

- Η RNA πολυμεράση προσθέτει νουκλεοτίδια RNA (τα οποία είναι συμπληρωματικά με τα νουκλεοτίδια ενός κλώνου DNA).

- Η ραχοκοκαλιά φωσφορικού σακχάρου RNA σχηματίζεται με τη βοήθεια της RNA πολυμεράσης για να σχηματίσει έναν κλώνο RNA.

- Οι δεσμοί υδρογόνου της έλικας RNA-DNA σπάνε, ελευθερώνοντας τον νεοσυντιθέμενο κλώνο RNA.

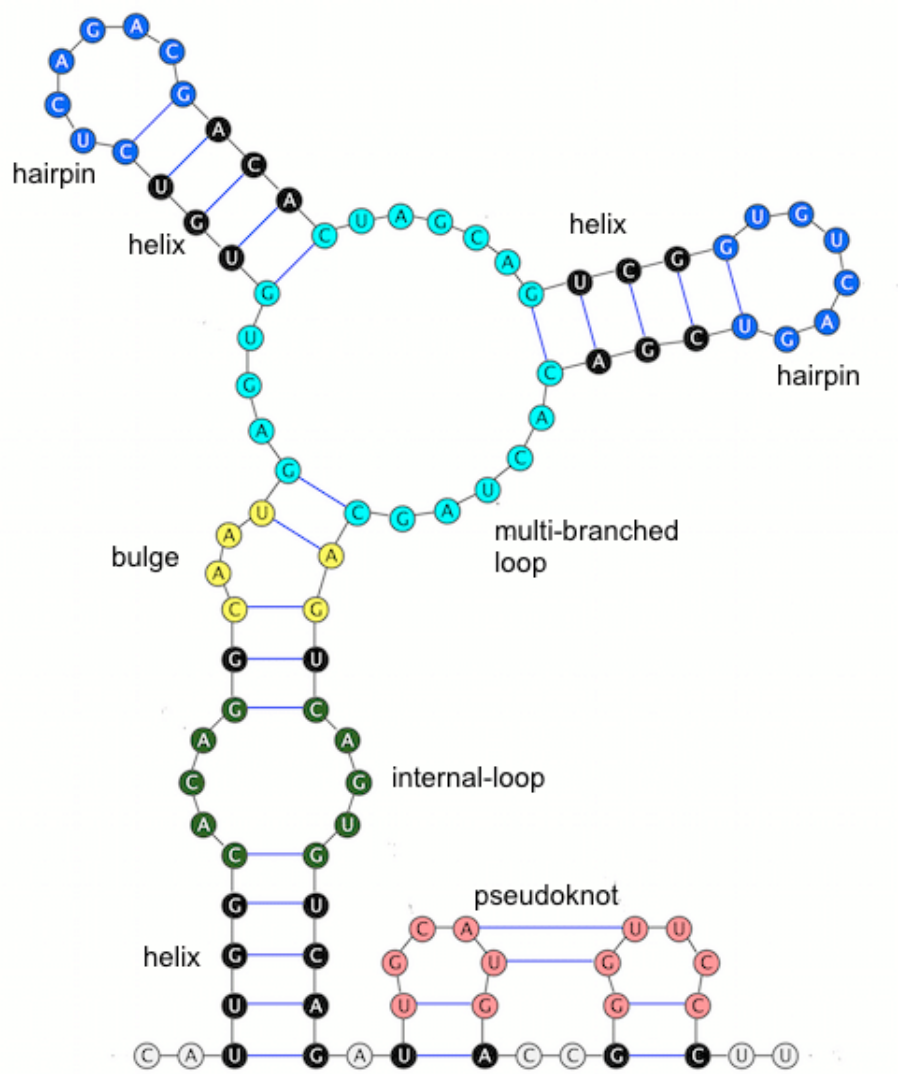
- Εάν το κύτταρο έχει πυρήνα, το RNA μπορεί να υποβληθεί σε περαιτέρω επεξεργασία. Αυτό μπορεί να περιλαμβάνει πολυαδενυλίωση, κάλυψη και μάτισμα.

- Το RNA μπορεί να παραμείνει στον πυρήνα ή να εξέλθει στο κυτταρόπλασμα μέσω του συμπλέγματος πυρηνικών πόρων.

Εάν η έκταση του DNA μεταγραφεί σε ένα μόριο RNA που κωδικοποιεί μια πρωτεΐνη, το RNA ονομάζεται αγγελιαφόρο RNA (mRNA). το mRNA, με τη σειρά του, χρησιμεύει ως πρότυπο για τη σύνθεση της πρωτεΐνης μέσω μετάφρασης. Άλλα τμήματα του DNA μπορούν να μεταγραφούν σε μικρά μη κωδικοποιητικά RNA, όπως microRNA, RNA μεταφοράς (tRNA), μικρό πυρηνικό RNA (snoRNA), μικρό πυρηνικό RNA (snRNA) ή ενζυμικά μόρια RNA, που ονομάζονται ριβοένζυμα, καθώς και μεγαλύτερα μη κωδικοποιητικά RNA, όπως το ριβοσωμικό RNA (rRNA) και το μακρύ μη κωδικοποιητικό RNA (lncRNA), όπως αναφέρθηκε και στην αμέσως προηγούμενη ενότητα. Τέλος, στην ιολογία, ο όρος μεταγραφή μπορεί επίσης να χρησιμοποιηθεί όταν αναφέρεται στη σύνθεση mRNA από ένα μόριο RNA (δηλαδή, ισοδύναμο με αντιγραφή RNA).

4.1.5 Δευτεροταγής δομή του RNA

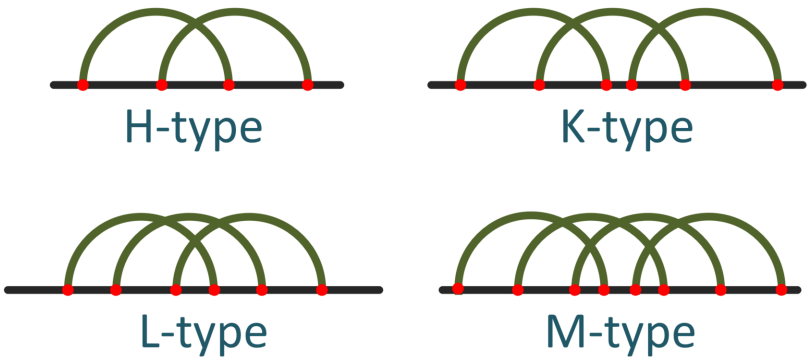
Τα ριβονουκλεϊκά οξέα είναι συνήθως μονές αλυσίδες. Σε μερικές περιπτώσεις, όπως σε RNA ιούς, βρίσκονται με μορφή διπλής έλικας. Οι μονές αλυσίδες των RNAs μπορούν να σχηματίσουν διπλοελικωμένα τμήματα, είτε μεταξύ τους, είτε συνηθέστερα τμήματα μεταξύ της ίδιας RNA αλυσίδας. Η δευτεροταγής δομή των νουκλεϊκών οξέων αναφέρεται σε αυτά τα τοπικά μοτίβα αναδίπλωσης και στις αλληλεπιδράσεις μεταξύ των νουκλεοτιδικών βάσεων μέσα σε μια ενιαία αλυσίδα RNA [57]. Η δευτεροταγής δομή του RNA είναι μεγάλης σημασίας καθώς επηρεάζει την αναδίπλωση, τη σταθερότητα και τη λειτουργία των μορίων RNA. Στα διπλοελικωμένα τμήματα βρίσκουμε τα ζεύγη των βάσεων A-U και G-C αλλά ακόμη παρατηρούνται και ασυνήθιστα ζεύγη βάσεων όπως G-U. Τα ελικωμένα τμήματα που σχηματίζονται με αυτόν τον τρόπο σπάνια είναι κανονικά, διότι τα τμήματα της RNA αλυσίδας που



Σχήμα 4.2: Οι πιο γνωστές δευτεροταγείς δομές ([55])

.

συμβάλλουν στο σχηματισμό τους δεν είναι τέλεια συμπληρωματικά μεταξύ τους. Έτσι υπολείμματα βάσεων που δε ζευγαρώνουν ξεφεύγουν έξω από το διπλοελικωμένο τμήμα σχηματίζοντας βρόχους ή ανακάμψεις (loops). Στο Σχήμα 4.2, παρουσιάζονται κάποια από τα σημαντικότερα μοτίβα που εντοπίζονται στις δευτεροταγείς δομές του RNA. Η αναλογία των ελικωμένων περιοχών προς τις μη ελικωμένες διαφοροποιείται στα διάφορα RNA και ποικίλλει πολύ. Σε ορισμένα μόρια του RNA, όπως του tRNA, μέχρι και 70% των υπολειμμάτων των βάσεων μπορεί να συμμετέχουν σε διπλοελικωμένα τμήματα. Οι αναδιπλούμενες αυτές δομές του μορίου, και κατ΄ επέκταση η δευτεροταγής και η τριτοταγής τους δομές, διαδραματίζουν κρίσιμο ρόλο στις ποικίλες



Σχήμα 4.3: Οι πιο κοινοί τύποι ψευδοκόμβων ([56]).

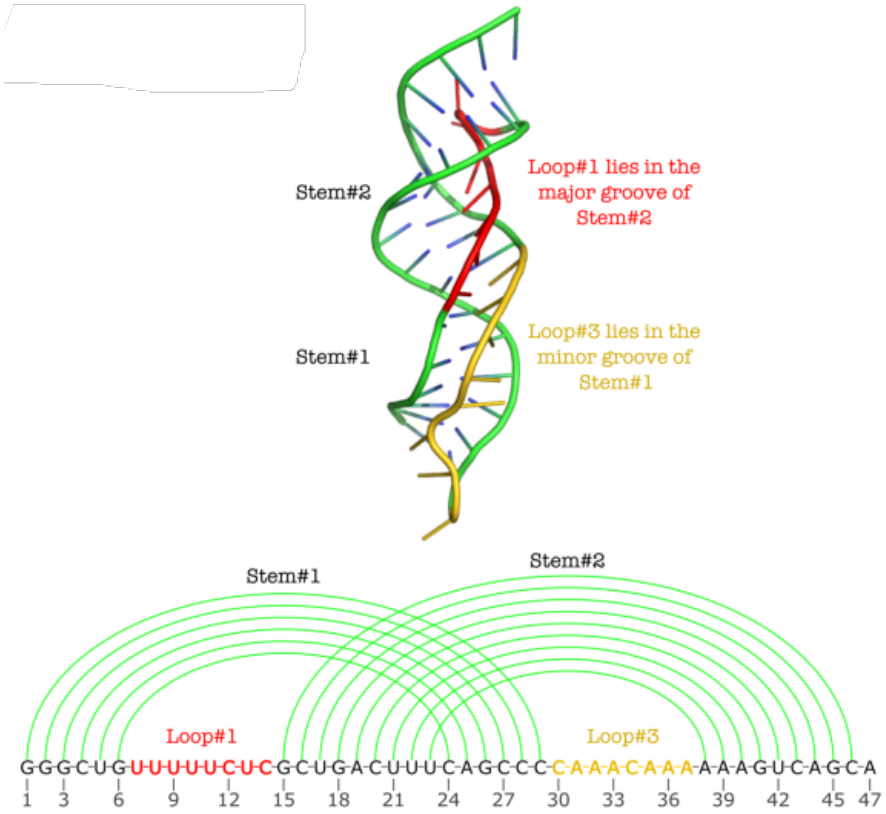
λειτουργίες του, με σηματικότερα τη συμμετοχή του σε ενζυμικές αντιδράσεις, στη γονιδιακή ρύθμιση, καθώς και στην αλληλεπίδραση του μορίου με άλλα μόρια.

4.1.6 Το μοτίβο του ψευδοκόμβου

Ένα από τα λιγότερο συχνά μοτίβα στις ακολουθίες RNA αλλά δύσκολο όσον αφορά την πρόβλεψη είναι ο ψευδοκόμβος. Θεωρείται ότι υπάρχει ψευδοκόμβος όταν δύο ζευγάρια βάσεων διασταυρώνονται. Αυτό το μοτίβο παρατηρήθηκε αρχικά στον ιό *Turnip Yellow Mosaic* [58]. Ο απλούστερος αλλά συχνότερος τύπος ψευδοκόμβου δημιουργείται από δύο ενότητες μονής αλυσίδας και δύο μη ζευγαρωμένες περιοχές (βρόχους). Πολλές παραλλαγές έχουν παρατηρηθεί, αλλά οι τέσσερις κύριοι τύποι είναι οι τύποι H, K, L και M, όπως φαίνονται στο Σχήμα 4.3 [56]. Συγκεκριμένα, ο τύπος H ψευδοκόμβου [17] αποτελείται από δύο τμήματα με ζεύγη βάσεων (stems) και δύο βρόχους (loops) αυθαίρετου μήκους. Η διασταύρωση ενός ζευγαριού βάσεων (ή κεντρικών βάσεων στην ορολογία μας) οδηγεί στη δημιουργία του. Για την καλύτερη οπτικοποίηση του ψευδοκόμβου παρουσιάζεται η τριτοταγής του δομή στο Σχήμα 4.4.

4.1.7 Ασύμμετροι βρόχοι ή Εξογκώματα (bulges) και εσωτερικοί βρόχοι (internal loops)

Οι ασύμμετροι βρόχοι/εξογκώματα δημιουργούνται από μη ζευγαρωμένες βάσεις (A, U, G, C) και το μέγεθός τους μπορεί να είναι από μία έως πολλές μη ζευγαρωμένες βάσεις. Η εμφάνισή τους σε όλους τους τύπους δομημένων λειτουργικών RNA [20] τονίζει τη σπουδαιότητά τους και οδήγησε στο να συμπεριληφθούν στην παρούσα έρευνα στο πλαίσιο πρόβλεψης ψευδοκόμβων. Για να απεικονίσουμε αυτό το μοτίβο, παρουσιάζουμε τις μη ζευγαρωμένες βάσεις που σχηματίζουν μια προεξοχή με κόκκινες τελείες στο Σχήμα 4.5 (α). Οι εσωτερικοί βρόχοι μπορεί να δημιουργηθούν σε μια ακολουθία RNA όταν το διπλό-αλυσιδωτό RNA διαχωρίζεται ως αποτέλεσμα της μη ζευγάρωσης μεταξύ των νουκλεοτιδίων. Η διαφορά μεταξύ των εσωτερικών

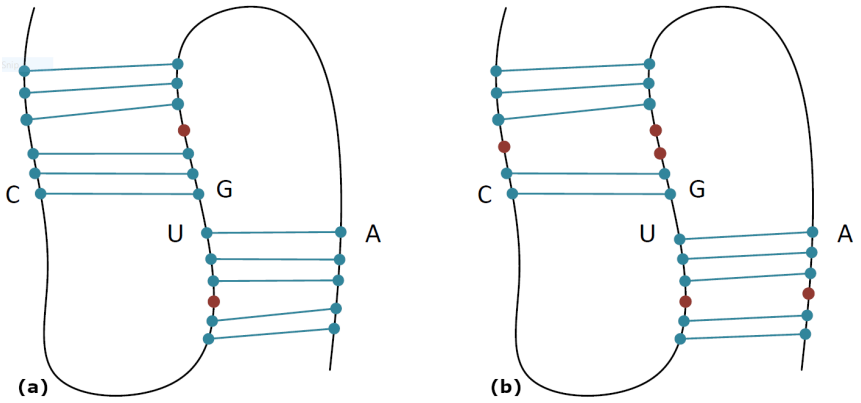


Σχήμα 4.4: Παράδειγμα τρισδιάστατης δομής ψευδοκόμβου τύπου H (πάνω μέρος) και η αντίστοιχη αναπαράσταση τόξου της δομής (κάτω μέρος) [59].

βρόχων και των βρόχων βλαστών είναι ότι οι εσωτερικοί βρόχοι υπάρχουν στη μέση μιας τεντωμένης διπλής αλυσίδας RNA. Για να απεικονίσουμε αυτό το μοτίβο, παρουσιάζουμε τις μη ζευγαρωμένες βάσεις που σχηματίζουν έναν εσωτερικό βρόχο με κόκκινες τελείες στο Σχήμα 4.5 (β).

4.1.8 Άλλα μοτίβα βρόχων/ανακάμψεων

Το RNA υφίσταται αναδίπλωση, με τις συμπληρωματικές του βάσεις να συνδέονται μέσω δεσμών υδρογόνου, δημιουργώντας μια διπλή έλικα. Αυτή η δομή αποτελείται κυρίως από περιοχές συμπληρωματικών βάσεων (helix ή stem) μέσω αντιστοίχισης βάσεων Watson-Crick (WC), καθώς και από βρόχους (loops). Οι βρόχοι είναι περιοχές του RNA με μη συζευγμένες βάσεις, ταξινομημένες βάσει του αριθμού, της διάταξης των γειτονικών τμημάτων του μορίου, και της θέσης τους στο μόριο. Εκτός από τους εσωτερικούς και ασσύμετρους βρόχους που αναφέραμε προηγουμένως, υπάρχουν και κάποια ακόμα σημαντικά μοτίβα βρόχων που παρουσιάζονται στις δομές RNA, με μεγάλη συχνότητα και συνεισφορά σε σημαντικές λειτουργίες. Ένα



Σχήμα 4.5: Η-τύπου ψευδοκόμβοι με εξογκώματα (α) και εσωτερικούς βρόχους (β). Οι μη ζευγαρωμένες βάσεις που σχηματίζουν εξογκώματα ή εσωτερικούς βρόχους αναπαρίστανται με κόκκινο.

βασικό δομικό στοιχείο στη δευτεροταγή δομή του RNA είναι ο βρόχος φουρκέτας (hairpin loop ή stem-loop). Πρόκειται για μια διπλή έλικα που καταλήγει σε έναν βρόχο. Αντίστοιχα, οι πολυδιακλαδισμένοι βρόχοι (multi-branch loop) αποτελούνται έναν βρόχο όπου τρία ή περισσότερα τμήματα stems συνδέονται με αυτόν. Αυτοί οι τύποι βρόχων είναι εξίσου συχνοί, όπως και άλλες δομές της δευτεροταγούς δομής του RNA, και η ανάλυσή τους έχει γίνει σε πλήθος ερευνητικών εργασιών.

4.2 Θερμοδυναμική RNA και η ελεύθερη ενέργεια

Η σταθερότητα μιας δευτεροταγούς δομής RNA ποσοτικοποιείται από τη μεταβολή της ελεύθερης ενέργειας ΔG, η οποία μετριέται σε kcal/mol. Η ελεύθερη ενέργεια G υποδεικνύει την κατεύθυνση μιας αυθόρμητης αλλαγής και εισήχθη από τον J.W.Gibbs το 1878. Η μεταβολή της ελεύθερης ενέργειας ΔG ποσοτικοποιεί τη διαφορά της ελεύθερης ενέργειας μεταξύ της αναδιπλωμένης κατάστασης του μορίου και της μη αναδιπλωμένης κατάστασης. Η ΔG αντιπροσωπεύει το έργο που επιτελείται από ένα σύστημα σε σταθερή θερμοκρασία και πίεση, όταν υφίσταται μια αντιστρεπτή διεργασία. Ένα διπλωμένο RNA έχει αρνητική μεταβολή ελεύθερης ενέργειας και όσο μικρότερη είναι, τόσο πιο σταθερή είναι η δομή. Τα ζεύγη βάσεων είναι συνήθως ευνοϊκά για τη σταθερότητα (δηλαδή συνεισφέρουν αρνητική μεταβολή της ελεύθερης ενέργειας), ενώ οι βρόχοι είναι συνήθως αποσταθεροποιητικοί (δηλαδή έχουν θετικές τιμές ενέργειας).

Η ενέργεια Gibbs (G) ορίζεται ως:

$$G = H - TS$$

όπου:

- H είναι η ενθαλπία του συστήματος,
- T είναι η απόλυτη θερμοκρασία,
- S είναι η εντροπία.

Για μια διαδικασία που συμβαίνει σε σταθερή θερμοκρασία και πίεση, η μεταβολή της ελεύθερης ενέργειας είναι συνάρτηση της μεταβολής της ενθαλπίας ΔH, της μεταβολής της εντροπίας ΔS και της θερμοκρασίας T (σε Kelvin) και δίνεται από τον τύπο:

$$\Delta G = \Delta H - T\Delta S$$

Η ενθαλπία (H) είναι μια μέτρηση της ροής θερμότητας που προκύπτει σε μια διεργασία. Η μεταβολή της ενθαλπίας (ΔH) για μια εξώθερμη αντίδραση, όπως η αναδίπλωση του RNA, (δηλαδή, η θερμότητα ρέει από το σύστημα προς το περιβάλλον) είναι αρνητική. Η ενθαλπία μετριέται σε kcal/mol. Ο σχηματισμός στελεχών RNA είναι ο κυρίαρχος ενθαλπικός παράγοντας, μέσω δεσμών υδρογόνου και αλληλεπιδράσεων στοίβαξης. Η εντροπία (S) είναι ευρέως αποδεκτή ως θερμοδυναμική συνάρτηση που μετρά την αταξία ενός συστήματος. Έτσι, η μεταβολή της εντροπίας ΔS μετρά την αλλαγή του βαθμού αταξίας. Εάν η ΔS είναι θετική, σημαίνει ότι υπήρξε αύξηση του βαθμού αταξίας. Μια αρνητική τιμή υποδηλώνει μείωση της αταξίας.
Ωστόσο, μια σύγχρονη θεώρηση της μεταβολής της εντροπίας την παρουσιάζει ως την ποσότητα διασποράς της ενέργειας ανά θερμοκρασία ή από τη μεταβολή του αριθμού των μικροκαταστάσεων: πόση ενέργεια διασπείρεται σε μια διεργασία ή πόσο ευρέως διαδεδομένη γίνεται σε μια συγκεκριμένη θερμοκρασία. Εάν το ΔS είναι αρνητικό, όπως για τους βρόχους RNA, σημαίνει ότι η ποσότητα της ενέργειας που διασκορπίζεται μειώνεται. Οι βρόχοι σε μια δομή RNA συμμετέχουν στην εντροπία περισσότερο από ό,τι στην ενθαλπία, επειδή η διαδικασία αναδίπλωσης περιορίζει τις μικροκαταστάσεις των νουκλεοτιδίων του βρόχου σε σύγκριση με την ξεδιπλωμένη αλυσίδα. Η εντροπία μετράται σε kcal/(mol K) ή μονάδες εντροπίας (1eu=1cal/(molK)). Στη διατριβή χρησιμοποιούμε τις μεταβολές της ελεύθερης ενέργειας για να ποσοτικοποιήσουμε τη σταθερότητα της δευτεροταγούς δομής του RNA.
Το εργαστήριο Turner και οι συνεργάτες του έχουν πραγματοποιήσει εκατοντάδες πειράματα, κυρίως με οπτική τήξη μικρών αλληλουχιών RNA, για τον προσδιορισμό της ελεύθερης ενέργειας των δομών που σχηματίζονται. Η συμβολή πολλών ερευνητών κατά τη διάρκεια περισσότερων από δύο δεκαετιών απέδωσε το μοντέλο Turner, το οποίο είναι ευρέως αποδεκτό ως βιολογικά ρεαλιστικό. Το μοντέλο Turner είναι ένα θερμοδυναμικό μοντέλο κοντινότερων γειτόνων, δηλαδή υποθέτει ότι η σταθερότητα ενός ζεύγους βάσεων ή ενός βρόχου εξαρτάται από την αλληλουχία του και την αλληλουχία του πιο γειτονικού ζεύγους βάσεων και τις αντίστοιχες στοιβάξεις. Στο Σχήμα 4.6 παρουσιάζονται οι κανόνες/συνεισφορές για την ελεύθερη ενέργεια με βάση το μοντέλο του Turner.

BOTTOM \ TOP	AU	CG	GC	UA	GU	UG
AU	-0.9	-1.8	-2.3	-1.1	-0.5	-0.7
CG	-2.1	-2.9	-3.4	-2.3	-1.5	-1.5
GC	-1.7	-2	-2.9	-1.8	-1.3	-1.5
UA	-0.9	-1.7	-2.1	-0.9	-0.7	-0.5
GU	-0.9	-1.7	-2.1	-0.9	-0.5	-0.5
UG	-0.9	-1.7	-2.1	-0.9	0.6	-0.5

Bases in Loop	Internal Loop	Bulge Loop	Hairpin Loop
1	0	3.3	0
2	0.8	5.2	0
3	1.3	6	7.4
4	1.7	6.7	5.9
5	2.1	7.4	4.4
6	2.5	8.2	4.3
7	2.6	9.1	4.1
8	2.8	10	4.1

Σχήμα 4.6: Οι κανόνες για την ελεύθερη ενέργεια με βάση το μοντέλο του Turner.

4.3 Ο Αλγόριθμος Nussinov

Ο αλγόριθμος Nussinov αναζητά μια δευτεροταγή δομή με τον μέγιστο αριθμό ζευγών βάσεων. Είναι ένας απλός και αποτελεσματικός αλγόριθμος δυναμικού προγραμματισμού. Ορίζει μια συνάρτηση $\delta(i, j)$, η οποία ισούται με 1 εάν οι βάσεις (νουκλεοτίδια) i^{th} και j^{th} είναι ένα συμπληρωματικό ζεύγος βάσεων, διαφορετικά $\delta(i, j) = 0$. Στη συνέχεια, υπολογίζει αναδρομικά τα στοιχεία ενός πίνακα αποτελεσμάτων $\gamma(i, j)$, ο οποίος είναι ο μέγιστος αριθμός ζευγών βάσεων που μπορούν να βρεθούν στην υποακολουθία $i \dots j$. Ο αλγόριθμος περιλαμβάνει ένα στάδιο συμπλήρωσης και ένα στάδιο εντοπισμού. Στο στάδιο συμπλήρωσης ο πίνακας δυναμικού προγραμματισμού συμπληρώνεται κατάλληλα. Η τελική λύση δίνεται από το $\gamma(1, L)$, όπου L είναι το μήκος της ακολουθίας. Αυτός ο αλγόριθμος παρουσιάζει πολυπλοκότητα $O(L^3)$ σε χρόνο και $O(L^2)$ σε μνήμη. Το στάδιο εντοπισμού έχει χρόνο $O(L)$ και εκεί εντοπίζεται η ακριβής ακολουθία που προβλέφθηκε. Μερικές φορές, υπάρχουν πολλές δομές με τον ίδιο αριθμό ζευγών βάσεων. Ωστόσο, αυτός ο αλγόριθμος εντοπισμού εντοπίζει μόνο μία από τις καλύτερες δομές. Ο αλγόριθμος Nussinov έχει διάφορους περιορισμούς όπως ότι δεν μπορεί να χειριστεί ψευδοκόμβους και η μεγιστοποίηση του αριθμού ζευγών βάσεων είναι ένα κριτήριο που από μόνο του δεν μπορεί να δώσει ακριβή πρόβλεψη. Ωστόσο, η ιδέα της μεγιστοποίησης των ζευγών βάσεων αποτέλεσε μία κεντρική ιδέα στην παρούσα διατριβή, που σε συνδυασμό με άλλες τεχνικές,

οδήγησαν στη δημιουργία συστημάτων με υψηλή προβλεπτική ικανότητα.
Στη συνέχεια ακολουθούν τα βήματα του αλγορίθμου Nussinov για το στάδιο της συμπλήρωσης της αρχικοποίησης και της συμπλήρωσης του πίνακα και στη συνέχεια για τον εντοπισμό της βέλτιστης δομής.

Αρχικοποίηση:

$$\gamma(i,i) = 0 \text{ for } 1 \leq i \leq L \tag{4.1}$$
$$\gamma(i,i-1) = 0 \text{ for } 2 \leq i \leq L \tag{4.2}$$

Αναδρομή:

$$\gamma(i,j) = \max \begin{cases} \gamma(i+1,j) \\ \gamma(i,j-1) \\ \gamma(i+1,j-1) + \delta(i,j) \\ \max_{i<k<j}(\gamma(i,k) + \gamma(k+1,j)) \end{cases} \tag{4.3}$$

Στάδιο Εντοπισμού της βέλτιστης δομής

Αρχικοποίηση:
push $(1, L)$ into the stack
Αναδρομή:
while stack is not empty
    Pop (i, j)
    if $i \geq j$ **then**
        continue
    else if $\gamma(i+1, j) = \gamma(i, j)$ **then**
        push$(i+1, j)$
    else if $\gamma(i, j-1) = \gamma(i, j)$ **then**
        push$(i, j-1)$
    else if $\gamma(i+1, j-1) = \gamma(i, j)$ **then**
        record i, j base-pair
        push$(i+1, j-1)$
    else
        for $k = i+1$ **to** $j-1$ **do**
            if $\gamma(i, k) + \gamma(k+1, j) = \gamma(i, j)$ **then**
                push$(k+1, j)$
                push(i, k)
                break **do**
            end if
        end for
    end if
end while

4.4 Συντακτική αναγνώριση προτύπων

Η συνακτική αναγνώριση προτύπων (syntactic pattern recognition) αποτελεί έναν κλάδο της τεχνητής νοημοσύνης και της επεξεργασίας σημάτων, ο οποίος στηρίζεται στη θεωρία γλωσσών και συντακτικών κανόνων για την αναγνώριση και την ανάλυση προτύπων. Στον πυρήνα αυτής της προσέγγισης βρίσκεται η ιδέα ότι κάθε πρότυπο μπορεί να περιγραφεί μέσω μιας γλώσσας, η οποία ορίζεται από ένα σύνολο συντακτικών κανόνων ή γραμματικής. Αυτή η γραμματική χρησιμοποιείται για την κατασκευή δέντρων συντακτικής ανάλυσης (parse trees), τα οποία περιέχουν τα ενδιαφέροντα σύμβολα ή τμήματα των δεδομένων στους τερματικούς κόμβους τους. Η γραμματική αποτελείται από το σύνολο των συντακτικών κανόνων και ένα λεξιλόγιο. Σύμφωνα με αυτά τα στοιχεία αναγνωρίζεται ή όχι αν μία ακολουθία συμβόλων ανήκει στη συγκεκριμένη γλώσσα. Η συντακτική αναγνώριση προτύπων έχει εφαρμογές σε διάφορους τομείς, όπως η επεξεργασία ομιλίας, η αναγνώριση χειρογράφων και η βιοπληροφορική. Σε αυτό το πλαίσιο, οι προσεγγίσεις που βασίζονται στη συντακτική αναγνώριση προτύπων έχουν αποδειχθεί ιδιαίτερα χρήσιμες, διότι επιτρέπουν την ακριβή και λεπτομερή ανάλυση των δεδομένων, ακόμα και σε περιπτώσεις όπου η συμβατική στατιστική προσέγγιση αποτυγχάνει. Στη βάση της συντακτικής αναγνώρισης προτύπων βρίσκεται η θεωρία των γλωσσών και της γραμματικής, όπως αναπτύχθηκε από τον Noam Chomsky. Σύμφωνα με την ιεραρχία του Chomsky, οι γραμματικές κατατάσσονται σε τέσσερις κατηγορίες, με τις context-free γραμματικές (CFG) να αποτελούν μία από αυτές. Οι CFG είναι ιδιαίτερα χρήσιμες στην ανάλυση και την επεξεργασία γλωσσικών δομών, καθώς επιτρέπουν την κατασκευή συντακτικών δέντρων για πολύπλοκες δομές, όπως είναι στην περίπτωση της παρούσας διατριβής τα πολύπλοκα μοτίβα των ψευδοκόμβων.

4.4.1 Ιεραρχία Chomsky

Η ιεραρχία Chomsky, που πήρε το όνομά της προς τιμήν του Αμερικανού γλωσσολόγου Noam Chomsky, αποτελεί μια σημαντική θεωρητική δομή στη μελέτη των γλωσσών και των γραμματικών στον τομέα της θεωρίας αυτομάτων και της επεξεργασίας φυσικών γλωσσών. Αυτή η ιεραρχία κατατάσσει τις γραμματικές σε τέσσερις επίπεδες κατηγορίες, βάσει της πολυπλοκότητας των κανόνων παραγωγής που χρησιμοποιούν. Η κάθε κατηγορία στην ιεραρχία αντιπροσωπεύει ένα διαφορετικό επίπεδο περιορισμού στους κανόνες που μπορεί να εφαρμόσει η γραμματική, και έτσι επηρεάζει την κλάση γλωσσών που μπορεί να περιγράψει. Η πρώτη κατηγορία, γνωστή ως Τύπος 3 ή γραμματικές τελεστών (regular grammars), είναι η πιο απλή και περιοριστική. Οι γραμματικές αυτού του τύπου μπορούν να παράγουν μόνο γλώσσες με πολύ απλές δομές και συχνά χρησιμοποιούνται για την περιγραφή των απλούστερων δομών στις φυσικές γλώσσες ή σε προγραμματιστικές γλώσσες. Στη συνέχεια, για πιο σύνθετα προβλήματα υπάρχει η κατηγορία Τύπου 2, ή context-free γραμματικές (CFG). Οι CFG είναι ικανές να παράγουν γλώσσες με πιο πολύπλοκες δομές και είναι ιδιαίτερα χρήσιμες στην ανάλυση και την παραγωγή δομών όπως αυτές που βρίσκονται στις

φυσικές γλώσσες και στις γλώσσες προγραμματισμού. Οι CFG αποτελούν τη βάση για πολλά εργαλεία ανάλυσης γλωσσών και επιτρέπουν την ανάπτυξη πολύπλοκων συντακτικών δομών. Η τρίτη κατηγορία, οι context-sensitive γραμματικές (CSG) ή Τύπος 1, επιτρέπει ακόμα πιο πολύπλοκες δομές. Οι CSG μπορούν να περιγράψουν γλώσσες που δεν μπορούν να αναπαρασταθούν με τις πιο απλές CFG. Ωστόσο, λόγω της αυξημένης πολυπλοκότητας, οι CSG είναι λιγότερο πρακτικές για χρήση σε πολλές εφαρμογές. Τέλος, ο Τύπος 0, ή απεριόριστες γραμματικές, περιλαμβάνει όλες τις δυνατές γραμματικές και μπορεί να παράγει οποιαδήποτε γλώσσα που είναι αναγνωρίσιμη από μία Turing machine. Αυτή η κατηγορία είναι η πιο γενική και περιλαμβάνει κάθε δυνατή γλώσσα, αλλά η χρήση της σε πρακτικές εφαρμογές είναι περιορισμένη λόγω της εξαιρετικά υψηλής πολυπλοκότητας. Η ιεραρχία Chomsky παρέχει ένα πλαίσιο για την κατανόηση των διαφορετικών επιπέδων πολυπλοκότητας στις γλωσσικές δομές και επιτρέπει στους ερευνητές και τους μηχανικούς να επιλέγουν το κατάλληλο επίπεδο γραμματικής για την ανάλυση ή την παραγωγή μιας συγκεκριμένης γλώσσας.

4.4.2 Γραμματικές Χωρίς Συμφραζόμενα (Context-free Grammars)

Οι Γραμματικές Χωρίς Συμφραζόμενα (Context-Free Grammars - CFG) είναι ένας τύπος γραμματικής στη θεωρία της συντακτικής ανάλυσης και επεξεργασίας γλωσσών. Μία CFG αποτελείται από ένα σύνολο κανόνων που καθορίζουν πώς συμβολοσειρές της γλώσσας μπορούν να παραχθούν από ένα αρχικό σύμβολο. Κάθε κανόνας αναφέρεται σε ένα μη τερματικό σύμβολο και ορίζει τον τρόπο παραγωγής συμβολοσειρών από αυτό το σύμβολο. Η βασική ιδέα πίσω από τις CFG είναι η απλοποίηση της αναπαράστασης της συντακτικής δομής μιας γλώσσας. Αυτό επιτυγχάνεται μέσω της χρήσης μη τερματικών συμβόλων (που αντιπροσωπεύουν κατηγορίες λέξεων ή φράσεων) και τερματικών συμβόλων (που αντιπροσωπεύουν τις ίδιες τις λέξεις της γλώσσας). Οι κανόνες της CFG περιγράφουν τον τρόπο με τον οποίο τα μη τερματικά σύμβολα μπορούν να αντικατασταθούν με συμβολοσειρές αποτελούμενες από άλλα μη τερματικά ή τερματικά σύμβολα. Οι CFG έχουν τέσσερα βασικά στοιχεία: τα σύνολα NT, T, R, και S. Το S είναι το αρχικό μη τερματικό σύμβολο, τα τερματικά και μη τερματικά σύμβολα απαρτίζουν τα σύνολα T και NT αντίστοιχα, ενώ οι συντακτικοί κανόνες αποτελούν το σύνολο R. Η σημειολογία των κανόνων είναι $L \to \delta$, όπου $L \in NT$ και $\delta \in (T \cup NT)^*$, ορίζοντας ότι το L μπορεί να παράγει μια σειρά από σύμβολα δ [67]. Λόγω της υψηλής εκφραστικής τους ικανότητας, υπάρχει ένας σημαντικός αριθμός αναλυτών στη βιβλιογραφία. Οι πιο πολύ αναφερόμενοι και ευρέως χρησιμοποιούμενοι αλγόριθμοι είναι ο CYK [69], που εισήχθη από τους Cocke, Younger, και Kasami, και ο αλγόριθμος του Earley [70]. Τροποποιήσεις των παραπάνω αναλυτών παρουσιάζονται στα [71, 72, 73] και ως παράλληλες εκδοχές στα [74, 75]. Τα συστήματα που αναπτύχθηκαν στην παρούσα διατριβή για την πρόβλεψη δευτεροταγών δομών RNA, ενσωματώνουν τον Yet Another Early Parser (YAEP) [76], ο οποίος αποτελεί μια αποδοτική υλοποίηση του αλγορίθμου Earley για ασαφείς γραμματικές. Μία περιγραφή Earley ακολουθεί στην επόμενη ενότητα, ενώ ο Yaep παρουσιάζεται στην ενότητα

4.4.4.

4.4.3 0 αλγόριθμος του Early

Ο αλγόριθμος ανάλυσης για τις Γραμματικές Χωρίς Συμφραζόμενα που παρουσιάστηκε από τον Earley το 1970 κατασκευάζει το δέντρο ανάλυσης χρησιμοποιώντας μια μεθοδολογία από πάνω προς τα κάτω. Ο αλγόριθμος του Earley τοποθετεί το σύμβολο τελείας ' • ' $\notin (N \cup NT)$, σε κάθε κανόνα παράγοντας dotted rules. Η ύπαρξη τελείας σε έναν κανόνα δηλώνει ότι το μέρος του κανόνα πριν από την τελεία έχει αναγνωριστεί, ενώ το μέρος του κανόνα μετά την τελεία δεν έχει ακόμη αναγνωριστεί. Σε περίπτωση που μια τελεία φτάσει στην τελευταία θέση ενός κανόνα που έχει το ριζικό σύμβολο (root symbol) στην αριστερή του πλευρά, τότε η είσοδος θεωρείται αναγνωρισμένη. Αυτός ο αλγόριθμος ορίζει και εφαρμόζει λειτουργίες, οι οποίες ονομάζονται (PARSER), (PREDICTOR), και (COMPLETER). Η είσοδος α=$\alpha_1\alpha_2\alpha_3 \ldots \alpha_n$ σαρώνεται από α_1 έως αn. Καθώς σαρώνεται κάθε σύμβολο εισόδου, κατασκευάζεται ένα σύνολο δεδομένων, αντιπροσωπεύοντας την κατάσταση της διαδικασίας αναγνώρισης σε αυτό το σημείο της σάρωσης. Ως εκ τούτου, ο αλγόριθμος κατασκευάζει $n+1$ σύνολα δεδομένων καταστάσεων. Μια κατάσταση είναι απλώς ένα σύνολο τριών ακεραίων SR, p, F. Το SR υποδηλώνει τον αριθμό του κανόνα· το p είναι η θέση του συμβόλου ' • '· και το F είναι η αρίθμηση του συνόλου όπου ο dotted rule αρχικά δημιουργήθηκε. Μια κατάσταση σε ένα σύνολο δεδομένων του Earley είναι της μορφής $i : {}_F Y \rightarrow \alpha \bullet Z\gamma$, σημαίνοντας τον συντακτικό κανόνα $Y \rightarrow \alpha Z\gamma$ με το σύμβολο ' • ' στην p^{th} θέση ($|\alpha| = p$), αρχικά δημιουργήθηκε στο F^{th} σύνολο δεδομένων και βρίσκεται στο σύνολο δεδομένων S_i. Καθώς η ανάγνωση των συμβόλων εισόδου συνεχίζεται, δημιουργούνται νέα σύνολα δεδομένων dotted rule. Οι τρεις λειτουργίες εφαρμόζονται διαδοχικά σε κάθε dotted rule όλων των συνόλων. Η παρουσία ενός ολοκληρωμένου dotted rule με ένα ριζικό σύμβολο στην αριστερή πλευρά του κανόνα, στο τελευταίο σύνολο δεδομένων, σηματοδοτεί την αναγνώριση της εισόδου.

Ο αλγόριθμος του συντακτικού αναλυτή Earley παρουσιάζεται στον Αλγόριθμο 1. Στην κύρια συνάρτηση EARLEY_PARSER, αρχικοποιείται ένας πίνακας συνόλων που περιέχουν καταστάσεις ανάλογα με το μήκος της συμβολοσειράς εισόδου (INITIALIZE(input_string)) και προστίθεται μια κατάσταση που έχει τον χαρακτήρα '•' στην αριστερή πλευρά του αρχικού συμβόλου S στο σύνολο με αρίθμηση 0 (συγκεκριμένα ADD_TO_SET((Start → • S, 0), Sets[0])). Στη συνέχεια, εκτελείται ένας διπλός εμφωλευμένος βρόχος. Ο εμφωλευμένος βρόχος εξετάζει κάθε κατάσταση σε κάθε σύνολο, και ένα σύνολο μπορεί να επεκταθεί κατά τη διάρκεια αυτού του βρόχου, καθώς οι τρεις λειτουργίες προσθέτουν καταστάσεις στα σύνολα. Σε κάθε κατάσταση, εξετάζεται αν στη δεξιά πλευρά του συμβόλου '•' υπάρχει ένα μη τελικό σύμβολο, ένα τελικό σύμβολο, ή αν η κατάσταση έχει ολοκληρωθεί (το '•' βρίσκεται στο τέλος του κανόνα) και καλούνται αντίστοιχα οι συναρτήσεις PREDICTOR, SCANNER, COMPLETER κατά περίπτωση. Στην περίπτωση που καλείται η συνάρτηση PREDICTOR, τότε για το μη τελικό σύμβολο που βρίσκεται δεξιά του '•' (μη

τελικό σύμβολο C στον ψευδοκώδικα ως τελεία κανόνας B $\rightarrow \alpha$ • C β), διατρέχονται όλοι οι γραμματικοί κανόνες για να επιλεγούν οι κανόνες που έχουν αυτό το σύμβολο στην αριστερή πλευρά του κανόνα (C $\rightarrow \delta$). Οι επιλεγμένοι κανόνες προστίθενται στο σύνολο αυτό, αφού τοποθετηθεί το '•' στην πρώτη θέση της δεξιάς πλευράς του κανόνα (C $\rightarrow$ • δ).
Όταν καλείται η συνάρτηση SCANNER, εάν το τελικό σύμβολο που βρίσκεται δεξιά του '•' (τελικό σύμβολο a στον ψευδοκώδικα ως τελεία κανόνας B $\rightarrow \gamma$ • a δ) είναι ίσο με το τρέχον εξεταζόμενο σύμβολο της συμβολοσειράς εισόδου (input_string[i]), αυτή η κατάσταση προστίθεται στο επόμενο σύνολο μετακινώντας το '•' μία θέση δεξιά (B $\rightarrow \gamma$ a • δ).
Όταν καλείται η συνάρτηση COMPLETER, διατρέχονται οι καταστάσεις στο σύνολο όπου η ολοκληρωμένη κατάσταση (A $\rightarrow \delta$ •) δημιουργήθηκε αρχικά (x στον ψευδοκώδικα) για να επιλεγούν οι καταστάσεις που έχουν το σύμβολο (A στον ψευδοκώδικα) στην αριστερή πλευρά του κανόνα, μία θέση μετά το '•' (B $\rightarrow \gamma$ •Aβ). Αυτές οι καταστάσεις προστίθενται στα Sets[i] μετακινώντας την τελεία μία θέση δεξιά (B $\rightarrow$ γ A • β).

4.4.4 0 αναλυτής Yaep

Το YAEP, συντομογραφία του Yet Another Earley Parser, αποτελεί μια αυτόνομη βιβλιοθήκη που έχει δημιουργηθεί και αποτελεί μία από τις ταχύτερες υλοποιήσεις του αναλυτή Earley. Μπορεί να αναλύσει 300.000 γραμμές προγράμματος σε γλώσσα C ανά δευτερόλεπτο και δεσμεύει περίπου 5MB μνήμης για ένα πρόγραμμα C 10.000 γραμμών. Οι δυνατότητές του περιλαμβάνουν την απλή κατευθείαν μετάφραση της σύνταξης, παράγοντας ένα δέντρο αφαιρετικής σύνταξης. Μπορεί να αναλύσει εισόδους που περιγράφονται από μια ασαφή γραμματική και να παράγει μια συμπαγή αναπαράσταση όλων των δυνατών δέντρων ανάλυσης χρησιμοποιώντας έναν κατευθυνόμενο ακυκλικό γράφο (Directed Acyclic Graph-DAG) αντί αναπαραστάσεων δέντρων. Επιπλέον, μπορεί να εκτελέσει τη συντακτική ανάκτηση λαθών, βρίσκοντας την ανάκτηση με τον ελάχιστο αριθμό αγνοημένων διατάξεων, επιτρέποντας την υλοποίηση αναλυτών με εξαιρετικά καλή ανάκτηση και αναφορά σφαλμάτων. Τέλος, έχει γρήγορη εκκίνηση και χρειάζεται ελάχιστη καθυστέρηση μεταξύ της επεξεργασίας της γραμματικής και της έναρξης της ανάλυσης. Η γραμματική για το YAEP μπορεί να δημιουργηθεί μέσω κλήσεων συναρτήσεων και η ενσωμάτωσή σου σε ένα ευρύτερο κώδικα είναι μια δομημένη και προσιτή διαδικασία.
Οι κατηγορίες των κόμβων στα συγκεκριμένα δέντρα περιέχονται στο enumeration yaep_tree_node_type που αναπαριστά όλους τους πιθανούς κόμβους του αφηρημένου συντακτικού δέντρου που προκύπτει από την ανάλυση της γραμματικής και είναι οι εξής:

- YAEP_NIL: Αναπαράσταση κόμβου με κενή μετάφραση.
- YAEP_ERROR: Αναπαράσταση κόμβου με μετάφραση λάθους.

- YAEP_TERM: Αναπαράσταση ενός τερματικού κόμβου,
- YAEP_ANODE: Αναπαράσταση ενός αφηρημένου κόμβου.
- YAEP_ALT: Αυτός ο τύπος κόμβου μας δείχνει ότι υπάρχουν περισσότερες από μία πιθανές μεταφράσεις του συγκεκριμένου κόμβου.

Προκειμένου να γίνεται αναφορά στον κάθε κόμβο ανάλογα με το είδος του, υπάρχει επίσης ένα union με το όνομα val, το οποίο περιέχει 5 μέλη, ένα για κάθε είδος κόμβου. Τα 5 αυτά μέλη είναι τα εξής: 1) nil για τον κόμβο YAEP_NIL, 2) error για τον κόμβο YAEP_ERROR, 3) term για τον κόμβο YAEP_TERM, 4) anode για τον κόμβο YAEP_ANODE, 5) alt για τον κόμβο YAEP_ALT.
Κάποια από τα παραπάνω είδη κόμβων περιέχουν κάποιες εσωτερικές μεταβλητές οι οποίες εκφράζουν ορισμένα διακριτά χαρακτηριστικά τους. Ο κόμβος τύπου term περιέχει την ακέραια μεταβλητή code, που δείχνει τον κωδικό του συγκεκριμένου τερματικού κόμβου, καθώς και τη μεταβλητή attr τύπου *void, η οποία είναι αναφορά σε ένα γνώρισμα του συγκεκριμένου κόμβου. Ο κόμβος τύπου alt περιέχει τη μεταβλητή node τύπου yaep_tree_node η οποία αναφέρεται στην πρώτη εναλλακτική μετάφραση του και τη μεταβλητή next, η οποία είναι του ίδιου τύπου και αναφέρεται στην επόμενη εναλλακτική μετάφραση. Τέλος, ο κόμβος τύπου anode περιέχει τη μεταβλητή name τύπου const char, που εκφράζει το όνομα του κόμβου, όπως δίνεται στη μετάφραση του κανόνα, τη μεταβλητή children που είναι τύπου yaep_tree_node και αναφέρεται στους κόμβους που είναι μεταφράσεις των συμβόλων που περιέχονται στον κανόνα της γραμματικής μαζί με τον αφηρημένο κόμβο και ουσιαστικά αποτελούν τα παιδιά του κόμβου σύμφωνα με τον κανόνα, καθώς και την ακέραια μεταβλητή cost η οποία εκφράζει ανάλογα με την τιμή μιας μεταβλητής σημαίας είτε το κόστος του συγκεκριμένου κόμβου μαζί με το κόστος όλων των παιδιών του είτε μόνο το κόστος του συγκεκριμένου αφηρημένου κόμβου.
Όλοι οι παραπάνω κόμβοι, οι μεταβλητές και οι διαθέσιμες συναρτήσεις προσαρμόστηκαν κατάλληλα και ενσωματώθηκαν στα προτεινόμενα συστήματα, ώστε να αναλυθούν σωστά οι ασαφείς γραμματικές που χρησιμοποιήθηκαν για τις διάφορες κατηγορίες ψευδοκόμβων.

Algorithm 1 Earley's Parser Algorithm

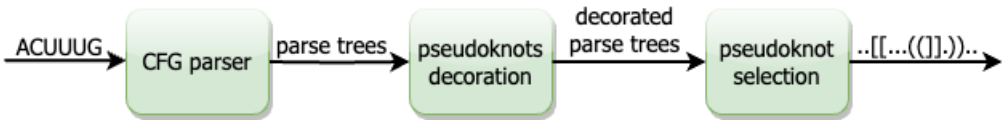
```
1  DECLARE ARRAY_OF_STATES Sets;
2
3  function INITIALIZE(input_string)
4      n ← LENGTH(input_string)
5      Sets ← CREATE_ARRAY(n + 1)
6      for i ← from 0 to n
7          Sets[i] ← EMPTY_SET
8      endfor
9
10 function EARLEY_PARSER(input_string, grammar)
11     INITIALIZE(input_string)
12     n ← LENGTH(input_string)
13     ADD_TO_SET((Start → •S, 0), Sets[0])
14     for i ← from 0 to n
15         for each state in Sets[i]
16           if (state is not completed)
17            if (RIGHT_TO_DOT(state) is a nonterminal)
18              PREDICTOR(state, i, grammar)
19            else
20              SCANNER(state, i, input_string)
21            endif
22           else
23              COMPLETER(state, i)
24           endif
25         endfor
26     endfor
27     return Sets
28
29 function PREDICTOR((B → α • C β, j), i, grammar)
30     for each (C → δ) in GRAMMAR_RULES
31         ADD_STATE_TO_SET((C → • δ, i), Sets[i])
32     endfor
33
34 function SCANNER((B → γ• a δ, j), i, input_string)
35     if (a is input_string[i])
36         ADD_STATE_TO_SET((B → γ a • δ, j), Sets[i+1])
37     endif
38
39 function COMPLETER((A → δ •, x), i)
40     for each (B → γ • A β, j) in Sets[x]
41         ADD_STATE_TO_SET((B → γ A • β, j), Sets[i])
42     endfor
```

Κεφάλαιο 5

Knotify

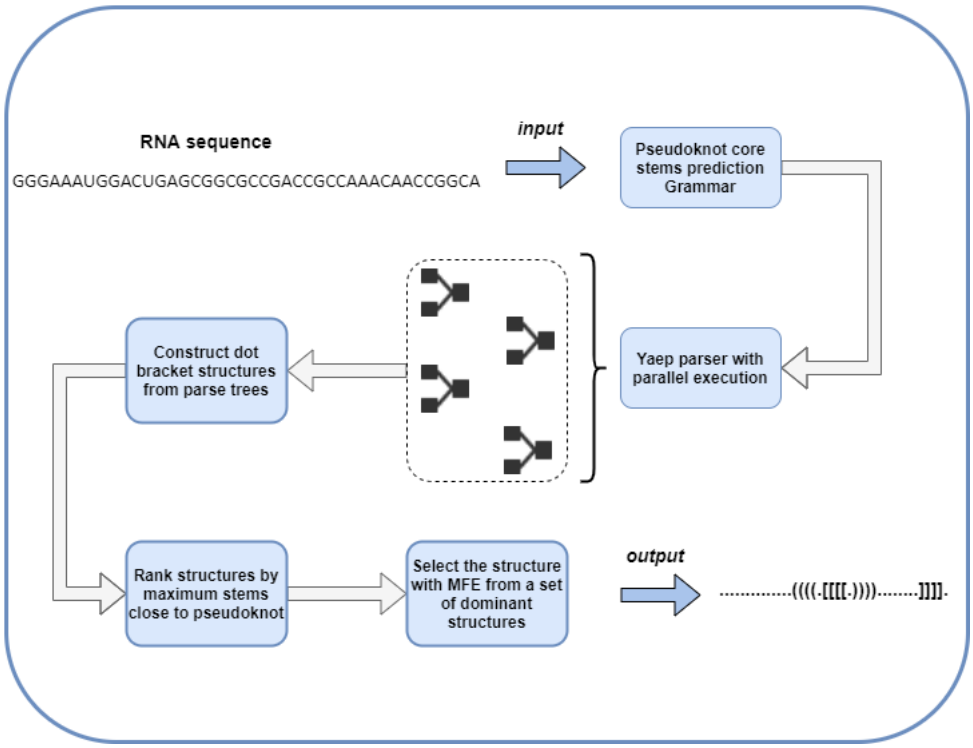
5.1 Προτεινόμενη μεθοδολογία – Ενδεικτικό παράδειγμα εφαρμογής

Στο Κεφάλαιο αυτό παρουσιάζεται μια επισκόπηση της προτεινόμενης μεθοδολογίας Knotify. Η διαδικασία πρόβλεψης των ψευδοκόμβων του RNA διαιρείται στις τρεις ακόλουθες εργασίες: (α) Η ακολουθία του RNA αναλύεται χρησιμοποιώντας έναν CFG αναλυτή, ώστε να παραχθούν όλα τα δέντρα που περιλαμβάνουν έναν ψευδοκόμβο τύπου Η, (β) όλα τα δέντρα, τα οποία έχουν παραχθεί στο προηγούμενο στάδιο, διατρέχονται για την αναγνώριση επιπλέον ζευγών βάσης γύρω από τον ψευδόκομβο και (γ) το βέλτιστο δέντρο επιλέγεται μέσω των κριτηρίων ελάχιστης ενέργειας και του μέγιστου αριθμού ζευγών βάσεων γύρω από τον ψευδοκόμβο. Αυτές οι τρεις βασικές εργασίες της προτεινόμενης μεθοδολογίας (βλ. Σχήμα 5.1) περιγράφονται εκτενώς στις Ενότητες 5.2, 5.3 και 5.4, αντίστοιχα. Όπως παρουσιάζεται



Σχήμα 5.1: Στάδια της προτεινόμενης μεθοδολογίας Knotify.

στο Σχήμα 5.1, η προτεινόμενη υλοποίηση δέχεται ως είσοδο μία ακολουθία RNA στη μορφή μιας συμβολοσειράς που αντιπροσωπεύει μια ακολουθία βάσεων, εκτελεί τα επιμέρους βήματα του αλγορίθμου Knotify και εξάγει την δευτεροταγή δομή σε μορφή κουκίδας-παρένθεσης (dot-bracket). Διακριτά τμήματα λογισμικού έχουν υλοποιηθεί για την πραγματοποίηση κάθε εργασίας/βήματος, και όλες οι λεπτομέρειες υλοποίησης περιγράφονται στην Ενότητα 5.5. Μια πιο εκτεταμένη αναπαράσταση της μεθοδολογίας παρουσιάζεται στο Σχήμα 5.2.



Σχήμα 5.2: Μια πιο εκτεταμένη αναπαράσταση της προτεινόμενης μεθοδολογίας.

5.2 CFG για την Αναγνώριση Ψευδοκόμβων τύπου H

Στην ενότητα αυτή περιγράφεται η γραμματική χωρίς συμφραζόμενα που υλοποιήθηκε για το πρώτο στάδιο του Knotify. Η προτεινόμενη μεθοδολογία ανίχνευσης ψευδοκόμβων σε ακολουθίες βάσεων, οι οποίες αντιπροσωπεύουν RNA, βασίζεται σε τεχνικές αναγνώρισης συντακτικών προτύπων και συγκεκριμένα σε έναν αποδοτικό CFG αναλυτή. Συνεπώς, είναι κρίσιμης σημασίας να επιλεγούν τα αρχικά πρότυπα για την γραμματική. Στην περίπτωση της αναγνώρισης των μοτίβων του RNA, η προφανής επιλογή είναι η αναπαράσταση των βάσεων αδενίνη, κυτοσίνη, γουανίνη και ουρακίλη ως A, C, G, U, αντίστοιχα. Αυτοί οι χαρακτήρες σε μια ακολουθία αποτελούν μια αναπαράσταση RNA. Προτείνεται επομένως, ένας CFG αναλυτής για την αναγνώριση ψευδοκόμβων στο RNA, όπου το προτεινόμενο λεξιλόγιο της γραμματικής περιλαμβάνει μόνο τέσσερα τερματικά σύμβολα T = {A, C, G, U}, με καθένα να αντιπροσωπεύει μια διακριτική βάση: αδενίνη, κυτοσίνη, γουανίνη και ουρακίλη, αντίστοιχα. Συνεπώς, κάθε ακολουθία RNA μπορεί γλωσσικά να αναπαρασταθεί ως συμβολοσειρά που περιέχει τα παραπάνω τερματικά σύμβολα, όπως για παράδειγμα, UAGGC ή AUGGCCGUACG. Η βασική ιδέα για την ανάπτυξη ενός τέτοιου συ-

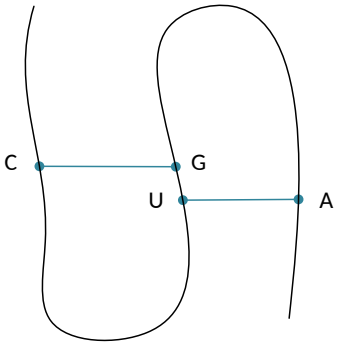
στήματος όπως το Knotify, βασίζεται στο γεγονός ότι η συντακτική αναγνώριση ενός δεδομένου μοτίβου μπορεί να μετατραπεί στη χρήση μιας κατάλληλης γραμματικής προτύπου, ώστε να αναλυθεί η γλωσσική αναπαράσταση των αρχικών μοτίβων. Με βάση αυτόν τον μετασχηματισμό είναι βέβαιο ότι η σχεδίαση της γραμματικής μπορεί να έχει σημαντική επίδραση στο αποτέλεσμα της αναγνώρισης. Συνεπώς, για το σύστημα Knotify, η δημιουργία της CFG που θα χρησιμοποιηθεί είναι το πρώτο σημαντικό βήμα για την αποτελεσματική υλοποίηση του συνολικού συστήματος και η σωστή σχεδίασή της είναι αναγκαία για να περιγράψει τη δομή του ψευδοκόμβου τύπου H μέσα σε μια οποιαδήποτε αυθαίρετη ακολουθία RNA.
Η γραμματική G_{RNA} που υλοποιήθηκε για τον σκοπό αυτό φαίνεται στον Πίνακα 5.1 με τους αντίστοιχους συντακτικούς κανόνες της.

Η δεύτερη στήλη του Πίνακα 5.1 αναδεικνύει όλους τους συντακτικούς κανόνες της γραμματικής για τον εντοπισμό του ψευδοκόμβου τύπου H. Η G_{RNA} περιλαμβάνει τα πέντε μη-τερματικά σύμβολα της συλλογής NT = {S, L, D, K, N}. Το S είναι το σύμβολο εκκίνησης και όλοι οι συντακτικοί κανόνες που έχουν το S στην αριστερή τους πλευρά, π.χ. ο κανόνας 0 έως τον κανόνα 15, στοχεύουν στον εντοπισμό ενός πιθανού ψευδοκόμβου στη συμβολοσειρά εισόδου. Ένας ψευδοκόμβος αποτελείται τουλάχιστον από δύο ζευγάρια βάσεων, στα οποία το μισό ενός ζευγαριού βάσης είναι ενδιάμεσο ανάμεσα στα δύο μισά ενός άλλου ζευγαριού βάσης. Για παράδειγμα, ο κανόνας 6: S → “C” L “U” D “G” L “A” καθορίζει την ύπαρξη ενός ψευδοκόμβου της μορφής C..U..G..A όπου τα ζευγάρια βάσεων C–G και U–A είναι ενδιάμεσα ζεύγη. Αυτά τα ζευγάρια βάσεων θα αναφέρονται ως **κεντρικές βάσεις** ή ζεύγη κεντρικών βάσεων στο υπόλοιπο μέρος του κειμένου και αποτελούν το πρώτο αποτέλεσμα της γραμματικής πάνω στο οποίο θα συνεχίσει τους υπολογισμούς του το Knotify. Το Σχήμα 5.3 απεικονίζει τις κεντρικές βάσεις C–G και U–A αυτού του παραδείγματος. Το μισό του ζευγαριού βάσεων U–A είναι ενδιάμεσο ανάμεσα στο ζευγάρι βάσεων C–G, δηλαδή η βάση U βρίσκεται μεταξύ του ζευγαριού βάσεων C–G. Αντίστοιχα, η βάση G που ανήκει στο ζευγάρι βάσεων C–G είναι, επίσης, ανάμεσα στο ζευγάρι βάσεων U–A. Αυτή η διαδικασία, στην απλοποιημένη της μορφή για λόγους κατανόησης του πρώτου βήματος, είναι ένα παράδειγμα ανίχνευσης της παρεμβολής των κεντρικών βάσεων, που οδηγεί στον εντοπισμό του ψευδοκόμβου τύπου H.

Συνεχίζοντας την περιγραφή της γραμματικής, επισημαίνεται ότι το L είναι το μη-τερματικό σύμβολο που θα δημιουργήσει ακολουθίες βάσεων που σχηματίζουν οι δύο επιμέρους βρόχοι του ψευδοκόμβου, δηλαδή ακολουθίες βάσεων μεταξύ των C και U, καθώς και μεταξύ των G και A. Το μη-τερματικό L μπορεί να δημιουργήσει συμβολοσειρές που ανήκουν στο σύνολο $(T)^* \neq \emptyset$, όπου T είναι το σύνολο των τερματικών συμβόλων, και $\emptyset$ είναι το κενό σύνολο. Το L επομένως, μπορεί να δημιουργήσει συμβολοσειρές με μήκος μεγαλύτερο ή ίσο με το μηδέν, όπως για παράδειγμα οι ακολουθίες A, UA, CCGGAU, κ.ο.κ. Το σύμβολο D είναι το μη-τερματικό σύμβολο που θα δημιουργήσει συμβολοσειρές βάσεων μεταξύ των δύο διασταυρούντων ζευγαριών βάσης (μερών των κεντρικών βάσεων), δηλαδή μεταξύ των βάσεων U και G στο παράδειγμα αυτό. Χρησιμοποιώντας τα μη-τερματικά σύμβολα K και N, το L μπορεί να αναγνωρίσει υποσυμβολοσειρές τερματικών συμβόλων μήκους μηδέν έως δύο

#	**Συντακτικοί Κανόνες**
0	S → "A" L "A" D "U" L "U"
1	S → "U" L "A" D "A" L "U"
2	S → "C" L "A" D "G" L "U"
3	S → "G" L "A" D "C" L "U"
4	S → "A" L "U" D "U" L "A"
5	S → "U" L "U" D "A" L "A"
6	S → "C" L "U" D "G" L "A"
7	S → "G" L "U" D "C" L "A"
8	S → "A" L "C" D "U" L "G"
9	S → "U" L "C" D "A" L "G"
10	S → "C" L "C" D "G" L "G"
11	S → "G" L "C" D "C" L "G"
12	S → "A" L "G" D "U" L "C"
13	S → "U" L "G" D "A" L "C"
14	S → "C" L "G" D "G" L "C"
15	S → "G" L "G" D "C" L "C"
16	L → "A" L
17	L → "U" L
18	L → "C" L
19	L → "G" L
20	L → "A"
21	L → "U"
22	L → "C"
23	L → "G"
24	D → K N
25	K → "A"
26	K → "U"
27	K → "C"
28	K → "G"
29	K → ϵ
30	N → "A"
31	N → "U"
32	N → "C"
33	N → "G"
34	N → ϵ

Πίνακας 5.1: Περιγραφή της G_{RNA} γραμματικής για ψευδοκόμβο τύπου H.

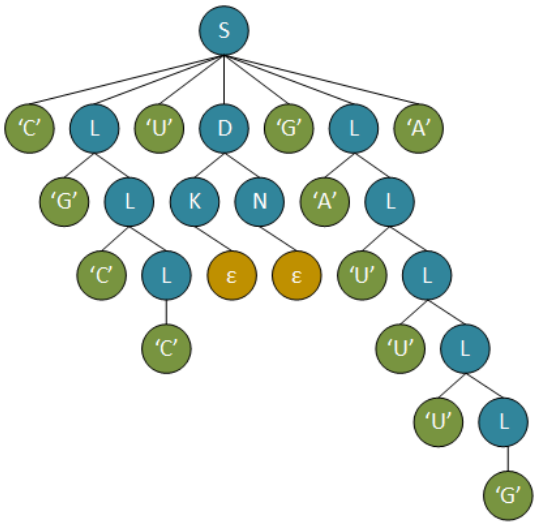


Σχήμα 5.3: Ψευδόκόμβος που εντοπίστηκε από τον κανόνα Σ → “C” L “U” D “G” L “A.”

βάσεων, δηλαδή, ϵ, A, U, C, G, AU, UA, AC, CA, κ.ο.κ., όπου το ϵ είναι η κενή συμβολοσειρά. Το μήκος της υποσυμβολοσειράς μεταξύ των διασταυρούντων ζευγαριών βάσεων είναι παραμετροποιήσιμο και συγκεκριμένα, το μέγιστο μήκος αυτής της υποσυμβολοσειράς μπορεί να καθοριστεί από τον χρήστη, όπως περιγράφεται και στο αποθετήριο Knotify στο GitHub [77]. Η βασική λογική είναι ότι με βάση την είσοδο από τον χρήστη μπορούν να σχηματιστούν αυτόματα παραπάνω από μία γραμματικές, που είναι τόσες όσες ορίζει η παράμετρος και οι αναλυτές τους εκτελούνται παράλληλα για τον εντοπισμό των ψευδοκόμβων με το αντίστοιχο μήκος της παραπάνω υποσυμβολοσειράς. Ειδικότερα, η μεταβλητή που καθορίζει το μήκος αυτό ονομάζεται dd και ανάλογα με την τιμή της θα δημιουργηθούν δυναμικά οι αντίστοιχες γραμματικές. Για παράδειγμα για $dd = 3$ θα δημιουργηθούν δυναμικά, μέσω μίας ειδικής συνάρτησης, τέσσερις γραμματικές για $dd = 0, 1, 2, 3$ αντίστοιχα, οι οποίες στη συνέχεια θα εκτελούνται παράλληλα στους διαθέσιμους επεξεργαστές της υπολογιστικής υποδομής για βελτιστοποίηση της απόδοσης του Knotify.

Όπως γίνεται αντιληπτό από τη λειτουργία ενός αναλυτή, η γραμματική G_{RNA} μπορεί να εντοπίσει ψευδοκόμβους σε συμβολοσειρές, όπου τα πρώτα και τα τελευταία σύμβολα της ακολουθίας ανήκουν στις κεντρικές βάσεις, δηλαδή κάθε φορά ο εντοπισμός αφορά τη συνολική ακολουθία εισόδου στον αναλυτή. Στο εξεταζόμενο παράδειγμα ενός ψευδοκόμβου που ανιχνεύεται από τον έκτο κανόνα, ο ψευδόκόμβος υπάρχει σε μια υποσυμβολοσειρά που ξεκινά από το τερματικό σύμβολο C και τελειώνει στο τερματικό σύμβολο A. Ωστόσο, αυτό δεν πρέπει να θεωρείται περιορισμός, καθώς στο Knotify ο αναλυτής εκτελείται σε όλες τις πιθανές υποακολουθίες των συμβολοσειρών χρησιμοποιώντας την τεχνική του ¨κινούμενου παραθύρου’ (sliding window), η οποία παρουσιάζεται στη συνέχεια.

Το δέντρο ανάλυσης που παράγεται από την ανάλυση της υποσυμβολοσειράς "C G C C U G A U U U G A" παρουσιάζεται στο Σχήμα 5.4. Επιστρέφοντας στο προηγούμενο παράδειγμα, ο συντακτικός κανόνας 6 χρησιμοποιήθηκε για να ανιχνευθεί ο ψευδοκόμβος της μορφής **C** ... **U G** ..... **A** (βλ. Σχήμα 5.3). Στη συνέχεια, οι κανόνες 19, 18 και 22 χρησιμοποιήθηκαν για την αναγνώριση των βάσεων μεταξύ των



Σχήμα 5.4: Συντακτικό Δέντρο για την αναγνώριση ψευδοκόμβου – υποακολουθία "C G C C U G A U U U G A."

C και U, δηλαδή C **G C C** U G ..... A. Ακολούθως, οι κανόνες 24, 29 και 34 χρησιμοποιήθηκαν για την αναγνώριση της κενής συμβολοσειράς μεταξύ των βάσεων U και G των κεντρικών βάσεων. Τέλος, οι κανόνες 16, 17, 17, 17 και 23 χρησιμοποιήθηκαν για την αναγνώριση των βάσεων μεταξύ των C και U, δηλαδή C G C C U G **A U U U G** A. Η ολοκλήρωση αυτής της υποσυμβολοσειράς στην αρχική ακολουθία RNA και η διαδικασία διακόσμησης (decoration) του ψευδοκόμβου με επιπλέον βάσεις αναλύεται στην Ενότητα 5.3.

Η προτεινόμενη μεθοδολογία αναλύει όλες τις υποσυμβολοσειρές, αρχίζοντας από αυτή που ξεκινά με το πρώτο σύμβολο της ακολουθίας και έχει το ελάχιστο δυνατό μήκος. Επαναληπτικά, το μήκος επεκτείνεται κατά ένα σύμβολο για να περιλάβει την πλήρη αρχική ακολουθία RNA. Με τον ίδιο επαναληπτικό τρόπο, τα σημεία έναρξης των συμβολοσειρών αυξάνονται για να αποκλείσουν το προηγούμενο σύμβολο έναρξης. Η ανάλυση ολοκληρώνεται όταν το μήκος της υποσυμβολοσειράς που πρέπει να αναλυθεί, υποβαθμίζεται περαιτέρω από ένα προκαθορισμένο όριο (δηλαδή, το ελάχιστο μήκος του ψευδοκόμβου). Αυτή η μεθοδολογία, εφόσον η γραμματική G_{RNA} είναι ασαφής, οδηγεί στη δημιουργία ενός σημαντικού αριθμού δέντρων ανάλυσης. Η μεθοδολογία επιλογής του βέλτιστου δέντρου ανάλυσης αναλύεται στην Ενότητα 5.4. Ο επιλεγμένος αναλυτής CFG είναι αυτός του YAEP [76], που είναι ένας αποδοτικός αναλυτής CFG βασισμένος στον αλγόριθμο του Earley [70], και σύμφωνα με τη βιβλιογραφία, μπορεί να χειριστεί ασαφείς γραμματικές.

Η επιλογή μιας συντακτικής γραμματικής χωρίς συμφραζόμενα βασίστηκε στην προοπτική επέκτασής της με γνωρίσματα (δημιουργίας μιας γραμματικής γνωρισμάτων) για την αποθήκευση πιθανοτήτων και τη δυνατότητα περικοπής δέντρων ανάλυσης κατά τη διάρκεια της διαδικασίας κατασκευής των δέντρων ανάλυσης, σε μελλοντικές

προσεγγίσεις και τροποποιήσεις του συστήματος. Για να ενισχυθεί η απόδοση του προτεινόμενου συστήματος, προτάθηκε και μια εναλλακτική υλοποίηση της πρώτης αυτής εργασίας εντοπισμού του ψευδοκόμβου τύπου H, χρησιμοποιώντας έναν άπληστο αλγόριθμο "brute-force". Αυτή η προσέγγιση διατρέχει τη συμβολοσειρά για να εντοπίσει όλα τα δυνατά ζευγάρια βάσης και στη συνέχεια διατρέχει όλα τα ζευγάρια βάσης για να εντοπίσει τα ζευγάρια βάσης που δυνητικά αποτελούν τις κεντρικές βάσεις ενός ψευδοκόμβου, δηλαδή αποτελούμενες από δύο τμήματα που συνδέονται με δύο μονές συμβολοσειρές ή βρόχους. Η άπληστη προσέγγιση ("brute-force") παρουσιάζει μία ταχύτερη εκτέλεση κατά 2,55 φορές σε σύγκριση με την υλοποίηση που χρησιμοποιεί τη γραμματική στην πρώτη εργασία του Knotify. Ωστόσο, η επιτάχυνση έρχεται με το κόστος της συνολικής επεκτασιμότητας που παρέχει η προσέγγιση που είναι βασισμένη στη γραμματική. Συγκρινόμενες βέβαια, οι δύο προτεινόμενες υλοποιήσεις με δύο από τις πιο γνωστές και αποδοτικές πλατφόρμες [32, 35] σε επίπεδο χρόνου εκτέλεσης, παρουσιάζουν μία αξιοσημείωτη επιτάχυνση. Η αξιολόγηση της απόδοσης αναφορικά με τον χρόνο εκτέλεσης αναλύεται στην Ενότητα 5.6.5.

5.3 Διαδικασία 'διακόσμησης' (decoration) των Κεντρικών Βάσεων

Αφού λοιπόν δημιουργηθούν τα δέντρα ανάλυσης σύμφωνα με τη μεθοδολογία που περιγράφεται στην Ενότητα 5.2, όλα τα δέντρα αναζητούνται προκειμένου να διακοσμηθεί ο ψευδόκόμβος με επιπλέον βάσεις. Η G_{RNA} σχεδιάστηκε με τέτοιο τρόπο ώστε να αναγνωρίζει μόνο τις κεντρικές βάσεις του ψευδοκόμβου προκειμένου να αξιοποιήσει την αποδοτικότητα του αναλυτή CFG. Αυτό το γεγονός οδηγεί σε μια γραμματική με λίγους συντακτικούς κανόνες και επιτρέπει στον αναλυτή να εκτελείται αποτελεσματικά. Ωστόσο, δημιουργείται η ανάγκη επεξεργασίας όλων των δέντρων, προκειμένου να ανιχνευθούν οι υπόλοιπες βάσεις που περιβάλλουν και πλαισιώνουν τις κεντρικές βάσεις του ψευδοκόμβου. Αυτές οι βάσεις θα καθορίσουν και όλα τα υπόλοιπα μοτίβα μαζί με αυτό του ψευδοκόμβου και κατ' επέκταση τη συνολική δευτεροταγή δομή του. Όλες οι βάσεις σε κάθε από τα δύο τμήματα βρόχους του ψευδοκόμβου εξετάζονται ακολουθιακά για να διαπιστωθεί αν μπορούν να σχηματίσουν ζευγάρι βάσεων με μια άλλη βάση που βρίσκεται σε κατάλληλη θέση.
Ο Πίνακας 5.2 παρουσιάζει το τμήμα του αλγορίθμου που αποκαλείται **decoration**. Έχοντας ανιχνευθεί οι κεντρικές βάσεις U–A και C–G στις θέσεις 9 και 11 και 5 και 10, αντίστοιχα, εξετάσθηκαν όλες οι βάσεις και στους δύο βρόχους του ψευδοκόμβου, δηλαδή, οι βάσεις στις θέσεις 6 έως 9 (αριστερός βρόχος) και 11 έως 15 (δεξιός βρόχος). Τα τμήματα αυτά εξετάσθηκαν προκειμένου να ανιχνευθεί η δυνατότητά τους να σχηματίσουν ζευγάρια βάσεων με βάσεις εκτός των βρόχων του ψευδοκόμβου. Έτσι, τα ζευγάρια βάσεων που ανήκουν στον αριστερό βρόχο ελέγχθηκαν για να διαπιστωθεί αν μπορούν να σχηματίσουν ζευγάρια βάσεων με βάσεις στις θέσεις 17 έως 19, ενώ οι βάσεις που ανήκουν στον δεξιό βρόχο εξετάσθηκαν για να ταιριάξουν με αυτές στις θέσεις 1 έως 4. Στα τμήματα αυτά του ψευδοκόμβου, ανιχνεύθηκαν

ακολουθιακά τα ζευγάρια βάσεων στις θέσεις 8–17, 7–18 και 4–11, καθώς και 3–12, αντίστοιχα. Ο Πίνακας 5.2 παρουσιάζει λεπτομερώς ολόκληρη τη διαδικασία.

String enumeration	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19
String	C	C	A	U	C	G	C	C	U	G	A	U	U	U	G	A	G	G	A
Parser output:	.	.	.	.	[	.	.	.	(	]	.	.	.	.	.	)	.	.	.
Step 1	.	.	.	.	[	.	.	(	(	]	.	.	.	.	.	)	)	.	.
Step 2	.	.	.	.	[	.	(	(	(	]	.	.	.	.	.	)	)	)	.
Step 3	.	.	.	[	[	.	(	(	(	]	]	.	.	.	.	)	)	)	.
step 4	.	.	[	[	[	.	(	(	(	]	]	]	.	.	.	)	)	)	.

Πίνακας 5.2: Η διαδικασία Decoration των κεντρικών βάσεων του ψευδοκόμβου.

Το προτεινόμενο σύστημα επιτρέπει στον χρήστη να επιλέξει προαιρετικά την ύπαρξη των βάσεων U–G στους δύο βρόχους του ψευδοκόμβου. Το ζεύγος βάσεων G-U (wobble pair) είναι μια θεμελιώδης μονάδα δευτεροταγούς δομής του RNA που υπάρχει σχεδόν σε κάθε κατηγορία RNA από οργανισμούς και των τριών φυλογενετικών περιοχών. Έχει συγκρίσιμη θερμοδυναμική σταθερότητα με τα ζεύγη βάσεων Watson-Crick και είναι σχεδόν ισομορφικό με αυτά. Ως εκ τούτου, συχνά υποκαθιστά τα ζεύγη βάσεων G-C ή A-U. Το ζεύγος βάσεων G-U έχει επίσης μοναδικές χημικές, δομικές, δυναμικές και συνδετικές ιδιότητες, οι οποίες μπορούν να μιμηθούν μόνο εν μέρει από τα ζεύγη βάσεων Watson-Crick. Επίσης, το σύστημα επιτρέπει σε μία διαφορετική του εκδοχη, την ύπαρξη ασύμμετρων βρόχων/εξογκωμάτων ή εσωτερικών βρόχων στους βρόχους του ψευδοκόμβου και περιγράφεται αναλυτικά στην Ενότητα 7.

5.4 Επιλογή Βέλτιστου Δέντρου

Σύμφωνα με τη βιβλιογραφία, έχουν παρουσιαστεί πολλές μεθοδολογίες που αντιμετωπίζουν το πρόβλημα της πρόβλεψης της δευτεροταγούς δομής RNA, με τις πιο διαδεδομένες να είναι η μέθοδος του ελάχιστης ελεύθερης ενέργειας (MFE) [78], η οποία προβλέπει την ακολουθία RNA που παρουσιάζει το χαμηλότερο επίπεδο ελεύθερης ενέργειας. Θεωρητικά, αυτή θα ήταν η πιο σταθερή δομή στη φύση, αλλά σε πραγματικές συνθήκες, όπως προκύπτει και από πειραματική αξιολόγηση, οι ακολουθίες που έχουν εντοπιστεί και καταγραφεί δεν έχουν πάντα τη δομή της ελάχιστης ενέργειας. Η αρχή της ελάχιστης ελεύθερης ενέργειας είναι βασικά μια επαναδιατύπωση του δευτέρου νόμου της θερμοδυναμικής προσαρμοσμένη στην περίπτωσή μας στην ελεύθερη ενέργεια της δομής του RNA. Μία επίσης σημαντική μέθοδος με εφαρμογή σε πολλά συστήματα της βιβλιογραφίας είναι η μέθοδος της μεγιστοποίησης των ζευγών βάσεων [8]. Πρόκειται για μια τεχνική η οποία βασίζεται στον αριθμό των ζευγών βάσεων που δημιουργούνται στο RNA και στοχεύει στην ακολουθία που στοχεύει στην έρευση αυτής με τα περισσότερα ζεύγη. Η αρχική θεωρία αναφέρεται σε όλη την ακολουθία, ωστόσο στην παρούσα διατριβή παρουσιάζεται μία διαφοροποιη-

μένη προσέγγιση που επικεντρώνεται στη μεγιστοποίηση των ζευγών βάσεων κοντά στον ψευδοκόμβο, δηλαδή γύρω από τις κεντρικές βάσεις. Η προσέγγιση αυτή με τον μέγιστο αριθμό ζευγών βάσεων γύρω από τις κεντρικές βάσεις του ψευδοκόμβου οδηγεί σε δομές με ελάχιστη ενέργεια ή σε δομές με ενέργεια πολύ κοντά σε αυτή που είναι επίσης πολύ πιθανό να εμφανιστούν σε πραγματικές συνθήκες στη φύση. Άλλες γνωστές μέθοδοι είναι η μέθοδος partition function [22], η οποία βασίζεται στο γεγονός ότι τα πραγματικά ζεύγη βάσεων έχουν μεγάλη πιθανότητα να βρίσκονται στην εκτιμώμενη κατανομή της ελάχιστης ελεύθερης ενέργειας. Η μέθοδος ενισχύει τη θετική προβλεπτική τιμή των πραγματικών ζευγών βάσεων μελετώντας τις παραμέτρους των πλησιέστερων γειτόνων τους για την ελεύθερη ενέργεια σε μια δεδομένη θερμοκρασία. Τέλος, η μέθοδος σύγκρισης αλληλουχιών [60] αφορά τον έλεγχο του μοτίβου των αντικαταστάσεων που παρατηρούνται σε μια διπλή ευθυγράμμιση δύο ομόλογων αλληλουχιών, ενώ εφαρμόζεται για την επίλυση του προβλήματος και η μέθοδος φυσικών πειραμάτων [61], η οποία επικεντρώνεται σε πραγματικά πειράματα με την εισαγωγή των μορίων σε κατάλληλα διαμορφωμένο υγρό περιβάλλον.
Το προτεινόμενο σύστημα χρησιμοποιεί ένα υβριδικό μοντέλο επιλογής του βέλτιστου δέντρου, συνδυάζοντας αρχές που προέρχονται από τις δύο πιο διαδεδομένες τεχνικές, δηλαδή τη μέθοδο της μεγιστοποίησης των ζευγών βάσεων και τη μέθοδο της ελάχιστης ελεύθερης ενέργειας MFE, για να προβλέψει το μοτίβο του ψευδοκόμβου της δευτεροταγούς δομής του RNA με ακρίβεια και αποδοτικότητα. Το MFE είναι οικονομικό από πλευράς απόδοσης όταν εφαρμόζεται σε μία συμβολοσειρά με γνωστή δομή κουκίδας-παρένθεσης, σε αντίθεση με τον υπολογισμό όλου του πίνακα της στρατηγικής δυναμικού προγραμματισμού. Αρχικά, όλα τα δέντρα ταξινομούνται ανάλογα με τον αριθμό των ζευγών βάσεων γύρω από τον εντοπισμένο ψευδόκομβο, στοιχείο που θεωρούμε ότι συμβάλει στη σταθερότητα της δομής, και το MFE εφαρμόζεται μόνο στα δέντρα που κατατάσσονται στην κορυφαία θέση της κατάταξης του αριθμού των ζευγών βάσεων. Αναζητείται με άλλα λόγια, η πιο σταθερή δομή, από την άποψη των ζευγών γύρω από τον ψευδοκόμβο και την ελάχιστη ελεύθερη ενέργεια.

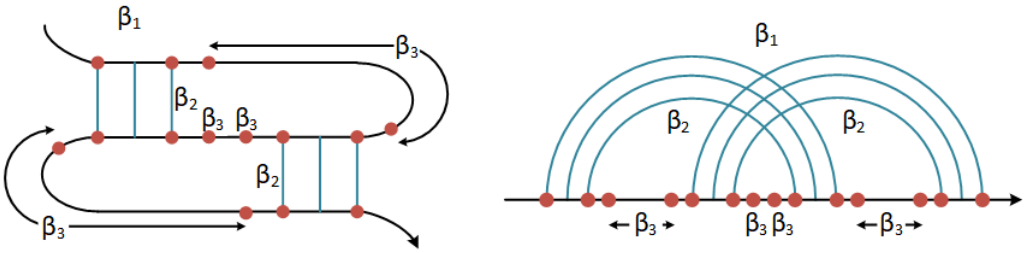
5.4.1 Υπολογισμός Ελάχιστης Ελεύθερης Ενέργειας

Προκειμένου να επιλεγεί η καλύτερη δομή από το σύνολο των υποψήφιων δευτεροταγών δομών, η μέθοδός μας επιλέγει τελικά, αυτή με την ελάχιστη ελεύθερη ενέργεια. Για την εκτέλεση αυτού του σημαντικού βήματος, ενσωματώθηκε ο αλγοριθμος υπολογισμού της ελεύθερης ενέργειας από το σύστημα HotKnots [81], προκειμένου να υπολογίσει την ενέργεια κάθε δομής για την τελική επιλογή της δομής με την ελάχιστη τιμή. Η υλοποίηση αυτή βασίζεται σε έναν αλγόριθμο που παρουσιάστηκε από τον Mathews [82] και επεκτάθηκε για τα ψευδοκόμβους από τον Dirks [62], που είναι και η κατάλληλη για την εργασία αυτή στο πλαίσιο του Knotify. Συγκεκριμένα, η ενέργεια του ψευδοκόμβου δίνεται από την ακόλουθη σχέση:

$$G^{pseudo} = \beta_1 + \beta_2 * B^p + \beta_3 * U^p \tag{5.1}$$

όπου το β_1 είναι το βάρος για την ύπαρξη του ψευδοκόμβου· το B^p είναι ο αριθμός των κεντρικών βάσεων και το U^p είναι ο αριθμός των ανεξάρτητων βάσεων μέσα

στον ψευδοκόμβο. Οι παράμετροι β_2 και β_3 τέθηκαν στο 0.1, όπως υπολογίστηκε πειραματικά στην εργασία του [81], και αναφέρονται στις κεντρικές βάσεις και τις μη ζευγαρωμένες βάσεις, αντίστοιχα, ενώ το βάρος β_1 τέθηκε στο 9.6 όπως προτείνεται και στην αρχική υλοποίηση. Η Εικόνα 5.5 παρέχει έναν παράδειγμα για τα βάρη β_1, β_2 και β_3 σε έναν ενδεικτικό ψευδόκομβο τύπου H.

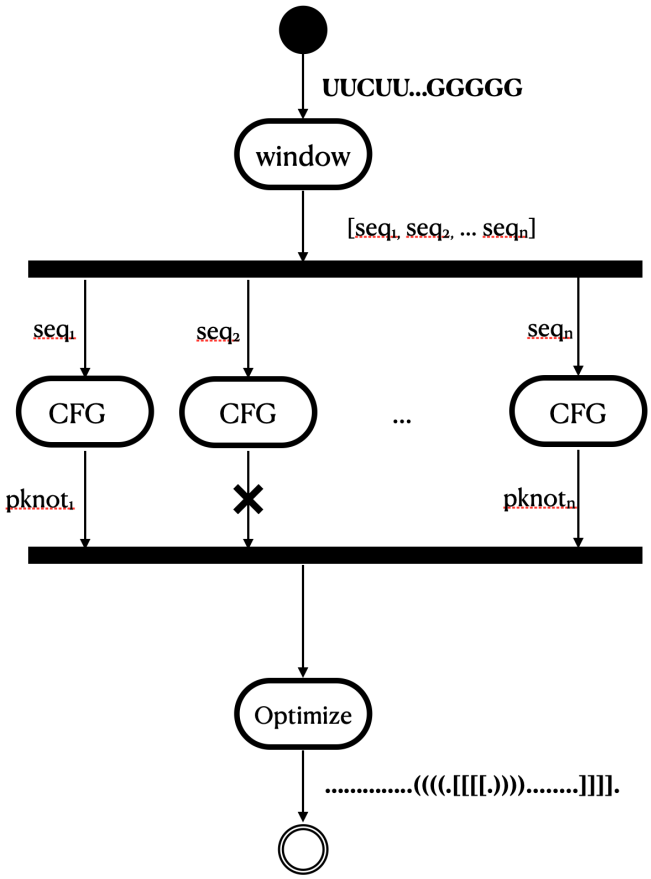


Σχήμα 5.5: Το βάρος για την ύπαρξη του ψευδοκόμβου είναι β_1, ενώ η συνεισφορά των κεντρικών βάσεων είναι β_2 και των μη ζευγαρωμένων βάσεων μέσα στον ψευδοκόμβο είναι β_3. Οι ενέργειες των στοιβαγμένων ζευγών βάσεων υπολογίζεται βάσει του μοντέλου του [82] (από [81]).

5.5 Μεθοδολογία Υλοποίησης του συστήματος και Εργαλεία

Σύμφωνα με τη βιβλιογραφία, η πρόβλεψη ψευδοκόμβων σε οποιαδήποτε αυθαίρετη ακολουθία RNA είναι ένα πρόβλημα NP-hard. Αφενός, οι αλγόριθμοι ελάχιστης ελεύθερης ενέργειας που προτείνονται για την πρόβλεψη ψευδοκόμβων χρησιμοποιούν δυναμικό προγραμματισμό για να κατατάξουν ψηλά όσον αφορά τον υπολογιστικό τους κόστος, ενώ η ακρίβειά τους μειώνεται αναλογικά με το μήκος της εισόδου. Επιπρόσθετα, οι υπάρχουσες ευρετικές προσεγγίσεις δεν έχουν ικανότητες γενίκευσης κατά τον έλεγχο με διάφορα σύνολα δεδομένων. Σε αυτό το πλαίσιο, η παρούσα εργασία παρουσιάζει μια νέα υβριδική στρατηγική για την επιλογή της ακολουθίας RNA που είναι η πιο πιθανή έκφραση ψευδοκόμβου. Σύμφωνα με το Σχήμα 5.2, η προτεινόμενη μεθοδολογία δημιουργεί αρχικά έναν υποχώρο όλων των πιθανών εκφράσεων ψευδοκόμβου, δηλαδή δέντρα που παράγονται με επιτυχία και πληρούν ορισμένα κριτήρια ελάχιστου μήκους, και στη συνέχεια λύνει ένα πρόβλημα βέλτιστης επιλογής, επιλέγοντας την αναπαράσταση ψευδοκόμβου που διαθέτει (i) το μέγιστο αριθμό ζευγών βάσεων γύρω από τον ψευδόκομβο και (ii) την ελάχιστη ελεύθερη ενέργεια. Η προτεινόμενη υλοποίηση είναι καθαυτή υβριδική. Χρησιμοποιώντας ρουτίνες κώδικα Python και C, υλοποιήθηκε ένα λογισμικό πακέτο αποτελεσματικό, ευέλικτο, εύκολο στη χρήση και επεκτάσιμο. Η γλώσσα Python χρησιμοποιήθηκε για να παρέχει υψηλό επίπεδο ευελιξίας και έτοιμες λειτουργίες όπως η δυνατότητα παραλληλοποίησης, η επίβλεψη δευτερευόντων διεργασιών και η διαχείριση αρχείων, ενώ η γλώσσα

C χρησιμοποιήθηκε για την διαχείριση του έργου της ανάλυσης και πολύπλολων εργασιών γύρω από τη γραμματική, όπως η διάσχιση γράφων. Συγκεκριμένα, η είσοδος χωρίστηκε σε πολλές υπο-ακολουθίδες (αυτή η διαδικασία περιγράφεται επίσης στην Ενότητα 5.2), οι οποίες εφόσον είναι ανεξάρτητες μεταξύ τους, παραλληλοποιήθηκαν για να επιταχύνουν τον χρόνο εκτέλεσης του συστήματος. Δημιουργείται ένα σύνολο από εργασίες που διαμοιράζονται σε αντίστοιχες παράλληλες διεργασίες του συστήματος στις οποιες εκτελείται ο CFG parser που αξιολογεί όλες τις παραγόμενες υπο-ακολουθίδες. Το μέγεθος των παράλληλων εργασιών είναι ανάλογο των πυρήνων της CPU για να αξιοποιηθεί στο μέγιστο, ενώ κάθε εργασία είναι μια περίπτωση YAEP-parser [76] υλοποιημένη σε C για να εξασφαλιστεί βέλτιστη διαχείριση πόρων και εξαιρετικά γρήγορη ανάλυση (Εικόνα 5.6).



Σχήμα 5.6: Παράλληλη ανάλυση της συμβολοσειράς με πολλαπλούς YAEP-parsers.

Κάθε παράλληλη διεργασία στην οποία εκτελείται ο CFG parser παράγει μια δομή ψευδοκόμβου που περιγράφει κάθε πιθανή δομή ψευδοκόμβου τύπου H. Εάν ορισμένες

υπο-ακολουθίδες δεν αντιπροσωπεύουν έναν ψευδόκομβο, ο CFG parser θα αποτύχει. Στη συνέχεια, όλοι ψευδοκόμβοι αποθηκεύονται σε κατάλληλες δομές για να αναλυθούν μέσω του πακέτου Pandas [68]. Δεδομένου ότι μέσα στις δομές αυτές περιλαμβάνονται όλες οι πιθανές λύσεις του προβλήματός μας (δηλαδή όλοι οι έγκυροι ψευδοκόμβοι), πρέπει να επιλέξουμε την βέλτιστη λύση με βάση τα κριτήρια επιλογής που επιλέχθηκαν, υποθέτοντας ότι η πιο κατάλληλη πρόβλεψη μπορεί να είναι αυτή που έχει την ελάχιστη ελεύθερη ενέργεια. Ωστόσο, ο υπολογισμός της ελάχιστης ελεύθερης ενέργειας για κάθε πιθανή δομή RNA είναι μια υπολογιστικά δύσκολη διαδικασία και απαιτεί παράλληλα μεγάλη υπολογιστική ισχύ και μνήμη. Η μεθοδολογία η οποία παρουσιάζεται, αντιμετωπίζει αυτήν την υπολογιστικά επίπονη εργασία προσαρμόζοντας την παρατήρηση ότι η ελαχιστοποίηση της ελεύθερης ενέργειας συνδέεται άμεσα με τον αυξημένο αριθμό των ζευγών βάσεων σε οποιοδήποτε δυνητικό ψευδόκομβο RNA. Συνεπώς, αντί να υπολογίσουμε την ελάχιστη ελεύθερη ενέργεια για όλες τις δομές ψευδοκόμβου, πραγματοποιήθηκε αναζήτηση για τον μέγιστο αριθμό ζευγών βάσεων και στη συνέχεια, ο υπολογισμός της ενέργειας για μικρό αριθμών δομών, διαδικασία που απαιτεί χρόνο $O(n)$ και χώρο $O(n)$, όπου n είναι το μήκος της εισόδου. Όπως αναφέρθηκε προηγουμένως, η πρώτη εργασία, δηλαδή η πρόβλεψη των κεντρικών βάσεων των ψευδοκόμβων, επιτελέστηκε με δύο διαφορετικές προσεγγίσεις: μία βασισμένη στον YAEP parser (knotify_yaep) και μία δεύτερη βασισμένη σε έναν εξαντλητικό αλγόριθμο (knotify_bruteforce). Η πρώτη υλοποίηση απαιτεί χρόνο $O(n^5)$: η πολυπλοκότητα του Earley parser [70] για τις ασαφείς γραμματικές συν ο χρόνο διάσχισης όλων των γραφημάτων (DAG), τα οποία παράγονται από τον YAEP parser. Από την άλλη, η δεύτερη υλοποίηση απαιτεί χρόνο $O(n^2) + O(n^4)$ $\approx O(n^4) : O(n^2)$ για την προσπέλαση της εισόδου προκειμένου να εντοπιστούν όλα τα πιθανά ζευγάρια βάσεων (το μέγιστο αριθμός ζευγών βάσεων είναι n^2) και στη συνέχεια $O(n^4)$ για τον εντοπισμό όλων των ζευγών βάσεων που μπορούν να αποτελέσουν τις κεντρικές βάσεις ενός ψευδοκόμβου. Ο πηγαίος κώδικας της υλοποίησης είναι διαθέσιμος στο αποθετήριο GitHub *knotify* [77].

5.6 Αξιολόγηση Επίδοσης του Knotify

5.6.1 Παρουσίαση Συνόλου Δεδομένων

Ένα σύνολο δεδομένων [83] από 262 ακολουθίες RNA χρησιμοποιήθηκε για να αξιολογηθεί η ακρίβεια της μεθοδολογίας μας έναντι άλλων μεθόδων. Αποτελείται από γνωστές ακολουθίες RNA προερχόμενων από ένα πλήθος βάσεων δεδομένων και ως εκ τούτου, θα πρέπει να θεωρηθεί ιδανικό για να συγκριθεί η προτεινόμενη μεθοδολογία με τα άλλα υψηλής επίδοσης συστήματα της βιβλιογραφίας, και συγκεκριμένα τα Hotknots, Iterative HFold (IHFold), IPknot και Knotty [81, 53, 35, 32]. Το σύνολο δεδομένων των 262 ακολουθιών RNA χωρίστηκε σε τέσσερις ομάδες ανάλογα με το μήκος τους. Συνεπώς, υπήρχε μια ομάδα 75 ακολουθιών RNA με μήκος μικρότερο από 30, μια ομάδα 68 ακολουθιών RNA με μήκος μεγαλύτερο από 30 και μικρότερο από 40, μια ομάδα 55 ακολουθιών με μήκος μεγαλύτερο από 40 και μικρότερο από

50, και μια ομάδα 64 ακολουθιών με μήκος μεγαλύτερο από 50. Οι προαναφερόμενες ομάδες σημειώνονται ως L<30 (#75), 30≤L<40 (#68), 40≤L<50 (#55), και L≥50 (#64), αντίστοιχα, στους πίνακες και τα διαγράμματα του Κεφαλαίου.

5.6.2 Μέθοδοι Αξιολόγησης

Για να αξιολογήσουμε τη μεθοδολογία μας, χρησιμοποιούμε τρεις μετρικές: i) την ακρίβεια της πρόβλεψης των κεντρικών βάσεων των ψευδοκόμβων (core stems prediction), ii) την ικανότητα πρόβλεψης των ζευγών βάσεων που υπάρχουν στην πραγματική δομή σε μορφή dot-bracket (πίνακας σύγχυσης – confusion matrix), και iii) τον χρόνο εκτέλεσης.

5.6.3 Πρόβλεψη της Θέσης των Ψευδοκόμβων

Ο Πίνακας 5.3 παρέχει μια συγκριτική ανάλυση μεταξύ του προτεινόμενου συστήματος Knotify και των προαναφερόμενων πλατφορμών, περιλαμβάνοντας συνοπτικά τη δυνατότητα πρόβλεψης των κεντρικών βάσεων των ψευδοκόμβων. Η σύγκριση πραγματοποιήθηκε τόσο για τις μεθόδους που υλοποιήθηκαν στο πλαίσιο της διατριβής αυτής, δηλαδή knotify_yaep και knotify_bruteforce, όσο και για τις υπόλοιπες μεθόδους, δηλαδή Knotty, HotKnots, IPknot και IHFold. Η μεθοδολογία Knotify, κατάφερε να ανιχνεύσει με ακρίβεια τις κεντρικές βάσεις των ψευδοκόμβων σε 143 από τις 262 ακολουθίες, ενώ το Knotty σε 121 ακολουθίες, το HotKnots σε 75, το IPknot σε 38 ακολουθίες και το IHFold σε 0 ακολουθίες. Για τον υπολογισμό της θέσης των κεντρικών βάσεων, επιτρέψαμε την πρόβλεψη μιας βάσης από κάθε ζευγάρι δεξιά ή αριστερά, δηλαδή το ζευγάρι *(i,j)* ισοδυναμεί με το *(i-1,j)*, *(i+1,j)*, *(i,j-1)* και *(i,j+1)*, όπως προτείνεται στο [82]. Αντίστοιχα, παρουσιάζονται τα ποσοστά επί τοις εκατό της πρόβλεψης της θέσης του ψευδοκόμβου στο σύνολο των ακολουθιών αξιολογησης. Μικρές αποκλίσεις σε όλες τις μετρικές μεταξύ των δύο προτεινόμενων μεθοδολογιών οφείλονται σε περιπτώσεις ισοβαθμίας ακολουθιών στην ταξινόμηση που γίνεται από το σύστημα.

Platform	**Exact Matches**	**Exact Matches (%)**
IHFold	0	0
HotKnots	75	28.6
IPknot	38	14.5
Knotty	121	46.1
knotify_yaep	143	54.5
knotify_bruteforce	144	54.9

Πίνακας 5.3: Πρόβλεψη της θέσης του ψευδοκόμβου σε ολόκληρο το σύνολο δεδομένων.

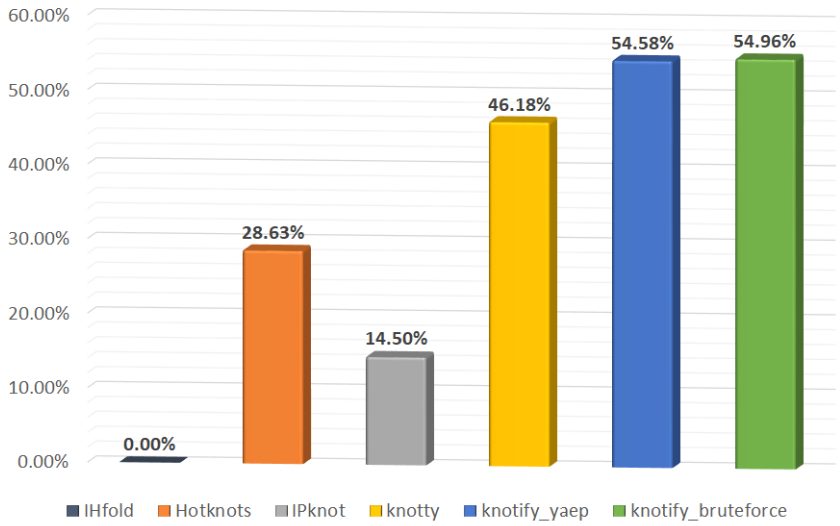
Οι μετρήσεις που διαιρούνται ανά μήκος ακολουθιών RNA παρουσιάζονται στον Πίνακα 5.4, όπου φαίνεται ότι η μεθοδολογία Knotify κατάφερε να προβλέψει ακριβώς

τις κεντρικές βάσεις σε περισσότερους ψευδοκόμβους σε σύγκριση με τις άλλες υλοποιήσεις σε τρεις από τις τέσσερις κατηγορίες, ενώ στις κατηγορίες όπου το μήκος είναι από 30 έως 40, η προτεινόμενη μεθοδολογία προέβλεψε μολις ένα λιγότερο από το Knotty.

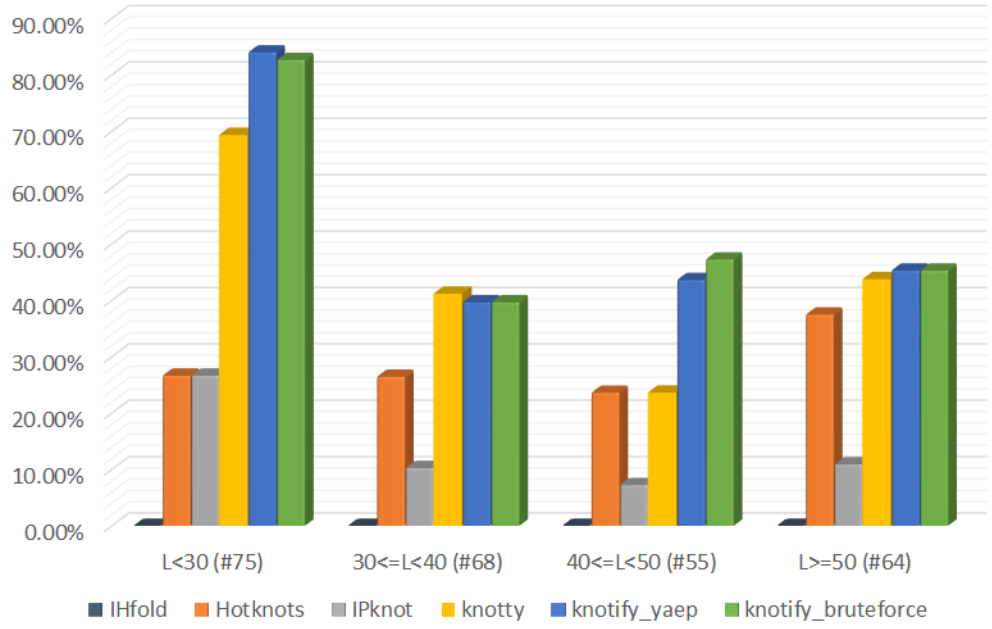
	L < 30 (#75)		**30 ≤ L < 40 (#68)**		**40 ≤ L <50 (#55)**		**L ≥ 50 (#64)**	
Platform	**Exact Matches**	**Exact Matches (%)**	**Exact Matches**	**Exact Matches (%)**	**Exact Matches**	**Exact Matches (%)**	**Exact Matches**	**Exact Matches (%)**
IHFold	0	0.00	0	0.00	0	0.00	0	0.00
Hotknots	20	26.67	18	26.47	13	23.64	24	37.5
IPknot	20	26.67	7	10.29	4	7.27	7	10.94
Knotty	52	69.33	28	41.18	13	23.64	28	43.75
knotify_yaep	63	84.00	27	39.71	24	43.64	29	45.31
knotify_bruteforce	62	82.67	27	39.71	26	47.27	29	45.31

Πίνακας 5.4: Πρόβλεψη της θέσης του ψευδοκόμβου με βάση το μήκος της ακολουθίας.

Το ποσοστό ακριβούς πρόβλεψης των κεντρικών βάσεων των ψευδοκόμβων ανά πλατφόρμα φαίνεται επίσης στο Σχήμα 5.7, ενώ το ποσοστό ακριβούς πρόβλεψης των κεντρικών βάσεων των ψευδοκόμβων ανά πλατφόρμα για όλες τις κατηγορίες μήκους των ακολουθιών RNA εμφανίζεται στο Σχήμα 5.8.



Σχήμα 5.7: Ποσοστό ακριβούς πρόβλεψης των κεντρικών βάσεων του ψευδοκόμβου ανά πλατφόρμα.



Σχήμα 5.8: Ποσοστό ακριβούς πρόβλεψης των κεντρικών βάσεων του ψευδοκόμβου ανά πλατφόρμα και μήκος ακολουθίας.

5.6.4 Πίνακας Σύγχυσης – Confusion Matrix

Το Knotify αξιολογείται στις ακολουθίες του Πίνακα 5.5 μαζί με τις μεθόδους IHFold, HotKnots, IPknot και Knotty. Ο Πίνακας 5.5 παρουσιάζει την απόδοση κάθε μεθόδου σε όρους ακρίβειας (precision ή positive predictive value–PPV), ανάκλησης (recall), F1-score και συντελεστή συσχέτισης Matthews (MCC). Οι εξισώσεις (5.2) - (5.5) παρέχουν τις ορισμούς, όπου οι TP–true positives αναφέρονται στον αριθμό των σωστά προβλεπόμενων βάσεων, οι FP–false positives αναφέρονται στον αριθμό των εσφαλμένα προβλεπόμενων βάσεων, οι FN–false negatives αναφέρονται στον αριθμό των βάσεων που δεν προβλέφθηκαν και οι TN–true negatives αναφέρονται στον αριθμό των βάσεων που δεν προβλέφθηκαν σωστά από το σύστημα.

$$PPV = \frac{TP}{TP + FP} \tag{5.2}$$

$$Recall = \frac{TP}{TP + FN} \tag{5.3}$$

$$F1 - score = \frac{2 \times PPV \times Recall}{PPV + Recall} \tag{5.4}$$

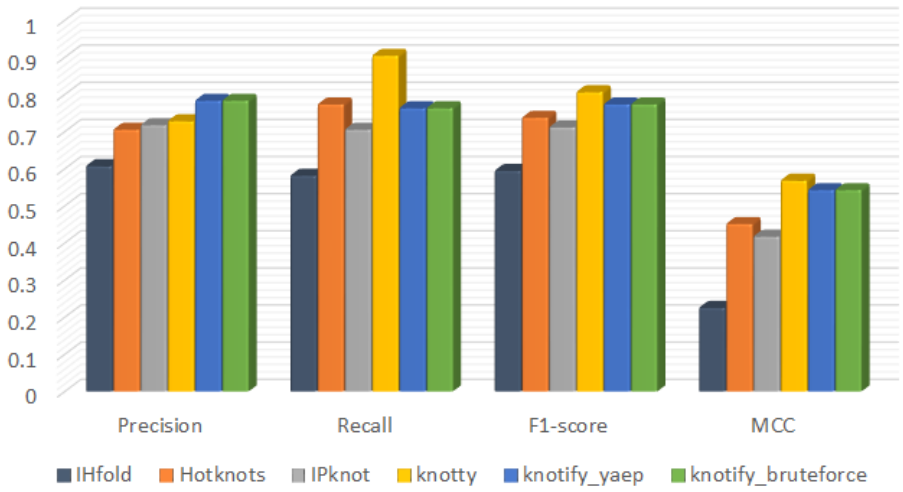
$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \tag{5.5}$$

Για να αξιολογηθεί η συνολική απόδοση, η έρευνα επικεντρώθηκε στην ακρίβεια, τον συντελεστή συσχέτισης Matthews (MCC) και το F1-score. Το τελευταίο είναι ο αρμονικός μέσος της ακρίβειας και της ανάκλησης. Το Knotify παρουσιάζει την υψηλότερη τιμή όσον αφορά τη μετρική της ακρίβειας, με τιμή 0.784, ενώ του Knotty ήταν 0.729, του IPknot 0.718, του Hotknots 0.706 και του IHFold 0.608. Όσον αφορά το F1-score και τη μετρική MCC, το Knotty είχε την καλύτερη επίδοση, με F1-score ίσο με 0.807 και MCC ίσο με 0.569. Η προτεινόμενη μεθοδολογία είχε την αμέσως καλύτερη επίδοση με τιμές πολύ κοντά στο Knotty (F1-score = 0.774, MCC = 0.543). Επιπλέον, το HotKnots σημείωσε F1-score ίσο με 0.738 και MCC ίσο με 0.452, ενώ το IPknot (F1-score = 0.712, MCC = 0.418) και το IHFold (F1-score = 0.595, MCC = 0.226) είχαν χαμηλότερη ακρίβεια τόσο στο F1-score όσο και στο MCC.

Platform	**tp**	**tn**	**fp**	**fn**	**Precision**	**Recall**	**F1-score**	**MCC**
IHFold	3056	3556	1968	2196	0.608	0.582	0.595	0.226
Hotknots	4180	3632	1744	1220	0.706	0.774	0.738	0.452
IPknot	3872	3767	1522	1615	0.718	0.706	0.712	0.418
Knotty	5026	3352	1870	528	0.729	0.905	0.807	0.569
knotify_yaep	4212	4102	1162	1300	0.784	0.764	0.774	0.543
knotify_bruteforce	4214	4101	1160	1301	0.784	0.764	0.774	0.543

Πίνακας 5.5: Precision, Recall, F1-score, και MCC ανά πλατφόρμα σε ολόκληρο το σύνολο δεδομένων.

Τα παραπάνω αποτελέσματα παρουσιάζονται επίσης στην Εικόνα 5.6.4.



Σχήμα 5.9: Precision, Recall, F1-score, και MCC ανά πλατφόρμα.

Στους Πίνακες 5.6 έως 5.9, παρουσιάζονται οι μετρικές ακρίβειας, ανάκλησης, F1-score και MCC ανά πλατφόρμα για τις τέσσερις κατηγορίες διαφορετικού μήκους ακολουθιών RNA. Σε αυτούς τους πίνακες φαίνεται ότι η προτεινόμενη μεθοδολογία ξεπέρασε όλες τις μεθόδους όσον αφορά τη μετρική της ακρίβειας για όλες τις κατηγορίες διαφορετικού μήκους, ενώ το Knotty ξεπέρασε την προτεινόμενη μεθοδολογία όσον αφορά το F1-score και τη μετρική MCC, κυρίως όταν οι ακολουθίες RNA ήταν μεγαλύτερες σε μέγεθος. Όταν το μήκος ήταν μικρότερο από 30, η μεθοδολογία Knotify είχε υψηλότερο F1-score και MCC από το Knotty. Όπως φαίνεται στον Πίνακα 5.3, η μεθοδολογία μας ήταν πιο ακριβής στην πρόβλεψη των κεντρικών βάσεων του ψευδοκόμβου σε όλο το σύνολο δεδομένων. Η αυξημένη τιμή MCC της πλατφόρμας Knotty σε μεγαλύτερες ακολουθίες RNA πιθανώς σχετίζεται με το γεγονός ότι οι μεγαλύτερες ακολουθίες RNA περιλαμβάνουν πολλές δομές (όπως hairpins, bulges) κ.α., τα οποία δεν σχετίζονται αναγκαστικά με τον ψευδοκόμβο και ενδέχεται να αυξήσουν τον συνολικό αριθμό των αληθώς θετικών (TP) προβλέψεων. Όπως θα δούμε στο Κεφάλαιο 7 η ενσωμάτωση τέτοιων δομών και στο Knotify οδηγεί στην περαιτέρω βελτίωση της προβλεπτικής του ικανότητας.

Platform	**tp**	**tn**	**fp**	**fn**	**Precision**	**Recall**	**F1-score**	**MCC**
IHFold	738	522	118	513	0.862	0.590	0.701	0.386
Hotknots	904	492	156	339	0.853	0.727	0.785	0.465
IPknot	916	514	124	337	0.881	0.731	0.799	0.510
Knotty	1196	469	146	80	0.891	0.937	0.914	0.722
knotify_yaep	1244	486	134	27	0.903	0.979	0.939	0.805
knotify_bruteforce	1242	485	136	28	0.901	0.978	0.938	0.802

Πίνακας 5.6: Precision, Recall, F1-score, και MCC ανά πλατφόρμα για ακολουθίες με μήκος <30.

Platform	**tp**	**tn**	**fp**	**fn**	**Precision**	**Recall**	**F1-score**	**MCC**
IHFold	550	832	352	587	0.610	0.484	0.539	0.191
Hotknots	922	851	294	254	0.758	0.784	0.771	0.528
IPknot	824	823	314	360	0.724	0.696	0.710	0.420
Knotty	1078	802	324	117	0.769	0.902	0.830	0.628
knotify_yaep	988	893	296	144	0.769	0.873	0.818	0.627
knotify_bruteforce	988	893	296	144	0.769	0.873	0.818	0.627

Πίνακας 5.7: Precision, Recall, F1-score, και MCC ανά πλατφόρμα για ακολουθίες με μήκος ≥30 και <40.

Platform	tp	tn	fp	fn	Precision	Recall	F1-score	MCC
IHFold	612	864	478	418	0.561	0.594	0.577	0.237
Hotknots	792	857	510	213	0.608	0.788	0.687	0.412
IPknot	764	911	410	287	0.651	0.727	0.687	0.414
Knotty	904	817	524	127	0.633	0.877	0.735	0.492
knotify_yaep	764	1010	298	300	0.719	0.718	0.719	0.490
knotify_bruteforce	772	1012	290	298	0.727	0.721	0.724	0.499

Πίνακας 5.8: Precision, Recall, F1-score, και MCC ανά πλατφόρμα για ακολουθίες με μήκος ≥40 και <50.

Platform	tp	tn	fp	fn	Precision	Recall	F1-score	MCC
IHFold	1156	1338	1020	678	0.531	0.63	0.577	0.196
Hotknots	1562	1432	784	414	0.666	0.790	0.723	0.439
IPknot	1368	1519	674	631	0.670	0.684	0.677	0.377
Knotty	1848	1264	876	204	0.678	0.901	0.774	0.515
knotify_yaep	1216	1713	434	829	0.737	0.595	0.658	0.402
knotify_bruteforce	1212	1711	438	831	0.735	0.593	0.656	0.398

Πίνακας 5.9: Precision, Recall, F1-score, και MCC ανά πλατφόρμα για ακολουθίες με μήκος ≥50.

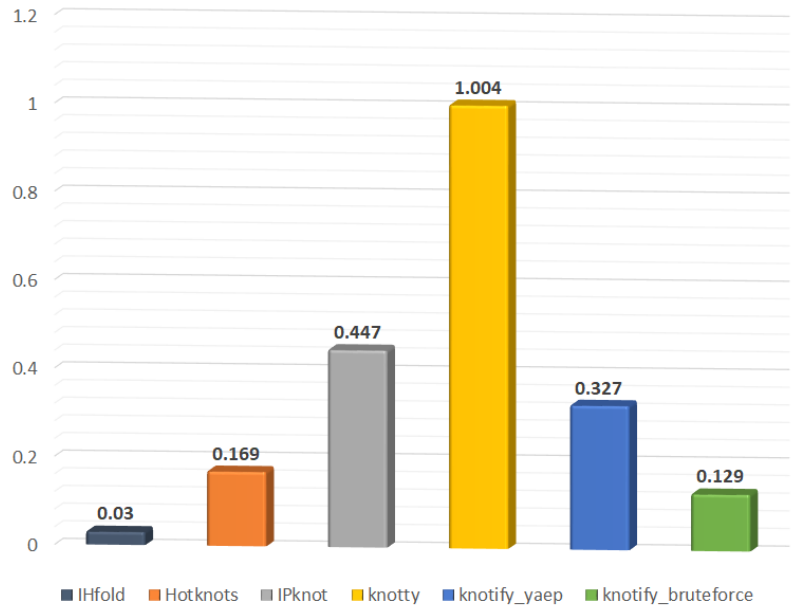
5.6.5 Σύγκριση χρόνου εκτέλεσης

Η τελευταία μετρική που χρησιμοποιήθηκε για τη σύγκριση της προτεινόμενης μεθοδολογίας με άλλες πλατφόρμες είναι αυτή του χρόνου εκτέλεσης. Στον Πίνακα 5.10 παρέχεται ο χρόνος εκτέλεσης που απαιτείται ανά πλατφόρμα για την πρόβλεψη της δευτεροταγούς δομής του RNA. Η τρίτη στήλη απεικονίζει τον συνολικό χρόνο εκτέλεσης που απαιτείται από κάθε πλατφόρμα για την ανάλυση όλων των 262 ακολουθιών RNA, ενώ η δεύτερη στήλη απεικονίζει τον μέσο χρόνο εκτέλεσης ανά ακολουθία RNA. Η μεθοδολογία μας ξεπέρασε το Knotty, το οποίο είχε χειρότερα αποτελέσματα όσον αφορά την πρόβλεψη των κυρίων κλαδικών βάσεων και την ακρίβεια, αλλά καλύτερα αποτελέσματα όσον αφορά το F1-score και τη μετρική MCC. Το Knotify_bruteforce απαιτούσε 33.894 δευτερόλεπτα, το knotify_yaep απαιτούσε 85.756 δευτερόλεπτα, και το Knotty απαιτούσε 263.303 δευτερόλεπτα. Η μεθοδολογία Knotify εκτέλεσε τις λειτουργίες της 7.76 (1.004/0.129) φορές ταχύτερα σε σύγκριση με την πλατφόρμα Knotty, με την οποία έχουν συγκρίσιμα ποσοστά στις υπόλοιπες μετρικές αξιολόγησης. Το IPknot και το Hotknots χρειάστηκαν κατά μέσο όρο ανά ακουλουθία 0.447 και 0.169 αντίστοιχα. Τέλος, το IHFold κατέγραψε τον χαμηλότερο χρόνο εκτέλεσης, αλλά είχε το χειρότερο προφίλ αξιολόγησης ακρίβειας.

Platform	Average Time (sec)	Total Time (sec)
IHFold	0.030	8.096
Hotknots	0.169	44.432
IPknot	0.447	117.246
Knotty	1.004	263.303
knotify_yaep	0.327	85.756
knotify_bruteforce	0.129	33.894

Πίνακας 5.10: Μέσος και συνολικός χρόνος εκτέλεσης που απαιτείται ανά πλατφόρμα για ολόκληρο το σύνολο δεδομένων.

Ο χρόνος εκτέλεσης που απαιτείται ανά πλατφόρμα παρουσιάζεται επίσης στο Σχήμα 5.10.



Σχήμα 5.10: Μέσος χρόνος εκτέλεσης που απαιτείται ανά πλατφόρμα σε δευτερόλεπτα

Στους Πίνακες 5.11–5.14, παρουσιάζεται ο μέσος και ο συνολικός χρόνος εκτέλεσης ανά πλατφόρμα για τέσσερις κατηγορίες διαφορετικών μηκών αλληλουχιών RNA. Αξίζει να σημειωθεί ότι ο χρόνος εκτέλεσης της πλατφόρμας Knotty αυξήθηκε σημαντικά ανάλογα με το μήκος της εισερχόμενης αλληλουχίας RNA, ενώ οι Hotknots, knotify_yaep, και knotify_bruteforce φαίνεται να αυξάνονται παρόμοια καθώς το μήκος της ακολουθίας RNA μεγαλώνει, διατηρώντας έναν αρκετά σταθερό ρυθμό.

Platform	Average Time (sec)	Total Time (sec)
IHFold	0.0233	1.748
Hotknots	0.0709	5.314
IPknot	0.0143	1.070
Knotty	0.0212	1.590
knotify_yaep	0.0697	5.226
knotify_bruteforce	0.0427	3.204

Πίνακας 5.11: Συνολικός χρόνος εκτέλεσης που απαιτείται ανά πλατφόρμα για ακολουθίες μήκους <30.

Platform	Average Time (sec)	Total Time (sec)
IHFold	0.0248	1.689
Hotknots	0.0982	6.680
IPknot	0.0408	2.777
Knotty	0.0692	4.703
knotify_yaep	0.1589	10.808
knotify_bruteforce	0.0964	6.555

Πίνακας 5.12: Συνολικός χρόνος εκτέλεσης που απαιτείται ανά πλατφόρμα για ακολουθίες μήκους ≥ 30 και < 40.

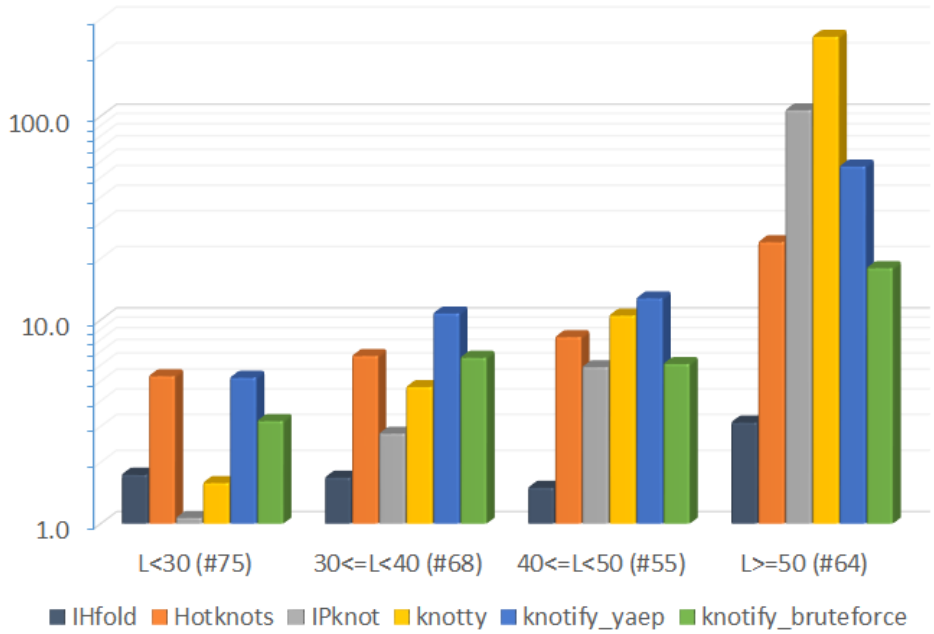
Platform	Average Time (sec)	Total Time (sec)
IHFold	0.0274	1.507
Hotknots	0.1503	8.264
IPknot	0.107	5.886
Knotty	0.1918	10.546
knotify_yaep	0.2331	12.821
knotify_bruteforce	0.111	6.103

Πίνακας 5.13: Συνολικός χρόνος εκτέλεσης που απαιτείται ανά πλατφόρμα για ακολουθίες μήκους ≥ 40 και < 50.

Ο συνολικός χρόνος εκτέλεσης που απαιτείται ανά πλατφόρμα για τις τέσσερις ομάδες διαφορετικού μήκους ακολουθιών RNA εμφανίζεται επίσης στο Σχήμα 5.11. Σημειώνεται επίσης ότι χρησιμοποιήθηκε λογαριθμική κλίμακα για την καλύτερη οπτικοποίηση των αποτελεσμάτων.

Platform	Average Time (sec)	Total Time (sec)
IHFold	0.0492	3.151
Hotknots	0.3777	24.172
IPknot	1.679	107.511
Knotty	3.851	246.462
knotify_yaep	0.8891	56.900
knotify_bruteforce	0.2817	18.030

Πίνακας 5.14: Συνολικός χρόνος εκτέλεσης που απαιτείται ανά πλατφόρμα για ακολουθίες μήκους ≥ 50.



Σχήμα 5.11: Μέσος χρόνος εκτέλεσης που απαιτείται ανά πλατφόρμα σε δευτερόλεπτα (σε λογαριθμική κλίμακα).

Κεφάλαιο 6

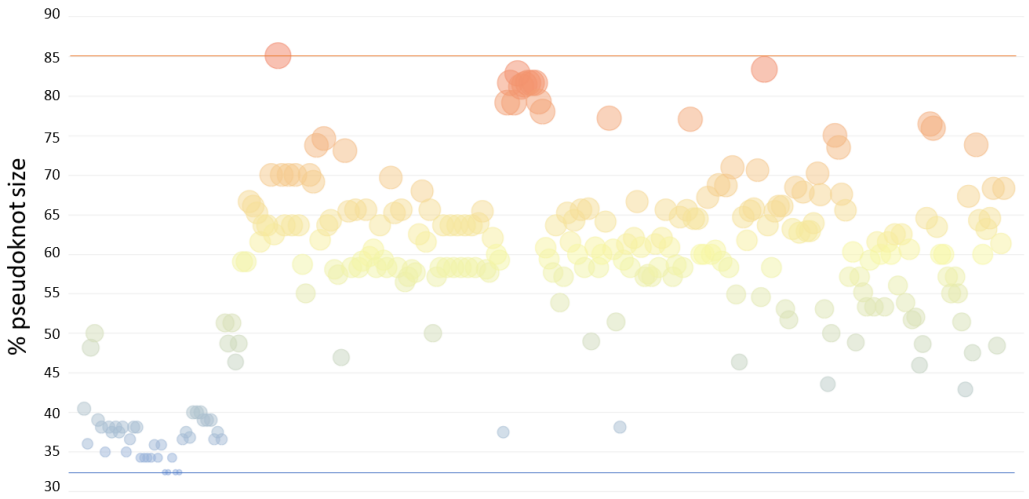
Εισαγωγή μεθόδου κλαδέματος για βελτιστοποίηση του χρόνου εκτέλεσης

Στο κεφάλαιο αυτό παρουσιάζεται μια βελτιωμένη έκδοση του Knotify με τη χρήση μιας τεχνικής κλαδέματος στο χώρο αναζήτησης της γραμματικής της μεθοδολογίας. Η βελτιωμένη μεθοδολογία καταφέρνει να αναγνωρίζει ψευδοκόμβους τύπου H μετά από τρεις εργασίες: (i) κατασκευή όλων των δέντρων ανάλυσης με τη χρήση ενός αναγνωριστή CFG, (ii) διάσχιση όλων των δημιουργημένων δέντρων για να αναγνωρίσει επιπλέον ζεύγη βάσης γύρω από τον ψευδοκόμβο, και (iii) επιλογή του βέλτιστου δέντρου με βάση τα κριτήρια ελάχιστης ελεύθερης ενέργειας και μέγιστου αριθμού ζευγών βάσης, με μια σημαντική βελτίωση του χρόνου εκτέλεσης. Η CFG G_{RNA} που προτάθηκε προηγουμένως είναι ικανή να ανιχνεύσει ψευδοκόμβους σε ακολουθίες όπου τα πρώτο και τελευταίο σύμβολο της ακολουθίας ανήκουν στην ομάδα των κεντρικών βάσεων του ψευδοκόμβου. Η μετέπειτα ανάλυση των ενδιάμεσων τμημάτων των ακολουθιών εκτελείται χρησιμοποιώντας την τεχνική του ΄κινούμενου παράθυρου΄. Η τεχνική του ΄κινούμενου παράθυρου΄ διαιρεί την αρχική είσοδο σε υποακολουθίες, ξεκινώντας από την υποακολουθία που ξεκινά με το πρώτο σύμβολο και έχει το ελάχιστο δυνατό μήκος. Στη συνέχεια, το μήκος της εξεταζόμενης υποακολουθίας αυξάνεται κατά ένα σύμβολο για να περιλαμβάνει τελικά ολόκληρη την αρχική ακολουθία RNA. Έπειτα, χρησιμοποιώντας την ίδια επαναληπτική διαδικασία, η θέση εκκίνησης του συμβόλου αυξάνεται κατά ένα για να εξαιρέσει το προηγούμενο σύνολο εκκίνησης, διαδικασία που αναλύθηκε στο προηγούμενο Κεφάλαιο.

Η νέα μεθοδολογία αποσκοπεί στη ενσωμάτωση μιας τεχνικής κλαδέματος στον χώρο αναζήτησης της γραμματικής με στόχο τη μείωση των παραγόμενων δέντρων. Αυτό επιτυγχάνεται μέσω μίας αποκοπής δέντρων, η οποία επηρεάζει μόνο την πρώτη εργασία της υλοποίησης του συστήματος.

6.1 Η μέθοδος του κλαδέματος στα παραγόμενα δέντρα της γραμματικής

Η τεχνική κλαδέματος βασίζεται στον καθορισμό δύο κατωφλίων όσον αφορά το ελάχιστο και το μέγιστο μήκος που μπορεί να έχει ένας ψευδοκόμβος σε σχέση με το μέγεθος της συνολικής ακολουθίας RNA. Αυτά τα δύο κατώφλια θα αναφέρονται ως *minimum_percentage* και *maximum_percentage*. Βάσει αυτών των δύο κατωφλίων, ο αριθμός των εξετασμένων υποακολουθιών κατά την εκτέλεση της τεχνικής του ΄κινούμενου παράθυρου΄ μειώνεται σημαντικά και κατ΄ επέκταση και ο χρόνος εκτέλεσης του συστήματος. Από την ανάλυση και το Σχήμα 6.1, διακρίνεται η δυνατότητα βελτίω-



Σχήμα 6.1: Ποσοστό του μήκους του ψευδοκόμβου σε σχέση με την συνολική ακολουθία.

σης όσον αφορά τον απαιτούμενο χρόνο για τον εντοπισμό και τη συνολική περιγραφή της δομής του ψευδοκόμβου. Εκμεταλλευόμενοι τα δύο όρια–κατώφλια, αρκούν για να βελτιώσουν τον χρόνο κατά 33% στη μεθοδολογία που βασίζεται σε γραμματική και 43% στη μεθοδολογία που βασίζεται σε αναζήτηση brute-force, χωρίς να θυσιάσουν την ακρίβεια των συστημάτων. Αυτό το εύρημα αποκαλύφθηκε παρατηρώντας τους στατιστικούς δείκτες, και συγκεκριμένα την πυκνότητα πιθανότητας, τη διασπορά και τη μέση τιμή της τυχαίας μεταβλητής που αναφέρεται στη σχέση του μήκους του ψευδοκόμβου με το συνολικό μήκος. Αυτή η αρχική ανάλυση υποδεικνύει ότι η συσχέτιση των χαρακτηριστικών θα μπορούσε περαιτέρω να βελτιώσει την απόδοση του συστήματος, με χρήση κάποιου πιο εξειδικευμένου αλγορίθμου, διαδικασία που θα μπορούσε να εξετασθεί σε μελλοντικές επεκτάσεις του συστήματος.

Η ίδια συλλογή δεδομένων [83] από 262 ακολουθίες RNA χρησιμοποιήθηκε για να εκτιμηθούν οι τιμές των ορίων *minimum_percentage* και *maximum_percentage*. Στο Σχήμα 6.1 παρουσιάζεται το ποσοστό του μήκους του ψευδοκόμβου στην ακολουθία RNA (μήκος ψευδοκόμβου/μήκος ακολουθίας RNA) για όλες τις 262 ακολουθίες RNA. Όπως φαίνεται σε αυτό το Σχήμα, για όλες τις 262 ακολουθίες, το μήκος του ψευδοκόμβου περιορίζεται μεταξύ του 32% και του 85% του μήκους της συνολικής ακολουθίας RNA.

6.2 Αξιολόγηση επίδοσης με προσαρμογή του κλαδέματος

Η ίδια συλλογή δεδομένων από 262 ακολουθίες RNA χρησιμοποιήθηκε για να αξιολογήσει την απόδοση της μεθοδολογίας με τη μέθοδο του κλαδέματος σε σύγκριση τα συστήματα Hotknots, Iterative HFold (IHFold), IPknot.

Platform	**Average Time**	**Total Time**
IHFold	0.030	8.096
Hotknots	0.169	44.432
IPknot	0.447	117.246
Knotty	1.004	263.303
Knotify_yaep	0.327	85.756
Knotify_bruteforce	0.129	33.894
Knotify_yaep_pruned	0.218	57.202
Knotify_bruteforce_pruned	0.073	19.377

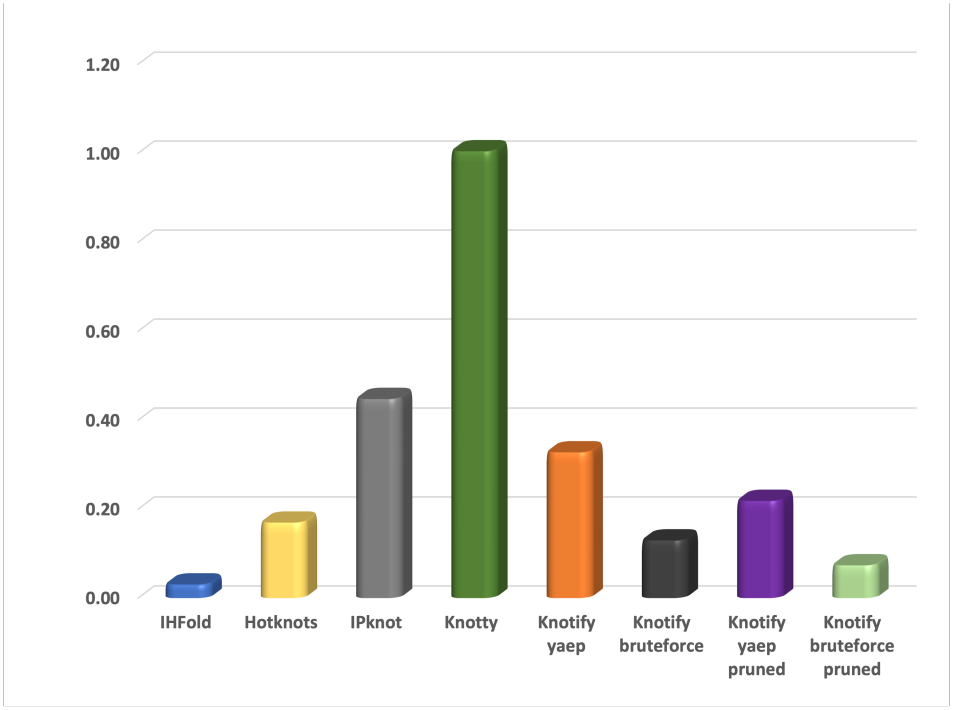
Πίνακας 6.1: Μέσος και συνολικός χρόνος εκτέλεσης που απαιτείται ανά πλατφόρμα για ολόκληρο το σύνολο δεδομένων, συμπεριλαμβανομένων των συστημάτων με τη μέθοδο κλαδέματος.

Όπως αναφέρθηκε παραπάνω, ο περιορισμός που επιβάλλεται στην τεχνική του ΄κινούμενου παραθύρου΄ με τη χρήση των κατωφλίων *minimum_percentage* και *maximum_percentage* δεν επηρεάζει την ακρίβεια της πρόβλεψης του συστήματος. Η προτεινόμενη μεθοδολογία προβλέπει τις κεντρικές βάσεις σε ποσοστό 76.4%, ενώ παρουσιάζει F1-score ίσο με 0.774 και MCC ίσο με 0.437, τα οποία αντιστοιχούν στην αρχική ακρίβεια της μεθοδολογίας Knotify, όπως φαίνεται στον Πίνακα 6.2. Σημαντική διαφορά, όμως, παρουσιάζει ο χρόνος εκτέλεσης του συστήματος με την ενσωμάτωση της μεθόδου κλαδέματος. Στον Πίνακα 6.1, παρουσιάζεται η αξιολόγηση του χρόνου εκτέλεσης ανά πλατφόρμα για την πρόβλεψη ψευδοκόμβου στο εξεταζόμενο σύνολο δεδομένων. Η τρίτη στήλη του πίνακα απεικονίζει τον συνολικό χρόνο εκτέλεσης που απαιτείται από κάθε πλατφόρμα για την ανάλυση όλων των 262 ακολουθιών RNA, ενώ η δεύτερη στήλη απεικονίζει τον μέσο χρόνο εκτέλεσης ανά ακολουθία

Platform	tp	tn	fp	fn	Precision	Recall	F1-score	MCC
IHFold	3056	3556	1968	2196	0.608	0.582	0.595	0.226
Hotknots	4180	3632	1744	1220	0.706	0.774	0.738	0.452
IPknot	3872	3767	1522	1615	0.718	0.706	0.712	0.418
Knotty	5026	3352	1870	528	0.729	0.905	0.807	0.569
Knotify_yaep	4212	4102	1162	1300	0.784	0.764	0.774	0.543
Knotify_bruteforce	4214	4101	1160	1301	0.784	0.764	0.774	0.543
Knotify_yaep_pruned	4212	4102	1162	1300	0.784	0.764	0.774	0.543
Knotify_bruteforce_pruned	4214	4101	1160	1301	0.784	0.764	0.774	0.543

Πίνακας 6.2: Precision, Recall, F1-score, και MCC ανά πλατφόρμα σε ολόκληρο το σύνολο δεδομένων, συμπεριλαμβανομένων των συστημάτων με τη μέθοδο κλαδέματος.

RNA. Σε αυτόν τον πίνακα, οι προτεινόμενες μεθοδολογίες παρουσιάζονται ως Knotify_yaep_pruned και Knotify_bruteforce_pruned, ενώ οι δύο προηγούμενες εκδόσεις του Knotify, χωρίς τη μέθοδο κλαδέματος, παρουσιάζονται ως Knotify_yaep και Knotify_bruteforce.



Σχήμα 6.2: Μέσος χρόνος εκτέλεσης που απαιτείται ανά πλατφόρμα σε δευτερόλεπτα, συμπεριλαμβανομένων των συστημάτων με τη μέθοδο κλαδέματος.

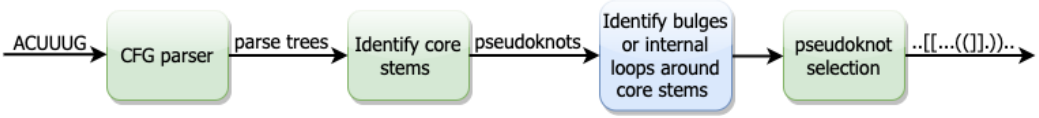
Η μεθοδολογία μας υπερτερεί σε σχέση με το Knotty, που είχε χειρότερα αποτελέσματα όσον αφορά την πρόβλεψη των κεντρικών βάσεων και την ακρίβεια, αλλά συνεχίζει να παρουσιάζει καλύτερες επιδόσεις όσον αφορά το F1-score και το MCC. Ο χρόνος εκτέλεσης που απαιτείται από το Knotify_bruteforce ήταν 33.894 δευτερόλεπτα, ενώ το Knotify_yaep χρειάστηκε 85.756 δευτερόλεπτα. Η Knotify_bruteforce_pruned απαιτούσε 19.377 δευτερόλεπτα, ενώ η Knotify_yaep_pruned χρειάστηκε 57.202 δευτερόλεπτα, και το Knotty απαιτούσε 263.303 δευτερόλεπτα. Η προτεινόμενη μεθοδολογία Knotify_yaep_pruned επιταχύνθηκε κατά 33% σε σύγκριση με την πλατφόρμα Knotify_yaep, ενώ η μεθοδολογία Knotify_bruteforce_pruned κατά 43% σε σύγκριση με την πλατφόρμα Knotify_bruteforce. Τέλος, η IHFold κατέγραψε τον χαμηλότερο χρόνο εκτέλεσης, παρόλο που είχε το χειρότερο προφίλ αξιολόγησης ακρίβειας, όπως αναφέρθηκε και στην Ενότητα 5.6.5. Ο απαιτούμενος χρόνος εκτέλεσης ανά πλατφόρμα παρουσιάζεται επίσης στο Σχήμα 6.2.
Η σημαντική συνεισφορά της μεθόδου κλαδέματος στη δραστική μείωση του χρόνου εκτέλεσης, οδήγησε στην επιλογή ενσωμάτωσής της σε όλα τα συστήματα (Knotify+ και Knotify για L-type) που υλοποιήθηκαν στη συνέχεια της διατριβής και παρουσιάζονται στα ακόλουθα Κεφάλαια.

Κεφάλαιο 7

Knotify+

7.1 Μεθοδολογία ανίχνευσης ψεδοκόμβων με ασύμμετρους βρόχους/εξογκώματα (bulges) και εσωτερικούς βρόχους (internal loops)

Σε αυτο το Κεφάλαιο παρουσιάζεται η πλατφόρμα Knotify+. Το Knotify+ αποτελεί μια επέκταση της πλατφόρμας Knotify με την ενσωμάτωση της μεθόδου κλαδέμος του Κεφαλαίου 6. Το Knotify+ είναι ικανό να προβλέπει ψευδοκόμβους τύπου H με ασσύμετρους βρόχους/εξογκώματα και εσωτερικούς βρόχους γύρω από τους κεντρικές βάσεις του ψευδοκόμβου, δηλαδή στους δύο βρόχους του ψευδοκόμβου. Το Knotify+ προσθέτει μια νέα εργασία (μπλε πλαίσιο στο Σχήμα 7.1) πριν από την επιλογή του ψευδοκόμβου, η οποία είναι υπεύθυνη για τον εντοπισμό των ασύμμετρων ή των εσωτερικών βρόχων γύρω από τις κεντρικές βάσεις. Συνεπώς, η βασική συνεισφορά του Knotify+ είναι η αύξηση της εκφραστικότητας της αρχικής μεθοδολογίας με την ενσωμάτωση στην διαδικασία πρόβλεψη της δευτεροταγούς δομής, ασύμμετρων και των εσωτερικών βρόχων στους δύο βρόχους του ψευδοκόμβου. Η διαδικασία αυτή παρουσιάζεται στην Ενότητα 7.2, ενώ μία λεπτομερής ανάλυση των επιμέρους εργασιών (Σχήμα 7.1) παρέχεται στις επόμενες υποενότητες.



Σχήμα 7.1: Επισκόπηση της προτεινόμενης μεθοδολογίας του Knotify+.

7.2 Διαδικασία διακόσμησης (decoration) για την ενσωμάτωση ασύμμετρων και εσωτερικών βρόχων

Κατά τη διάρκεια της πρώτης εργασίας, όλα τα δέντρα ανάλυσης κατασκευάστηκαν από τον CFG αναλυτή. Στη συνέχεια, ακολουθεί από το σύστημα η διάσχιση όλων αυτών των δέντρων για τον εντοπισμό των κεντρικών βάσεων. Όλες οι βάσεις που βρίσκονται σε κάθε έναν από τους δύο βρόχους ελέγχονται διαδοχικά για το εάν μπορούσαν να δημιουργήσουν ζεύγη με μια βάση σε μια κατάλληλη θέση του ψευδοκόμβου ή για την ύπαρξη ασύμμετρου η συμμετρικού βρόχου. Η διαδικασία αυτή αποτελεί το νέο, μετασχηματισμένο decoration για το σύστημα Knotify+.
Στον Πίνακα 7.1, παρουσιάζεται η διαδικασία διακόσμησης των κεντρικών βάσεων. Αφού ο αναλυτής ανιχνεύσει τις κεντρικές βάσεις U-A και C-G στις θέσεις 10-17 και 5-11 αντίστοιχα, καθορίζονται οι δύο βρόχοι του ψευδοκόμβου. Ο αριστερός βρόχος βρίσκεται στις θέσεις 6 έως 9, και ο δεξιός βρόχος βρίσκεται στις θέσεις 12 έως 16. Οι βάσεις σε αυτούς τους βρόχους ελέγχθηκαν αρχικά για το εάν μπορούσαν να συνδεθούν με βάσεις εκτός των βρόχων του ψευδοκόμβου. Τα ζευγάρια βάσεων στον αριστερό βρόχο ελέγχθηκαν για ταιριάσματα/ζεύγη με βάσεις στις θέσεις 18-22, ενώ οι βάσεις στο δεξιό βρόχο ελέγχθηκαν για ταιριάσματα/ζεύγη με βάσεις στις θέσεις 1-4. Και στους δύο βρόχους του ψευδοκόμβου, τα ζευγάρια των βάσεων στις θέσεις 9-18, 8-19, 4-12 και 3-14 ανιχνεύθηκαν στα βήματα 1 έως 4, αντίστοιχα. Ο Πίνακας 7.1 παρουσιάζει αυτήν τη διαδικασία λεπτομερώς, με τις αγκύλες και τις παρενθέσεις, που καθορίζουν τα εκάστοτε ζεύγη βάσεων σε κάθε ένα βήμα της μεθοδολογιας.

Position	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22
String	A	C	A	U	C	C	G	C	C	U	G	A	U	U	U	G	A	G	C	A	C	A
Core stems:	.	.	.	.	[	.	.	.	.	(	]	.	.	.	.	.	)	.	.	.	.	.
Stage 1	.	.	.	.	[	.	.	.	(	(	]	.	.	.	.	.	)	)	.	.	.	.
Stage 2	.	.	.	.	[	.	.	(	(	(	]	.	.	.	.	.	)	)	)	.	.	.
Stage 3	.	.	.	[	[	.	.	(	(	(	]	]	.	.	.	.	)	)	)	.	.	.
Stage 4	.	.	[	[	[	.	.	(	(	(	]	]	]	.	.	.	)	)	)	.	.	.
Stage 5	[	.	[	[	[	.	(	(	(	(	]	]	]	.	]	.	)	)	)	.	)	.

Πίνακας 7.1: Η διαδικασία decoration του ψευδοκόμβου τύπου Η. Οι μη ζευγαρωμένες βάσεις γύρω από τις κεντρικές βάσεις του ψευδοκόμβου σχηματίζουν ασύμμετρους ή εσωτερικούς βρόχους και συμβολίζονται με κόκκινες τελείες.

Μόλις δεν ήταν δυνατή η δημιουργία περισσότερων συνεχόμενων ζευγών βάσεων, ο έλεγχος για την ύπαρξη ασσύμετρων βρόχων/εξογκωμάτων και εσωτερικών βρόχων πραγματοποιήθηκε (βήμα 5) από το Knotify+. Για κάθε πλευρά, αριστερά και δεξιά, εξετάστηκαν οι μη ζευγαρωμένες βάσεις για το εάν μπορούσαν να σχηματίσουν ζεύγος βάσεων μετά τη δημιουργία εξογκώματος ή εσωτερικού βρόχου. Στο παράδειγμά μας, στην αριστερή πλευρά, το σύνολο στις θέσεις 6-7 μπορεί να σχηματίσει ζεύγη

βάσεων με το σύνολο στις θέσεις 20-22 μετά τη δημιουργία εξογκώματος. Αυτά τα δύο σύνολα ονομάστηκαν το αριστερό ζευγάρι συνόλων. Στη δεξιά πλευρά, το σύνολο στις θέσεις 1-2 μπορεί να σχηματίσει ζεύγη βάσεων με το σύνολο στις θέσεις 14-16 μετά τη δημιουργία μοτίβου εξογκώματος και αυτά τα δύο σύνολα ονομάστηκαν το δεξί ζευγάρι συνόλων.
Οι χρήστες μπορούν να καθορίσουν το μέγιστο μέγεθος εξογκωμάτων, το οποίο δίνεται ως παράμετρος κατά την εκτέλεση του προγράμματος από τη γραμμή εντολών. Αυτή η παράμετρος ονομάζεται *maximum_bulge_size*. Για κάθε ζεύγος συνόλων, μπορεί να υπάρχει εξόγκωμα μήκους από 0 έως *maximum_bulge_size* σε κάθε ένα από τα δύο σύνολα. Στην περίπτωση όπου το μήκος του ασύμμετρου βρόχου/εξογκώματος είναι μηδέν από τη μία πλευρά και μεγαλύτερο από το μηδέν από την άλλη πλευρά, τότε δημιουργείται ένας ασύμμετρος βρόχος. Διαφορετικά, εάν το μήκος τους είναι μεγαλύτερο από το μηδέν από τις δύο πλευρές, τότε προκύπτει ένας εσωτερικός βρόχος. Για την επιλογή του κατάλληλου μοτίβου εξογκώματος ή εσωτερικού βρόχου, χρησιμοποιείται το καρτεσιανό γινόμενο αυτών των περιπτώσεων, και παράγεται ένα πλήθος συμβολοσειρών τύπου dot-bracket.
Εφαρμόζοντας στη συνέχεια, τα κριτήρια της ελάχιστης ελευθέρης ενέργειας και του μεγαλύτερου αριθμού ζευγών βάσεων του ψευδοκόμβου, επιλέγεται η βέλτιστη δομή. Το αποτέλεσμα αυτής της διαδικασίας παρουσιάζεται στο βήμα 5 του Πίνακα 7.1. Όσον αφορά το αριστερό ζευγάρι συνόλων, μπορεί να υπάρχει ένα ζεύγος βάσεων στις θέσεις 7-21 μετά τη δημιουργία εξογκώματος στη θέση 20. Όσον αφορά το δεξί ζευγάρι συνόλων, μπορεί να υπάρχει ένα ζεύγος βάσεων στις θέσεις 2-16 μετά τη δημιουργία εξογκώματος στις θέσεις 14-15, μπορεί να υπάρχει ένα ζεύγος βάσεων στις θέσεις 1-14 μετά τη δημιουργία εξογκώματος στη θέση 2, ή μπορεί να υπάρχει ένα ζεύγος βάσεων στις θέσεις 1-15 μετά τη δημιουργία εξογκώματος στη θέση 2 και ένα στη θέση 14, δημιουργώντας έτσι έναν εσωτερικό βρόχο. Η τελευταία περίπτωση ήταν αυτή που επιλέχθηκε τελικά, όπως φαίνεται στο βήμα 5, όπου ο εσωτερικός βρόχος υπογραμμίζεται με κόκκινο.
Το Knotify+ επιτρέπει στον χρήστη να επιλέξει την επιλογή των ζευγών βάσεων U-G ως παράμετρο από τη γραμμή εντολών, καθώς και την τιμή του *maximum_bulge_size*. Ο πηγαίος κώδικας και οι λεπτομέρειες υλοποίησης είναι διαθέσιμοι προς χρήση και επέκταση στο GitHub [80].

7.3 Αξιολόγηση της απόδοσης του Knotify+

7.3.1 Παρουσίαση Συνόλου Δεδομένων

Για να αξιολογηθεί η ακρίβεια του Knotify+ έναντι άλλων μεθοδολογιών, κατασκευάστηκε ένα σύνολο δεδομένων [83] που αποτελείται από 260 γνωστές αλληλουχίες RNA που περιλαμβάνουν ψευδοκόμβους. Αυτό το σύνολο δεδομένων είναι ελαφρώς τροποποιημένο από αυτό που παρουσιάστηκε στο Κεφάλαιο για το Knotify. Ένα μεγάλο αριθμό αυτών των ακολυθιών περιλαμβάνει εξογκώματα ή εσωτερικούς βρόχους μετά τις κεντρικές βάσεις του ψευδοκόμβου. Το σύνολο δεδομένων χωρίστη-

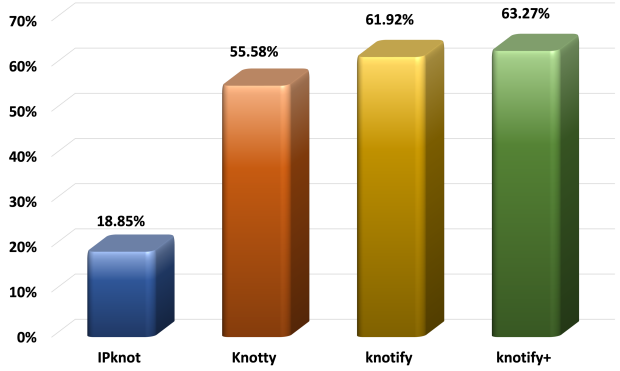
κε σε τέσσερα σύνολα ανάλογα με το μήκος. Το πρώτο σύνολο αποτελείται από 75 ακολουθίες RNA με μήκος μικρότερο από 30, το δεύτερο αποτελείται από 67 ακολουθίες με μήκος μεταξύ 30 και 40, το τρίτο αποτελείται από 55 ακολουθίες με μήκος μεταξύ 40 και 50, και το τελευταίο σύνολο αποτελείται από 63 ακολουθίες με μήκος μεγαλύτερο από ή ίσο με 50. Οι αλληλουχίες επιλέχθηκαν από πλατφόρμες βάσεων RNA δεδομένων [84, 85] που παρέχουν διαθέσιμα για δημόσια χρήση τα συγκεκριμένα δεδομένα. Η προτεινόμενη μεθοδολογία συγκρίθηκε με τα δύο πιο αποδοτικά εργαλεία του προηγούμενου Κεφαλαίου, δηλαδή το IPknot και το Knotty [35, 32], καθώς και με την προηγούμενη έκδοση της υλοποίησης, το Knotify. Επομένως, κατά τη διαδικασία αξιολόγησης της απόδοσης χρησιμοποιήθηκαν τέσσερεις πλατφόρμες, το IPknot, το Knotty, το Knotify και το Knotify+.

7.3.2 Μέθοδοι Αξιολόγησης του Knotify+

Στη μέτρηση της απόδοσης του Knotify+, επιλέχθηκαν οι ίδιες τρεις μεθόδοι που χρησιμοποιήθηκαν στο 5.6.2: α) το ποσοστό πρόβλεψης των κεντρικών βάσεων του ψευδοκόμβου, β) ο πίνακας σύγχυσης που περιλαμβάνει την ακρίβεια (PPV), την ανάκληση (Recall), το F1-score και το MCC (συντελεστής συσχέτισης Matthews), και γ) ο χρόνος εκτέλεσης. Όσον αφορά την πλατφόρμα Knotify+, όλα τα πειράματα υλοποιήθηκαν με την παράμετρο *maximum_ bulge_ size* ίση με 3.

7.3.3 Knotify+: Πρόβλεψη κεντρικών βάσεων του ψευδοκόμβου

Στον Πίνακα 7.2, παρουσιάζεται η δυνατότητα κάθε πλατφόρμας να προβλέπει τις κεντρικές βάσεις των ψευδοκόμβων, δηλαδή αυτό που αναφέρεται και ως θέση του ψευδοκόμβου. Η δεύτερη στήλη παρουσιάζει τον αριθμό των ψευδοκόμβων για τους οποίους μια πλατφόρμα κατάφερε να προβλέψει και τις δύο κεντρικές βάσεις, ενώ η τέταρτη στήλη παρουσιάζει τον αριθμό των ψευδοκόμβων για τους οποίους μια πλατφόρμα κατάφερε να προβλέψει μόνο μία κεντρική βάση. Επειδή κάποιες ακολουθίες του συνόλου των δεδομένων, περιέχουν αρκετά περίπλοκα μοτίβα στους ψευδοκόμβους, κάνοντας ακόμα πιο σύνθετη την πρόβλεψη της θέσης του, όλες οι μέθοδοι που χρησιμοποιήθηκαν εξετάστηκαν και ως προς την πρόβλεψη τουλάχιστον ενός ζεύγους κεντρικών βάσεων.
Η προτεινόμενη μεθοδολογία, Knotify+, όπως και το Knotify, εντόπισε τις δύο κεντρικές βάσεις του ψευδοκόμβου τέλεια σε 142 από τις 260 ακολουθίες, ενώ το IPknot σε 38 ακολουθίες και το Knotty σε 121 ακολουθίες. Επιπλέον, το Knotify+ κατάφερε να εντοπίσει τουλάχιστον ένα ζεύγος κεντρικών βάσεων του ψευδοκόμβου σε 45 ακολουθίες, ενώ το IPknot το έκανε σε 22 ακολουθίες, το Knotty σε 47 ακολουθίες και το Knotify σε 38 ακολουθίες. Ως εκ τούτου, το Knotify+ υπερτερεί των άλλων πλατφορμών, καταφέρνοντας να προβλέψει τουλάχιστον ένα ζεύγος κεντρικών βάσεων στο 63.27% των ακολουθιών του συνόλου δεδομένων, έναντι 18.85% του IPknot, 55.58% του Knotty και 61.92% του Knotify. Αυτό το εύρημα δείχνει ότι ακόμα και



Σχήμα 7.2: Ποσοστό πρόβλεψης τουλάχιστον ενός ζεύγους κεντρικών βάσεων ανά πλατφόρμα

στις περιπτώσεις όπου η ακριβής πρόβλεψη δεν ήταν δυνατή, το Knotify+ προέβλεψε τουλάχιστον ένα ζεύγος κεντρικών βάσεων καλύτερα από την προηγούμενη υλοποίηση Knotify και τις άλλες δύο γνωστές πλατφόρμες, αναδεικνύοντας τη βελτίωση που προκύπτει από την αύξηση της εκφραστικότητας του νέου συστήματος.

Platform	2 Matches	2 Matches (%)	1 Match	At least 1 Match (%)
IPknot	38	14.62	22	18.85
Knotty	121	46.54	47	55.58
Knotify	142	54.62	38	61.92
Knotify+	142	54.62	45	63.27

Πίνακας 7.2: Πρόβλεψη της θέσης του ψευδοκόμβου σύμφωνα με τις κεντρικές βάσεις σε ολόκληρο το σύνολο δεδομένων

7.3.4 Knotify+: Πίνακας Σύγχυσης, Ακρίβεια, Ανάκληση, F1-score, και MCC

Η επίδοση όλων των εξεταζόμενων πλατφορμών αναφορικά με την ακρίβεια, την ανάκληση, το F1-score, και το MCC παρουσιάζονται στον Πίνακα 7.3.
Η προτεινόμενη μεθοδολογία υπερτερεί της προηγούμενης έκδοσης του Knotify όσον αφορά την ανάκληση, το F1-score και το MCC, και μειώνει επίσης την απόσταση από το Knotty, το οποίο εξακολουθεί να έχει καλύτερη απόδοση στις μετρικές αυτές στο σύνολο των δεδομένων. Επιπλέον, όσον αφορά τη μετρική της ακρίβειας, το Knotify+ διατήρησε την καλύτερη απόδοση από το Knotty, όπως και το Knotify. Το Knotify+ κατάφερε να επιτύχει έναν υψηλότερο αριθμό tp από το Knotify, ένα γεγονός που αποδεικνύει τη βελτίωση στην προβλεπτική ικανότητα του συστήματος.

Platform	**tp**	**tn**	**fp**	**fn**	**Precision**	**Recall**	**F1-score**	**MCC**
IPknot	3850	3746	1488	1606	0.721	0.706	0.713	0.421
Knotty	5006	3331	1836	517	0.732	0.906	0.810	0.574
Knotify	4170	4061	1154	1305	0.783	0.762	0.772	0.540
Knotify+	4342	3975	1306	1053	0.769	0.805	0.786	0.558

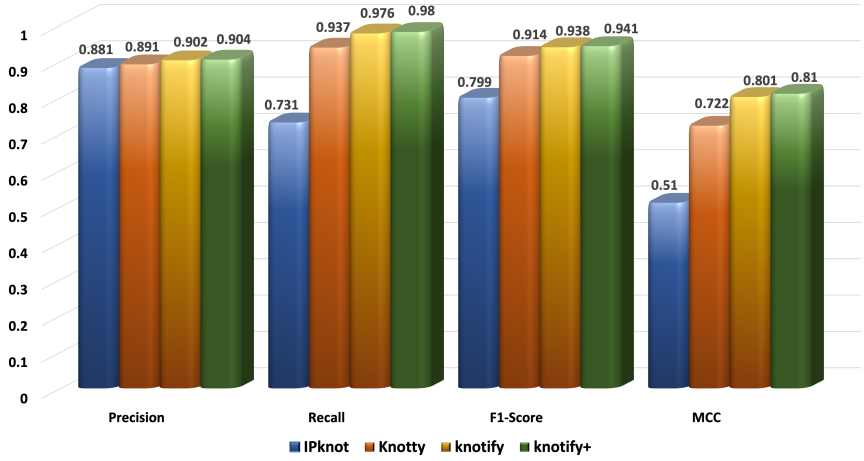
Πίνακας 7.3: Ο Πίνακας Σύγχυσης για κάθε πλατφόρμα στο σύνολο των δεδομένων.

Από την άλλη πλευρά, η προσπάθειά του να προσθέσει μοτίβα εξογκωμάτων και εσωτερικών βρόχων, αυξάνει τον αριθμό των fp και, συνεπώς, μειώνει την ακρίβεια. Παρά τη μείωση αυτή στην ακρίβεια, το F1-score, ο αρμονικός μέσος της ακρίβειας και της ανάκλησης, η οποία αποτελεί τελικά και τη μετρική που περιγράφει το συνολικό ποσοστό πρόβλεψης, ήταν υψηλότερο στο Knotify+ από ότι στο Knotify. Τέλος, το IPknot παρουσίασε τη χαμηλότερη απόδοση σε όλες τις μετρικές αξιολόγησης, όπως και προηγουμένως.
Στον Πίνακα 7.4, παρουσιάζονται οι πίνακες σύγχυσης, για τα τέσσερα σύνολα για κάθε πλατφόρμα, παρέχοντας λεπτομερώς τις τιμές tp, tn, fp και fn. Το Knotify+ κατέγραψε περισσότερα tp και ίσα ή λιγότερα fp και fn για τις ακολουθίες που είναι μικρότερες από 40 ($L < 30$ και $30 \leq L < 40$) σε σύγκριση με τις αξιολογούμενες πλατφόρμες. Η ικανότητά του για πρόβλεψη σε ακολουθίες μεγαλύτερες από 40 ($40 \leq L < 50$ και $L > 50$) ήταν καλύτερη από αυτή του Knotify και συγκρίσιμη, αν και εξακολουθούσε να είναι χαμηλότερη από αυτή του Knotty, το οποίο αυξάνει την ικανότητά του για πρόβλεψη όταν το μήκος αυξάνεται.

Length	L < 30				30 ≤ L < 40				40 ≤ L < 50				L ≥ 50			
Platform	**tp**	**tn**	**fp**	**fn**	**tp**	**tn**	**fp**	**fn**	**tp**	**tn**	**fp**	**fn**	**tp**	**tn**	**fp**	**fn**
IPknot	916	514	124	337	824	810	294	355	754	897	396	284	1368	1519	674	631
Knotty	1196	469	146	80	1064	786	316	117	894	803	510	124	1848	1264	876	204
Knotify	1230	490	132	39	748	991	288	304	748	991	288	304	1218	1723	420	831
Knotify+	1248	486	132	25	1010	847	316	110	798	1004	328	242	1286	1638	530	676

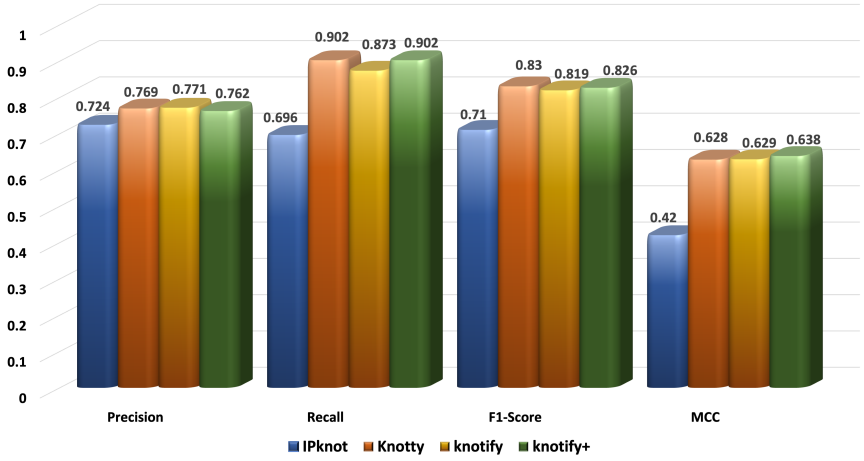
Πίνακας 7.4: Ο Πίνακας Σύγχυσης για κάθε πλατφόρμα ανά σύνολο δεδομένων.

Τα Σχήματα 7.3 - 7.6 παρουσιάζουν τα αποτελέσματα για κάθε μετρική ανά σύνολο δεδομένων. Αξιολογώντας αυτά τα σχήματα, το Knotify+ υπερισχύει όλων των μεθόδων σε όλες τις μετρικές, όταν το μήκος ήταν μικρότερο από 30. Στις ακολουθίες μεταξύ 30 και 40, ήταν ακόμα πιο αποτελεσματικό από το Knotify όσον αφορά το F1-score και το MCC λόγω της υψηλής ανάκλησης του και του συγκρίσιμου ποσοστού ακρίβειας. Στις ακολουθίες μεταξύ 40 και 50, το Knotify+ υπερτερούσε του Knotify σε όλες τις μετρικές και ήταν ισοδύναμο με το Knotty όσον αφορά το F1-score και το MCC.
Τέλος, για τις ακολουθίες μεγαλύτερες από 50, το Knotty ξεπερνούσε τις άλλες



Σχήμα 7.3: Μετρικές για ακολουθίες μήκους < 30.

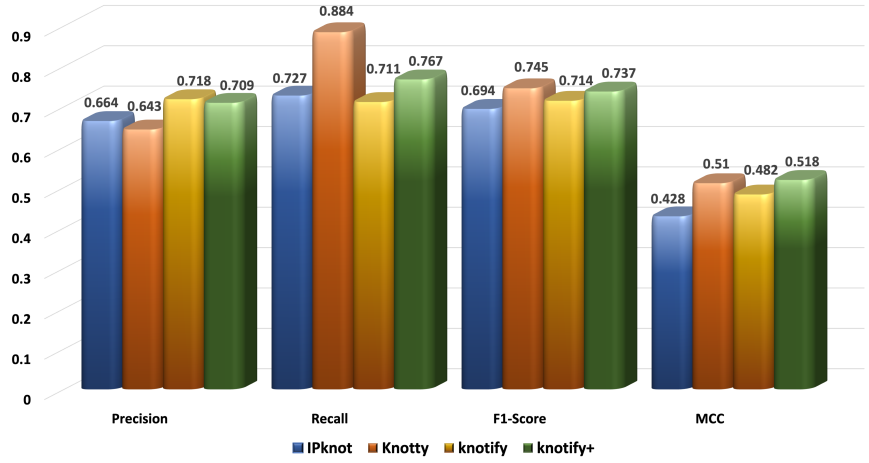
μεθόδους. Ο κύριος λόγος για αυτήν την υπεροχή είναι ότι καθώς το μήκος της ακολουθίας αυξάνεται, υπάρχουν περισσότερα μοτίβα εκτός από τους ψευδοκόμβους, όπως hairpins, multiloops κ.α., τα οποία το Knotify+ δεν μπορεί να προβλέψει σε αυτήν του την έκδοση. Αυτές οι δομές μπορούν να εντοπιστούν από το Knotty, αυξάνοντας τον αριθμό tp, οδηγώντας σε υψηλότερες μετρικές ανάκλησης και F1-score.



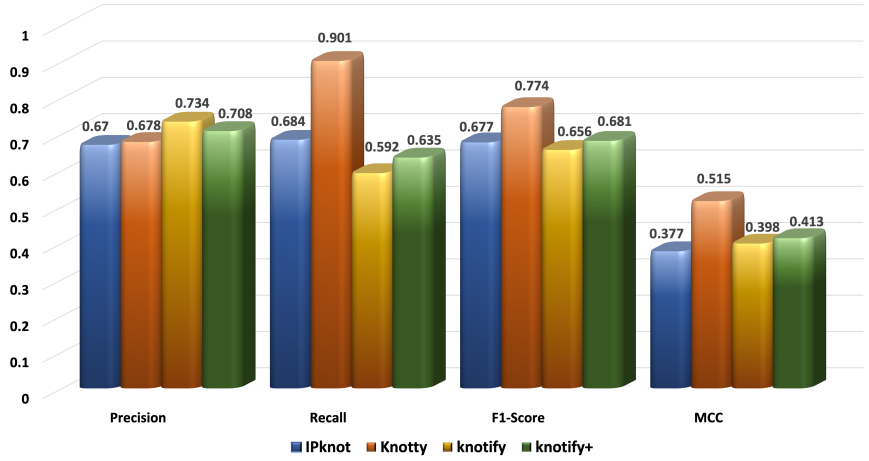
Σχήμα 7.4: Μετρικές για ακολουθίες μήκους ≥ 30 και < 40.

7.3.5 Knotify+: Σύγκριση του χρόνου εκτέλεσης

Η τρίτη μετρική αξιολόγησης είναι ο χρόνος εκτέλεσης, όπου συγκρίνονται τα αποτελέσματα της προτεινόμενης μεθόδου με τις υπόλοιπες πλατφόρμες. Ο Πίνακας



Σχήμα 7.5: Μετρικές για ακολουθίες μήκους ≥ 40 και < 50.



Σχήμα 7.6: Μετρικές για ακολουθίες μήκους ≥ 50.

7.5 παρουσιάζει τον χρόνο εκτέλεσης που απαιτείται από την κάθε πλατφόρμα για την πρόβλεψη της δευτεροταγούς δομής.
Η δεύτερη στήλη του Πίνακα 7.5 παρουσιάζει τον χρόνο εκτέλεσης ανά πλατφόρμα για το σύνολο των δεδομένων. Το Knotify+ χρειάστηκε 74,05 δευτερόλεπτα, το IPknot χρειάστηκε 117.02 δευτερόλεπτα. Το Knotty χρειάστηκε 582.91 δευτερόλεπτα. Μόνο το Knotify, είναι ταχύτερο, γεγονός που αναμενόταν αφού προσθέθηκε στην προτεινόμενη μεθοδολογία μία ακόμη αλγοριθμική διαδικασία για την έρευση των εσωτερικών και των ασσύμετρων βρόχων μέσα στους δύο βρόχους του ψευδοκόμβου, με την ταυτόχρονη όμως αύξηση στην ακρίβεια των προβλέψεων. Παρόλη βέβαια, αυτή την μικρή αύξηση στο χρόνο εκτέλεσης σε σχέση με το Knotify, το Knotify+ ήταν περίπου οκτώ φορές ($582.9/74.05 = 7.87$) πιο γρήγορο από το Knotty, υπερτερώντας

Platform	Total Time (sec)	Average Time (sec)
IPknot	117.02	0.45
Knotty	582.91	2.24
Knotify	56.43	0.22
Knotify+	74.05	0.28

Πίνακας 7.5: Ο απαιτούμενος χρόνος εκτέλεσης για κάθε πλατφόρμα στο σύνολο των δεδομένων.

σημαντικά από τις υπόλοιπες πλατφόρμες. Η τρίτη στήλη παρουσιάζει τον μέσο χρόνο εκτέλεσης για κάθε πλατφόρμα, στοιχεία που δείχνουν τη μικρότερη τάξη μεγέθους στη μετρική του χρόνου εκτέλεσης για την προτεινόμενη μεθοδολογία.

7.4 Συμπεράσματα

Αξιολογώντας τη συνολική απόδοση του Knotify+, η ενσωμάτωση στους ψευδοκόμβους τύπου H ασύμμετρους και εσωτερικούς βρόχους οδήγησε σε μια αποτελεσματική πρόβλεψη της δευτεροταγούς δομής με μια ακρίβεια συγκρίσιμη με τις γνωστές πλατφόρμες της βιβλιογραφίας. Ειδικά για ακολουθίες που είναι μικρότερες από 30 βάσεις, ξεπέρασε όλες τις εξεταζόμενες μεθόδους, δείχνοντας ότι η ενίσχυση της εκφραστικότητάς του οδήγησε σε μια σημαντική πρόοδο σε σχέση με την προηγούμενη έκδοσή του. Το πιο σημαντικό εύρημα ήταν ότι η προτεινόμενη μεθοδολογία υπερτερούσε της προηγούμενης έκδοσης του Knotify όσον αφορά την ανάκληση, το F1-score και το MCC σε όλα τα σύνολα, δείχνοντας σημαντική βελτίωση για ακολουθίες μεγαλύτερες από 40. Επιπλέον, το Knotify+ συνέχισε να ξεπερνά το Knotty για μικρές ακολουθίες, ενώ ήταν συγκρίσιμο για ακολουθίες μεταξύ 30 και 50, και μειώνει σημαντικά τη διαφορά στην ακρίβεια από το Knotty για ακολουθίες με μήκος μεγαλύτερο από 50. Ταυτόχρονα, το Knotify+ διατήρησε το υψηλότερο ποσοστό πρόβλεψης των κεντρικών βάσεων σε σύγκριση με όλες τις εξεταζόμενες μεθόδους και ήταν περίπου οκτώ φορές ταχύτερο από το Knotty, το οποίο είναι ο κύριος ανταγωνιστής του.

Κεφάλαιο 8

Μια επέκταση του Knotify για ψευδοκόμβους τύπου L

8.1 Προτεινόμενη CFG για την ανίχνευση ψευδοκόμβων τύπου L

Αυτό το Κεφάλαιο παρουσιάζει μια επέκταση του Knotify που σχεδιάστηκε για να προβλέπει ψευδοκόμβους τύπου L σε ακολουθίες RNA. Εισάγεται μια νέα CFG γραμματική, η οποία ενσωματώνεται στο πρώτο στάδιο της πλατφόρμας Knotify, επιτρέποντας την πρόβλεψη των ψευδοκόμβων τύπου L. Τα επόμενα δύο βήματα για την τελική πρόβλεψη των ψευδοκόμβων ακολουθούν τις ίδιες αρχές με το Knotify, αλλά τροποποιούνται κατάλληλα για να ταιριάζουν με τον τύπο L. Όπως και στο Knotify, σε αυτό το κεφάλαιο χρησιμοποιείται η $G_{Lpseudo}$ γραμματική που παρουσιάζεται στον Πίνακα 8.1 για την εργασία της πρόβλεψης των ψευδοκόμβων τύπου L. Σε αυτό το Κεφάλαιο, παρόμοια με τις προηγούμενες εκδόσεις του Knotify και του Knotify+, χρησιμοποιείται μια ελαφρώς τροποποιημένη υβριδική μέθοδος επιλογής βέλτιστου δέντρου, η οποία συνδυάζει τις δύο πιο κοινές αρχές, της ελάχιστης ελεύθερης ενέργειας (MFE) και του μέγιστου ταιριάσματος βάσεων (maximum base pairing).

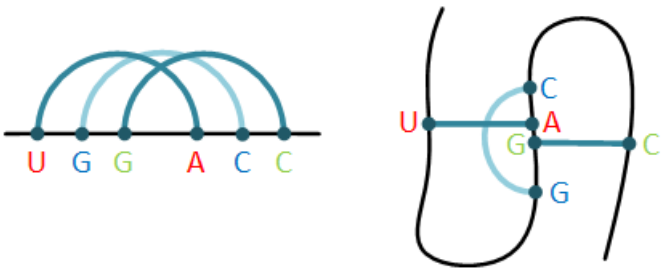
Στη γραμματική $G_{Lpseudo}$ που παρουσιάζεται στον Πίνακα 8.1, η δεύτερη στήλη περιέχει τους συντακτικούς κανόνες. Η γραμματική αποτελείται από τέσσερα μη τερματικά σύμβολα, δηλαδή NT = {S, L, D, K}, με το S ως το σύμβολο εκκίνησης. Όλοι οι συντακτικοί κανόνες με το S στο αριστερό μέρος (δηλαδή κανόνες 0 έως 63) στοχεύουν στον εντοπισμό ενός πιθανού ψευδοκόμβου τύπου L στην είσοδο. Ένας ψευδόκομβος τύπου L ορίζεται από τουλάχιστον τρία ζεύγη κεντρικών βάσεων, όπως φαίνεται στο Σχήμα 8.1.

Για τη συντακτική αναγνώριση ενός ψευδοκόμβου τύπου L, πρέπει να χρησιμοποιηθεί μια κατάλληλη συντακτική γραμματική για να αναλύσει τη γλωσσική αναπαράσταση της ακολουθίας RNA. Η CFG $G_{Lpseudo}$ σχεδιάστηκε για να περιγράψει αποτελεσμα-

Enumeration	**Syntactic Rules**
0	S → "A" L "A" L "A" D "U" L "U" L "U"
1	S → "A" L "A" L "U" D "U" L "U" L "A"
2	S → "A" L "A" L "G" D "U" L "U" L "C"
3	S → "A" L "A" L "C" D "U" L "U" L "G"
4	S → "A" L "U" L "A" D "U" L "A" L "U"
5	S → "A" L "U" L "U" D "U" L "A" L "A"
6	S → "A" L "U" L "G" D "U" L "A" L "C"
7	S → "A" L "U" L "C" D "U" L "A" L "G"
8	S → "A" L "G" L "A" D "U" L "C" L "U"
9	S → "A" L "G" L "U" D "U" L "C" L "A"
10	S → "A" L "G" L "G" D "U" L "C" L "C"
11	S → "A" L "G" L "C" D "U" L "C" L "G"
	.
	.
	.
60	S → "C" L "A" L "C" D "G" L "G" L "U"
61	S → "C" L "U" L "C" D "G" L "G" L "A"
62	S → "C" L "G" L "C" D "G" L "G" L "C"
63	S → "C" L "C" L "C" D "G" L "G" L "G"
64	L → "A" L
65	L → "U" L
66	L → "C" L
67	L → "G" L
68	L → "A"
69	L → "U"
70	L → "C"
71	L → "G"
72	D → K K
73	K → "A"
74	K → "U"
75	K → "C"
76	K → "G"
77	K → ϵ

Πίνακας 8.1: Συνταχτικοί κανόνες $G_{Lpseudo}$.

τικά τη σύνταξη ενός ψευδοκόμβου τύπου L σε μια ακολουθία RNA, με τις κεντρικές βάσεις να σχηματίζουν τα εναλλασσόμενα αντιπαράθεσης ζευγάρια βάσεων. Για παράδειγμα, ο κανόνας S → "U" L "G" L "G" D "A" L "C" L "C" υποδηλώνει την ύπαρξη ενός ψευδοκόμβου τύπου L, όπως U..G..G..A..C..C, όπου τα U–A, G–C και G–C εναλλάσσονται, με τα χρώματα να υποδεικνύουν τα ζευγάρια βάσεων. Αυτά τα εναλλασσόμενα ζευγάρια βάσεων ονομάζονται **κεντρικές βάσεις** για τους ψευδοκόμβους τύπου L και αντιπροσωπεύονται από τα σύμβολα ′(′, ′)′, ′[′, ′]′, ′{′ και ′}′ στην αναπαράσταση dot-bracket (Πίνακες 8.2 και 8.3 στις επόμενες ενότητες). Εφόσον υπάρχουν τέσσερα δυνατά ζευγάρια βάσεων (A–U, U–A, C–G, G–C) και τρεις δυνατές εναλλαγές, υπάρχουν 64 (4^3) συντακτικοί κανόνες με S στο αριστερό μέρος, με τους κανόνες 12 έως 59 να παραλείπονται από τον Πίνακα 8.1, καθώς είναι εύκολα κατανοητοί. Το Σχήμα 8.1, όπως αναφέρθηκε, παρέχει ένα παράδειγμα του τρόπου με τον οποίο οι κεντρικές βάσεις εναλλάσσονται μεταξύ τους για να δημιουργήσουν έναν ψευδόκομβο τύπου L.



Σχήμα 8.1: Ο κανόνας S → "U" L "G" L "G" D "A" L "C" L "C" που εντοπίζει την ύπαρξη ενός ψευδοκόμβου τύπου L

8.2 Decoration των κεντρικών βάσεων για ψευδόκομβους τύπου L

Στην ενότητα αυτή παρουσιάζεται ένα παράδειγμα της διαδικασίας διακόσμησης των κεντρικών βάσεων του ψευδοκόμβου, με σκοπό την εύρεση της δευτεροταγούς δομής της ακολουθίας. Αφού δημιουργηθούν τα συντακτικά δέντρα, όπως περιγράφηκε στην ενότητα 8.1, ο ψευδόκομβος διακοσμείται με επιπλέον ζευγάρια βάσεων εξερευνώντας όλα τα δημιουργημένα δέντρα. Ο αναλυτής εξετάζει σειριακά κάθε βάση εντός των βρόχων του ψευδοκόμβου για να καθορίσει εάν μπορεί να σχηματίσει ζευγάρι βάσης με άλλη βάση που βρίσκεται στη σωστή θέση. Ο αλγόριθμος του **decoration** παρουσιάζεται αναλυτικά στον Πίνακα 8.2. Αφού εντοπίσει τις κύριες βάσεις στις θέσεις 2-10, 9-16 και 5-14 για U–A, G–C και G–C αντίστοιχα, εξετάζονται οι βάσεις στο βρόχο στις θέσεις 11-13 και στον βρόχο στις θέσεις 6-8 για πιθανή δημιουργία ζεύγους με βάσεις εκτός των βρόχων (θέσεις 0-1 και 17-19 αντίστοιχα). Στη συνέχεια, εξετάζονται οι βάσεις στον εσωτερικό βρόχο στις θέσεις 3-4 και στον

εσωτερικό βρόχο στη θέση 15 για πιθανή δημιουργία ζεύγους βάσεων. Με τη σειρά, αναγνωρίζονται ζευγάρια βάσης στις θέσεις 1–11 (βήμα 1), 8–17 και 7–18 (βήμα 2), καθώς και 4–15 (βήμα 3).

String enumeration	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19
String	U	U	U	A	C	G	G	C	C	G	A	A	U	U	C	G	C	G	G	A
Parser output:	.	.	(	.	.	[	.	.	.	{	)	.	.	.	]	.	}	.	.	.
Step 1	.	(	(	.	.	[	.	.	.	{	)	)	.	.	]	.	}	.	.	.
Step 2	.	(	(	.	.	[	.	{	{	{	)	)	.	.	]	.	}	}	}	.
Step 3	.	(	(	.	[	[	.	{	{	{	)	)	.	.	]	]	}	}	}	.

Πίνακας 8.2: Εντοπισμός ζευγών βάσεων γύρω από τις κεντρικές βάσεις του ψευδοκόμβου.

Εφαρμόζοντας στη συνέχεια, τα κριτήρια της ελάχιστης ελευθέρης ενέργειας και του μεγαλύτερου αριθμού ζευγών βάσεων γύρω από τον ψευδοκόμβο, επιλέγεται η βέλτιστη δομή. Ο αλγόριθμος της ελεύθερης ενέργειας που χρησιμοποιήθηκε είναι αυτός που προτείνεται για το συγκεκριμένο τύπο ψευδοκόμβου, ενώ για το κριτήριο του μέγιστου αριθμού ζευγών έγιναν οι κατάλληλες τροποποιήσεις, ώστε να συνυπολογιζονται στην απόφαση όλα τα ζεύγη βάσεων γύρω από τα τρία πλέον, ζεύγη κεντρικών βάσεων του ψευδοκόμβου.
Επίσης, η συγκεκριμένη υλοποίηση του Knotify επιτρέπει στον χρήστη να προχωρήσει στην επιλογή των ζευγών βάσεων U-G ως παράμετρο από τη γραμμή εντολών, επιλογή η οποία εκκινεί μία διαδικασία δυναμικής υλοποίησης μια νέας γραμματικής με επιπλέον κανόνες. Συγκεκριμένα επιτρέπονται πλέον άλλα δύο ζευγάρια U–G και G–U, με αποτέλεσμα οι κανόνες της νέας γραμματικής να γίνονται 6^3, δηλαδή 216.

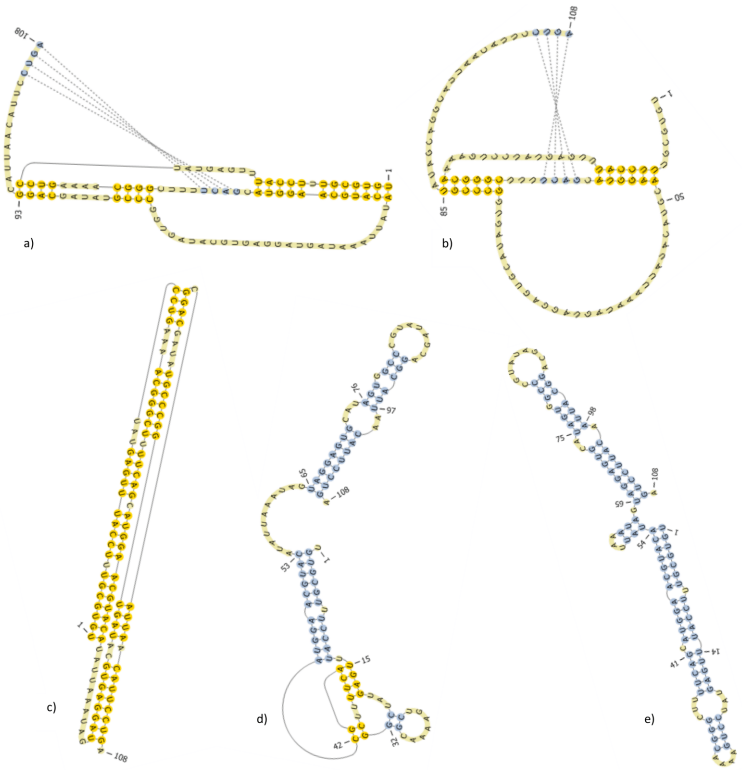
8.3 Πρόβλεψη ενός γνωστού ψευδοκόμβου τύπου L

Για να επιβεβαιώσουμε την αποτελεσματικότητα της νέας γραμματικής και του συνολικού τροποποιημένου προτεινόμενου συστήματος, είναι κοινή πρακτική να κατασκευάζουμε ένα κατάλληλο σύνολο δεδομένων και να συγκρίνουμε τις προβλεπόμενες δευτεροταγείς δομές, διαδικασία που πραγματοποιήθηκε και στα προηγούμενα Κεφάλαια. Δυστυχώς, στην περίπτωση των ψευδοκόμβων τύπου L, το διαθέσιμο σύνολο δεδομένων είναι εξαιρετικά περιορισμένο. Μέχρι σήμερα, έχει παρατηρηθεί μόνο ένας ψευδόκομβος τύπου L, ο οποίος παρουσιάστηκε αρχικά στην εργασία [87]. Αυτή η ακολουθία RNA χρησιμοποιήθηκε ως είσοδος στην προτεινόμενη πλατφόρμα, καθώς και σε τρεις άλλες προηγμένες μεθόδους, IPknot [35], Probknot [34] και Knotty [32]. Οι αντίστοιχες αναπαραστάσεις με τη μορφή dot-bracket παρουσιάζονται στον Πίνακα 8.3, δείχνοντας ότι η πλατφόρμα μας, που αναφέρεται ως "Knotify", προέβλεψε με ακρίβεια τις κεντρικές βάσεις του ψευδοκόμβου τύπου L. Αντίθετα, οι Knotty και

IPknot προέβλεψαν ψευδοκόμβους τύπου H, ενώ το Probknot προέβλεψε μια δομή με βρόχους, αποτυγχάνοντας να προβλέψει τις κεντρικές βάσεις του ψευδοκόμβου τύπου L.

Platform	**RNA/ Dot Bracket**
	UGUGCGUUUCCAUUUGAGUAUCCUGAAAACGGGCUUUUCAGCAUGGAACGUACAUAUUAAAUAGUAGGAGUGCAUAGUGGCCCGUAUAGCAGGCAUUAACAUUCCUGA
Ground truth	(((((((.(((((........[[[[....[[[[....{{{{.)))))))))))).....]]]].....]]]]..........}}}}
Knotify	.......((((((..............[[[[[[...{{{{.))))))..............................]]]]]]..................}}}}
IPknot	.((((((.(((((.[[[[...((.......))((..]]]])))))))))))))..........(((((((...(((.(((........))))))..))))))))).
Knotty	(((((((.((((((((((...[[[[...[[[[[[[.)))))).))))))))))).........(((((((..((((]]]]]]]....]]]].)))))..)))))))).
Probknot	(((((((.((((((((((...((((....))))...)))))).))))))))))))(((...))).(((((((..((((.(((........))))))).)))))))).

Πίνακας 8.3: Πρόβλεψη του ψευδοκόμβου τύπου L



Σχήμα 8.2: Ground truth (a), η πρόβλεψη της προτεινόμενης πλατφόρμας (b), η πρόβλεψη της πλατφόρμας Knotty (c), η πρόβλεψη της πλατφόρμας IPknot (d), και η πρόβλεψη της πλατφόρμας Probknot (e) του ψευδοκόμβου που παρουσιαστηκε στο [87].

Το Σχήμα 8.2 παρουσιάζει την πραγματική δομή της ακολουθίας στο υποσχήμα (α), την πρόβλεψη της πλατφόρμας μας στο υποσχήμα (β), και τις προβλέψεις των Knotty, IPknot και Probknot στα υποσχήματα (γ), (δ) και (ε) αντίστοιχα.

Όπως προκύπτει, η πρόβλεψη του προτεινόμενου συστήματος προσεγγίζει σημαντικά την πραγματική δομή του ψευδοκόμβου, σε αντίθεση με τα άλλα συστήματα που προβλέπουν τελείως διαφορετικές δομές όπως ψευδοκόμβους τύπου H ή απλούς βρόχους. Η διαδικασία οπτικοποίησης των δομών ψευδοκόμβων RNA που φαίνονται στο Σχήμα 8.2 πραγματοποιήθηκε μέσω της χρήσης του εργαλείου pseudoviewer [88].
Ο Πίνακας 8.4 παρουσιάζει τα αποτελέσματα για κάθε μέθοδο αναφορικά με την ακρίβεια, την ανάκληση, το F1-score, και το MCC. Επίσης, παρέχονται λεπτομερώς οι τιμές tp, tn, fp και fn για όλες τις μεθόδους.

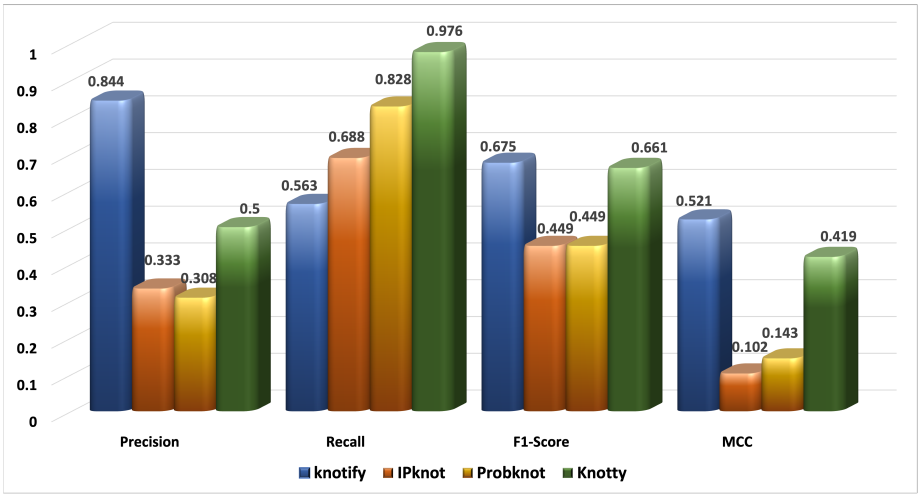
Platform	**tp**	**tn**	**fp**	**fn**	**Precision**	**Recall**	**F1-score**	**MCC**
Knotify	27	55	5	21	0.844	0.563	0.675	0.521
IPknot	22	32	44	10	0.333	0.688	0.449	0.102
Probknot	24	25	54	5	0.308	0.828	0.449	0.143
Knotty	40	27	40	1	0.500	0.976	0.661	0.419

Πίνακας 8.4: Οι βασικές μετρικές αξιολόγησης των υπό αξιολόγηση πλατφορμών.

Η προτεινόμενη μεθοδολογία Knotify L-τύπου ξεπέρασε τις άλλες μεθόδους όσον αφορά την ακρίβεια, με τιμή ίση με 0.844, ενώ το Knotty είχε 0.500, το IPknot 0.333 και το Probknot τιμή ίση με 0.308. Όσον αφορά το F1-score, η πλατφόρμα Knotify πέτυχε ένα σκορ 0.675, ενώ το Knotty 0.661, το οποίο ήταν πολύ κοντά στην προτεινόμενη πλατφόρμα και οι άλλες δύο πλατφόρμες είχαν τιμή ίση με 0.449. Επίσης, ξεπέρασε όλες τις μεθόδους όσον αφορά το MCC, με τιμή ίση με 0.521. Τέλος, το Knotty πέτυχε μία πολύ υψηλή τιμή ανάκλησης (0.976) έχοντας μόνο ένα ψευδώς αρνητικό αποτέλεσμα, το οποίο μπορεί να οφείλεται στο γεγονός ότι μεγάλες RNA ακολουθίες συχνά περιέχουν πολλές δομές (όπως εξογκώματα και άλλους βρόχους) που δεν σχετίζονται απαραίτητα άμεσα με τον ψευδόκομβο. Αυτό μπορεί να αυξήσει το συνολικό αριθμό των σωστών θετικών και να μειώσει τον αριθμό των ψευδώς αρνητικών. Στη μελλοντική επέκταση του συστήματος, σχεδιάζεται η βελτίωση ακόμη περισσότερο της μεθοδολογίας για να αντιμετωπίζει ακόμη πιο πολύπλοκα μοτίβα, όπως οι ψευδοκόμβοι L-τύπου που περιέχουν ασύμμετρους βρόχους (bulges), εσωτερικούς βρόχους ή hairpins, στα πρότυπα του Knotify+. Δυστυχώς, με βάση όσα στοιχεία γνωρίζουμε από τη βιβλιογραφία, κανένας άλλος ψευδοκόμβος τύπου L δεν έχει καταγραφεί, ώστε να τον συμπεριλάβουμε στο σύνολο δεδομένων προς αξιολόγηση. Τα παραπάνω αποτελέσματα παρουσιάζονται επίσης στο Σχήμα 8.3 για καλύτερη αντιπαραβολή των στοιχείων μέσω της οπτικοποίησης.

8.4 Συμπεράσματα

Στο παρόν Κεφάλαιο παρουσιάστηκε μια νέα μέθοδος για τον εντοπισμό των ψευδοκόμβων τύπου L, έναν σπάνιο τύπο ψευδοκόμβων. Η προτεινόμενη υβριδική στρατη-



Σχήμα 8.3: Precision, Recall, F1-score και MCC ανά πλατφόρμα.

γική επιλέγει την ακολουθία RNA με την πιο πιθανή έκφραση ψευδόκομβου. Αυτό περιλαμβάνει τη δημιουργία ενός συνόλου δομών ψευδοκόμβων τύπου L και στη συνέχεια την επίλυση ενός προβλήματος βελτιστοποίησης για την επιλογή του ψευδόκομβου με τον μέγιστο αριθμό ζευγών βάσεων γύρω από τον ψευδόκομβο και την ελάχιστη ελεύθερη ενέργεια. Οι ρουτίνες κώδικα υλοποιήθηκαν σε γλώσσα C, από Python που χρησιμοποιήθηκε στην αρχική έκδοση του Knotify, για περαιτέρω βελτιστοποίηση του χρόνου εκτέλεσης κατά τη εργασία της αρχικής ανάλυσης της ακολουθίας. Η ακολουθία εισόδου διαιρείται σε πολλαπλές υποακολουθίες για παραλληλοποίηση του φόρτου εργασίας και χρησιμοποιείται παράλληλη ανάλυση με πολλαπλούς CFGs για την αξιολόγηση όλων των υποακολουθιών. Κάθε περίπτωση ανάλυσης δημιουργεί μια δομή ψευδοκόμβου που περιγράφει πιθανούς ψευδόκομβους εντός του πεδίου της CFG. Όλοι οι ψευδοκόμβοι αποθηκεύονται σε μια δομή δεδομένων, και η πλέον πιθανή λύση επιλέγεται με βάση την ελάχιστη ελεύθερη ενέργεια. Το Knotify για τον L-τύπο ψευδοκόμβου ξεπέρασε τις άλλες τρεις μεθόδους όσον αφορά την ακρίβεια, με τιμή 0.844, ενώ οι άλλες μεθόδοι σημείωσαν 0.500, 0.333 και 0.308. Όσον αφορά το F1-score, η προτεινόμενη μεθοδολογία σημείωσε επίδοση ίση με 0.671, ενώ οι άλλες μεθόδοι 0.661, 0.449 και 0.449 αντίστοιχα. Η προτεινόμενη μεθοδολογία ξεπέρασε όλες τις μεθόδους όσον αφορά τη μετρική MCC, επιτυγχάνοντας τιμή ίση με 0.521, ενώ η πλατφόρμα Knotty πέτυχε ένα πολύ υψηλό σκορ ανάκλησης σε σχέση με τις υπόλοιπες μεθόδους. Όλα αυτά τα ευρήματα έδειξαν ότι ο Knotify για τον L-τύπο αποτελεί ένα χρήσιμο εργαλείο που παρέχει ελπιδοφόρα αποτελέσματα και πρέπει να εξεταστεί περαιτέρω προκειμένου να ανιχνεύσει πιο πολύπλοκα συνδυαστικά μοτίβα.

Κεφάλαιο 9

Συμπεράσματα και Μελλοντικές Επεκτάσεις

Ο ακριβής προσδιορισμός των δευτεροταγών δομών του RNA, συμπεριλαμβανομένων των ψευδοκόμβων, αποτελεί μια πρόκληση με σημαντικά αποτελέσματα στον τομέα της γενετικής και της υγείας. Σε αυτήν τη διατριβή, αναπτύχθηκαν τρεις καινοτόμες μεθοδολογίες, οι Knotify, Knotify+ και Knotify για L-τύπο, για τον εντοπισμό και την πρόβλεψη διαφόρων τύπων ψευδοκόμβων στο RNA. Αυτά τα συστήματα αξιολογήθηκαν βάσει διαφόρων μετρικών, συμπεριλαμβανομένων του ποσοστού πρόβλεψης των κεντρικών βάσεων ψευδοκόμβων, του χρόνου εκτέλεσης, της ακρίβειας, της ανάκλησης, του F1-score και του συντελεστή συσχέτισης Matthews (MCC).
Για την πρόβλεψη των κεντρικών βάσεων, το Knotify παρουσίασε μία υψηλή προβλεπτική ικανότητα, ανιχνεύοντας με επιτυχία τις κεντρικές βάσεις σε 143 από τις 262 ακολουθίες. Ξεπέρασε άλλες ευρέως χρησιμοποιούμενες μεθόδους, όπως τα συστήματα Knotty, HotKnots, IPknot και IHFold. Ο Knotify έδειξε υψηλότερη ακρίβεια στην πρόβλεψη των κετρικών βάσεων των ψευδοκόμβων σε τρία από τα τέσσερα σύνολα ακολουθιών RNA, με συγκρίσιμη απόδοση με τον Knotty στο υπόλοιπο σύνολο. Όσον αφορά την πρόβλεψη της συνολικής δομής, το Knotify παρουσιάζει εξαιρετική απόδοση σε όρους ακρίβειας, επιτυγχάνοντας τιμή ίση με 0.784. Από την άλλη μεριά, η πλατφόρμα Knotty είχε υψηλότερο F1-score και MCC, αλλά το Knotify είναι πολύ κοντά στις παραπάνω επιδόσεις. Το εντυπωσιακό είναι ότι ο Knotify είχε καλύτερη ακρίβεια σε όλα τα εύρη του μήκους ακολουθίας RNA, ενώ ο Knotty ξεχώριζε κυρίως σε μεγαλύτερες ακολουθίες RNA. Αυτό υποδηλώνει ότι το Knotify είναι μια πολύ αξιόπιστη μέθοδος για την πρόβλεψη των βάσεων, ιδίως για μικρότερες ακολουθίες RNA. Αναφορικά με τον χρόνο εκτέλεσης, το Knotify αποδείχθηκε αποτελεσματικό, ξεπερνώντας τον Knotty κατά 7.76 φορές. Επίσης, επέδειξε συγκρίσιμους ή καλύτερους χρόνους εκτέλεσης από το IPknot και το HotKnots, από τους οποίους βέβαια είχε πολύ καλύτερες προβλέψεις, και έτσι το Knotify αναδειχθηκε ως μια αποδοτική λύση αναφορικά με τον χρόνο εκτέλεσης για την πρόβλεψη των ψευδοκόμβων στις ακολουθίες RNA, σημαντικό επίσης στοιχείο για τις αναλύσεις των βιολόγων.

Συνοψίζοντας τα αποτελέσματα για το Knotify+, τη βελτιωμένη έκδοση του Knotify, επισημαίνουμε ότι παρατηρήθηκαν ακόμα καλύτερα αποτελέσματα, και ειδικά στην πρόβλεψη των κεντρικών βάσεων των ψευδοκόμβων. Το Knotify+ προέβλεψε με επιτυχία τις κεντρικές βάσεις των ψευδοκόμβων σε 142 από τις 260 ακολουθίες, ξεπερνώντας το IPknot και το Knotty όσον αφορά την μετρική αυτή. Επιπλέον, το Knotify+ ξεπέρασε το Knotify, το IPknot και το Knotty στην πρόβλεψη τουλάχιστον μίας κεντρικής βάσης, υποδηλώνοντας τη βελτιωμένη δυνατότητα του να προβλέπει τις κεντρικές βάσεις των ψευδοκόμβων, ακόμα και σε περιπτώσεις όπου η ακριβής πρόβλεψη είναι ιδιαίτερα δύσκολη.

Κατά την εξέταση των μετρικών αξιολόγησης για το Knotify+, αυτό έδειξε ανώτερη απόδοση σε σύγκριση με το Knotify συνολικά. Το Knotify+ πέτυχε υψηλότερη ανάκληση, F1-score και MCC, μειώνοντας τη διαφορά με το Knotty, που παραμένει η μέθοδος με την καλύτερη απόδοση σε αυτές τις μετρικές. Το Knotify+ διατήρησε καλύτερη ακρίβεια από το Knotty, όπως και από το Knotify. Αυτό υποδεικνύει ότι ο Knotify+ βελτιώνει τη συνολική ακρίβεια των προβλέψεων, ιδίως όσον αφορά την ανάκληση, το F1-score και το MCC. Ωστόσο, το Knotty ξεπέρασε όλες τις μεθόδους σε μεγαλύτερες ακολουθίες RNA, πιθανότατα λόγω της ικανότητάς του να ανιχνεύει πρόσθετα περίπλοκα μοτίβα, όπως οι βρόχοι και τα hairpins, που ο Knotify+ δεν ανιχνεύει σε αυτήν την έκδοση. Οι μελλοντικές επεκτάσεις αποσκοπούν στη βελτίωση του Knotify+ για τη διαχείριση πιο πολύπλοκων μοτίβων εντός των βρόχων των ψευδοκόμβων, ώστε να βελτιώσει την απόδοση των προβλέψεών του και σε μεγαλύτερες ακολουθίες.

Τέλος, για την πρόβλεψη των ψευδοκόμβων τύπου L, το Knotify για L-τύπο έδειξε ελπιδοφόρα αποτελέσματα, ξεπέρασε άλλες μεθόδους όσον αφορά την ακρίβεια με ένα σκορ 0.844. Το F1-score και το MCC ήταν υψηλότερα για τον Knotify από τις άλλες μεθόδους, ενώ το Knotty παρόλο που παρουσίασε μία υψηλή τιμή ανάκλησης, δεν έδειξε αντίστοιχα καλές τιμές στις υπόλοιπες μετρικές αξιολόγησης, ούτε στον συνολικό εντοπισμό του μοτίβου. Με βάση τα παραπάνω, προκύπτει ότι το Knotify για τους ψευδοκόμβους τύπου L είναι αποτελεσματικό και ακριβές στην πρόβλεψη αυτού του συγκεκριμένου τύπου ψευδοκόμβου και παρά τον μικρό αριθμό δεδομένων, είναι μία έτοιμη πλατφόρμα για την πρόβλεψη τέτοιων μοτίβων, όταν αυτά εντοπιστούν στη φύση και απαιτηθεί η περαιτέρω ανάλυσή τους.

Συνολικά, τα αποτελέσματα δείχνουν την αποτελεσματικότητα των αναπτυγμένων συστημάτων (Knotify, Knotify+ και Knotify για τους ψευδοκόμβους τύπου L) στην πρόβλεψη των ψευδοκόμβων του RNA. Αυτά τα συστήματα παρέχουν αξιόπιστες προβλέψεις για τις κεντρικές βάσεις (core stems) των ψευδοκόμβων, τη συνολική δομή και την ακρίβεια των ζευγών βάσεων, διατηρώντας παράλληλα εξαιρετικούς χρόνους εκτέλεσης. Περαιτέρω βελτιώσεις μπορούν να πραγματοποιηθούν με στόχο την αυξημένη ακρίβεια πρόβλεψης για μεγαλύτερες ακολουθίες RNA και πιο πολύπλοκα μοτίβα ψευδοκόμβων. Συμπερασματικά, η εισαγωγή των Knotify, Knotify+ και Knotify για τους ψευδοκόμβους τύπου L αποτελεί σημαντικό βήμα προόδου στον τομέα της πρόβλεψης της δευτεροταγούς δομής του RNA και ειδικότερα στον εντοπισμό των ψευδοκόμβων. Αυτές οι μεθοδολογίες δείχνουν βελτιωμένη ακρίβεια, απόδοση

και δυνατότητες σε σύγκριση με τις υπάρχουσες μεθόδους. Η επιτυχημένη υλοποίηση των Knotify, Knotify+ και Knotify για τους ψευδοκόμβους τύπου L ανοίγει νέους δρόμους για μελλοντικές έρευνες, συμπεριλαμβανομένης της εξερεύνησης πιο πολύπλοκων μοτίβων ψευδοκόμβων με π.χ. hairpins και ψευδοκόμβων τύπου K και M. Παράλληλα, είναι σημαντική η ανάπτυξη ενός online εργαλείου με όλες τις διαθέσιμες λειτουργικότητες, για τη βελτίωση της προσβασιμότητας των συστημάτων για τους ερευνητές που ασχολούνται με τον συγκεκριμένο τομέα της δομικής βιολογίας. Η σύγχρονη διαδικτυακή πλατφόρμα, η οποία βρίσκεται σε τελικό στάδιο υλοποίησης, παρέχει με ένα εύχρηστο γραφικό περιβάλλον στους χρήστες ώστε να επιλέξουν τη γραμματική που επιθυμούν, να τροποποιήσουν τις παραμέτους του μεγέθους των ασύμμετρων και εσωτερικών βρόχων, καθώς και άλλα στοιχεία με στόχο την ευρεία χρήση του από την κοινότητα.

Όλα τα συστήματα που υλοποιήθηκαν και οι προτάσεις για μελλοντικές επεκτάσεις αποτελούν ένα εγχείρημα με μεγάλη δυναμική, που συμβάλουν δραστικά στον τομέα της βιοπληροφορικής και της βιολογίας, και προάγουν την ανάπτυξη συνεργασιών και συνεργιών μεταξύ των αντίστοιχων διεπιστημονικών ομάδων, επιστημόνων στον τομέα της υγείας, βιολόγων και ειδικών πληροφορικής.

Παράρτημα Α΄

Συντομογραφίες

AI	Artificial Intelligence
CCJ	Chen–Condon–Jabbari
CFG	Context-Free Grammar
CNN	Convolutional Neural Network
CYK	Cocke–Younger–Kasami
DAG	Directed Acyclic Graph
DNA	Deoxyribonucleic Acid
IBPMP	Improved Base-Pair Maximization Principle
lncRNA	long non-coding RNA
LSTM	Long Short-Term Memory
mRNA	messenger RNA
MCC	Matthews Correlation Coefficient
MFE	Minimum Free Energy
miRNA	microRNA
ncRNA	non-coding RNA
NP	Nondeterministic Polynomial
piRNA	piwi-interacting RNA
pre-mRNA	precursor mRNA
rRNA	ribosomal RNA
RNA	Ribonucleic Acid
scRNA	small conditional RNA
siRNA	small interfering RNA
snoRNA	small nucleolar RNA
srRNA	small regulatory RNA
tRNA	transfer RNA
tsRNA	tRNA-derived small RNA
YAEP	Yet Another Early Parser

Bibliography

[1] Crick, F. Central Dogma of Molecular Biology. *Nature*. **1970**, *227*, 561–563

[2] Wu, L.; Belasco, J. Let Me Count the Ways: Mechanisms of Gene Regulation by miRNAs and siRNAs. *Mol. Cell* **2008**, *29*, 1–7.

[3] Rossi J. Ribozyme diagnostics comes of age. *Chem. Biol.*. **2004** *11*, 894–895

[4] Shi, Y. A Glimpse of Structural Biology through X-Ray Crystallography. *Cell* **2014**, *159*, 995–1014.

[5] Barnwal, R., Yang, F. & Varani, G. Applications of NMR to structure determination of RNAs large and small. Archives Of Biochemistry And Biophysics. **2017**, *628*, 42–56

[6] Makris E, Kolaitis A, Andrikos C, Moulos V, Tsanakas P, Pavlatos C. Knotify+: Toward the Prediction of RNA H-Type Pseudoknots, Including Bulges and Internal Loops. Biomolecules. 2023; 13(2):308. https://doi.org/10.3390/biom13020308

[7] Zuker, M. Calculating nucleic acid secondary structure. *Curr. Opin. Struct. Biol.* **2000**, *10*, 303–10.

[8] Nussinov, R.; Jacobson, A.B. Fast algorithm for predicting the secondary structure of single-stranded RNA. *Proc. Natl. Acad. Sci. USA* **1980**, *77*, 6309–6313.

[9] Wang, L.; Liu, Y.; Zhong, X.; Liu, H.; Lu, C.; Li, C.; Zhang, H. DMfold: A novel method to predict RNA secondary structure with pseudoknots based on deep learning and improved base pair Maximization Principle. *Front. Genet.* **2019**, *10*, 143.

[10] Rastogi, T.; Beattie, T.L.; Olive, J.E.; Collins, R.A. A long-range pseudoknot is required for activity of the Neurospora VS ribozyme. *EMBO J.* **1996**, *15*, 2820–2825.

[11] Ke, A.; Zhou, K.; Ding, F.; Cate, J.H.; Doudna, J.A. A conformational switch controls hepatitis delta virus ribozyme catalysis. *Nature* **2004**, *429*, 201–205.

[12] Adams, P.L.; Stahley, M.R.; Kosek, A.B.; Wang, J.; Strobel, S.A. Crystal structure of a self-splicing group I intron with both exons. *Nature* **2004**, *430*, 45–50.

[13] Theimer, C.A.; Blois, C.A.; Feigon, J. Structure of the human telomerase RNA pseudoknot reveals conserved tertiary interactions essential for function. *Mol. Cell* **2005**, *17*, 671–682.

[14] Shen, L.X.; Tinoco, I., Jr. The structure of an RNA pseudoknot that causes efficient frameshifting in mouse mammary tumor virus. *J. Mol. Biol.* **1995**, *247*, 963–978.

[15] Nixon, P.L.; Rangan, A.; Kim, Y.G.; Rich, A.; Hoffman, D.W.; Hennig, M.; Giedroc, D.P. Solution structure of a luteoviral P1–P2 frameshifting mRNA pseudoknot. *J. Mol. Biol.* **2002**, *322*, 621–633.

[16] Michiels, P.J.; Versleijen, A.A.; Verlaan, P.W.; Pleij, C.W.; Hilbers, C.W.; Heus, H.A. Solution structure of the pseudoknot of SRV-1 RNA, involved in ribosomal frameshifting. *J. Mol. Biol.* **2001**, *310*, 1109–1123.

[17] Staple, D.W.; Butcher, S.E. Pseudoknots: RNA structures with diverse functions. *PLoS Biol.* **2005**, *3*, e213.

[18] Wyatt, J., Puglisi, J. & Tinoco, I. RNA folding: pseudoknots, loops and bulges. *Bioessays*. **1989**, *11*, 100–106

[19] Turner, D. Bulges in nucleic acids. *Current Opinion In Structural Biology*. **1992**, *2*, 334–337

[20] Hermann, T. & Patel, D. RNA bulges as architectural and recognition motifs. *Structure*. **2000**, *8*, R47-R54

[21] Wu, H. & Uhlenbeck, O. Role of a bulged A residue in a specific RNA-protein interaction. *Biochemistry*. **1987**, *26*, 8221–8227

[22] Mathews, D.H. Using an RNA secondary structure partition function to determine confidence in base pairs predicted by free energy minimization. *RNA* **2004**, *10*, 1178–1190.

[23] Woese, C. & Gutell, R. Evidence for several higher order structural elements in ribosomal RNA. *Proc. Natl. Acad. Sci. U. S. A.*. **1989**, *86*, 3119–3122

[24] Andrikos, C., Makris, E., Kolaitis, A., Rassias, G., Pavlatos, C. & Tsanakas, P. Knotify: An Efficient Parallel Platform for RNA Pseudoknot Prediction Using Syntactic Pattern Recognition. *Methods Protoc.* **2022**, *5*, 14

[25] Lorenz, R.; Bernhart, S.; Höner zu Siederdissen, C.; Tafer, H.; Flamm, C.; Stadler, P.; Hofacker, I. ViennaRNA package 2.0. *Algorithms Mol. Biol. AMB* **2011**, *6*, 26. https://doi.org/10.1186/1748-7188-6-26.

[26] Zuker, M. Mfold web server for nucleic acid folding and hybridization prediction. *Nucleic Acids Res.* **2003**, *31*, 3406–15. https://doi.org/10.1093/nar/gkg595.

[27] Cao, S. & Chen, S. Predicting structures and stabilities for H-type pseudoknots with interhelix loops. *RNA (New York, N.Y.).* **15**, 696-706 (2009,4), https://pubmed.ncbi.nlm.nih.gov/19237463

[28] Akutsu, T. Dynamic programming algorithms for RNA secondary structure prediction with pseudoknots. *Discret. Appl. Math.* **2000**, *104*, 45–62.

[29] Meyer, I.M.; Miklos, I. SimulFold: simultaneously inferring RNA structures including pseudoknots, alignments, and trees using a Bayesian MCMC framework. *PLoS Comput. Biol.* **2007**, *3*, 149.

[30] Van Batenburg, F.; Gultyaev, A.P.; Pleij, C.W. An APL-programmed genetic algorithm for the prediction of RNA secondary structure. *J. Theor. Biol.* **1995**, *174*, 269–280.

[31] Isambert, H.; Siggia, E.D. Modeling RNA folding paths with pseudoknots: application to hepatitis delta virus ribozyme. *Proc. Natl. Acad. Sci. USA* **2000**, *97*, 6515–6520.

[32] Jabbari, H.; Wark, I.; Montemagno, C.; Will, S. Knotty: efficient and accurate prediction of complex RNA pseudoknot structures. *Bioinformatics* **2018**, *34*, 3849–3856.

[33] Chen, H. L.; Condon, A.; Jabbari, H. An O(n(5)) algorithm for MFE prediction of kissing hairpins and 4-chains in nucleic acids. *Journal of computational biology* **2009**, *16(6)*, 803–815. https://doi.org/10.1089/cmb.2008.0219.

[34] Bellaousov, S.; Mathews, D.H. ProbKnot: fast prediction of RNA secondary structure including pseudoknots. *RNA* **2010**, *16*, 1870–80.

[35] Sato, K.; Kato, Y.; Hamada, M.; Akutsu, T.; Asai, K. IPknot: fast and accurate prediction of RNA secondary structures with pseudoknots using integer programming. *Bioinformatics* **2011**, *27*, 85–93.

[36] Sato, K. ; Kato, Y. Prediction of RNA secondary structure including pseudoknots for long sequences. *Briefings In Bioinformatics* 2021 **23**

[37] Knudsen, B.; Hein, J. RNA secondary structure prediction using stochastic context-free grammars and evolutionary history. *Bioinformatics* **1999**, *15*, 446–454.

[38] Knudsen, B.; Hein, J. Pfold: RNA secondary structure prediction using stochastic context-free grammars. *Nucleic Acids Res.* **2003**, *31*, 3423–3428. https://doi.org/10.1093/nar/gkg614.

[39] Sukosd, Z.; Knudsen, B.; Vaerum, M.; Kjems, J.; Andersen, E.S. Multithreaded comparative RNA secondary structure prediction using stochastic context-free grammars. *BMC Bioinform.* **2011**, *12*, 103.

[40] Pedersen, J.S.; Meyer, I.M.; Forsberg, R.; Simmonds, P.; Hein, J. A comparative method for finding and folding RNA secondary structures within protein-coding regions. *Nucleic Acids Res.* **2004**, *32*, 4925–4936.

[41] Do, C.B.; Woods, D.A.; Batzoglou, S. CONTRAfold: RNA secondary structure prediction without physics-based models. *Bioinformatics* **2006**, *22*, e90–e98.

[42] Pedersen, J.S.; Bejerano, G.; Siepel, A.; Rosenbloom, K.; Lindblad-Toh, K.; Lander, E.S.; Kent, J.; Miller, W.; Haussler, D. Identification and classification of conserved RNA secondary structures in the human genome. *PLoS Comput. Biol.* **2006**, *2*, e33.

[43] Nawrocki, E.P.; Kolbe, D.L.; Eddy, S.R. Infernal 1.0: inference of RNA alignments. *Bioinformatics* **2009**, *25*, 1335–1337.

[44] Anderson, J.W.; Haas, P.A.; Mathieson, L.A.; Volynkin, V.; Lyngsø, R.; Tataru, P.; Hein, J. Oxfold: kinetic folding of RNA using stochastic context-free grammars and evolutionary information. *Bioinformatics* **2013**, *29*, 704–710.

[45] Singh, J.; Hanson, J.; Paliwal, K.; Zhou, Y. RNA secondary structure prediction using an ensemble of two-dimensional deep neural networks and transfer learning. *Nat. Commun.* **2019**, *10*, 1–13.

[46] Zakov, S., Goldberg, Y., Elhadad, M. & Ziv-Ukelson, M. Rich Parameterization Improves RNA Structure Prediction. *Springer Berlin Heidelberg* **2011**

[47] Wang, Y.; Liu, Y.; Wang, S.; Liu, Z.; Gao, Y.; Zhang, H.; Dong, L. ATTfold: RNA secondary structure prediction with pseudoknots based on attention mechanism. *Front. Genet.* **2020**, *11*, 1564.

[48] Kangkun, M.; Jun, W.; Yi, X. Prediction of RNA secondary structure with pseudoknots using coupled deep neural networks. *Biophys. Rep.* **2020**, *6*, 146–154.

[49] Reuter, J. & Mathews, D. RNAstructure: software for RNA secondary structure prediction and analysis. *BMC Bioinformatics.* **2010** *11*, 129

[50] Bernhart, S.; Hofacker, I.; Will, S.; Gruber, A.; Stadler, P. RNAalifold: Improved Consensus Structure Prediction for RNA Alignments. *BMC BioInform.* **2008**, *9(474).*

[51] Sato, K., Hamada, M., Asai, K. & Mituyama, T. CENTROIDFOLD: a web server for RNA secondary structure prediction. *Nucleic Acids Research.* **2009**, *5*, W277-W280

[52] Available online: https://www.genome.gov/genetics-glossary/RNA-Ribonucleic-Acid

[53] Jabbari, H.; Condon, A. A fast and robust iterative algorithm for prediction of RNA pseudoknotted secondary structures. *MC Bioinform.* **2014**, *15*, 147.

[54] Watson, J.; Crick, F. Molecular Structure Of Nucleic Acids. *Am. J. Psychiatry* **2003**, *160*, 623–624. https://doi.org/10.1176/appi.ajp.160.4.623

[55] Mamuye, A., Merelli, E. & Tesei, L. A Graph Grammar for Modelling RNA Folding. *Electronic Proceedings In Theoretical Computer Science.* **231** pp. 31-41 (2016,12)

[56] Makris, E., Kolaitis, A., Andrikos, C., Moulos, V., Tsanakas, P., Pavlatos, C. An intelligent grammar-based platform for RNA H-type pseudoknot prediction. *Artificial Intelligence Applications and Innovations.* AIAI 2022 IFIP WG 12.5 International Workshops. IFIP Advances in Information and Communication Technology **2022**, *vol 652*, Springer.

[57] Wikipedia contributors Nucleic acid secondary structure — Wikipedia, The Free Encyclopedia. (2023), https://en.wikipedia.org/w/index.php?title=Nucleic_acid_secondary _structure&oldid=1188087483 [Online; accessed 14-January-2024]

[58] Rietveld, K.; Van Poelgeest, R.; Pleij, C.W.; Van Boom, J.; Bosch, L. The tRNA-Uke structure at the 3' terminus of turnip yellow mosaic virus RNA. Differences and similarities with canonical tRNA. *Nucleic Acids Res.* **1982**, *10*, 1929–1946.

[59] Lu, X. & Olson, W. 3DNA: a software package for the analysis, rebuilding and visualization of three-dimensional nucleic acid structures. *Nucleic Acids Res.* **2003**, *31*, 5108–5121.

[60] Rivas, E.; Eddy, S.R. Noncoding RNA gene detection using comparative sequence analysis. *BMC Bioinform.* **2001**, *2*, 8.

[61] Chu, Y.; R.Corey, D. RNA Sequencing: Platform Selection, Experimental Design, and Data Interpretation. *Nucleic Acid Ther.* **2012**, *22*, 271–274.

[62] Dirks, R.; Pierce, N. Introduction A Partition Function Algorithm for Nucleic Acid Secondary Structure Including Pseudoknots. *J. Comput. Chem.* **2003**, *24*, 1664–77.

[63] Kucharík, M.; Hofacker, I.L.; Stadler, P.F.; Qin, J. Pseudoknots in RNA folding landscapes. *Bioinformatics* **2016**, *32*, 187–194.

[64] Hopcroft, J.E.; Ullman, J.D. *Formal languages and their relation to automata*; Addison-Wesley Longman Publishing Co., Inc.: Boston, MA, USA, 1969.

[65] Chomsky, N. Three models for the description of language. *IRE Trans. Inf. Theory* **1956**, *2*, 113 –124. https://doi.org/10.1109/TIT.1956.1056813

[66] Sipser, M. *Introduction to the theory of computation*; Thomson Course Technology: Boston, MA, USA, 2006; Volume 2.

[67] Aho, A.V.; Lam, M.S.; Sethi, R.; Ullman, J.D. *Compilers: principles, techniques, and tools*, 2nd ed.; Addison Wesley: London, UK 2006.

[68] McKinney, W.; Pandas: a foundational Python library for data analysis and statistics. *Python High Perform. Sci. Comput.* **2011**, *14*, 1–9.

[69] Younger, D.H. Recognition and parsing of context-free languages in n^3. *Inf. Control.* **1967**, *10*, 189–208.

[70] Earley, J. An efficient context-free parsing algorithm. *Commun. ACM* **1970**, *13*, 94–102. https://doi.org/http://doi.acm.org/10.1145/362007.362035

[71] Graham, S.L.; Harrison, M.A.; Ruzzo, W.L. An improved context-free recognizer. *ACM Trans. Program. Lang. Syst.* **1980**, *2*, 415–462.

[72] Ruzzo, W.L. General context-free language recognition. PhD Thesis, University of California, Berkeley, CA, USA, 1978.

[73] Geng, T.; Xu, F.; Mei, H.; Meng, W.; Chen, Z.; Lai, C. A practical GLR parser generator for software reverse engineering. *JNW*, **2014**, *9(3)*, 769–776.

[74] Pavlatos, C.; Dimopoulos, A.C.; Koulouris, A.; Andronikos, T.; Panagopoulos, I.; Papakonstantinou, G. Efficient reconfigurable embedded parsers. *Comput. Lang. Syst. Struct.* **2009**, *35*, 196–215. https://doi.org/10.1016/j.cl.2007.08.001

[75] Chiang, Y.; Fu, K. Parallel parsing algorithms and VLSI implementations for syntactic pattern recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* **1984**, *6*, 302–314.

[76] Available online: https://github.com/vnmakarov/yaep (accessed on 25 March 2020).

[77] Available online: https://github.com/ntua-dslab/knotify/tree/02-mdpi-2021-r2 (accessed on 30 January 2022).

[78] Trotta, E. On the normalization of the minimum free energy of RNAs by sequence length. *PLoS ONE* **2014**, *9*, e113380.

[79] Available online: https://github.com/chriskor1 (accessed on 9 March 2023).

[80] Available online: https://github.com/ntua-dslab/Knotify/releases/tag/04-Knotify+ (accessed on 17 December 2022).

[81] Ren, J.; Rastegari, B.; Condon, A.; Hoos, H.H. HotKnots: heuristic prediction of RNA secondary structures including pseudoknots. *RNA* **2005**, *11*, 1494–1504.

[82] Mathews, D.; Sabina, J.; Zuker, M.; Turner, D. Expanded sequence dependence of thermodynamic parameters improves prediction of RNA secondary structure1. *J. Mol. Biol.* **1999**, *288*, 911–40. https://doi.org/10.1006/jmbi.1999.2700.

[83] Available online: https://bit.ly/Knotify_plus_dataset_mdpi

[84] Taufer, M., Licon, A., Araiza, R., Mireles, D., Van Batenburg, F., Gultyaev, A. & Leung, M. PseudoBase++: an extension of PseudoBase for easy searching, formatting and visualization of pseudoknots. *Nucleic Acids Research.* **2009**, *37*, D127-D135.

[85] Danaee, P., Rouches, M., Wiley, M., Deng, D., Huang, L. & Hendrix, D. bpRNA: large-scale automated annotation and analysis of RNA secondary structure. *Nucleic Acids Research.* **2018**, *46*, 5381–5394.

[86] Available online: https://https://github.com/chriskor1.

[87] Bon, M.; Vernizzi, G.; Orland, H.; Zee, A. Topological classification of RNA structures. *Journal of molecular biology*, *379(4)*, **2008**, 900–911.

[88] Byun, Y.;Han K. PseudoViewer3: generating planar drawings of large-scale RNA structures with pseudoknots. Bioinformatics, Vol. 25, 1435-1437, 2009.

Παράρτημα Β΄

Ενδεικτικοί Κώδικες των Συστημάτων

Algorithm 2 Apply Free Energy and Stems Criterion Algorithm

```
import pandas as pd
from knotify.energy.base import BaseEnergy
def apply_free_energy_and_stems_criterion(
    data: pd.DataFrame,
    sequence: str,
    max_stem_allow_smaller: int,
    energy: BaseEnergy,):
    """
    This function selects and sorts the best RNA structures based on energy and number of stems.
    Args:
    data (DataFrame): Data containing RNA structures.
    sequence (str): The RNA biological sequence.
    max_stem_allow_smaller (int): The maximum allowable number of stems that can be smaller than the maximum.
    energy (BaseEnergy): An object that calculates the energy of structures.
    Returns:
    DataFrame: The modified data with the best structures.
    """

    # Calculate the total number of stems for each row of data.
    data["stems"] = data["left_loop_stems"] + data["right_loop_stems"]
    # Filter rows based on the number of stems.
    data = data[data["stems"] >= data["stems"].max() - max_stem_allow_smaller].reset_index(drop=True)
    # Calculate the actual number of stems based on dot-bracket representation.
    data["real_stems"] = data["dot_bracket"].apply(lambda r: sum(x != "." for x in r))
    # Calculate the energy for each RNA structure in the DataFrame.
    data["energy"] = data["dot_bracket"].apply(lambda r: energy.eval(sequence, r))
    # Sort data based on energy, actual number of stems, and 'dd' dimension.
    data.sort_values(["energy", "real_stems", "dd"], ascending=(True, False, True), inplace=True)
    # Reset the DataFrame index.
    data = data.reset_index(drop=True)

    return data
```

Algorithm 3 Grammar Generation for H-type Pseudoknot Detection

```
import jinja2

TEMPLATE = """
S : 'a' L 'a' D 'u' L 'u' # R01 (1 3 5)
  | 'a' L 'g' D 'u' L 'c' # R02 (1 3 5)
  ...
  | 'u' L 'u' D 'g' L 'g' # R36 (1 3 5)
{% endif %}
  ;
L : 'a' L # L1 (1)
    | 'u' L # L2 (1)
    | 'c' L # L3 (1)
    | 'g' L # L4 (1)
    | 'a'   # 0
    | 'u'   # 0
    | 'c'   # 0
    | 'g'   # 0
D : {% for x in range(max_dd_size) %}D{{ x }} {%
    endfor %} # M1 ({% for x in range(max_dd_size) %}{{
     x }} {% endfor %})
E : 'a' # 0
    | 'u' # 0
    | 'c' # 0
    | 'g' # 0
    | ;
{% for x in range(max_dd_size) %}
D{{ x }} : E # N{{ x }} (0)
{% endfor %}
"""

def generate_grammar(allow_ug: bool, max_dd_size: int)
    -> str:
    """
    Generate grammar for pseudoknot detection, based
       on parameters.
    """
    return (
        jinja2.Environment()
        .from_string(TEMPLATE)
        .render(
            allow_ug=allow_ug,
            max_dd_size=max_dd_size,
        )
    )
```

Algorithm 4 Struct Definitions for Pseudoknot Detection

```
// Structure representing a pseudoknot in an RNA
   sequence
struct pseudoknot {
    int left_loop_size;      // Size of the left loop
    int dd_size;             // Size of the D-D (
       directed distance) segment
    struct pseudoknot *next; // Pointer to the next
       pseudoknot in the list
};

// Structure for maintaining a list of integer values
struct size {
    int value;        // Integer value
    struct size *next; // Pointer to the next size
       node in the list
};
```

Algorithm 5 Struct Management for Pseudoknot Detection (Part 1)

```
1  struct pseudoknot *create_pseudoknot(int
       left_loop_size, int dd_size) {
2    struct pseudoknot *ps =
3        (struct pseudoknot *)malloc(sizeof(struct
           pseudoknot));
4    ps->left_loop_size = left_loop_size;
5    ps->dd_size = dd_size;
6    ps->next = NULL;
7    return ps;
8  };
9
10 struct size *create_size(int size) {
11   struct size *size_node = (struct size *)malloc(
       sizeof(struct size));
12   size_node->value = size;
13   size_node->next = NULL;
14   return size_node;
15 };
16
17 struct pseudoknot *
18 append_pseudoknot_if_not_exists(struct pseudoknot *
     list,
19                                 struct pseudoknot *
                                      upcoming_knot) {
20   for (struct pseudoknot *it = list; it != NULL; it =
       it->next) {
21     if (it->left_loop_size == upcoming_knot->
         left_loop_size &&
22         it->dd_size == upcoming_knot->dd_size) {
23       return list;
24     }
25   }
26   upcoming_knot->next = list;
27   return upcoming_knot;
28 }
```

Algorithm 6 Struct Management for Pseudoknot Detection (Part 2)

```
1  struct size *append_size_if_not_exists(struct size *
       list, int size) {
2    struct size *size_node;
3    struct size *iter = list;
4    while (iter != NULL) {
5      if (iter->value == size) {
6        return list;
7      }
8      iter = iter->next;
9    }
10   size_node = create_size(size);
11   size_node->next = list;
12   return size_node;
13 }
14
15 struct pseudoknot *concatenate_punique(struct
       pseudoknot *list1,
16                                        struct
                                             pseudoknot *
                                             list2) {
17   struct pseudoknot *iter2 = list2, *next = NULL;
18   struct pseudoknot *new_list_head = list1;
19
20   if (list1 == NULL) {
21     return list2;
22   }
23   if (list2 == NULL) {
24     return list1;
25   }
26   while (iter2 != NULL) {
27     next = iter2->next;
28     new_list_head = append_pseudoknot_if_not_exists(
           new_list_head, iter2);
29     iter2 = next;
30   }
31
32   return new_list_head;
33 }
```

Algorithm 7 Struct Management for Pseudoknot Detection (Part 3)

```
1  struct size *concatenate_sunique(struct size *list1,
       struct size *list2) {
2    struct size *iter2 = list2;
3    struct size *new_list_head = list1;
4
5    if (list1 == NULL) {
6      return list2;
7    }
8    if (list2 == NULL) {
9      return list1;
10   }
11   while (iter2 != NULL) {
12     new_list_head = append_size_if_not_exists(
           new_list_head, iter2->value);
13
14     iter2 = iter2->next;
15   }
16   return new_list_head;
17 }
```

Algorithm 8 Helper Functions for Pseudoknot Detection (Part 1)

```
1  void print_node_type(struct yaep_tree_node *node) {
2    switch (node->type) {
3    case YAEP_ANODE:
4      printf("YAEP_NODE\n");
5      break;
6    case YAEP_NIL:
7      printf("YAEP_NIL\n");
8      break;
9    case YAEP_ERROR:
10     printf("YAEP_ERROR\n");
11     break;
12   case YAEP_TERM:
13     printf("YAEP_TERM\n");
14     break;
15   case YAEP_ALT:
16     printf("YAEP_ALT\n");
17     break;
18   case _yaep_VISITED:
19     printf("_yaep_VISITED\n");
20     break;
21   case _yaep_MAX:
22     printf("_yaep_MAX\n");
23     break;
24   default:
25     printf("NULL\n");
26     break;
27   }
28 }
29 static int read_token_func(void **attr) {
30   *attr = NULL;
31   if (s_input[s_ntok]) {
32     return s_input[s_ntok++];
33   } else {
34     return -1;
35   }
36 }
37 static void syntax_error_func(int err_tok_num,
38                       void *err_tok_attr,
39                       int start_ignored_tok_num,
40                       void *start_ignored_tok_attr,
41                       int start_recovered_tok_num,
42                       void *start_recovered_tok_attr) {}
43 static void *parse_alloc_func(int size) {
44   void *result;
45   result = malloc(size);
46   return result;
47 }
```

Algorithm 9 Helper Functions for Pseudoknot Detection (Part 2)

```
struct yaep_tree_node *parse(const char *description)
   {
  struct grammar *g;
  struct yaep_tree_node *root;
  int ambiguous_p;

  // Grammar parsing and tree creation logic
  if ((g = yaep_create_grammar()) == NULL) {
    fprintf(stderr, "yaep_create_grammar: No memory\n
       ");
    exit(1);
  }
  yaep_set_debug_level(g, 0);
  yaep_set_one_parse_flag(g, 0);
  yaep_set_cost_flag(g, 0);
  yaep_set_lookahead_level(g, 2);
  if (yaep_parse_grammar(g, TRUE, description) != 0) {
    fprintf(stderr, "%s\n", yaep_error_message(g));
    exit(1);
  }
  int parsed = yaep_parse(g, read_token_func,
     syntax_error_func,
                          parse_alloc_func, NULL, &
                             root, &ambiguous_p);
  if (parsed) {
    printf("this should not be accepted\n");
    fprintf(stderr, "yaep_parse: %s\n",
       yaep_error_message(g));
  }
  return root;
}

char *terminal_codes_to_char(int code) {
  // Convert terminal codes to their respective
     character representation
  switch (code) {
  case 97:
    return "a";
  case 99:
    return "c";
  case 103:
    return "g";
  case 117:
    return "u";
  default:
    return "ERROR";
  }
}
```

Algorithm 10 YAEP Graph Traversal Functions for H-type Pseudoknot Detection (Part 1)

```
1  int traverse_parse_tree_for_loop(struct yaep_tree_node
       *node) {
2    switch (node->type) {
3    case YAEP_ANODE:
4      return traverse_parse_tree_for_loop(node->val.
         anode.children[0]) + 1;
5    case YAEP_TERM:
6      return 1;
7    default:
8      return -1;
9    }
10 }
11
12 void dump(struct size *s) {
13   for (struct size *a = s; a != NULL; a = a->next) {
14     printf("%d ->", a->value);
15   }
16   printf("\n");
17 }
18 struct size *cartesianProduct(struct size *A, struct
     size *B) {
19   struct size *sizes = NULL;
20   if (A == NULL)
21     return B;
22   if (B == NULL)
23     return A;
24   for (struct size *iterA = A; iterA != NULL; iterA =
       iterA->next) {
25     for (struct size *iterB = B; iterB != NULL; iterB
         = iterB->next) {
26       sizes = append_size_if_not_exists(sizes, iterA->
           value + iterB->value);
27     }
28   }
29   return sizes;
30 }
31
32 struct size *multiCartesianProduct(struct size **A,
     int count) {
33   struct size *sizes = NULL;
34   for (int i = 0; i < count; i++) {
35     sizes = cartesianProduct(sizes, A[i]);
36   }
37   return sizes;
38 }
```

Algorithm 11 YAEP Graph Traversal Functions for H-type Pseudoknot Detection (Part 2)

```
1  struct size *traverse_parse_tree_for_dd(struct
       yaep_tree_node *node) {
2    switch (node->type) {
3    case YAEP_ANODE:
4      if ((node->val.anode.name)[0] == 'M') {
5        struct size **childSizes = malloc(s_max_dd_size
           * sizeof(struct size *));
6        for (int i = 0; i < s_max_dd_size; i++) {
7          childSizes[i] = traverse_parse_tree_for_dd(
             node->val.anode.children[i]);
8        }
9        return multiCartesianProduct(childSizes,
           s_max_dd_size);
10     } else if ((node->val.anode.name)[0] == 'N') {
11       return traverse_parse_tree_for_dd(node->val.
           anode.children[0]);
12     }
13     break;
14   case YAEP_TERM:
15     return create_size(1);
16   case YAEP_NIL:
17     return create_size(0);
18   case YAEP_ALT:
19     struct size *anode_dds =
         traverse_parse_tree_for_dd(node->val.alt.node);
20     struct size *alt_dds = NULL;
21     if (node->val.alt.next != NULL) {
22       alt_dds = traverse_parse_tree_for_dd(node->val.
           alt.next);
23     }
24     struct size *new_list = concatenate_sunique(
         anode_dds, alt_dds);
25     return new_list;
26   default:
27     printf("Error in node type\n");
28   }
29   return NULL;
30 }
```

Algorithm 12 YAEP Graph Traversal Functions for H-type Pseudoknot Detection (Part 3)

```
1  struct pseudoknot *traverse_parse_tree(struct
       yaep_tree_node *node) {
2    int loop_size;
3    struct size *dd_sizes;
4    struct pseudoknot *pknots = NULL;
5    struct pseudoknot *pseudoknots = NULL;
6    if (!node) {
7      return pseudoknots;
8    }
9    switch (node->type) {
10   case YAEP_ERROR:
11     return NULL;
12   case YAEP_NIL:
13     return NULL;
14   case YAEP_TERM:
15     return NULL;
16   case YAEP_ANODE:
17     if ((node->val.anode.name)[0] != 'R') {
18       return NULL;
19     }
20     loop_size = traverse_parse_tree_for_loop(node->val
         .anode.children[0]);
21     dd_sizes = traverse_parse_tree_for_dd(node->val.
         anode.children[1]);
22
23     while (dd_sizes != NULL) {
24       pseudoknots = append_pseudoknot_if_not_exists(
25           pseudoknots, create_pseudoknot(loop_size,
               dd_sizes->value));
26       dd_sizes = dd_sizes->next;
27     }
28     return pseudoknots;
29   case YAEP_ALT:
30     pknots = traverse_parse_tree(node->val.alt.node);
31     pseudoknots = NULL;
32     if (node->val.alt.next != NULL) {
33       pseudoknots = traverse_parse_tree(node->val.alt.
           next);
34     }
35     pseudoknots = concatenate_punique(pseudoknots,
         pknots);
36     return pseudoknots;
37   default:
38     printf("Unhandled node type\n");
39     return pseudoknots;
40   }
41 }
```

Algorithm 13 YAEP H-type Pseudoknot Detection and Procedure Initialization

```
1  void detect_pseudoknots(char *sequence, void (*cb)(int
       , int, int, int)) {
2    s_input = sequence;
3    int len = strlen(sequence);
4    int min_window_size =
5        s_min_window_size ? s_min_window_size : (len *
           s_min_window_size_ratio);
6    int max_window_size =
7        s_max_window_size ? s_max_window_size : (len *
           s_max_window_size_ratio);
8
9    for (int right = len - 1; right >= min_window_size -
         1; right--) {
10     for (int left = right - min_window_size + 1;
11          left > right - max_window_size && left >= 0;
               left--) {
12       s_ntok = left;
13       struct yaep_tree_node *root = parse(s_definition
           );
14       struct pseudoknot *ps = traverse_parse_tree(root
           );
15
16       for (struct pseudoknot *i = ps; i != NULL; i = i
           ->next) {
17         if (i->dd_size < s_min_dd_size) {
18           continue;
19         }
20         cb(left, right - left + 1, i->left_loop_size,
              i->dd_size);
21       }
22     }
23     s_input[right] = '\0';
24   }
25 }
26
27 void initialize(char *_grammar, int _allow_ug, int
     _min_dd_size,
28                 int _max_dd_size, int _min_window_size
                      , int _max_window_size,
29                 float _min_window_size_ratio, float
                      _max_window_size_ratio) {
30   s_definition = strdup(_grammar);
31   s_min_dd_size = _min_dd_size;
32   s_max_dd_size = _max_dd_size;
33   s_min_window_size = _min_window_size;
34   s_max_window_size = _max_window_size;
35   s_min_window_size_ratio = _min_window_size_ratio;
36   s_max_window_size_ratio = _max_window_size_ratio;
37 }
```

Algorithm 14 Grammar Generation for L-type Pseudoknot Detection

```
import jinja2

TEMPLATE = """
S : 'a' L 'a' L 'a' D 'u' L 'u' L 'u' # R01 (1 3 5 7 9)
    | 'a' L 'a' L 'u' D 'u' L 'u' L 'a' # R02 (1 3 5 7 9)
    | 'a' L 'a' L 'g' D 'u' L 'u' L 'c' # R03 (1 3 5 7 9)
    ...
    | 'u' L 'u' L 'u' D 'g' L 'g' L 'g' # R215 (1 3 5 7 9)
    | 'u' L 'g' L 'a' D 'g' L 'u' L 'u' # R216 (1 3 5 7 9)
L : 'a' L # L1 (1)
    | 'u' L # L2 (1)
    | 'c' L # L3 (1)
    | 'g' L # L4 (1)
    | 'a'   # 0
    | 'u'   # 0
    | 'c'   # 0
    | 'g'   # 0
D : {% for x in range(max_dd_size) %}D{{ x }} {% endfor %} # M1 ({% for x in range(max_dd_size) %}{{ x }} {% endfor %})
E : 'a' # 0
    | 'u' # 0
    | 'c' # 0
    | 'g' # 0
    | ;
{% for x in range(max_dd_size) %}
D{{ x }} : E # N{{ x }} (0)
{% endfor %}
"""
def generate_grammar(allow_ug: bool, max_dd_size: int) -> str:
    return (
        jinja2.Environment()
        .from_string(TEMPLATE)
        .render(
            allow_ug=allow_ug,
            max_dd_size=max_dd_size,
        )
    )
```

Algorithm 15 YAEP Graph Traversal Functions for L-type Pseudoknot Detection

```
1  struct pseudoknot *traverse_parse_tree(struct
       yaep_tree_node *node) {
2    int loop_size;
3    struct size *dd_sizes;
4    struct pseudoknot *pknots = NULL;
5    struct pseudoknot *pseudoknots = NULL;
6    if (!node) {
7      return pseudoknots;}
8    switch (node->type) {
9    case YAEP_ERROR:
10     return NULL;
11   case YAEP_NIL:
12     return NULL;
13   case YAEP_TERM:
14     return NULL;
15   case YAEP_ANODE:
16     if ((node->val.anode.name)[0] != 'R') {
17       return NULL;}
18     left_left_loop_size =
19         traverse_parse_tree_for_loop(node->val.anode.
              children[0]);
20     left_right_loop_size =
21         traverse_parse_tree_for_loop(node->val.anode.
              children[1]);
22     dd_sizes = traverse_parse_tree_for_dd(node->val.
          anode.children[2]);
23     right_left_loop_size =
24         traverse_parse_tree_for_loop(node->val.anode.
              children[3]);
25     while (dd_sizes != NULL) {
26       pseudoknots = append_pseudoknot_if_not_exists(
27           pseudoknots,
28           create_pseudoknot(left_left_loop_size,
                 left_right_loop_size,
29                             right_left_loop_size,
                                dd_sizes->value));
30       dd_sizes = dd_sizes->next;}
31     return pseudoknots;
32   case YAEP_ALT:
33     pknots = traverse_parse_tree(node->val.alt.node);
34     pseudoknots = NULL;
35     if (node->val.alt.next != NULL) {
36       pseudoknots = traverse_parse_tree(node->val.alt.
             next);}
37     pseudoknots = concatenate_punique(pseudoknots,
          pknots);
38     return pseudoknots;
39   default:
40     printf("Unhandled node type\n");
41     return pseudoknots;}}
```

Algorithm 16 YAEP L-type Pseudoknot Detection and Procedure Initialization

```
1  void detect_pseudoknots(char *sequence, void (*cb)(int
       , int, int, int, int)) {
2    s_input = sequence;
3    int len = strlen(sequence);
4    int min_window_size =
5        s_min_window_size ? s_min_window_size : (len *
             s_min_window_size_ratio);
6    int max_window_size =
7        s_max_window_size ? s_max_window_size : (len *
             s_max_window_size_ratio);
8
9    for (int right = len - 1; right >= min_window_size -
         1; right--) {
10     for (int left = right - min_window_size + 1;
11          left > right - max_window_size && left >= 0;
                left--) {
12       s_ntok = left;
13       struct yaep_tree_node *root = parse(s_definition
             );
14       struct pseudoknot *ps = traverse_parse_tree(root
             );
15       for (struct pseudoknot *i = ps; i != NULL; i = i
             ->next) {
16         if (i->dd_size < s_min_dd_size) {
17           continue;
18         }
19         cb(left, right - left + 1, i->
               left_left_loop_size,
20            i->left_right_loop_size, i->
                  right_left_loop_size, i->dd_size);
21       }
22     }
23     s_input[right] = '\0';
24   }
25 }
26 void initialize(char *_grammar, int _allow_ug, int
       _min_dd_size,
27                 int _max_dd_size, int _min_window_size
                       , int _max_window_size,
28                 float _min_window_size_ratio, float
                       _max_window_size_ratio) {
29   s_definition = strdup(_grammar);
30   s_min_dd_size = _min_dd_size;
31   s_max_dd_size = _max_dd_size;
32   s_min_window_size = _min_window_size;
33   s_max_window_size = _max_window_size;
34   s_min_window_size_ratio = _min_window_size_ratio;
35   s_max_window_size_ratio = _max_window_size_ratio;
36 }
```