ΧΑΡΟΚΟΠΕΙΟ ΠΑΝΕΠΙΣΤΗΜΙΟ

ΣΧΟΛΗ Ψηφιακής Τεχνολογίας

ΤΜΗΜΑ Πληροφορικής & Τηλεματικής

ΠΡΟΓΡΑΜΜΑ ΜΕΤΑΠΤΥΧΙΑΚΩΝ ΣΠΟΥΔΩΝ Πληροφορική και Τηλεματική

ΚΑΤΕΥΘΥΝΣΗ Τηλεπικοινωνιακά Δίκτυα και Υπηρεσίες Τηλεματικής

Μέθοδοι πρόβλεψης της δευτεροταγούς δομής του RNA - μια επισκόπηση με έμφαση σε γραμματικές προσεγγίσεις

Διπλωματική Εργασία

Βασιλική Λιάγκου

Αθήνα, 2018

ΧΑΡΟΚΟΠΕΙΟ ΠΑΝΕΠΙΣΤΗΜΙΟ

ΣΧΟΛΗ Ψηφιακής Τεχνολογίας

ΤΜΗΜΑ Πληροφορικής & Τηλεματικής

ΠΡΟΓΡΑΜΜΑ ΜΕΤΑΠΤΥΧΙΑΚΩΝ ΣΠΟΥΔΩΝ Πληροφορική και Τηλεματική

ΚΑΤΕΥΘΥΝΣΗ Τηλεπικοινωνιακά Δίκτυα και Υπηρεσίες Τηλεματικής

Τριμελής Εξεταστική Επιτροπή

Αλέξανδρος Δημόπουλος (Επιβλέπων)

Δρ. Ηλεκτρολόγος Μηχανικός & Μηχανικός Υπολογιστών, Εξωτερικός Διδάσκων

Κωνσταντίνος Τσερπές

Επίκ. Καθηγητής, Τμ. Πληροφορικής και Τηλεματικής,
Χαροκόπειο Πανεπιστήμιο

Ηρακλής Βαρλάμης

Επίκ. Καθηγητής, Τμ. Πληροφορικής και Τηλεματικής,
Χαροκόπειο Πανεπιστήμιο

Η Λιάγκου Βασιλική

δηλώνω υπεύθυνα ότι:

1) Είμαι ο κάτοχος των πνευματικών δικαιωμάτων της πρωτότυπης αυτής εργασίας και από όσο γνωρίζω η εργασία μου δε συκοφαντεί πρόσωπα, ούτε προσβάλει τα πνευματικά δικαιώματα τρίτων.

2) Αποδέχομαι ότι η ΒΚΠ μπορεί, χωρίς να αλλάξει το περιεχόμενο της εργασίας μου, να τη διαθέσει σε ηλεκτρονική μορφή μέσα από τη ψηφιακή Βιβλιοθήκη της, να την αντιγράψει σε οποιοδήποτε μέσο ή/και σε οποιοδήποτε μορφότυπο καθώς και να κρατά περισσότερα από ένα αντίγραφα για λόγους συντήρησης και ασφάλειας.

ΕΥΧΑΡΙΣΤΙΕΣ

Θα ήθελα να ευχαριστήσω τον επιβλέποντα καθηγητή μου, κ. Αλέξανδρο Δημόπουλο, για την ανάθεση του θέματος και την άριστη καθοδήγηση και βοήθεια όποτε την χρειαζόμουν.

Πίνακας περιεχομένων

Περίληψη στα Ελληνικά 7
Abstract 9
Κατάλογος εικόνων 11
Κατάλογος πινάκων 13
Συντομογραφίες 15
1 Εισαγωγή 17
1.1 Βιολογικό υπόβαθρο 17
1.1.1 Κύριες διαφορές μεταξύ του DNA και του RNA 20
1.2 Δομές του RNA 21
1.2.1 Δευτεροταγής δομή του RNA 23
1.3 Μέθοδοι προσδιορισμού δευτεροταγούς δομής RNA 25
1.3.1 Πειραματική επαλήθευση 27
1.3.2 Πρόβλεψη δευτεροταγούς δομής - Υπολογιστικές μέθοδοι 29
2 Επισκόπηση γραμματικών και αλγορίθμων 35
2.1 Τυπικές γραμματικές 35
2.2 Η ιεραρχία Chomsky 36
2.3 Στοχαστικές γραμματικές χωρίς συμφραζόμενα 38
2.3.1 Στοχαστικές γραμματικές χωρίς συμφραζόμενα στην πρόβλεψη δομής του RNA 38
2.3.2 Στοχαστικές γραμματικές χωρίς συμφραζόμενα και ψευδοκόμβοι του RNA 42
2.4 Διφορούμενες γραμματικές 42
2.5 Αλγόριθμοι πρόβλεψης δομής με βάση SCFG γραμματικές 44
2.5.1 Αλγόριθμος CYK 44
2.5.2 Αλγόριθμος Inside 45
2.5.3 Αλγόριθμος Inside-outside 46
2.5.4 Chomsky Normal Form 47
3 Υλοποιήσεις Λογισμικού με βάση SCFG γραμματικές 49
3.1 Automatic RNA Secondary Structure Determination with Stochastic Context-Free Grammars 49
3.2 Pfold: RNA secondary structure prediction using stochastic context-free grammars 51
3.3 Multithreaded comparative RNA secondary structure prediction using stochastic context-free grammars 54
3.4 Parallel RNA secondary structure prediction using stochastic context-free grammars 56
3.5 Load Balancing Algorithm in Cluster-based RNA secondary structure Prediction 60
3.6 Evolving stochastic context–free grammars for RNA secondary structure prediction 63

3.7 Research on folding diversity in statistical learning methods for RNA secondary structure prediction 66
3.8 Μέθοδοι δομικής στοίχισης – Stemloc & Consan 69
3.8.1 Accelerated probabilistic inference of RNA structure evolution 69
3.8.2 Efficient pairwise RNA structure prediction and alignment using sequence alignment constraints 71
3.9 Προσεγγίσεις SCFG γραμματικών για πρόβλεψη RNA δευτεροταγούς δομής συμπεριλαμβανομένων ψευδοκόμβων. 74
3.9.1 RNA Pseudoknotted structure prediction using stochastic multiple context free grammar 75
3.9.2 A Grammar-Based Approach to RNA Pseudoknotted Structure Prediction for Aligned Sequences 78
4 Υλοποιήσεις Υλικού με βάση SCFG γραμματικές 81
4.1 Fine-grained Parallel Application Specific Computing for RNA Secondary Structure Prediction Using SCFGs on FPGA 81
4.2 FPQRNA: Hardware-Accelerated QRNA Package For Noncoding RNA Gene Detecting On FPGA 85
4.3 Parallel Hardware Stochastic Context-Free Parsers 90
5 Συμπεράσματα 95
5.1 Προσεγγίσεις με βάση των αριθμών των ακολουθιών εισόδου 95
5.2 Προσεγγίσεις με βάση την υλοποίηση 97
5.3 Μια αξιολόγηση των συγκριτικών προσεγγίσεων πρόβλεψης της δομής του RNA 99
5.4 Μια συγκριτική αξιολόγηση των παράλληλων υλοποιήσεων του CYK/inside αλγορίθμου 100
5.5 Επίλογος 102
Βιβλιογραφία 105

Περίληψη στα Ελληνικά

Το RNA ανήκει στα νουκλεινικά οξέα που είναι μία από τις τέσσερις κύριες τάξεις βιομορίων που είναι απαραίτητα για κάθε μορφή ζωής. Η δομή του, και πιο συγκεκριμένα η δευτεροταγής δομή του, καθορίζει και τη βιολογική του δράση. Η δομή του RNA μπορεί να προσδιοριστεί πειραματικά. Ωστόσο, οι πειραματικές μέθοδοι είναι δαπανηρές και χρονοβόρες και ως εκ τούτου η πρόβλεψη της δομής με υπολογιστικά εργαλεία είναι αναγκαία. Η επιστήμη των υπολογιστών μπορεί και έχει συμβάλλει σημαντικά στην πρόγνωση της δευτεροταγούς δομής με την ανάπτυξη αλγορίθμων που μοντελοποιούν την αναδίπλωση του RNA.

Σκοπός αυτής της εργασίας είναι να παρουσιάσει το πρόβλημα της πρόγνωσης της δευτεροταγούς δομής του RNA και των διαφορετικών υλοποιήσεων που έχουν προταθεί για την λύση αυτού του προβλήματος με βάση τις γραμματικές προσεγγίσεις. Για το λόγο αυτό, μελετήθηκαν περίπου σαράντα αναφορές που προσεγγίζουν το πρόβλημα της πρόγνωσης της δευτεροταγούς δομής.

Οι αλγόριθμοι που θα μας απασχολήσουν στο πλαίσιο αυτής της εργασίας βασίζονται σε γραμματικές και πιο συγκεκριμένα σε στοχαστικές γραμματικές χωρίς συμφραζόμενα (SCFG). Ξεκινώντας από τη δεκαετία του '90, ο Sakakibara (1994) και οι Eddy και Durbin (1994) χρησιμοποίησαν τις SCFG γραμματικές για ανάλυση και πρόβλεψη της δομής του RNA. Ακολούθησε ένας μεγάλος αριθμός διαφορετικών προσεγγίσεων για την επίλυση αυτού του προβλήματος. Η αύξηση των υπολογιστικών απαιτήσεων σε χώρο και μνήμη καθώς και η ταυτόχρονη ανάπτυξη των πολυπύρηνων (multicore) επεξεργαστών και παράλληλων αρχιτεκτονικών οδήγησε σε παράλληλες υλοποιήσεις αυτών των αλγορίθμων σε επίπεδο λογισμικού και υλικού. Η παρούσα εργασία πραγματοποιεί μια ανασκόπηση των διαφορετικών υλοποιήσεων που έχουν προταθεί μέχρι σήμερα. Παρ’ όλα αυτά, το πρόβλημα της ακριβούς και γρήγορης πρόγνωσης της δευτεροταγούς δομής του RNA παραμένει ενεργό και αποτελεί μια υπολογιστική πρόκληση για τους ερευνητές.

Λέξεις κλειδιά: πρόβλεψη, δευτεροταγής δομή, στοχαστικές γραμματικές χωρίς συμφραζόμενα, RNA

Abstract

RNA belongs to the class of nucleic acid which is one of the four main classes of biomolecules that are necessary for every form of life. Its structure, and in particular its secondary structure, also determines its biological function. The structure of RNA can be determined experimentally. However, experimental methods are costly and time-consuming, and therefore the prediction of the structure with computational tools is necessary. Computer science has significantly contributed to the secondary structure prediction by the development of algorithms that model the folding of RNA.

In this thesis the problem of RNA secondary structure prediction is presented and different proposed solutions based on grammatical approaches among recent years are reviewed. Indeed, about 40 references were reviewed.

The reviewed algorithms in this thesis are based on formal grammars and, more specifically, on stochastic context-free grammars (SCFG). Starting in the 1990s, Sakakibara (1994) and Eddy and Durbin (1994) used SCFG grammars to analyze and predict the RNA secondary structure. A large number of different approaches to solving this problem were followed. Increasing computational requirements in space and memory as well as the simultaneous development of multicore processors and parallel architectures led to software and/or hardware approaches of these algorithms on parallel machines. The current thesis reviews the different implementations that have been proposed so far. However, the problem of accurate and fast RNA secondary structure prediction remains active and is a computational challenge for researchers.

Keywords: SCFG, RNA, prediction, secondary structure, RNA

Κατάλογος εικόνων

Εικόνα 1.1: Το μονόκλωνο RNA με τις νιτρογενείς βάσεις....18
Εικόνα 1.2: Νουκλεοτίδια RNA....18
Εικόνα 1.3: Συμπληρωματικότητα βάσεων....19
Εικόνα 1.4: Το κεντρικό δόγμα της Μοριακής Βιολογίας....19
Εικόνα 1.5: Η μεταγραφή του DNA σε RNA και η μετάφραση του σε πρωτεΐνη....20
Εικόνα 1.6: Το μονόκλωνο RNA στα αριστερά και το δίκλωνο DNA στα δεξιά....21
Εικόνα 1.7: Πρωτοταγής δομή του RNA....22
Εικόνα 1.8: Δευτεροταγής & τριτοταγής δομή του RNA....22
Εικόνα 1.9: Στοιχεία δευτεροταγούς δομής του RNA....23
Εικόνα 1.10: Διακλαδισμένος βρόχος (Multi-branch loop)....24
Εικόνα 1.11: Ψευδοκόμβος....25
Εικόνα 1.12: Α. Η ακολουθία RNA Β. Δευτεροταγής δομής σε αναπαράσταση παρενθέσεων Γ. Δευτεροταγής δομής σε αναπαράσταση γραφήματος....26
Εικόνα 1.13: Είσοδος: ένα σύνολο ακολουθιών με την ίδια βιολογική λειτουργία που θεωρείται ότι έχουν περίπου την ίδια δομή, Έξοδος: τα κοινά δομικά στοιχεία, οι ευθυγραμμισμένες ακολουθίες και μια φυλογενία που εξηγεί καλύτερα τα παρατηρούμενα δεδομένα....31
Εικόνα 2.1: Ιεραρχία των γραμματικών κατά Chomsky....37
Εικόνα 2.2: Δέντρο ανάλυσης (αριστερά) και η αντίστοιχη δευτεροταγής δομή (δεξιά)....40
Εικόνα 2.3: Δέντρα ανάλυσης για την συμβολοσειρά ab....43
Εικόνα 2.4: AND-or δέντρο για την συμβολοσειρά ab....43
Εικόνα 2.5: Εφαρμογή του inside αλγορίθμου για τον υπολογισμό της πιθανότητας a(i, j, u)....45
Εικόνα 2.6: Εφαρμογή του outside αλγορίθμου για τον υπολογισμό της πιθανότητας β(i, j, u)....46
Εικόνα 3.1: Μέθοδος εύρεσης της SCFG γραμματικής Α) Πολλαπλή στοίχιση ενός RNA βρόγχου φουρκέτας. Υποδεικνύονται οι βασικές ζευγαρωμένες περιοχές Β) Δευτεροταγής δομή της πρώτης ακολουθίας στην πολλαπλή στοίχιση Γ) Μια SCFG γραμματική που περιγράφει την ίδια δομή. Οι έλικες είναι ορθογώνιοι, οι βρόχοι είναι κύκλοι και τα τρίγωνα είναι σημεία διακλάδωσης που δεν αντιπροσωπεύουν βάσεις.....50
Εικόνα 3.2: Μια πρόβλεψη δομής για τρεις υποθετικές ακολουθίες....53
Εικόνα 3.3: Προκύπτουσα δομή με την εισαγωγή πιθανότητα λάθους....53
Εικόνα 3.4: Μείωση χρόνου εκτέλεσης για το φυλογενετικό τμήμα του αλγορίθμου.....55
Εικόνα 3.5: Μείωση χρόνου εκτέλεσης για το κομμάτι της SCFG γραμματικής.....56
Εικόνα 3.6: Παράλληλη υλοποίηση του CYK αλγορίθμου χρησιμοποιώντας 4 επεξεργαστές....58
Εικόνα 3.7: Κάθε κελί M(i,j,k)εξαρτάται από τα κελιά M(i, mid, kleft) και M(mid + 1, j, kright)....59
Εικόνα 3.8: Ο κυματοειδής υπολογισμός στην στατική προσέγγιση....61
Εικόνα 3.9: Ο κυματοειδής υπολογισμός στην δυναμική προσέγγιση....62
Εικόνα 3.10: Επιτάχυνση για τέσσερα μήκη αλληλουχιών RNA: 120 (5Secoli), 377 (RnaseP), 1542 (lsuecoli) και 2904 (ssuecoli). Ο αριθμός επεξεργαστών είναι 2, 4, 8, 16 και 32....62
Εικόνα 3.11: Διάγραμμα ταξινόμησης κανόνων αναδίπλωσης....67
Εικόνα 3.12: Δομή του Focusfold....68
Εικόνα 3.13: Σύγκριση του FocusFold και του RGRNA A. Το TrainSetA αποτελεί τα δεδομένα εκπαίδευσης και toTestSetA αποτελεί το σύνολο δοκιμών B. Το TrainSetA αποτελεί τα δεδομένα εκπαίδευσης και toTestSetB αποτελεί το σύνολο δοκιμών....68
Εικόνα 3.14: Πλαίσια αναδίπλωσης και στοίχισης - Fold και aligned envelopes....71
Εικόνα 3.15: Συγκριτική απόδοση ως προς την ακρίβεια στοίχισης....73
Εικόνα 3.16: Συγκριτική απόδοση ως προς την ακρίβεια πρόβλεψης....73
Εικόνα 3.17: Διαδικασία υπολογισμού του γ....76

Εικόνα 3.18: Σύγκριση προβλεπόμενης δομής της Corona_pk_3 με την καταχωρημένη της δομή στην RFam 78
Εικόνα 3.19: Επισκόπηση της μεθόδου [30] 79
Εικόνα 4.1: Η κατανομή εργασιών και η σειρά υπολογισμού με 4-PE 82
Εικόνα 4.2: Η δομή του επιταχυντή του CYK/inside αλγορίθμου 84
Εικόνα 4.3: Δείκτης απόδοσης / κατανάλωσης ενέργειας (FPGA VS. CPU) 84
Εικόνα 4.4: Απόδοση / κατανάλωση ενέργειας και λόγος απόδοσης / κόστους (FPGA VS. cluster) ...85
Εικόνα 4.5: Παράδειγμα κατανομής εργασιών και σειρά υπολογισμού για 4 PE 87
Εικόνα 4.6: Αρχιτεκτονική του επιταχυντή του QRNA αλγορίθμου 88
Εικόνα 4.7: Ο λόγος απόδοσης / κατανάλωσης ενέργειας σε σύγκριση με τους μικροεπεξεργαστές γενικής χρήσης 90
Εικόνα 4.8: Επισκόπηση της προσέγγισης [18] 91
Εικόνα 4.9: Διάσχιση του δέντρου ανάλυσης 92
Εικόνα 4.10: Αρχιτεκτονική σημασιολογικού αξιολογητή 93
Εικόνα 5.1: Συσχέτιση μεθόδων μονής ακολουθίας 99
Εικόνα 5.2: Συσχέτιση συγκριτικών μεθόδων σε δεδομένα υψηλής ομοιότητας 100
Εικόνα 5.3: Χρόνος εκτέλεσης σε δευτερόλεπτα στη δυναμική προσέγγιση 101
Εικόνα 5.4: Χρόνος εκτέλεσης σε δευτερόλεπτα στην προσέγγιση 34 με Xeon επεξεργαστές 101
Εικόνα 5.5: Απόδοση / κατανάλωση ενέργειας και λόγος απόδοσης / κόστους (FPGA VS. cluster) .102

Κατάλογος πινάκων

Πίνακας 3.1:Σύγκριση ακρίβειας αποτελεσμάτων με είσοδο 60 E. Coli SSU rRNA ακολουθίες και 20 E. coli LSU rRNA ακολουθίες 60
Πίνακας 3.2: Τα αποτελέσματα των μέτρων αξιολόγησης για τις επτά γραμματικές που προέκυψαν .. 65
Πίνακας 3.3: Αποτελέσματα πρόγνωσης της προσέγγισης [27] 77
Πίνακας 3.4: Συγκριτική ακρίβεια πρόβλεψης μεταξύ της [29] [30] και [31] 80
Πίνακας 4.1: χρόνοι εκτέλεσης μεταξύ της [33] και τριών υλοποιήσεων λογισμικού 89

Συντομογραφίες

AG	Attribute Grammar
BNF	Backus Normal Form
CM	Covariance Model
CNF	Chomsky Normal Form
EM	Expectation Maximization
FPGA	Field Programmable Gate Array
HDL	Hardware Design Language
HMM	Hidden Markov Model
MEA	Maximum Expected Accuracy
MFE	Minimum Free Energy
ML	Maximum Likelihood
NMR	Nuclear Magnetic Resonance
PE	Processing Element
PPV	Positive Predictive Value
SCFG	Stochastic Context-Free Grammar
SMCFG	Stochastic Multiple Context Free Grammar
SMP	Symmetric Multiprocessing

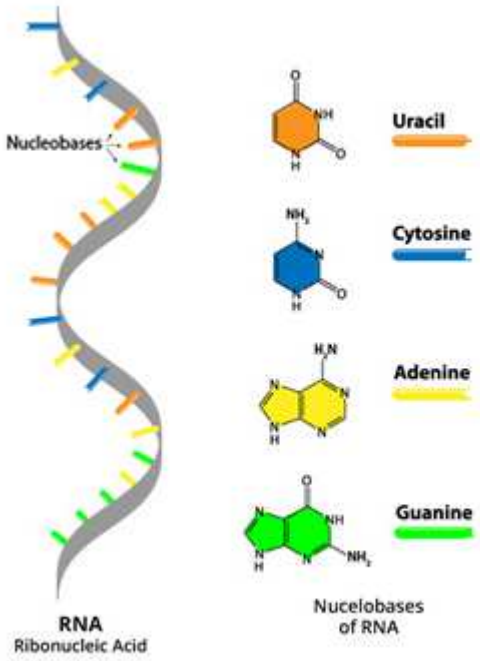
1 Εισαγωγή

Η Βιοπληροφορική είναι ένας διεπιστημονικός κλάδος όπου συνυπάρχουν και συνεργάζονται η επιστήμη των υπολογιστών και η επιστήμη της βιοχημείας και της μοριακής βιολογίας. Χρησιμοποιείται δε για την επίλυση αρκετών βιοχημικών και βιολογικών προβλημάτων. Σκοπός αυτής της εργασίας είναι να παρουσιάσει ένα υφιστάμενο πρόβλημα της μοριακής βιολογίας, αυτό της πρόγνωσης της δευτεροταγούς δομής του RNA, και των διαφορετικών λύσεων και υλοποιήσεων που έχουν προταθεί με την βοήθεια της επιστήμης των υπολογιστών και πιο συγκεκριμένα με την χρήση γραμματικών προσεγγίσεων.

Στο πρώτο κεφάλαιο παρατίθεται μια εισαγωγή στον ορισμό και την δομή του RNA καθώς και στις μεθόδους προσδιορισμού της δευτεροταγούς δομής του. Στο δεύτερο κεφάλαιο παρουσιάζονται οι γραμματικές και η χρήση τους στην ανάλυση και πρόγνωση της δομής του RNA καθώς και οι βασικοί αλγόριθμοι που αναπτύχθηκαν για τον σκοπό αυτό. Στα κεφάλαια 3 και 4 γίνεται ανασκόπηση των υλοποιήσεων των παραπάνω αλγορίθμων σε επίπεδο λογισμικού και υλικού, αντίστοιχα. Τέλος στο κεφάλαιο 5 αναφέρονται τα συμπεράσματα που προέκυψαν από τη συγκεκριμένη εργασία.

1.1 Βιολογικό υπόβαθρο

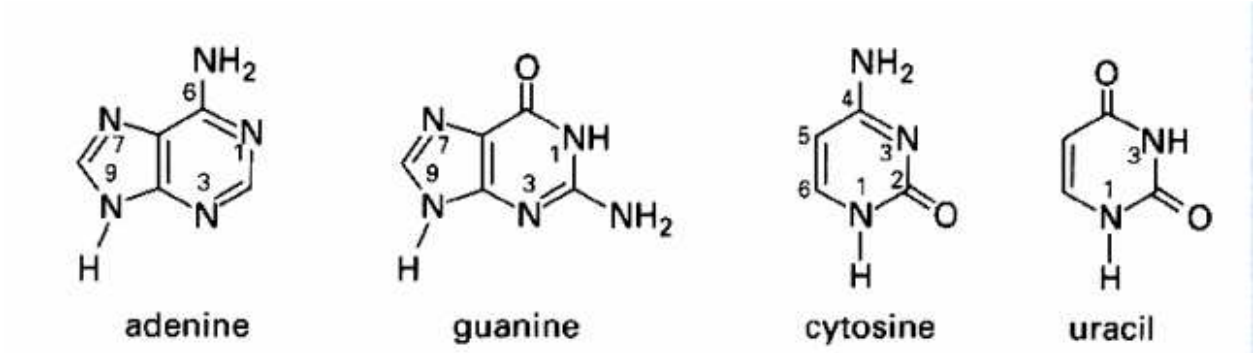
Το RNA (Ribonucleic acid), ριβοζονουκλεϊκό οξύ, από χημικής άποψης είναι ένα μακρομόριο που αποτελείται από μία αλυσίδα ριβονουκλεοτιδίων που συνδέονται μεταξύ τους με ομοιοπολικούς χημικούς δεσμούς. Το μόριο του RNA περιλαμβάνει, τέσσερις τύπους νουκλεοτιδίων που συνδέονται μεταξύ τους με 3'-5' φωσφοδιεστερικούς δεσμούς, και περιέχουν τις νουκλεινικές βάσεις αδενίνη (A), γουανίνη (G), κυτοσίνη (C) και ουρακίλη (U).



Εικόνα 1.1: Το μονόκλωνο RNA με τις νιτρογενείς βάσεις

(Πηγή: https://www.technologynetworks.com/)

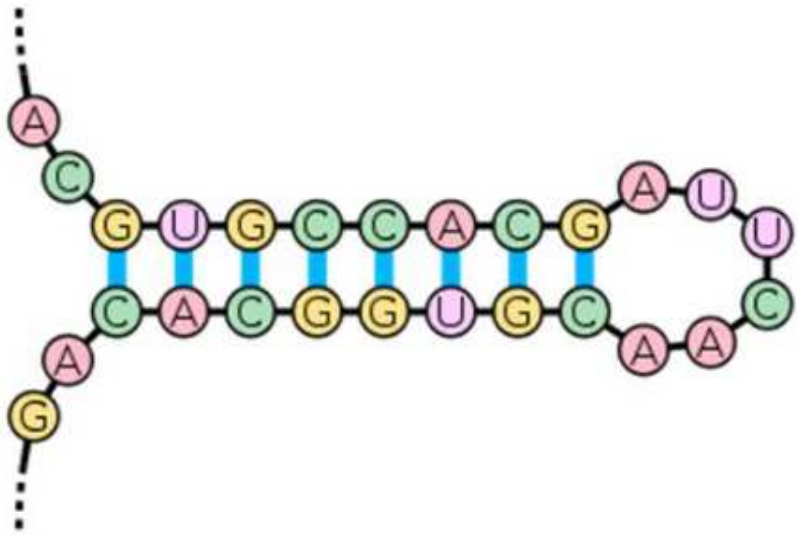
Η αδενίνη σχηματίζει δεσμούς υδρογόνου με την ουρακίλη και η γουανίνη σχηματίζει δεσμούς με την κυτοσίνη. Με αυτό τον τρόπο, λέμε ότι η αδενίνη είναι συμπληρωματική νουκλεινική βάση της ουρακίλης και ότι η γουανίνη είναι συμπληρωματική της κυτοσίνης [1].



Εικόνα 1.2: Νουκλεινικές βάσεις RNA

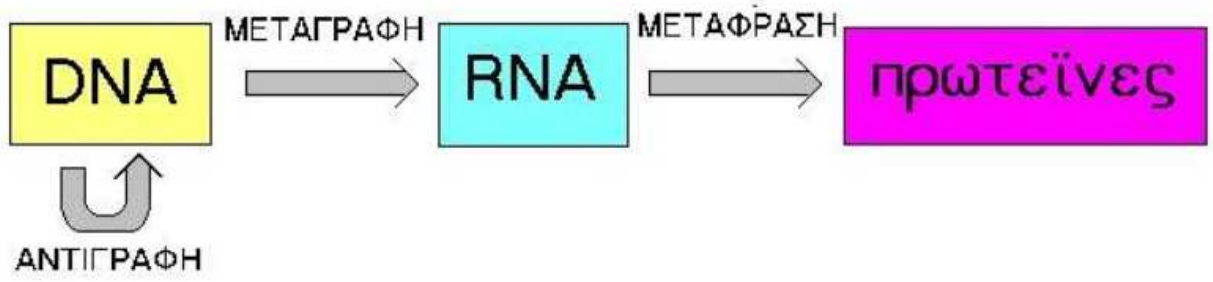
(Πηγή: https://en.wikipedia.org)

Μεταξύ γουανίνης και κυτοσίνης σχηματίζονται τρεις δεσμοί υδρογόνου, ενώ μεταξύ αδενίνης και ουρακίλης δύο δεσμοί υδρογόνου.



Εικόνα 1.3: Συμπληρωματικότητα βάσεων
(Πηγή: https://en.wikipedia.org)

Το RNA αποτελεί μαζί με το DNA, τις πρωτεΐνες, τους υδατάνθρακες και τα λιποειδή τα πιο σημαντικά βιολογικά μακρομόρια (οι τέσσερεις κύριες τάξεις βιομορίων) που είναι απαραίτητα για όλες τις γνωστές μορφές ζωής. Σύμφωνα με το κεντρικό δόγμα της Μοριακής Βιολογίας [6] όπως διατυπώθηκε αρχικά, η ροή της γενετικής πληροφορίας σε ένα κύτταρο είναι από το DNA μέσω του RNA σε πρωτεΐνες. Αν και σήμερα είναι γνωστό ότι ρέει και προς την αντίθετη κατεύθυνση.



Εικόνα 1.4: Το κεντρικό δόγμα της Μοριακής Βιολογίας
(Πηγή: https://el.wikipedia.org/wiki/Κεντρικό_δόγμα_της_μοριακής_βιολογίας)

Ιστορικά οι πρωτεΐνες θεωρήθηκαν ότι παίζουν πρωταγωνιστικούς ρόλους στο κύτταρο (εξ ου και στο όνομα τους το συνθετικό «πρώτος») ως ένζυμα, ως δομικά στοιχεία, κτλ. Το DNA (δεοξυριβονουκλεϊνικό οξύ) φέρει όλες τις γενετικές πληροφορίες που απαιτούνται για την ανάπτυξη του κυττάρου. Το RNA είναι το «φωτοαντίγραφο DNA» του κυττάρου. Όταν το κύτταρο

χρειάζεται να προβεί σε μια δράση, παράγει μια συγκεκριμένη πρωτεΐνη που σαν ένζυμο ή με άλλο τρόπο ξεκινά και βοηθά να πραγματοποιηθεί η εν λόγω δράση. Η παραγωγή της εν λόγω πρωτεΐνης γίνεται με ενεργοποίηση του ανάλογου γονιδίου για την πρωτεΐνη αυτή. Το γονίδιο είναι το τμήμα του DNA που κωδικοποιεί αυτή την πρωτεΐνη - και παράγει πολλαπλά αντίγραφα αυτού του τεμαχίου DNA με τη μορφή αγγελιαφόρου RNA, του mRNA. Τα πολλαπλά αντίγραφα του mRNA στη συνέχεια χρησιμοποιούνται για να μεταφράσουν τον γενετικό κώδικα και να βιοσυντεθεί η πρωτεΐνη [2]. Ειδικότερα το αγγελιαφόρο mRNA είναι ο φορέας πληροφοριών από το DNA στον τόπο της πρωτεϊνοσύνθεσης. Αλλά υπάρχουν και μόρια RNA με διαφορετικές δράσεις όπως τα μεταφορικά (tranfer-tRNA) και τα ριβοσωματικά (ribosomal-rRNA) που συμμετέχουν στην πρωτεϊνοσύνθεση. Περιληπτικά αναφέρουμε ότι τα tRNA επιλέγουν το κατάλληλο αμινοξύ και το μεταφέρουν σε κυτταρικές δομές, που αποτελούνται από νουκλεϊκά οξέα (rRNA), όπου και με την βοήθεια του mRNA φέρνουν τα αμινοξέα το ένα δίπλα στο άλλο με την κατάλληλη σειρά για να συνδεθούν με χημικούς (πεπτιδικούς) δεσμούς και να προκύψει η ορθή πρωτοταγής δομή της πρωτεΐνης. Τα rRNA έχουν και καταλυτική δράση σαν ένζυμα στα μόρια των νουκλεικών οξέων και θεωρούνται τα πρώτα ένζυμα που υπήρξαν στη φύση και μετά αντικαταστάθηκαν στη φύση από τα ένζυμα πρωτεϊνικής δομής.

Τα διάφορα είδη RNA είναι γνωστό, ότι συμμετέχουν επιπλέον και σε άλλες κυτταρικές διαδικασίες, κυρίως, σε συνδυασμό με ορισμένες πρωτεΐνες.

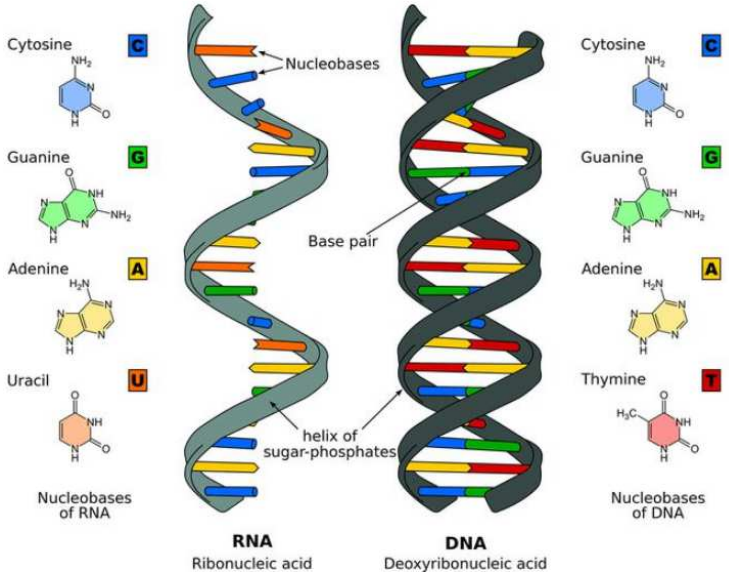
DNA:	TAC	CGC	GGC	TAT	TAC	TGC	CAG	GAA	GGA	ACT
RNA:	AUG	GCG	CCG	AUA	AUG	ACG	GUC	CUU	CCU	UGA
Protein:	Met	Ala	Pro	Ile	Met	Thr	Val	Leu	Pro	Stop

Εικόνα 1.5: Η μεταγραφή του DNA σε RNA και η μετάφραση του σε πρωτεΐνη
(Πηγή: http://bix.ucsd.edu/bioalgorithms)

1.1.1 Κύριες διαφορές μεταξύ του DNA και του RNA

Μεταξύ του DNA και του RNA, δυο είναι οι βασικότερες διαφορές. Η πρώτη διαφορά είναι ότι το RNA είναι μονόκλωνο μόριο που μπορεί και να σχηματίσει δίκλωνα τμήματα στο μόριό του, ενώ το

DNA αποτελείται από δυο έλικες. Μια επίσης σημαντική διαφορά είναι ότι το RNA περιέχει την ουρακίλη αντί της θυμίνης που περιέχεται στο DNA.



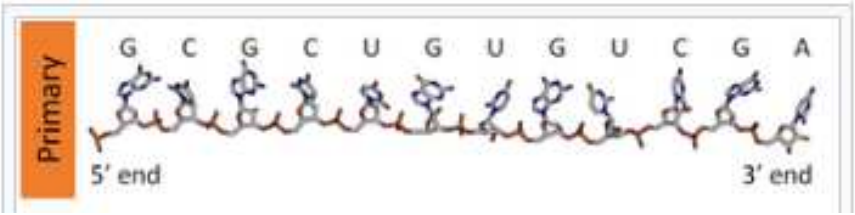
Εικόνα 1.6: Το μονόκλωνο RNA στα αριστερά και το δίκλωνο DNA στα δεξιά
(Πηγή: https://www.thoughtco.com/dna-versus-rna-608191)

Όπως ήδη προαναφέρθηκε, η λειτουργία των δυο αυτών μαρκομορίων είναι επίσης διαφορετική. Το DNA αποθηκεύει την γενετική πληροφορία ενώ το RNA λειτουργεί κυρίως ως εκμαγείο για την πρωτεϊνοσύνθεση.

1.2 Δομές του RNA

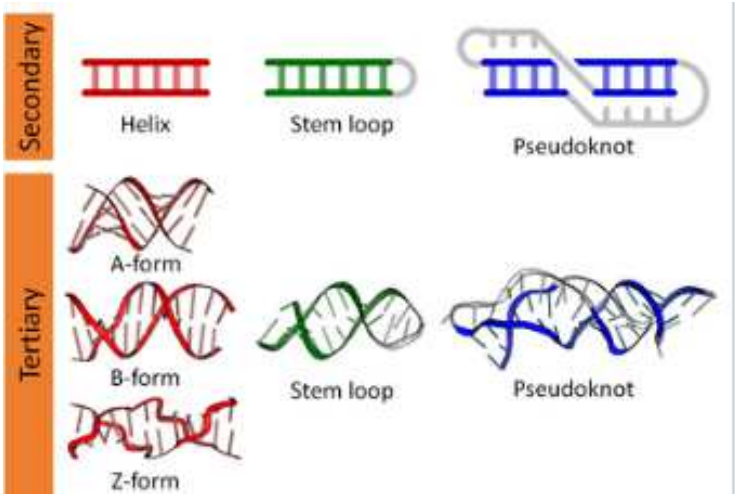
Οι λειτουργίες του RNA προσδιορίζονται από την δομή του. Όπως και στις πρωτεΐνες, τα διάφορα επίπεδα οργάνωσης του μορίου του RNA είναι η πρωτοταγής, η δευτεροταγής και η τριτοταγής δομή.

Η πρωτοταγής δομή του RNA αναφέρεται στην αλληλουχία των νουκλεοτιδίων αδενίνης, γουανίνης, κυτοσίνης και ουρακίλης.



Εικόνα 1.7: Πρωτοταγής δομή του RNA
(Πηγή: https://en.wikipedia.org)

Η δευτεροταγής δομή του RNA αναφέρεται στην διαμόρφωση-αναδίπλωση τμημάτων του μονόκλωνου μορίου RNA στον χώρο. Το RNA αναδιπλώνεται με τρόπο ώστε να σχηματίζονται δεσμοί υδρογόνου μεταξύ των συμπληρωματικών βάσεων. Η συνολική διαμόρφωση ολόκληρου του μορίου στο χώρο (η τρισδιάστατη δομή), αποτελεί την τριτοταγή δομή του RNA.



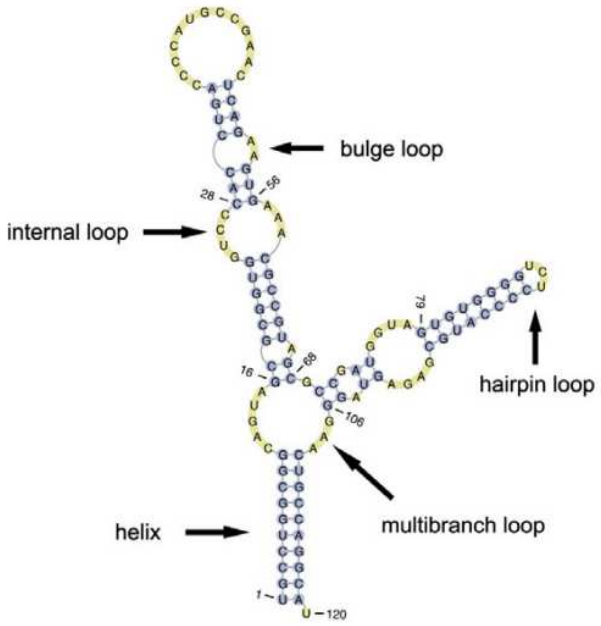
Εικόνα 1.8: Δευτεροταγής & τριτοταγής δομή του RNA
(Πηγή: https://en.wikipedia.org)

Η τριτοταγής δομή είναι ενεργειακά λιγότερο σταθερή σε σχέση με την δευτεροταγή δομή. Όπως συμβαίνει και με τις πρωτεΐνες, η αναδίπλωση του RNA κατά την οποία σχηματίζεται η δευτεροταγής δομή γίνεται πριν από την τριτοταγή δομή, δηλαδή την τελική διαμόρφωση όλου του μορίου στο χώρο. Επειδή ο σχηματισμός τριτοταγούς δομής συνήθως δεν προκαλεί αλλαγές στη δευτεροταγής δομή, οι δύο διαδικασίες μπορούν να περιγραφούν ανεξάρτητα. Λειτουργικά μόρια RNA (tRNAs, ριβοσωμικά RNA, κλπ.), έχουν συνήθως χαρακτηριστικές δομές περιοχές στον χώρο (domains) και συνεπώς χαρακτηριστικές δευτεροταγείς δομές που είναι προϋποθέσεις για την λειτουργία τους. Ως συνέπεια, οι δευτεροταγείς δομές διατηρούνται σε μεγάλο βαθμό στην εξέλιξη

για πολλές κατηγορίες μορίων RNA. Στο πλαίσιο αυτής της εργασίας θα μας απασχολήσει μόνο η δευτεροταγής δομή του RNA, η οποία αναλύεται εκτενέστερα παρακάτω.

1.2.1 Δευτεροταγής δομή του RNA

Όπως προαναφέρθηκε, το RNA αναδιπλώνεται ώστε οι συμπληρωματικές του βάσεις να ενώνονται με δεσμούς υδρογόνου σχηματίζοντας μια διπλή έλικα. Η δομή όμως που προκύπτει αποτελείται από κομμάτια συμπληρωματικών περιοχών (helix ή stem) - αντιστοίχιση βάσεων Watson-Crick (WC) - και από βρόγχους (loops), όπως φαίνεται στην *Εικόνα 1.9*. Οι βρόχοι είναι περιοχές του RNA που περιέχουν μη συζευγμένες βάσεις. Οι βρόχοι κατηγοριοποιούνται βάσει του αριθμού και του τρόπου που είναι στοιβαγμένα τα παραπλήσια τμήματα του μορίου και της θέσης τους εντός του μορίου του RNA.



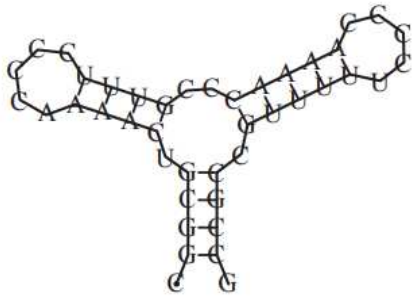
Εικόνα 1.9: Στοιχεία δευτεροταγούς δομής του RNA
(Πηγή: [8])

Το βασικό και πιο συνηθισμένο δομικό στοιχείο της δευτεροταγούς δομής του RNA είναι ο **βρόγχος φρουρκέτας** (Hairpin loop or stem-loop). Αποτελείται μια διπλή έλικα που καταλήγει σε έναν μη συζευγμένο βρόχο. Η δευτεροταγής δομή του RNA μπορεί να κατευθύνει την αναδίπλωση

του RNA, να προστατεύσει την δομική σταθερότητα για το αγγελιαφόρο RNA (mRNA), να παρέχει θέσεις αναγνώρισης για πρωτεΐνες δέσμευσης RNA και να χρησιμεύσει ως υπόστρωμα για ενζυμικές αντιδράσεις [1].

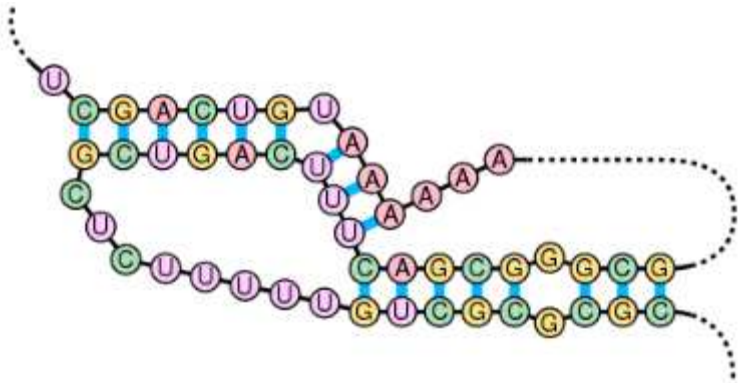
Οι **εσωτερικοί βρόχοι** (internal loops) στο RNA εντοπίζονται στο μέσο, στο εσωτερικό της αναδιπλωμένης αλυσίδας RNA. Οι εσωτερικοί βρόχοι είναι είτε συμμετρικοί είτε ασύμμετροι. Οι **ασύμμετροι βρόχοι** ονομάζονται εξογκώματα (bulge loops) και είναι ένας τύπος βρόχου όπου μόνο ένας κλώνος περιέχει μη συζευγμένες βάσεις.

Οι **πολλαπλοί διακλαδισμένοι βρόχοι** (multi-branch loop) είναι βρόχοι που αποτελούνται από τρεις ή περισσότερους βρόγχους. Αυτές οι δομές είναι επίσης αρκετά συχνές όπως και οι υπόλοιπες δομές δευτεροταγούς δομής. Ένα παράδειγμα διακλαδισμένου βρόχου φαίνεται στην Εικόνα 1.10.



Εικόνα 1.10: Διακλαδισμένος βρόχος (Multi-branch loop)
(Πηγή: [8])

Εκτός από τις προαναφερθείσες δομές, το RNA μπορεί να σχηματίζει πιο περίπλοκες δευτερεύουσες δομές οι οποίες αποτελούνται από μη ζευγαρωμένα ζεύγη βάσεων και ονομάζονται **ψευδοκόμβοι** (pseudoknots). Ψευδοκόμβος (Εικόνα 1.11) είναι μια τριτοταγής δομή των νουκλεϊκών οξέων στην οποία οι βρόχοι διασταυρώνονται και διακλαδώνονται με συνέπεια η μία πλευρά (το στέλεχος) του ενός να κάνει δεσμούς υδρογόνου με τη μία πλευρά του άλλου [7].



Εικόνα 1.11: Ψευδοκόμβος
(Πηγή: https://en.wikipedia.org/wiki/Pseudoknot)

1.3 Μέθοδοι προσδιορισμού δευτεροταγούς δομής RNA

Η πρωτεύουσα σημασία του RNA για κάθε τομέα της ζωής είναι αναμφισβήτητη. Το RNA κωδικοποιεί και χειρίζεται γενετική πληροφορία, αναστέλλει την αντιγραφή του DNA, χρησιμεύει ως πρότυπο για τη πρωτεϊνοσύνθεση, μεταφράζει DNA κώδικα σε πεπτίδια, καταλύει ενζυματικά αντιδράσεις αλλά και πολλές περισσότερες λειτουργίες. Επιπρόσθετα, η δομή του RNA μπορεί να λειτουργήσει ως στόχος φαρμάκου. Το πόσο σημαντική είναι η δομή σε ένα βιολογικό μόριο έχει αποδειχθεί στη δομική βιοχημεία όπου φαίνεται ότι οι βιολογικές δράσεις των μορίων εξαρτώνται και είναι απόρροια της δομής. Οι διαφορετικοί ρόλοι του RNA εξαρτώνται από την ικανότητά του να αναδιπλώνεται για να σχηματίσει τις δομές οι οποίες αναφέρθηκαν σε προηγούμενη παράγραφο. Συνεπώς για να καθοριστεί η εκάστοτε λειτουργία ενός μορίου RNA απαιτείται ο προσδιορισμός της δομής του κι αυτό είναι ένα σημαντικό θέμα που έχει απασχολήσει τη μοριακή βιολογία αλλά και την επιστήμη της πληροφορικής.

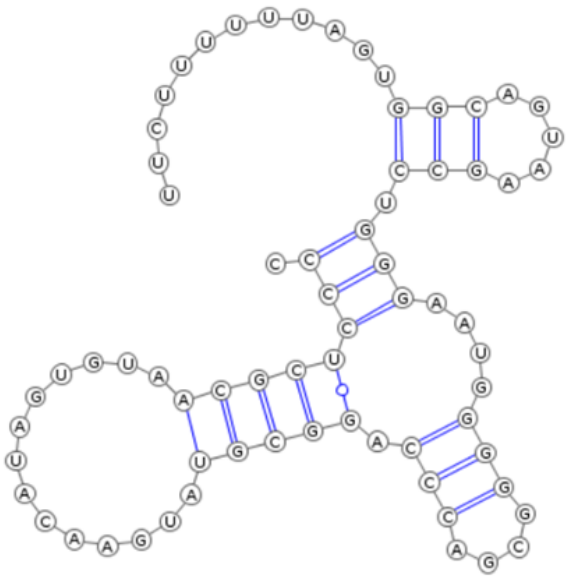
Το πρώτο βήμα για τον προσδιορισμό της δομής είναι ο προσδιορισμός της πρωτοταγούς δομής του, δηλαδή της αλληλουχίας των νουκλεοτιδίων στο μόριο του RNA. Μέχρι σήμερα έχουν γίνει γνωστές δισεκατομμύρια αλληλουχίες νουκλεοτιδίων. Το δεύτερο βήμα είναι ο προσδιορισμός της δευτεροταγούς δομής με βάση την αλληλουχία RNA. Πρέπει να σημειωθεί ότι η δευτεροταγής και τριτοταγής δομή του μορίου εξαρτώνται από την πρωτοταγή δομή. Στην εικόνα Εικόνα *1.12*.Α φαίνεται μια ακολουθία RNA και στην Εικόνα *1.12*.Β φαίνεται η αντίστοιχη δευτεροταγής δομή της σε απεικόνιση παρενθέσεων, όπου οι παρενθέσεις υποδηλώνουν ζεύγη βάσεων και οι κουκίδες

απεικονίζουν ελεύθερες – μη ζευγαρωμένες βάσεις. Στην Εικόνα *1.12*.Γ παρουσιάζεται η δευτεροταγής δομή της ακολουθίας RNA σε απεικόνιση γραφικού διαγράμματος.

```
A.
UUCUUUUUUAGUGGCAGUAAGCCUGGGAAUGGGGGCGACCCAGGCGUAUGAACAUAG
UGUAACGCUCCCC

B.
.............(((.....))).(((....(((....))).(((((..........
....)))))))

Γ.
```



Εικόνα 1.12: Α. Η ακολουθία RNA Β. Δευτεροταγής δομής σε αναπαράσταση παρενθέσεων Γ. Δευτεροταγής δομής σε αναπαράσταση γραφήματος

(Πηγή: https://rtips.dna.bio.keio.ac.jp/ipknot/process.php)

Οι διαθέσιμες γνωστές δευτεροταγείς δομές είναι πολύ λιγότερες σε σχέση με τις διαθέσιμες γνωστές αλληλουχίες και οι τριτοταγείς δομές είναι ακόμα πιο λίγες. Ο προσδιορισμός της δευτεροταγούς δομής μπορεί να πραγματοποιηθεί είτε με πειραματική επαλήθευση είτε με μεθόδους πρόβλεψης χρησιμοποιώντας υπολογιστικά εργαλεία.

1.3.1 Πειραματική επαλήθευση

Η κρυσταλλογραφία με χρήση ακτίνων-Χ (X-ray crystallography), η φασματοσκοπία NMR (Nuclear magnetic resonance spectroscopy) και η χημική ανίχνευση είναι οι πιο ευρέως διαδεδομένες εργαστηριακές μέθοδοι που χρησιμοποιούνται για τον καθορισμό της δομής του RNA και τα πειραματικά αποτελέσματα επιδεικνύουν την μεγαλύτερη ακρίβεια καθορισμού της δομής, από όλες τις μεθόδους που έχουν αναπτυχθεί για τον ίδιο σκοπό. Είναι όμως χρονοβόρες και κοστίζουν αρκετά στην εργαστηριακή εκτέλεση και πολλάκις είναι αδύνατον να εφαρμοστούν.

Τα αποτελέσματα των μεθόδων αυτών αποθηκεύονται σε βάσεις δεδομένων. Υπάρχουν δεκάδες διαθέσιμες εξειδικευμένες βάσεις δεδομένων, όπως η RFAM (http://rfam.xfam.org/) [28], η **MiRBase (**http://www.mirbase.org) [62], η **MirTarBase** (http://mirtarbase.mbc.nctu.edu.tw) [63] και η **TarBase** (http://diana.imis.athena-innovation.gr/DianaTools/index.php?r=tarbase/index) [49].

1.3.1.1 Κρυσταλλογραφία με χρήση ακτίνων Χ

Η κρυσταλλογραφία με χρήση ακτίνων Χ (X-ray crystallography) είναι μια τεχνική για τον προσδιορισμό της λεπτομερούς δομής των μορίων, συμπεριλαμβανομένων πολύπλοκων βιολογικών μακρομορίων όπως οι πρωτεΐνες και τα νουκλεϊνικά οξέα. Η τεχνική αυτή χρησιμοποιείται ήδη από τις αρχές του εικοστού αιώνα.

Η βασική προσέγγιση είναι η ίδια με αυτή της κρυσταλλογραφίας των πρωτεϊνών στο ότι ένα μακρομόριο υποβάλλεται σε μια ποικιλία συνθηκών και θερμοκρασιών διαλύματος, αλλιώς γνωστό ως «χώρος συνθηκών», σε μια προσπάθεια να αναπτυχθεί ένας κρύσταλλος διάθλασης. Τα άτομα του κρυστάλλου κάνουν μια προσπίπτουσα δέσμη ακτίνων Χ να διαθλάται με συγκεκριμένο τρόπο. Μετρώντας τις γωνίες και τις εντάσεις αυτών των διαφωτισμένων δομών, μπορεί κανείς να παράγει μια τρισδιάστατη εικόνα της πυκνότητας των ηλεκτρονίων μέσα στον κρύσταλλο. Από αυτή την πυκνότητα ηλεκτρονίων, μπορούν να προσδιοριστούν οι μέσες θέσεις των ατόμων στον κρύσταλλο, καθώς και οι χημικοί δεσμοί τους, η κρυσταλλογραφική τους διαταραχή και διάφορες άλλες πληροφορίες [50]. Το RNA έχει χαρακτηριστικά που μπορούν να αξιοποιηθούν στην κρυσταλλογραφία συμπεριλαμβανομένου του μειωμένου αλφαβήτου δομικών στοιχείων (4 βάσεις

σε αντίθεση με 20 αμινοξέα στις πρωτεΐνες) και της κυριαρχίας του ζεύγους βάσεων και της έλικας μορφής Α. Επιπλέον, σχεδόν οποιαδήποτε αλληλουχία RNA μικρότερη από 400 νουκλεοτίδια σε μήκος είναι εύκολο να παρασκευαστεί σε μεγάλες ποσότητες, αναδιπλώνεται στην φυσική κατάσταση και είναι διαλυτή στις υψηλές συγκεντρώσεις (10-25 mg/ml) που είναι αναγκαία για την ανάπτυξη κρυστάλλων [8]. Τα μειονεκτήματα αυτής της μεθόδου είναι ότι οι διαδικασίες κρυσταλλοποίησης είναι ακριβές, χρονοβόρες και όχι πλήρως επαναλήψιμες. Επιπλέον, η διαμορφωτική πλαστικότητα του RNA περιορίζει την εφαρμογή της κρυσταλλογραφίας ακτίνων Χ.

1.3.1.2 Φασματοσκοπία NMR

Μια άλλη εργαστηριακή μέθοδος που χρησιμοποιείται για τον προσδιορισμό της δομής του RNA είναι η φασματοσκοπία πυρηνικού μαγνητικού συντονισμού (nuclear magnetic Resonance NMR), η χρήση δηλαδή φασματοσκοπίας πυρηνικού μαγνητικού συντονισμού για τη λήψη πληροφοριών σχετικά με τη δομή και τη δυναμική των μορίων νουκλεϊκού οξέος, όπως το DNA ή το RNA [56]. Η φασματοσκοπία NMR, με την ικανότητά της να εκτελεί δομικές μελέτες ατομικής ανάλυσης σε διάλυμα, είναι ένα ισχυρό εργαλείο για τη μελέτη νουκλεϊνικών οξέων αφού σχεδόν οι μισές από όλες τις τρέχουσες γνωστές δομές RNA προσδιορίστηκαν χρησιμοποιώντας τεχνικές NMR.

Η φασματοσκοπιά NMR χρησιμοποιείται, κυρίως, για την εύρεση της δομής μικρών μορίων RNA μεγέθους έως και 100 νουκλεοτιδίων. Έχει πλεονεκτήματα έναντι της κρυσταλλογραφίας ακτίνων Χ, καθώς τα μόρια παρατηρούνται στην κατάσταση του φυσικού διαλύματός τους παρά σε ένα κρυσταλλικό πλέγμα που μπορεί να επηρεάσει τις δομικές ιδιότητες του μορίου. Επιπλέον το κόστος είναι μικρότερο σε σχέση με την κρυσταλλογραφία ακτίνων Χ αλλά οι δομές που προκύπτουν στην φασματοσκοπία είναι λιγότερο λεπτομερείς.

1.3.1.3 Χημική ανίχνευση

Η χημική ανίχνευση είναι μια άλλη πειραματική μέθοδος που χρησιμοποιείται ευρέως ως μια γρήγορη προσέγγιση για τον καθορισμό 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 σύγκριση με την κρυσταλλογραφία ακτίνων Χ και την φασματοσκοπία NMR, η χημική ανίχνευση είναι γρήγορη και εύκολη και προσφέρει έναν απλό και ακριβή τρόπο ανάλυσης της δομής του RNA. Επιπλέον, η χημική ανίχνευση συμπλόκων RNA και RNA/πρωτεΐνης μπορεί να δώσει μεγάλα σύνολα δεδομένων και επομένως λεπτομερείς πληροφορίες σχετικά με τη δομή του RNA. Όπως και σε άλλες πειραματικές προσεγγίσεις, η χημική ανίχνευση είναι εγγενώς θορυβώδης και τα πειραματικά αποτελέσματα εξαρτώνται από τις συνθήκες (π.χ. το PH, η ιοντική ισχύς κ.α.) κάτω από τις οποίες λαμβάνει χώρα το πείραμα. Ως εκ τούτου, η δομή που προκύπτει μπορεί να αποκλίνει από την πραγματική δομή του μορίου.

1.3.2 Πρόβλεψη δευτεροταγούς δομής - Υπολογιστικές μέθοδοι

Σύμφωνα με τα προαναφερθέντα, οι πλήρως ακριβείς μέθοδοι προσδιορισμού της δομής του RNA είναι η κρυσταλλογραφία με τη χρήση ακτίνων Χ και η φασματοσκοπία NMR. Ωστόσο, αυτές οι μέθοδοι είναι χρονοβόρες, πολύ δαπανηρές και πολλάκις ανεφάρμοστες αφού συνήθως τα βιολογικά δείγματα προς μελέτη είναι σε ελάχιστες ποσότητες. Επομένως έχει δημιουργηθεί η ανάγκη για αντικατάσταση αυτών των χρονοβόρων και ακριβών εργαστηριακών μεθόδων, προκειμένου να επιτευχθεί γρήγορη εξαγωγή αποτελεσμάτων. Για τον σκοπό αυτό, έχουν αναπτυχθεί υπολογιστικές μέθοδοι πρόβλεψης. Η πρόβλεψη δευτεροταγούς δομής του RNA είναι η διαδικασία πρόβλεψης της θέσης των δεσμών υδρογόνου σε ένα μόριο RNA που βασίζεται μόνο στην αλληλουχία νουκλεοτιδίων του. Οι κυριότερες υπολογιστικές μέθοδοι που χρησιμοποιούνται ευρέως στον τομέα της πρόβλεψης της δευτεροταγής δομής του RNA είναι η συγκριτική ανάλυση ακολουθίας, οι μέθοδοι πρόβλεψης χρησιμοποιώντας κανόνες θερμοδυναμικής και οι μέθοδοι γραμματικών προσεγγίσεων.

Μια αλληλουχία RNA μπορεί να αναδιπλωθεί σε μεγάλο αριθμό θεωρητικά πιθανών δευτεροταγών δομών. Το πλήθος τους αυξάνεται εκθετικά με το μήκος της αλληλουχίας. Για μια αλληλουχία με n νουκλεοτίδια, οι πιθανές δευτεροταγής δομές είναι περισσότερες από $1{,}8^n$, σύμφωνα με τους Zuker και Sankoff [51], υποθέτοντας ότι και τα τέσσερα νουκλεοτίδια (A, C, G και U) εμφανίζονται με την ίδια πιθανότητα και τυχαία στην αλληλουχία RNA. Για παράδειγμα, ένα RNA των 200 νουκλεοτιδίων έχει πάνω από 10^{50} πιθανές δευτεροταγείς δομές. Χρειαζόμαστε επομένως έναν

αλγόριθμο δυναμικού προγραμματισμού για να διακρίνουμε τη βιολογικά σωστή δομή από τις υπόλοιπες.

Σημαντική παράμετρος της πρόβλεψης της δευτεροταγούς δομής του RNA είναι ότι υπάρχουν πολλές αλληλουχίες των οποίων οι δομές δεν έχουν ακόμη προσδιοριστεί πειραματικά και για τις οποίες δεν υπάρχουν εγγραφές στις βάσεις δεδομένων. Γι' αυτό καθίσταται αναγκαία η πρόβλεψη δομής. Από την άλλη πλευρά δεν υπάρχει μέθοδος που να προβλέπει τέλεια τη δομή. Οι υπολογιστικές μέθοδοι που έχουν αναπτυχθεί παρουσιάζουν καλές επιδόσεις, παρ' όλα αυτά δεν έχουν φτάσει σε τέτοιο επίπεδο, ώστε να μπορούν να επιτύχουν την ακρίβεια των εργαστηριακών μεθόδων. Είναι όμως φθηνές και πολύ γρήγορες στην εκτέλεση.

Γενικά, η επιτυχία μιας μεθόδου πρόγνωσης εξαρτάται από το είδους του αλγορίθμου, από το μέγεθος του συνόλου - οι μέθοδοι που χρησιμοποίησαν μεγαλύτερα σύνολα αποδίδουν καλύτερα - και από το αν χρησιμοποιεί πολλαπλές στοιχίσεις (οι μέθοδοι που χρησιμοποιούν πολλαπλές στοιχίσεις αποδίδουν πάντα καλύτερα). Μελέτες ανασκόπησης βασισμένες στα δημοσιευμένα αποτελέσματα μεθόδων πρόγνωσης έχουν δείξει ότι με τις παρούσες μεθοδολογίες, οι μέθοδοι πρόγνωσης έχουν ένα ανώτατο όριο στην αναμενόμενη ακρίβεια και, μάλιστα, από ένα σημείο και μετά η απόδοση δεν αυξάνει [7].

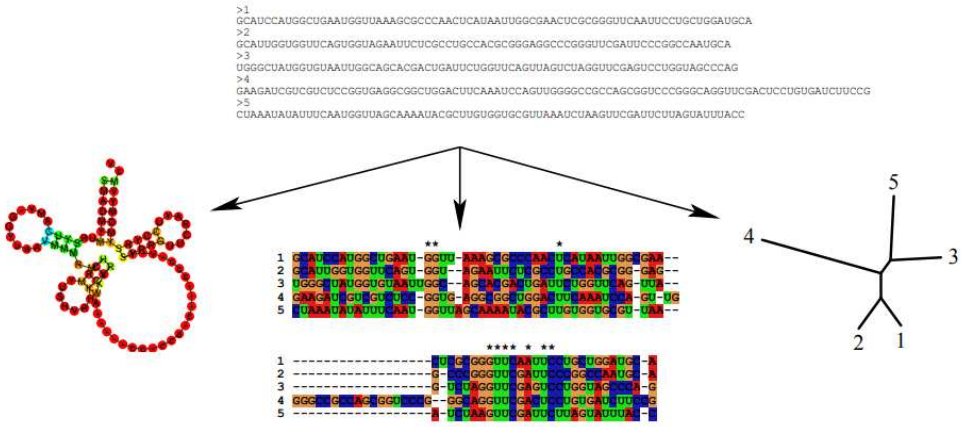
Ωστόσο, οι συνδυαστικές μέθοδοι μπορούν να αυξήσουν την απόδοση οποιασδήποτε μεθόδου πρόγνωσης. Έχουν αναπτυχθεί αλγόριθμοι που συνδυάζουν την θερμοδυναμική και την σύγκριση αλληλουχιών με τα πειραματικά αποτελέσματα που προκύπτουν από την χημική ανάλυση, και/ή την φασματοσκοπία NMR. Για παράδειγμα έχουν προταθεί διάφορες προσεγγίσεις για την ενσωμάτωση μελέτης χημικών δεδομένων σε υπολογιστικά εργαλεία με βάση τη θερμοδυναμική.

1.3.2.1 Συγκριτική ανάλυση αλληλουχίας

Όταν είναι διαθέσιμος ένας μεγάλος αριθμός ακολουθιών RNA, η πρότυπη τεχνική για τον προσδιορισμό της δευτεροταγούς δομής είναι η συγκριτική ανάλυση της πρωτοταγούς δομής δηλαδή της αλληλουχίας των βάσεων. Η μέθοδος αυτή βασίζεται σε δύο θεμελιώδεις υποθέσεις: πρώτον ότι διαφορετικές, ομόλογες αλληλουχίες RNA είναι ικανές να διπλώνονται στην ίδια

δευτεροταγή και τριτοταγή δομή και δεύτερον ότι κατά τη διάρκεια της εξέλιξης, η τριτοταγής δομή ενός μορίου RNA παραμένει ως επί το πλείστον αμετάβλητη, ενώ η πρωτοταγής δομή μπορεί να αλλάξει σημαντικά.

Οι μέθοδοι ανάλυσης συγκριτικής αλληλουχίας RNA προβλέπουν μια δομή αναζητώντας μια στοίχιση για ζεύγη βάσεων που είναι κοινές σε όλες τις αλληλουχίες στο σύνολο δεδομένων. Μπορεί να προβλέψει με ακρίβεια κανονικά και μη κανονικά ζεύγη βάσεων που εμφανίζονται σε δευτερεύουσες και τριτογενείς δομές. Μια προσέγγιση της μεθόδου αυτής φαίνεται στην Εικόνα 1.13.



Εικόνα 1.13: Είσοδος: *ένα σύνολο ακολουθιών με την ίδια βιολογική λειτουργία που θεωρείται ότι έχουν περίπου την ίδια δομή,* ***Έξοδος:*** *τα κοινά δομικά στοιχεία, οι ευθυγραμμισμένες ακολουθίες και μια φυλογενία που εξηγεί καλύτερα τα παρατηρούμενα δεδομένα*

(Πηγή:RNA structure analysis)

Η συγκριτική ανάλυση συμβάλλει στη μοριακή βιολογία με τους εξής τρόπους: (1) στον καθορισμό των περιορισμών και των μοτίβων από τα οποία μπορεί να συναχθεί η μοριακή δομή (2) ως απαραίτητο υπόβαθρο και συμπληρωματικό στοιχείο της πειραματικής ανάλυσης της μοριακής λειτουργίας και δομής (3) ως ουσία της γενεαλογικής ανάλυσης και ταξινόμησης και (4) στην παροχή ενός γενικού εννοιολογικού και οργανωτικού πλαισίου για μεγάλο μέρος της μελλοντικής βιολογικής έρευνας.

Καμία από τις φυσικοχημικές προσεγγίσεις δεν μπορεί να παρέχει τόσο λεπτομερή εικόνα των σχέσεων μεταξύ των βάσεων σε μεγάλα RNAs όσο η συγκριτική ανάλυση. Η μέθοδος αυτή μπορεί να διευκολυνθεί με προβλέψεις δευτεροταγών δομών με βάση την ελαχιστοποίηση της ελεύθερης

ενέργειας. Όταν υπάρχουν μόνο μία ή λίγες γνωστές αλληλουχίες για ένα RNA, η ελαχιστοποίηση ελεύθερης ενέργειας μπορεί επίσης να χρησιμοποιηθεί για την πρόβλεψη μοντέλων δευτεροταγούς δομής που μπορούν να εξεταστούν σε σχέση με πειραματικά δεδομένα όπως παραδείγματος χάριν αυτά που προκύπτουν από την χημική ανίχνευση.

Η συγκριτική ανάλυση RNA, ωστόσο, είναι μια επαναληπτική διαδικασία που απαιτεί σημαντικά δεδομένα αλληλουχίας, ακριβείς στοιχίσεις αλληλουχιών και την ανάλυση μιας δομής που είναι κοινή σε όλες τις αλληλουχίες στο σύνολο δεδομένων. Η προσέγγιση αυτή είναι εξαιρετικά ακριβής, αλλά σε πολλές περιπτώσεις δεν είναι εφικτή. Οι μέθοδοι που εφαρμόζονται πραγματικά εξαρτώνται από τα διαθέσιμα δεδομένα, το επιθυμητό επίπεδο ακρίβειας (συμβιβασμός μεταξύ ποιότητας και κόστους) και τα συγκεκριμένα ζητήματα που πρέπει να αντιμετωπιστούν από τον ερευνητή. Οι αλληλουχίες που πρέπει να συγκριθούν πρέπει να είναι επαρκώς παρόμοιες ώστε να μπορούν να στοιχηθούν, αλλά πρέπει να είναι αρκετά ανόμοιες ώστε να μπορεί να ανιχνευθεί ένας αριθμός υποκαταστάσεων (substitutions). Επιπλέον προϋποθέτει μεγάλη υπολογιστική ισχύ και αποθηκευτικό χώρο.

1.3.2.2 Θερμοδυναμικές προσεγγίσεις

Η πρόβλεψη δευτεροταγούς δομής του RNA, χρησιμοποιώντας τους νόμους της θερμοδυναμικής, μπορεί να χρησιμοποιηθεί για την ανάπτυξη υποθέσεων σχετικά με τη δομή μιας αλληλουχίας βάσεων στο RNA. Η πρόβλεψη της δομής επιτυγχάνεται με την εύρεση της χαμηλότερης ενεργειακά δομής, ελάχιστη ελεύθερη ενέργεια (Minimum Free Energy - MFE), η οποία σύμφωνα με τη θερμοδυναμική είναι η πιο σταθερή και η πιο πιθανή δομή στην θερμοδυναμική ισορροπία.

Η βέλτιστη δομή ανάμεσα σε ένα σύνολο πιθανών δευτεροταγών δομών μιας δεδομένης αλληλουχίας RNA προσδιορίζεται με την χρήση αλγορίθμων δυναμικού προγραμματισμού. Οι αλγόριθμοι δυναμικού προγραμματισμού είναι υπολογιστικά αποτελεσματικοί αφού συνήθως επιστρέφουν το ίδιο αποτέλεσμα και εγγυώνται ότι θα βρεθεί η δομή χαμηλότερης ελεύθερης ενέργειας, λαμβάνοντας υπόψη τους κανόνες για τον προσδιορισμό της σταθερότητας. Η πρόβλεψη δευτεροταγών δομών RNA χρησιμοποιώντας υπολογιστικές μεθόδους που προσπαθούν να ελαχιστοποιήσουν την ελεύθερη ενέργεια δεν είναι καινούργια. Ο πρώτος αλγόριθμος δυναμικού

προγραμματισμού για τον υπολογισμό της χαμηλότερης ενεργειακά δομής ενός RNA δημοσιεύθηκε το 1981 από τους Zuker και Stiegler [12]. Το πιο ευρέως χρησιμοποιούμενο λογισμικό που ενσωματώνει αλγορίθμους ελάχιστης ελεύθερης ενέργειας είναι τα Mfold [64] και το RNAfold στο ViennaRNA λογισμικό [65].

Ωστόσο, ένα από τα μειονεκτήματα της θερμοδυναμικής μεθόδου είναι ότι οι περισσότεροι αλγόριθμοι δεν μπορούν να προβλέψουν μη κανονικά ζεύγη βάσεων και ζεύγη βάσεων που δεν βρίσκονται σε τυπικές έλικες. Επιπλέον, τα αποτελέσματα των υπολογιστικών προγραμμάτων δεν είναι πάντα αξιόπιστα. Υπάρχει πολύ μεγαλύτερη τάση για τη σωστή πρόβλεψη δομών φουρκέτας παρά για ζεύγη μεγάλων αποστάσεων.

Τέλος, ένα πρόβλημα που κάνει το όλο θέμα ακόμα πιο πολύπλοκο και δεν έχει ακόμα αντιμετωπιστεί - αναφερθεί στη μελέτη της δευτεροταγούς δομής του RNA στη βιβλιογραφία, είναι και κάποιες νουκλεικές βάσεις - πλην των 4 που αναφέρθηκαν - που υπάρχουν σε μικρά ποσοστά στο μόριο του RNA. Και η έρευνα στη Βιοχημεία έχει δείξει ότι τα βιολογικά μόρια που υπάρχουν σε πολύ μικρά ποσοστά συνήθως έχουν πολύ σημαντικούς ρόλους στη φύση.

1.3.2.3 Γραμματικές προσεγγίσεις

Η χρήση των γραμματικών για την ανάλυση της δομής του RNA έχει δύο μορφές. Μια γραμματική μπορεί να χρησιμοποιηθεί είτε για την πρόβλεψη της πιο πιθανής δομής δεδομένης μιας αλληλουχίας RNA είτε για την κατασκευή ενός μοντέλου οικογένειας αλληλουχιών RNA με διατηρημένη δομή. Η κατηγορία των γραμματικών που χρησιμοποιούνται κυρίως είναι οι στοχαστικές γραμματικές χωρίς συμφραζόμενα (stochastic context free grammars - SCFG) αλλά και παραλλαγές αυτών όπως είναι οι phylo-SCFGS [21], [60], SCFG γραμματικές σε συνδυασμό με μοντέλα μοριακής εξέλιξης, και οι profile-SCFGS [34], [61], γνωστά και ως μοντέλα συνδιακύμανσης (covariance models). Επίσης μπορούν να χρησιμοποιηθούν σε συνδυασμό με τις άλλες δυο προαναφερόμενες μεθόδους όπως η θερμοδυναμική και η συγκριτική ανάλυση ακολουθιών.

Οι γραμματικές προσεγγίσεις χρησιμοποιήθηκαν για την πρόβλεψη της δευτεροταγούς δομής του RNA στις αρχές της δεκαετίας του '90 από τον Sakakibara (1994) [40] και τους Eddy και Durbin (1994) [41]. Στα επόμενα κεφάλαια ακολουθεί εκτενής ανάλυση των γραμματικών προσεγγίσεων στην πρόβλεψη της δομής του RNA.

2 Επισκόπηση γραμματικών και αλγορίθμων

Σκοπός αυτού του κεφαλαίου είναι μια σύντομη επισκόπηση των γραμματικών και πώς αυτές μπορούν να εφαρμοστούν στην μοντελοποίηση της δευτεροταγούς δομής του RNA.

Σύμφωνα με την θεωρία των τυπικών γλωσσών (formal language theory), μια γραμματική είναι ένα σύνολο κανόνων παραγωγής συμβολοσειρών μιας επίσημης γλώσσας [55]. Οι κανόνες περιγράφουν τον τρόπο σχηματισμού συμβολοσειρών από το αλφάβητο της γλώσσας που ισχύουν σύμφωνα με τη σύνταξη της γλώσσας. Μια γραμματική δεν περιγράφει την έννοια των συμβολοσειρών – μόνο την μορφή τους [54].

2.1 Τυπικές γραμματικές

Μια τυπική γραμματική συμβολίζεται με G (N, Σ, P, S) και αποτελείται από τα ακόλουθα στοιχεία [54]:

- Ένα πεπερασμένο σύνολο N από μη τερματικά σύμβολα
- Ένα πεπερασμένο σύνολο Σ από τερματικά σύμβολα, όπου $\Sigma \cap N = \varnothing$
- Ένα πεπερασμένο σύνολο P από κανόνες παραγωγής
- Ένα αρχικό σύμβολο S

Για την δημιουργία μιας συμβολοσειράς, ξεκινάμε από ένα αρχικό σύμβολο S (start symbol) και εφαρμόζουμε τους κανόνες παραγωγής μέχρι να δημιουργηθεί μια συμβολοσειρά που περιέχει μόνο τερματικά σύμβολα. Τα σύμβολα χωρίζονται σε δυο κατηγορίες, στα μη-τερματικά σύμβολα, τα οποία συνήθως συμβολίζονται με κεφαλαία γράμματα και τα τερματικά σύμβολα που συμβολίζονται με πεζά γράμματα. Τα μη-τερματικά σύμβολα δεν εμφανίζονται στις τελικές συμβολοσειρές ενώ τα τερματικά εμφανίζονται. Σε έναν κανόνα παραγωγής της μορφής $A \rightarrow xB$, το μη-τερματικό σύμβολο A, στην αριστερή πλευρά του κανόνα, αντικαθίσταται από ένα σύνολο τερματικών και/ή μη-τερματικών συμβόλων στη δεξιά πλευρά του κανόνα. Στην μοντελοποίηση προτάσεων τα τερματικά είναι οι λέξεις. Αντίστοιχα, στην μοντελοποίηση των RNA ακολουθιών τα τερματικά είναι τα νουκλεοτίδια. Παρακάτω παρατίθεται ένα παράδειγμα από την αναφορά [54] που δείχνει την δημιουργία συμβολοσειρών με βάση τους κανόνες παραγωγής μιας γραμματικής.

Έστω ότι το αλφάβητο αποτελείται από τα σύμβολα a και b, το σύμβολο έναρξης S και τους παρακάτω κανόνες παραγωγής:

1. $S \rightarrow aSb$
2. $S \rightarrow ba$

Ξεκινάμε με το αρχικό σύμβολο *S* και επιλέγουμε έναν κανόνα που θα εφαρμοστεί σε αυτό. Εάν επιλέξουμε τον κανόνα 1, παράγουμε τη συμβολοσειρά aSb. Αν στη συνέχεια επιλέξουμε ξανά τον κανόνα 1, θα αντικαταστήσουμε το S με aSb και θα πάρουμε τη συμβολοσειρά aaSbb. Αν τώρα επιλέξουμε τον κανόνα 2, αντικαθιστούμε το S με ba και παίρνουμε τη συμβολοσειρά aababb, η οποία αποτελείται μόνο από τερματικά σύμβολα. Η δημιουργία των συμβολοσειρών που προκύπτουν από αυτή τη διαδικασία ονομάζεται παραγωγή (derivation) από τη γραμματική και μπορεί να παρασταθεί και ως εξής:

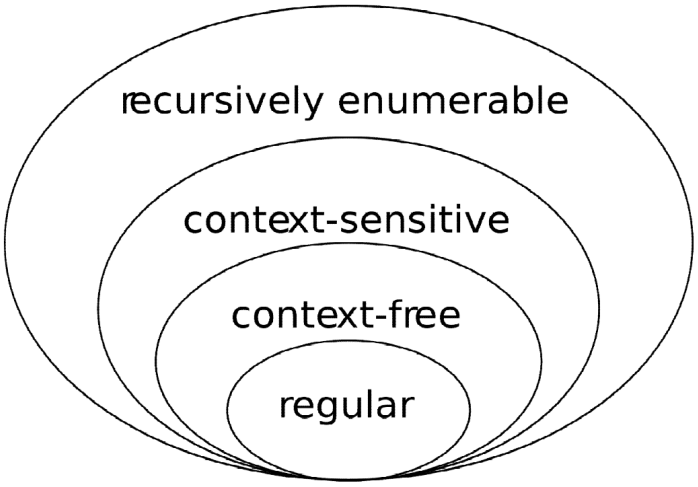
$$S \rightarrow aSb \rightarrow aaSbb \rightarrow aababb$$

Η γλώσσα της τυπικής γραμματικής είναι το άπειρο σύνολο $\{a^n bab^n \mid n \geq 0\}$ = {ba, abab, aababb, aaababbb, …}.

Στην επόμενη ενότητα παρατίθεται η κατηγοριοποίηση των γραμματικών που προτάθηκε από τον γλωσσολόγο Chomsky [55].

2.2 Η ιεραρχία Chomsky

Το 1956 ο Chomsky κατηγοριοποίησε τις γραμματικές με βάση τους τύπους των κανόνων παραγωγής [55]. Η διαφορά μεταξύ αυτών των τύπων είναι ότι έχουν ολοένα αυστηρότερους κανόνες παραγωγής και συνεπώς μπορούν να εκφράσουν λιγότερες τυπικές γλώσσες. Οι τέσσερις τύποι με αυξανόμενη σειρά πολυπλοκότητας είναι: κανονικές γραμματικές (regular grammars), γραμματικές χωρίς συμφραζόμενα (context free grammars), γραμματικές με συμφραζόμενα (context sensitive grammars) και γενικές γραμματικές (unrestricted grammars). Η ιεραρχία κατά Chomsky φαίνεται στην Εικόνα 2.1.



Εικόνα 2.1: Ιεραρχία των γραμματικών κατά Chomsky
(Πηγή: https://en.wikipedia.org/wiki/Chomsky_hierarchy)

Παρακάτω ακολουθείται ο συμβολισμός της αναφοράς [3], όπου το γράμμα W συμβολίζει μη-τερματικό σύμβολο, το γράμμα a συμβολίζει τερματικό σύμβολο, τα α και γ συμβολίζουν συμβολοσειρές με μη-τερματικά και/ή τερματικά σύμβολα συμπεριλαμβανομένης της κενής συμβολοσειράς και το β συμβολίζει συμβολοσειρές με μη-τερματικά και/ή τερματικά σύμβολα εξαιρουμένης της κενής συμβολοσειράς.

Στις κανονικές γραμματικές οι κανόνες παραγωγής είναι της μορφής W → aW ή W → a, και παράγουν προτάσεις από τα αριστερά προς τα δεξιά. Το πρώτο μέλος του κανόνα παραγωγής αποτελείται μόνο από ένα μη τερματικό σύμβολο, ενώ το δεύτερο μέλος περιέχει μια ακολουθία τερματικών συμβόλων και ένα μη τερματικό σύμβολο. Σύμφωνα με τον Chomsky, μια κανονική γραμματική δεν μπορεί να περιγράψει γλώσσες με συσχετισμούς απομακρυσμένων τερματικών συμβόλων, όπως είναι οι παλίνδρομες γλώσσες – η συμβολοσειρά μια παλίνδρομης γλώσσας είναι της μορφής abba, abaaba κ.τ.λ. - και οι γλώσσες αντιγραφής - η συμβολοσειρά μια τέτοιας γλώσσας είναι της μορφής aa, abab, abbabb κ.τ.λ..

Στις γραμματικές χωρίς συμφραζόμενα οι κανόνες παραγωγής είναι της μορφής W → β. Το πρώτο μέλος του κανόνα παραγωγής αποτελείται μόνο από ένα μη τερματικό σύμβολο, ενώ το δεύτερο μέλος μπορεί περιέχει οποιαδήποτε ακολουθία τερματικών και μη- τερματικών συμβόλων. Τέτοιες γραμματικές μπορούν να περιγράψουν παλίνδρομες γλώσσες. Αυτός είναι και ο λόγος που οι

γραμματικές χωρίς συμφραζόμενα χρησιμοποιούνται στην μοντελοποίηση των ζευγών βάσεων του RNA, όπως περιγράφεται εκτενέστερα στο κεφάλαιο 2.3.

Στις γραμματικές με συμφραζόμενα οι κανόνες παραγωγής είναι της μορφής $\alpha_1 W \alpha_2 \rightarrow \alpha_1 b \alpha_2$. Το πρώτο και το δεύτερο μέλος του κανόνα παραγωγής αποτελούνται από τουλάχιστον ίσο αριθμό συμβόλων ή μεγαλύτερο, αφού το b δεν μπορεί να είναι η κενή συμβολοσειρά. Τέτοιες γραμματικές μπορούν να περιγράψουν γλώσσες αντιγραφής.

Στις γενικές γραμματικές οι κανόνες παραγωγής είναι της μορφής $a_1 W a_2 \rightarrow \gamma$. Τα δυο μέλη των κανόνων παραγωγής μπορεί να είναι οποιοσδήποτε συνδυασμός συμβόλων. Από μια οποιαδήποτε συμβολοσειρά μπορεί να παραχθεί οποιαδήποτε άλλη (ή και η ίδια) συμβολοσειρά.

2.3 Στοχαστικές γραμματικές χωρίς συμφραζόμενα

Οι στοχαστικές γραμματικές χωρίς συμφραζόμενα και η χρήση τους στην πρόγνωση της δομής του RNA είναι το κυρίως θέμα της παρούσας εργασίας. Ωστόσο, δεν είναι αυτονόητος ο τρόπος που αυτές οι γραμματικές βρίσκουν εφαρμογή στο πρόβλημα της πρόγνωσης. Αυτός είναι και ο σκοπός αυτού του κεφαλαίου.

2.3.1 Στοχαστικές γραμματικές χωρίς συμφραζόμενα στην πρόβλεψη δομής του RNA

Όπως αναφέρθηκε στο κεφάλαιο 2.2, οι γραμματικές χωρίς συμφραζόμενα μπορούν να περιγράψουν την δευτεροταγή δομή του RNA. Οι παρακάτω κανόνες παραγωγής μπορεί να ανήκουν σε μια γραμματική χωρίς συμφραζόμενα:

1. $S \rightarrow aSa'$
2. $S \rightarrow aa'$

Οι κανόνες αυτοί εξασφαλίζουν την παραγωγή συμβολοσειρών που περιέχουν τόσα a όσα και a'. Για κάθε παραγωγή ενός a εξασφαλίζεται και η παραγωγή ενός a'. Αν τα a όσα και a'

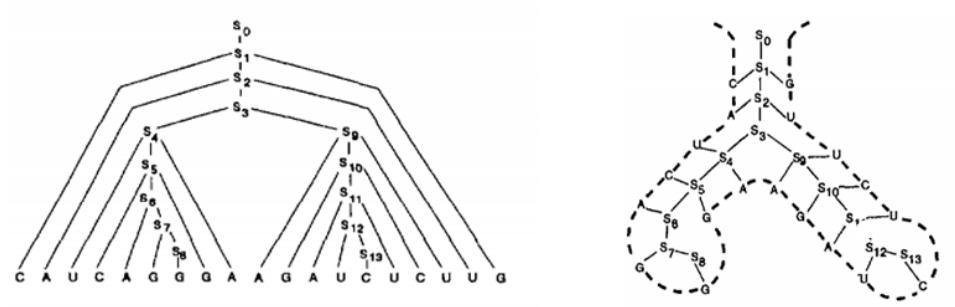
αντικατασταθούν με τα συμπληρωματικά ζεύγη του RNA, δηλαδή A-U και/ή C-G, τότε μια τέτοια γραμματική θα μπορούσε να περιγράψει μια RNA ακολουθία. Έτσι για κάθε παραγωγή μιας RNA βάσης θα μπορούσε να παραχθεί και η συμπληρωματική της. Στην πραγματικότητα χρησιμοποιούνται αρκετά πιο πολύπλοκες γραμματικές ώστε να μπορούν να περιγράψουν μεγαλύτερες RNA ακολουθίες. Παρακάτω παρατίθεται ένα παράδειγμα [58] για την καλύτερη απεικόνιση της χρήσης των CFG γραμματικών στην μοντελοποίηση της RNA δευτεροταγής δομής. Έστω η RNA ακολουθία και οι παρακάτω κανόνες γραμματικής:

$$
\begin{aligned}
P = \{ \quad & S_0 \rightarrow S_1, & & S_7 \rightarrow \mathtt{G}\ S_8, \\
& S_1 \rightarrow \mathtt{C}\ S_2\ \mathtt{G}, & & S_8 \rightarrow \mathtt{G}, \\
& S_2 \rightarrow \mathtt{A}\ S_3\ \mathtt{U}, & & S_9 \rightarrow \mathtt{A}\ S_{10}\ \mathtt{U}, \\
& S_3 \rightarrow S_4\ \ S_9, & & S_{10} \rightarrow \mathtt{G}\ S_{11}\ \mathtt{C}, \\
& S_4 \rightarrow \mathtt{U}\ S_5\ \mathtt{A}, & & S_{11} \rightarrow \mathtt{A}\ S_{12}\ \mathtt{U}, \\
& S_5 \rightarrow \mathtt{C}\ S_6\ \mathtt{G}, & & S_{12} \rightarrow \mathtt{U}\ S_{13}, \\
& S_6 \rightarrow \mathtt{A}\ S_7, & & S_{13} \rightarrow \mathtt{C} \qquad \}
\end{aligned}
$$

Η ακολουθία μπορεί να παραχθεί από την εφαρμογή των κανόνων με την παρακάτω σειρά:

$$
\begin{aligned}
S_0 \ & \Rightarrow S_1 \Rightarrow \mathtt{C}S_2\mathtt{G} \Rightarrow \mathtt{CA}S_3\mathtt{UG} \\
& \Rightarrow \mathtt{CA}S_4S_9\mathtt{UG} \Rightarrow \mathtt{CAU}S_5\mathtt{A}S_9\mathtt{UG} \\
& \Rightarrow \mathtt{CAUC}S_6\mathtt{GA}S_9\mathtt{UG} \\
& \Rightarrow \mathtt{CAUCA}S_7\mathtt{GA}S_9\mathtt{UG} \\
& \Rightarrow \mathtt{CAUCAG}S_8\mathtt{GA}S_9\mathtt{UG} \\
& \Rightarrow \mathtt{CAUCAGGGA}S_9\mathtt{UG} \\
& \Rightarrow \mathtt{CAUCAGGGAA}S_{10}\mathtt{UUG} \\
& \Rightarrow \mathtt{CAUCAGGGAAG}S_{11}\mathtt{CUUG} \\
& \Rightarrow \mathtt{CAUCAGGGAAGA}S_{12}\mathtt{UCUUG} \\
& \Rightarrow \mathtt{CAUCAGGGAAGAU}S_{13}\mathtt{UCUUG} \\
& \Rightarrow \mathtt{CAUCAGGGAAGAUCUCUUG}.
\end{aligned}
$$

Η προέλευση της ακολουθίας από τους παραπάνω κανόνες μπορεί να αναπαρασταθεί σε μια δομή δέντρου, που αντιστοιχεί στην δευτερεύουσα δομή της RNA ακολουθίας. Η παραγωγή της δευτεροταγούς δομής της παραπάνω RNA ακολουθίας και το δέντρο ανάλυσης φαίνονται στην *Εικόνα 2.2*.



Εικόνα 2.2: Δέντρο ανάλυσης (αριστερά) και η αντίστοιχη δευτεροταγής δομή (δεξιά)
(Πηγή: [58])

Με την χρήση των CFG γραμματικών, μπορούμε να ελέγξουμε αν μια συμβολοσειρά ανήκει σε μια συγκεκριμένη γραμματική καθώς παράγουμε όλες τις πιθανές προελεύσεις της συμβολοσειράς, τους διαφορετικούς κανόνες με τους οποίους μπορεί να παραχθεί η συγκεκριμένη συμβολοσειρά. Για την εύρεση της πιο πιθανής προέλευσης, μπορούμε να προσθέσουμε πιθανότητες σε κάθε κανόνα παραγωγής. Οι γραμματικές χωρίς συμφραζόμενα όπου κάθε κανόνας παραγωγής για ένα μη-τερματικό σύμβολο έχει μια πιθανότητα ονομάζονται στοχαστικές γραμματικές χωρίς συμφραζόμενα (stochastic context free grammars - SCFG). Μια SCFG γραμματική παράγει αλληλουχίες, αποδίδει μια πιθανότητα σε κάθε παραγόμενη αλληλουχία και βρίσκει το δέντρο ανάλυσης με την μεγαλύτερη πιθανότητα. Αυτή η ιδιότητα βρίσκει εφαρμογή στο πρόβλημα πρόβλεψης της δευτεροταγούς δομής του RNA, όπου με βάση μια RNA ακολουθία εισόδου x, πρέπει να προβλεφθεί η καλύτερη δομή y ή αλλιώς το βέλτιστο δέντρο ανάλυσης (parse tree).

Μια SCFG γραμματική για την πρόβλεψη της δευτεροταγούς δομής ορίζει (1) ένα σύνολο κανόνων παραγωγής (transformation rules), (2) μια κατανομή πιθανότητας πάνω στους κανόνες παραγωγής που εφαρμόζονται σε κάθε μη τερματικό σύμβολο και (3) μια αντιστοίχιση (mapping) από τις προελεύσεις στις δευτεροταγείς δομές [57]. Παρακάτω παρατίθεται ένα παράδειγμα [57] μιας SCFG γραμματικής:

- **Κανόνες παραγωγής**
  S → aSu | uSa | cSg | gSc | gSu | uSg | aS | cS | gS | uS | ϵ :
- **Πιθανότητες κανόνων**. Η πιθανότητα μετατροπής ενός μη-τερματικού S σε aSu είναι pS → aSu, και ομοίως για τους άλλους κανόνες μετατροπής.

- **Αντιστοίχιση από τις προελεύσεις στις δευτεροταγείς δομές**. Η δευτεροταγής δομή y που αντιστοιχεί σε μια προέλευση σ περιέχει ένα ζεύγος βάσεων μεταξύ δύο γραμμάτων αν και μόνο αν τα δύο γράμματα δημιουργήθηκαν στο ίδιο βήμα της προέλευσης για το σ.

  Για μια ακολουθία x = agucu με δευτεροταγή δομή y = ((.)), η μοναδική προέλευση s που αντιστοιχεί στο y είναι:

  $$S \rightarrow aSu \rightarrow agScu \rightarrow aguScu \rightarrow agucu$$

  Η SCFG γραμματική μοντελοποιεί την κοινή πιθανότητα δημιουργίας της προέλευσης s και της ακολουθίας x όπως παρακάτω:

  $$P(x, \sigma) = PS\rightarrow aSu \cdot PS\rightarrow gSc \cdot PS\rightarrow uS \cdot PS$$

  Συνεπώς:

  $$P(y \mid x) = \sum_{\sigma \epsilon y} P(\sigma \mid x) = \frac{\sum_{\sigma \epsilon y} P(x \mid \sigma)}{\sum_{\sigma' \epsilon \Omega(x)} P(x \mid \sigma')}$$

  Όπου Ω(x) είναι όλες οι πιθανές προελεύσεις της ακολουθίας x.

Ένα πλεονέκτημα των SCFG γραμματικών είναι ότι μπορούν να παραμετροποιηθούν με βάση γνωστές δομές του RNA. Με άλλα λόγια, ένα σύνολο γνωστών RNA δομών μπορεί να χρησιμοποιηθεί για να «εκπαιδεύσει» καλύτερα μια SCFG γραμματική ώστε να παράγει καλύτερες προβλέψεις.

Οι στοχαστικές γραμματικές χωρίς συμφραζόμενα (SCFG) χρησιμοποιήθηκαν για την πρόβλεψη της δευτεροταγούς RNA δομής στις αρχές της δεκαετίας του '90 από τον Sakakibara (1994) [40] και τους Eddy και Durbin (1994) [41]. Στη συνέχεια προτάθηκαν αρκετές προσεγγίσεις για πρόβλεψη της RNA δευτεροταγούς δομής με βάση SCFG γραμματικές. Στα κεφάλαια 0 και 4 παρατίθενται εκτενέστερα ορισμένες από αυτές τις προσεγγίσεις.

2.3.2 Στοχαστικές γραμματικές χωρίς συμφραζόμενα και ψευδοκόμβοι του RNA

Ένας από τους περιορισμούς των SCFG γραμματικών στην πρόβλεψη της δευτεροταγούς δομής του RNA είναι ότι δεν μπορούν να μοντελοποιήσουν RNA ψευδοκόμβους. Η δομή ενός ψευδοκόμβου δεν μπορεί να είναι χωρίς συμφραζόμενα [42]. Για την SCFG προσέγγιση μιας τέτοιας δομής έχουν προταθεί αρκετές υλοποιήσεις που αναφέρονται εκτενέστερα στο κεφάλαιο 3.9.

2.4 Διφορούμενες γραμματικές

Στα επόμενα κεφάλαια χρησιμοποιείται αρκετές φορές ο όρος «διφορούμενες γραμματικές». Γι' αυτό κρίνεται σκόπιμος ο ορισμός του.

Διφορούμενες (ambiguous grammars) ονομάζονται οι γραμματικές χωρίς συμφραζόμενα στις οποίες μια συμβολοσειρά μπορεί να παραχθεί από διαφορετικούς κανόνες παραγωγής της γραμματικής. Μη-διφορούμενες (unambiguous grammars) ονομάζονται οι γραμματικές χωρίς συμφραζόμενα στις οποίες κάθε συμβολοσειρά έχει μια μοναδική προέλευση. Παρακάτω παρατίθεται ένα παράδειγμα μιας διφορούμενης γραμματικής [18].

Έστω μια γραμματική G με μη τερματικά σύμβολα {R, A, B, C, D}, τερματικά σύμβολα {a, b, c} και τους παρακάτω κανόνες παραγωγής:

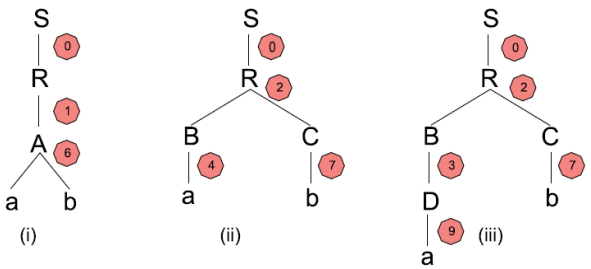
Rule #	Rule
0	$S \rightarrow$ R
1	$R \rightarrow$ A
2	$R \rightarrow$ BC
3	$B \rightarrow$ D
4	$B \rightarrow$ a
5	$A \rightarrow$ a
6	$A \rightarrow$ ab
7	$C \rightarrow$ b
8	$C \rightarrow$ c
9	$D \rightarrow$ a

(Πηγή: [18])

Η συμβολοσειρά ab μπορεί να παραχθεί από τρεις διαφορετικούς συνδυασμούς των κανόνων παραγωγής:

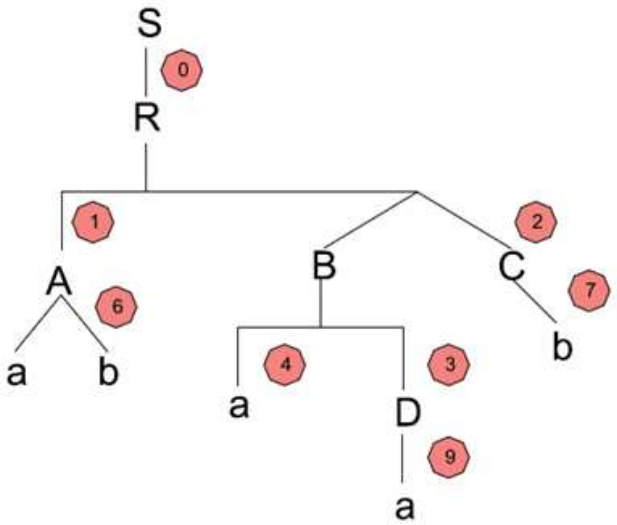
1. S → R → A → ab
2. S → R → BC → aC → ab
3. S → R → BC → DC → aC → ab

Τα αντίστοιχα δέντρα ανάλυσης φαίνονται στην Εικόνα 2.3.



Εικόνα 2.3: Δέντρα ανάλυσης για την συμβολοσειρά ab
(Πηγή: [18])

Μια απεικόνιση όλων των δέντρων ανάλυσης σε μια μοναδική δομή που ονομάζεται AND-or δέντρο φαίνεται στην Εικόνα 2.4.



Εικόνα 2.4: AND-or δέντρο για την συμβολοσειρά ab
(Πηγή: [18])

2.5 Αλγόριθμοι πρόβλεψης δομής με βάση SCFG γραμματικές

Σκοπός αυτού του κεφαλαίου είναι μια συνοπτική παρουσίαση των πιο ευρέως χρησιμοποιούμενων αλγορίθμων στο πεδίο της ανάλυσης και πρόβλεψης της δευτεροταγούς δομής μιας RNA ακολουθίας με βάση τις SCFG γραμματικές.

Σύμφωνα με τον Durbin [3], οι αλγόριθμοι που βασίζονται σε SCFG γραμματικές για την ανάλυση/πρόγνωση μιας RNA ακολουθίας/δομής λύνουν τα παρακάτω προβλήματα:

- Το πρόβλημα της στοίχισης (alignment problem) - οι αλγόριθμοι αυτοί ελέγχουν αν μια ακολουθία x ανήκει σε μια στοχαστική γραμματική G. Επιπλέον υπολογίζουν την βέλτιστη στοίχιση ή αλλιώς το δέντρο ανάλυσης (parse tree) μιας ακολουθίας x σε μια στοχαστική γραμματική G.
- Το πρόβλημα της βαθμολόγησης (scoring problem) – οι αλγόριθμοι αυτοί υπολογίζουν την πιθανότητα μιας ακολουθίας x να ανήκει σε μια στοχαστική γραμματική G.
- Το πρόβλημα της εκπαίδευσης (training problem) - οι αλγόριθμοι αυτοί υπολογίζουν τις βέλτιστες παραμέτρους των κανόνων παραγωγής για μια μη παραμετροποιημένη στοχαστική γραμματική G, δεδομένου ενός συνόλου ακολουθιών/δομών.

2.5.1 Αλγόριθμος CYK

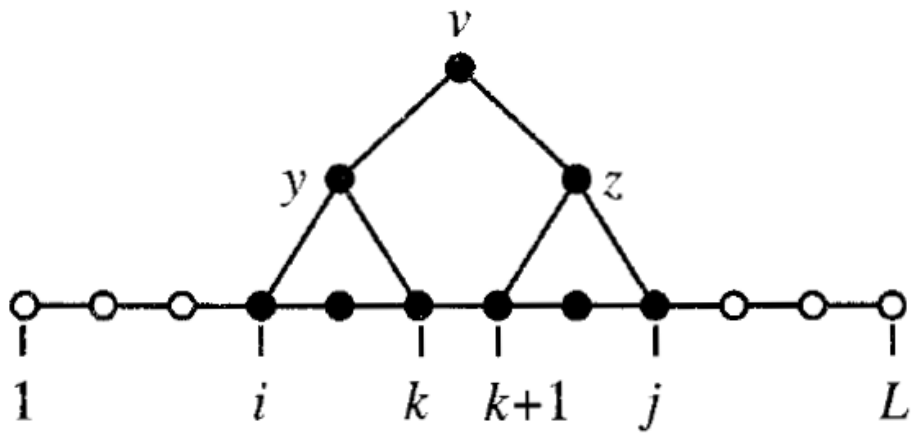
Ένας από τους πιο γνωστούς αλγορίθμους γραμματικής ανάλυσης που βρίσκει εφαρμογή στο πρώτο πρόβλημα είναι ο αλγόριθμος CYK (Coche-Younger-Kasami) [59]. Ο CYK αλγόριθμος προτάθηκε ξεχωριστά από τους Coche, Younger και Kasami το 1967 και ελέγχει αν μια συμβολοσειρά x μήκους n ανήκει σε μια δεδομένη CFG γραμματική G. Η χρονική πολυπλοκότητα είναι $O(n^3|G|)$. Να σημειώσουμε εδώ ότι ο CYK αλγόριθμος απαιτεί η γραμματική να είναι σε CNF μορφή (βλ. κεφάλαιο 2.5.4). Επίσης, έχει επεκταθεί ώστε να χρησιμοποιείται και για στοχαστικές γραμματικές χωρίς συμφραζόμενα. Πιο συγκεκριμένα, ο CYK δημιουργεί έναν τρισδιάστατο πίνακα και ξεκινάει να ελέγχει αν κάθε υποσύνολο μιας συμβολοσειράς μπορεί να παραχθεί από ένα μη τερματικό σύμβολο της γραμματικής ξεκινώντας από το αρχικό σύμβολο. Η ανάλυση ξεκινά από τα υποσύνολα μήκους ‘1’, στη συνέχεια ελέγχει τα υποσύνολα μήκους ‘2’ και ούτε καθ’ εξής. Η

συμβολοσειρά ανήκει στην γραμματική εάν μπορεί να παραχθεί από το σύμβολο έναρξης S. Το δέντρο ανάλυσης ανακτάται με μια traceback διαδικασία του πίνακα. Στην περίπτωση των SCFG γραμματικών, ο CYK υπολογίζει το βέλτιστο δέντρο ανάλυσης.

Ένας άλλος δημοφιλής αλγόριθμος που έχει υλοποιηθεί για την πρώτη περίπτωση είναι ο αλγόριθμος Earley [16]. Ο αλγόριθμος Earley είναι παρόμοιας πολυπλοκότητας με τον CYK με την κύρια διαφορά ότι δεν απαιτεί η γραμματική να είναι σε CNF μορφή.

2.5.2 Αλγόριθμος Inside

Ο αλγόριθμος inside βρίσκει εφαρμογή στο δεύτερο πρόβλημα υπολογίζοντας την πιθανότητα μια δεδομένη συμβολοσειρά να ανήκει στην SCFG γραμματική G. Πιο συγκεκριμένα, ο inside αλγόριθμος υπολογίζει την πιθανότητα α(i, j, u) ενός υποσυνόλου X_i, …., X_j μιας συμβολοσειράς μήκους L να προκύψει από το μη τερματικό σύμβολο W_u. Η πιθανότητα αυτή υπολογίζεται από το άθροισμα των υποδέντρων ανάλυσης για τα μη τερματικά σύμβολα W_y και W_z όπως φαίνεται και στην εικόνα Εικόνα *2.5*.



Εικόνα 2.5: Εφαρμογή του inside αλγορίθμου για τον υπολογισμό της πιθανότητας α(i, j, u)
(Πηγή: [3])

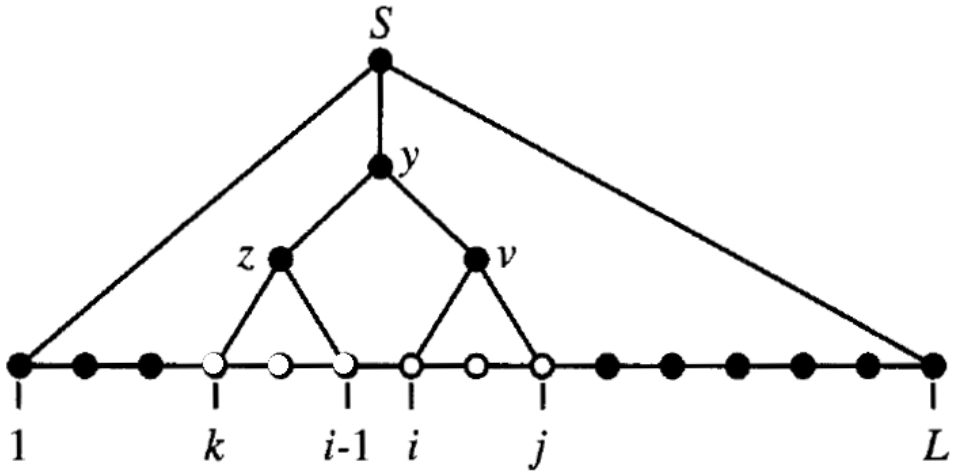
Όπως και στον CYK αλγόριθμο, ο υπολογισμός γίνεται μέσω ενός τρισδιάστατου πίνακα και η ανάλυση ξεκινά από τα υποσύνολα μήκους ‘1’, όπου i=j, και ούτε καθ’ εξής. Στον CYK αλγόριθμο υπολογίζεται η δομή με την μεγαλύτερη πιθανότητα (βέλτιστο δέντρο ανάλυσης) ενώ στον inside

αλγόριθμο αθροίζονται όλες οι πιθανότητες των δέντρων για κάθε υποσύνολο της συμβολοσειράς ή μιας RNA ακολουθίας.

2.5.3 Αλγόριθμος Inside-outside

Ο αλγόριθμος inside-outside [23] βρίσκει εφαρμογή στο τρίτο πρόβλημα. Χρησιμοποιείται για την εκτίμηση των βέλτιστων παραμέτρων μιας γραμματικής που προκύπτει από την ανάλυση ενός συνόλου γνωστών ακολουθιών και δομών RNA. Ως παράμετροι ορίζονται οι πιθανότητες των κανόνων γραμματικής. Πιο συγκεκριμένα, με βάση τις πιθανότητες α(i, j, u) και β(i, j, u), υπολογίζεται, μέσω ενός τρισδιάστατου πίνακα, πόσο συχνά χρησιμοποιείται ένα μη τερματικό σύμβολο και/ή ένας κανόνας παραγωγής. Με αυτό τον τρόπο είναι δυνατή η δημιουργία μιας γραμματικής και στη συνέχεια η εκτίμηση και βελτιστοποίηση των κανόνων παραγωγής με βάση ήδη γνωστές δομές. Η πιθανότητα α(i, j, u) προκύπτει από τον αλγόριθμο inside και η πιθανότητα β(i, j, u) προκύπτει από τον αλγόριθμο outside ο οποίος εξηγείται παρακάτω.

Ο αλγόριθμος outside υπολογίζει την πιθανότητα β(i, j, u) του δέντρου ανάλυσης ολόκληρης της συμβολοσειράς μήκους L να προκύψει από το αρχικό σύμβολο S εξαιρώντας τα υποδέντρα που υπολογίστηκαν στον inside αλγόριθμο, όπως φαίνεται στην εικόνα Εικόνα *2.6*.



Εικόνα 2.6: Εφαρμογή του outside αλγορίθμου για τον υπολογισμό της πιθανότητας β(i, j, u)
(Πηγή: [3])

Όπως και στον CYK αλγόριθμο, ο υπολογισμός γίνεται μέσω ενός τρισδιάστατου πίνακα. Η ανάλυση ξεκινά από τα υποσύνολα μήκους L, όπου L το μήκος της συμβολοσειράς, και συνεχίζει με υποσύνολα L – 1, L – 2 και ούτω καθ' εξής.

2.5.4 Chomsky Normal Form

Όπως αναφέρθηκε παραπάνω, οι αλγόριθμοι CYK και inside-outside προϋποθέτουν η γραμματική να είναι σε CNF μορφή. Η Chomsky normal form (CNF) μορφή προϋποθέτει όλοι οι κανόνες παραγωγής της γραμματικής να είναι της μορφής [69]:

$$A \rightarrow BC,$$
$$A \rightarrow a,$$
$$S \rightarrow \varepsilon,$$

Τα A, B και C είναι μη τερματικά σύμβολα, το a είναι τερματικό σύμβολο, το S είναι το αρχικό σύμβολο και το ε είναι η κενή συμβολοσειρά. Από την παραπάνω μορφή κανόνων προκύπτει ότι η CNF μορφή περιορίζει τους κανόνες γραμματικής να έχουν είτε μόνο μη τερματικά είτε μόνο τερματικά σύμβολα στην δεξιά πλευρά του κανόνα. Κάθε γραμματική σε CNF μορφή είναι γραμματική χωρίς συμφραζόμενα και κάθε γραμματική χωρίς συμφραζόμενα μπορεί να μετατραπεί σε CNF μορφή.

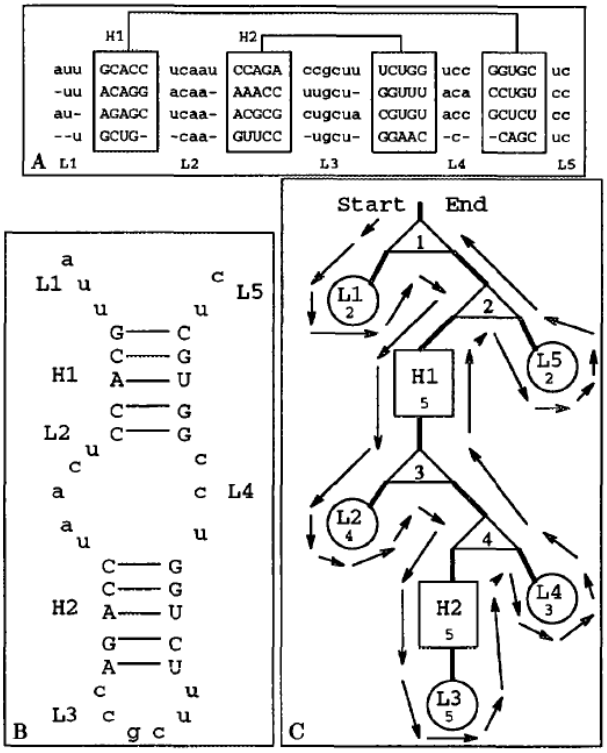
3 Υλοποιήσεις Λογισμικού με βάση SCFG γραμματικές

Σκοπός αυτού του κεφαλαίου είναι να γίνει ανασκόπηση υλοποιήσεων σε λογισμικό επίπεδο που έχουν προταθεί στη βιβλιογραφία για την χρήση των SCFG γραμματικών στην ανάλυση και πρόβλεψη της δευτεροταγούς δομής του RNA. Οι προσεγγίσεις αυτές είναι είτε σειριακές με την χρήση ενός υπολογιστή είτε παράλληλες με την χρήση συστοιχίας υπολογιστών. Παρουσιάζονται, επίσης, στο κεφάλαιο 3.9 ορισμένες προσεγγίσεις που έχουν προταθεί για την πρόβλεψη της δευτεροταγούς δομής του RNA συμπεριλαμβανομένων ψευδοκόμβων.

3.1 Automatic RNA Secondary Structure Determination with Stochastic Context-Free Grammars

Η προσέγγιση του Grate L. [26] το 1995 είναι μια από τις πρώτες προσεγγίσεις όπου οι SCFG γραμματικές χρησιμοποιήθηκαν για την πρόβλεψη της δευτεροταγούς δομής RNA. Η προσέγγιση του Grate είναι η εξαγωγή της SCFG γραμματικής από μεγάλες ακολουθίες στοιχισμένου/ χαρτογραφημένου RNA (aligned/mapped) RNA και η εύρεση της κοινής δομής με βάση αυτή την γραμματική.

Με τη μέθοδο αυτή χρησιμοποιείται ένας αλγόριθμος, ο αλγόριθμος εύρεσης έλικας (helix finding algorithm), σε ένα σύνολο βάσεων για να εξάγει πιθανές υποψήφιες έλικες. Στη συνέχεια, στα αποτελέσματα του παραπάνω αλγορίθμου εφαρμόζεται επαναληπτικά ένας αλγόριθμος δυναμικού προγραμματισμού, ο αλγόριθμος κατασκευής δέντρου (Tree Construction Algorithm), για την εύρεση του βέλτιστου δέντρου και από το δέντρο που προκύπτει εξάγεται η SCFG γραμματική. Η Εικόνα 3.1 δείχνει πολύ χαρακτηριστικά την μέθοδο που μόλις περιγράφηκε. Στην Εικόνα 3.1Α φαίνονται οι στοιχισμένες ακολουθίες με τα ζεύγη βάσεων που προκύπτουν. Στην Εικόνα 3.1Β φαίνεται η δευτεροταγής δομή της πρώτης ακολουθίας και στην Εικόνα 3.1Γ φαίνεται η SCFG γραμματική που περιγράφει την δομή της Εικόνα 3.1Β.



Εικόνα 3.1: Μέθοδος εύρεσης της SCFG γραμματικής Α) Πολλαπλή στοίχιση ενός RNA βρόγχου φουρκέτας. Υποδεικνύονται οι βασικές ζευγαρωμένες περιοχές Β) Δευτεροταγής δομή της πρώτης ακολουθίας στην πολλαπλή στοίχιση Γ) Μια SCFG γραμματική που περιγράφει την ίδια δομή. Οι έλικες είναι ορθογώνιοι, οι βρόχοι είναι κύκλοι και τα τρίγωνα είναι σημεία διακλάδωσης που δεν αντιπροσωπεύουν βάσεις.

(Πηγή [26])

Οι δυο αλγόριθμοι εφαρμόστηκαν παράλληλα σε έναν Maspar MP-2204 επεξεργαστή και σειριακά σε ένα Sun Sparc 2. Η επιτάχυνση στην παράλληλη υλοποίηση αυξήθηκε 40 φορές σε σχέση με την σειριακή υλοποίηση.

Για την πειραματική αξιολόγηση, η παραπάνω μέθοδος εφαρμόστηκε για την εύρεση δομής στις ακολουθίες βάσεων των rRNA 16S και 23S, οι οποίες αποτελούνται από 2.688 και 7.977 βάσεις αντίστοιχα. Τα πειράματα έλαβαν χώρα σε σειριακή υλοποίηση σε έναν Unix HP/Apollo 68040. Ο συνολικός χρόνος εκτέλεσης για την ακολουθία 16S ήταν δέκα λεπτά ενώ για την 23S ήταν δεκαεπτά λεπτά.

Παρ’ όλα αυτά, η παραπάνω μέθοδος παρουσίαζε ελλείψεις όπως ότι δεν συμπεριλαμβάνονται οι ψευδοκόμβοι στις προβλέψεις, ότι παράγεται μόνο μια βέλτιστη δομή και τέλος ότι λανθασμένες προβλέψεις είναι πάντα πιθανό να προκύψουν.

3.2 Pfold: RNA secondary structure prediction using stochastic context-free grammars

Το 2003 οι Knudsen & Hein δημιούργησαν το Pfold [22] το οποίο αποτελεί την πρώτη εφαρμογή που βασίζεται σε στοχαστικές γραμματικές χωρίς συμφραζόμενα (SCFG grammar) για την πρόβλεψη της δομής RNA από μια σειρά RNA ακολουθιών. Η εφαρμογή αυτή αποτελεί την πιο επιτυχημένη εφαρμογή για πρόγνωση δομής RNA και είναι διαθέσιμη μέσω ενός web server στην διεύθυνση http://www.daimi.au.dk/~compbio/pfold. Ο αλγόριθμος του Pfold έχει σχεδιαστεί για να παράγει ένα εξελικτικό δέντρο, το φυλογενετικό δέντρο, και μια δευτεροταγή δομή από ένα ευθυγραμμισμένο σύνολο αλληλουχιών RNA.

Το Pfold λαμβάνει ένα σύνολο αλληλουχιών RNA ως είσοδο και προβλέπει μια κοινή δευτεροταγή δομή για όλες τις RNA αλληλουχίες. Το Pfold είναι βασισμένο στον αλγόριθμο που εισήγαγαν οι ίδιοι ερευνητές το 1999, και ονομάζεται από τους ίδιους ως αλγόριθμος KH-99 [21], και χρησιμοποιεί μια SCFG γραμματική που έχει σχεδιαστεί ειδικά για πρόβλεψη δευτεροταγούς δομής RNA.

Ο αλγόριθμος KH-99 [21], δεδομένης μιας σειράς ακολουθιών παράγει μια κοινή δομική πρόβλεψη για όλες τις ακολουθίες και ένα φυλογενετικό δέντρο. Η πιο πιθανή δομή υπολογίζεται χρησιμοποιώντας τον αλγόριθμο CYK. Το φυλογενετικό δέντρο, το οποίο είναι ένα διάγραμμα διακλάδωσης ή «δέντρο» που δείχνει τις εξελικτικές σχέσεις [67] μεταξύ των ακολουθιών που λαμβάνει ως είσοδο, υπολογίζεται χρησιμοποιώντας μια προσέγγιση μέγιστης πιθανότητας (maximum likelihood - ML) της παρακάτω SCFG γραμματικής:

$$S \rightarrow L \mid LS$$
$$L \rightarrow s \mid dFd$$
$$F \rightarrow dFd \mid LS$$

Τα κύρια μέρη είναι η SCFG γραμματική και το εξελικτικό μοντέλο. Ο αλγόριθμος KH-99, με βάση την SCFG γραμματική, παράγει την κατανομή πιθανοτήτων των δομών RNA. Οι πιθανότητες υπολογίζονται χρησιμοποιώντας τον αλγόριθμο inside–outside [23]. Το μοντέλο υπολογίζεται με

βάση τα παρακάτω βήματα: ***α.*** δημιουργείται μια βάση δεδομένων με τις γνωστές δομές, ***β***. υπολογίζονται οι συχνότητες των απλών βάσεων και των ζευγών βάσεων, ***γ***. υπολογίζεται ο ρυθμός μετάλλαξης *και* ***δ***. υπολογίζονται οι πιθανότητες παραγωγής της γραμματικής.

Ο αλγόριθμος Pfold, δεδομένης μιας σειράς ακολουθιών, δημιουργεί το φυλογενετικό δέντρο χρησιμοποιώντας τον αλγόριθμο γειτονικής σύνδεσης (neighbour - joining algorithm) και στη συνέχεια βελτιστοποιεί τα μήκη των κλαδιών χρησιμοποιώντας την προσέγγιση μέγιστης πιθανότητας. Οι πιθανότητες για ζεύγη (και μη) βάσεων υπολογίζονται σαρώνοντας ξανά το φυλογενετικό δέντρο. Η δομή που επιστρέφεται από τον αλγόριθμο Pfold είναι η δομή με τον μέγιστο αριθμό αναμενόμενων σωστά προβλεπόμενων δομών. Ο χρόνος εκτέλεσης είναι $O(L^3)$, όπου L το μήκος των ακολουθιών.

Ο KH-99 είναι χρήσιμος για έναν μικρό αριθμό ακολουθιών λόγω του μεγάλου χρόνου υπολογισμού του. Ο Pfold είναι πιο βελτιωμένος σε σχέση με τον KH-99 ως προς τον χρόνο υπολογισμού και τα λάθη στοίχισης (alignment errors).

Όσον αφορά τον χρόνο υπολογισμού, ο αλγόριθμος KH-99 υπολογίζει το φυλογενετικό δέντρο χρησιμοποιώντας την προσέγγιση μέγιστης πιθανότητας (ML) της SCFG γραμματικής. Το μειονέκτημα εδώ είναι ότι η μέθοδος είναι αργή. Στο Pfold, το δέντρο υπολογίζεται πριν την πρόβλεψη της δομής χρησιμοποιώντας τον αλγόριθμο γειτονικής σύνδεσης (neighbour - joining algorithm). Αυτό έχει ως αποτέλεσμα οι υπολογισμοί για την εκτίμηση του δέντρου να εκτελούνται μόνο μία φορά, σε αντίθεση με τις πολλαπλές επαναλήψεις που χρησιμοποιούνται στον αλγόριθμο KH-99. Έτσι επιτυγχάνεται αύξηση στην ταχύτητα.

Όσον αφορά τα λάθη στοίχισης, το Pfold είναι πιο βελτιωμένο ως προς τον χειρισμό των κενών (gaps) μιας ακολουθίας. Ο αλγόριθμος KH-99 χειρίζεται τα κενά σαν άγνωστες βάσεις. Το μειονέκτημα αυτής της προσέγγισης είναι ότι όταν αντιμετωπίζονται τα κενά ως άγνωστα νουκλεοτίδια, τα κενά μπορούν να σχηματίσουν ζεύγη με νουκλεοτίδια, το οποίο δεν είναι επιτρεπτό. Το Pfold αντιμετωπίζει αυτό το πρόβλημα με την απομάκρυνση των στηλών όπου λιγότερο από το 75% των αλληλουχιών έχουν νουκλεοτίδια. Ένα παράδειγμα φαίνεται στην Εικόνα 3.2, όπου στο πάνω κομμάτι της τα κενά αντιμετωπίζονται ως άγνωστα νουκλεοτίδια. Η δομή, που εμφανίζεται ως παρενθέσεις, περιλαμβάνει λανθασμένα ζεύγη μεταξύ νουκλεοτιδίων και κενών.

Στην κάτω στοίχιση, οι στήλες με κενά δεν έχουν συμπεριληφθεί στην πρόβλεψη, επειδή το λιγότερο από 75% των αλληλουχιών έχουν νουκλεοτίδια σε αυτές τις θέσεις. Αυτό έχει ως συνέπεια να σχηματίζονται ζεύγη μόνο μεταξύ νουκλεοτιδίων.

```
Seq 1       CGAC - - - - - AGCUGAGUGUGACUUUAGAAU
Seq 2       UGACGGUCUAGCUGACUGAUACUUCAGAGU
Seq 3       CGAC - - - - - AGCUGAAUGAGACUUCAGAAU
Structure   . . . ( ( ( ( ( . . . . . ) ) ) ) ) . . . . . . . . . . . .

Seq 1       CGAC - - - - - AGCUGAGUGUGACUUUAGAAU
Seq 2       UGACGGUCUAGCUGACUGAUACUUCAGAGU
Seq 3       CGAC - - - - - AGCUGAAUGAGACUUCAGAAU
Structure   . . . . . . . . . . . ( ( ( ( . . . . . . . . . ) ) ) ) . . .
```

Εικόνα 3.2: Μια πρόβλεψη δομής για τρεις υποθετικές ακολουθίες

(Πηγή [21])

Επιπλέον το Pfold εισάγει σε κάθε νουκλεοτίδιο μια μικρή πιθανότητα 1% να είναι ένα οποιοδήποτε άλλο νουκλεοτίδιο. Έτσι ο Pfold γίνεται πιο ισχυρός αφού ένα μόνο σφάλμα δεν μπορεί να αλλάξει σημαντικά μια πρόβλεψη. Ένα παράδειγμα φαίνεται στην Εικόνα 3.3, στο πάνω μέρος φαίνεται η δομή τριών υποθετικών ακολουθιών εφαρμόζοντας τον KH-99, ενώ στο κάτω η προκύπτουσα δομή είναι μια μεγαλύτερη φουρκέτα όταν όλα τα νουκλεοτίδια έχουν πιθανότητα 1% να είναι οποιοδήποτε άλλο νουκλεοτίδιο.

```
Seq 1       UGGCG - - CUAGCCAUCUGAUACUUCAGAUU
Seq 2       UGACGGACUAGCCAACUGAUACUUCAGAUA
Seq 3       U - ACGUAGUAGCCAUCUGAUACUUCAGAU -
Structure   . . . . . . . . . . . . . . . ( ( ( ( . . . . ) ) ) ) . . .

Seq 1       UGGCG - - CUAGCCAUCUGAUACUUCAGAUU
Seq 2       UGACGGACUAGCCAACUGAUACUUCAGAUA
Seq 3       U - ACGUAGUAGCCAUCUGAUACUUCAGAU -
Structure   . . . . . . . . . . . . . ( ( ( ( ( ( . . . . ) ) ) ) ) ) .
```

Εικόνα 3.3: Προκύπτουσα δομή με την εισαγωγή πιθανότητα λάθους

(Πηγή [21])

Ο χρήστης μέσω της ιστοσελίδας εισάγει ένα σύνολο έως και 40 ακολουθιών και 500 RNA νουκλεοτιδίων σε FASTA μορφή και λαμβάνει μια αναφορά στην ηλεκτρονική του διεύθυνση. Το Pfold αναφέρει τη δομή με τον υψηλότερο αναμενόμενο αριθμό σωστών προβλέψεων. Μόλις επιλεγεί η καλύτερη δομή, αξιολογείται η αξιοπιστία της πρόβλεψης για κάθε θέση. Επιπλέον, παράγει μια γραφική παράσταση (dot plot) για να αποδώσει γραφικά μια γενική εικόνα της πρόβλεψης. Με βάση αυτά τα αποτελέσματα, η αναφορά προς τον χρήστη περιλαμβάνει μια

σύνοψη της εισόδου, το δέντρο που προέκυψε, την προβλεπόμενη δομή, την αξιοπιστία των μεμονωμένων προβλέψεων και την γραφική παράσταση (dot plot).

Το Pfold είναι θεωρητικά ικανό να επιτύχει υψηλής ποιότητας προβλέψεις για μεγάλες και βιολογικά σημαντικές δομές RNA, όπως τα γονιδιώματα των ιών RNA. Παρ' όλα αυτά, είναι «single-threaded» και έχει ως αποτέλεσμα να χρειάζεται αρκετό χρόνο για την πρόγνωση μεγάλων ακολουθιών. Επίσης, και οι δυο αλγόριθμοι, KH-99 και Pfold, δεν είναι ικανοί να προβλέψουν ψευδοκόμβους (pseudoknots).

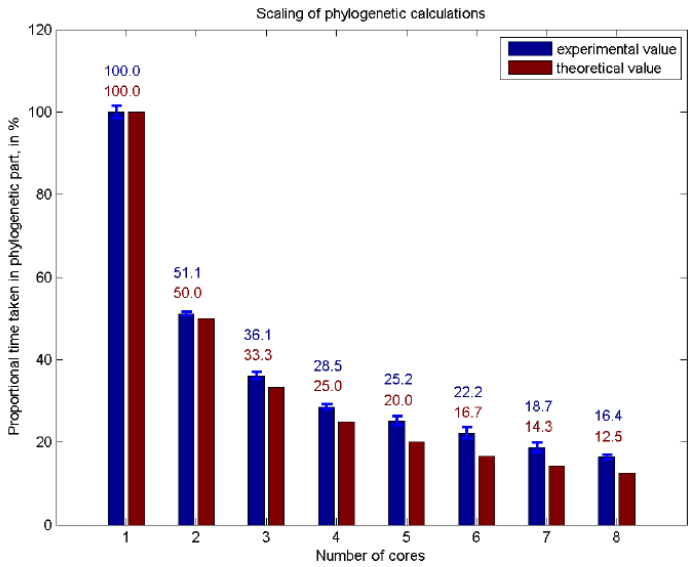
3.3 Multithreaded comparative RNA secondary structure prediction using stochastic context-free grammars

Οκτώ χρόνια αργότερα, το 2011, η ομάδα των Sükösd Z., Knudsen B., Værum M., Kjems J. και Andersen E. παρουσίασαν μια βελτιωμένη έκδοση του Pfold, το PPfold [24]. Το PPfold είναι η multithreaded έκδοση του αλγόριθμου Pfold και ο πρώτος παράλληλος συγκριτικός αλγόριθμος πρόβλεψης δομής RNA μέχρι τότε με τόσο μεγάλη ακρίβεια πρόβλεψης για μεγάλες ακολουθίες RNA.

Το PPfold παρουσιάζει, σε σχέση με το Pfold, πιο βελτιωμένο περιβάλλον χρήστη (user interface) και είναι πιο ευέλικτο αφού είναι διαθέσιμο και ως δωρεάν plugin για τα CLC workbences (Workbences που χρησιμοποιούνται για ανάλυση DNA/RNA δεδομένων). Το PPfold είναι υλοποιημένο σε γλώσσα Java και ο κώδικας είναι διαθέσιμος στην ιστοσελίδα http://www.daimi.au.dk/~compbio/pfold/downloads.html. Η PPfold έκδοση 3.0 προβλέπει την δομή του RNA με βάση την SCFG γραμματική του αλγορίθμου KH-99 αλλά και εισάγοντας δεδομένα από πειράματα χημικής ανίχνευσης.

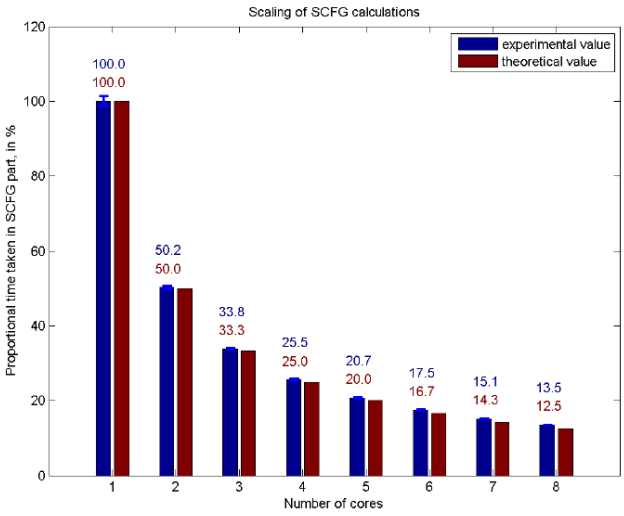
Το PPfold, λόγω της multithreaded φύσης του, είναι ικανό να παράγει την πρόβλεψη δομής μεγάλων ακολουθιών RNA σε πολύ λιγότερο χρόνο σε σχέση με το Pfold. Το PPfold υλοποιεί παράλληλα threads για τον υπολογισμό του φυλογενετικού δέντρου και την εκτέλεση του inside–outside αλγορίθμου [23].

Όσον αφορά τον υπολογισμό του φυλογενετικού δέντρου, το PPfold δημιουργεί ομάδες στηλών με ζεύγη βάσεων και υπολογίζει την μέγιστη πιθανότητα εκτελώντας ένα thread για την κάθε ομάδα. Έτσι επιτυγχάνει την μείωση του χρόνου εκτέλεσης αφού ο υπολογισμός της πιθανότητας είναι το πιο χρονοβόρο μέρος της διαδικασίας. Στην Εικόνα 3.4 φαίνεται η μείωση του χρόνου υπολογισμού του φυλογενετικού δέντρου καθώς οι πυρήνες αυξάνονται από ένας έως οκτώ.



Εικόνα 3.4: Μείωση χρόνου εκτέλεσης για το φυλογενετικό τμήμα του αλγορίθμου.
(Πηγή: [24])

Για την εκτέλεση του inside–outside αλγορίθμου, το PPfold υλοποιεί την ασύγχρονη εκτέλεση εργασιών (jobs) επιτυγχάνοντας έτσι την ολοκλήρωση της εκτέλεσης των threads σε χρόνο 1/n, όπου n ο αριθμός των threads. Στην Εικόνα 3.5 φαίνεται η μείωση του χρόνου εκτέλεσης του inside–outside αλγορίθμου καθώς οι πυρήνες αυξάνονται από ένας έως οκτώ.



Εικόνα 3.5: Μείωση χρόνου εκτέλεσης για το κομμάτι της SCFG γραμματικής.
(Πηγή: [24])

Και στα δυο παραπάνω διαγράμματα παρατηρούμε ότι ο θεωρητικός χρόνος εκτέλεσης σχεδόν συμπίπτει με τον πειραματικό χρόνο εκτέλεσης που χρειάστηκε στην πράξη.

Για την αξιολόγηση της απόδοσης του αλγορίθμου, οι ερευνητές σύγκριναν τους χρόνους εκτέλεσης μεταξύ του Pfold και του PPfold στο ίδιο σύνολο ακολουθιών. Το PPfold είναι αναμφισβήτητα πολύ πιο γρήγορο. Επιπλέον, χρησιμοποίησαν το PPfold για την πρόβλεψη της δομής 24 γονιδιωμάτων του ιού HIV-1 σε έναν οκταπύρηνο υπολογιστή με 6 GB μνήμης. Η πρόβλεψη ολοκληρώθηκε σε 65 λεπτά με μεγάλα ποσοστά ακρίβειας.

3.4 Parallel RNA secondary structure prediction using stochastic context-free grammars

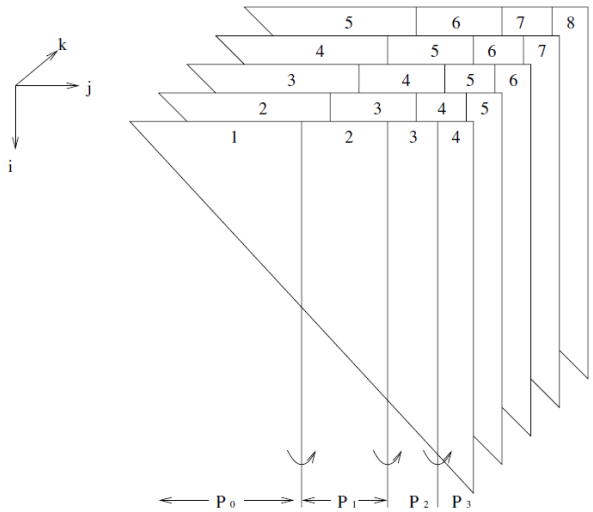
Το 2005 η ομάδα των Liu T. και Schmidt B. πρότεινε μια παράλληλη υλοποίηση του CYK αλγορίθμου για την στοίχιση μιας αλληλουχίας RNA χρησιμοποιώντας SCFG γραμματική και την coarse-grained (χονδρόκοκκη) υλοποίηση του αλγορίθμου σε ένα σύμπλεγμα υπολογιστών (PC cluster) και μια συστοιχία SMPs [34].

Τα μοντέλα συνδιακύμανσης (covariance model -CM), είναι μια ειδική περίπτωση γραμματικών χωρίς συμφραζόμενα, κατάλληλα φτιαγμένα για να περιγράφουν μια πολλαπλή στοίχιση RNA. Το

όνομα τους προκύπτει από τη συνδιακύμανση των διαδοχικών στηλών μιας πολλαπλής στοίχισης [7]. Ένα μοντέλο συνδιακύμανσης μπορεί να χρησιμοποιηθεί για να προσδιοριστεί η πιθανότητα μια αλληλουχία RNA να ανήκει στην RNA οικογένεια. Το πιο πιθανό δέντρο ανάλυσης που παράγει (ή εκπέμπει) την ακολουθία καθορίζει τη βαθμολογία ομοιότητας. Ο τρισδιάστατος CYK/inside αλγόριθμος δυναμικού προγραμματισμού υπολογίζει αυτό το σκορ ως εξής: υπολογίζει επαναληπτικά έναν τρισδιάστατο πίνακα M(i, j, k) για όλα τα i = 1, . . . , j + 1, j = 0, . . .,L, k = 1, . . . , K , όπου L το μήκος της RNA ακολουθίας και K οι καταστάσεις του μοντέλου συνδιακύμανσης. Η τιμή M(i, j, k) αντιπροσωπεύει την τιμή log-odds του πιο πιθανού δέντρου ανάλυσης του μοντέλου συνδιακύμανσης στην κατάσταση *k* που δημιουργεί την υποακολουθία x_i . . .x_j [34]. Το τέλος της επανάληψης περιέχει το αποτέλεσμα της καλύτερης ανάλυσης της αλληλουχίας. Το βέλτιστο δέντρο ανάλυσης ανακτάται με μια traceback διαδικασία, χρησιμοποιώντας τον CYK/outside αλγόριθμο, που ακολουθεί τη διαδρομή του πίνακα με την μέγιστη βαθμολογία.

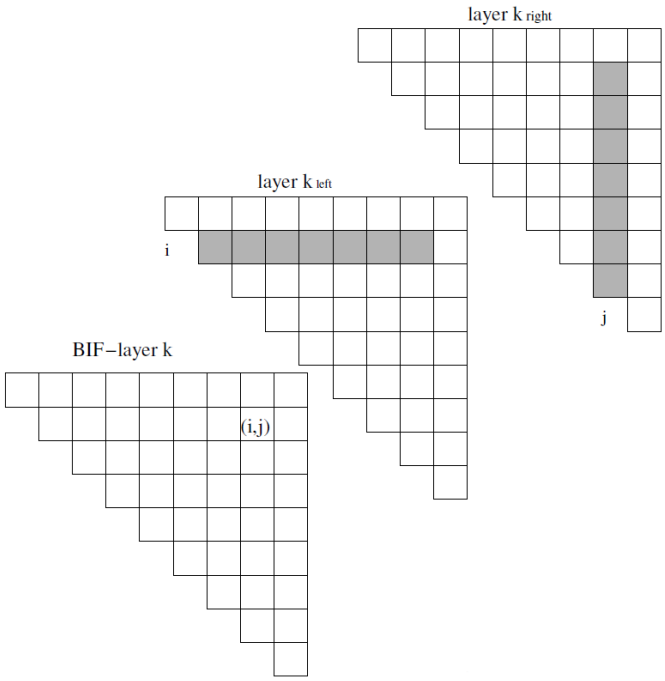
Οι ερευνητές πρότειναν την παράλληλη υλοποίηση του παραπάνω αλγορίθμου (parallel alignment algorithm) αλλά και του traceback αλγορίθμου (parallel traceback algorithm) για να μειώσουν τον χρόνο εκτέλεσης.

Η παράλληλη υλοποίηση του αλγορίθμου φαίνεται στην Εικόνα 3.6. Σε fine-grained (λεπτόκοκκες) αρχιτεκτονικές, τα κελιά του πίνακα σε κάθε διαγώνιο υπολογίζονται παράλληλα. Αυτό όμως απαιτεί γρήγορη επικοινωνία μεταξύ των μονάδων επεξεργασίας PE. Για την coarse-grained αρχιτεκτονική σε ένα σύμπλεγμα υπολογιστών είναι πιο αποτελεσματικό να ανατίθεται ένας αριθμός γειτονικών στηλών σε κάθε επεξεργαστή P_x, όπως φαίνεται και στην Εικόνα 3.6. Ο κάθε επεξεργαστής υπολογίζει τα κελιά του πίνακα, λαμβάνοντας τα δεδομένα από τον προηγούμενο επεξεργαστή. Με παρόμοιο τρόπο εφαρμόζεται και η παράλληλη υλοποίηση του traceback αλγορίθμου.



Εικόνα 3.6: Παράλληλη υλοποίηση του CYK αλγορίθμου χρησιμοποιώντας 4 επεξεργαστές
(Πηγή [34])

Επιπλέον, οι ερευνητές διαίρεσαν τις εργασίες (tasks) σε άνισα μεγέθη προκειμένου να επιτευχθεί καλύτερη ισορροπία στον φόρτο υπολογισμού. Ακολούθησαν αυτό τον τρόπο ανάθεσης εργασιών σε κάθε μονάδα επεξεργασίας P_x, επειδή κάθε κελί του πίνακα δεν απαιτεί ίσο πλήθος υπολογισμών με τα υπόλοιπα κελιά. Η διαδικασία αυτή ονομάζεται ανομοιόμορφος κυματοειδής υπολογισμός (irregular wavefront computation). Όπως φαίνεται και στην Εικόνα 3.7, κάθε κελί M(i,j,k) εξαρτάται από τα κελιά M(i, mid, kleft) και M(mid + 1, j, kright) (σκιασμένη περιοχή) για ένα επίπεδο διακλάδωσης (BIF-layer). Αυτό έχει ως αποτέλεσμα να αυξάνεται ο φόρτος υπολογισμού για κάθε P_{i+1} σε σχέση με το P_i.



Εικόνα 3.7: Κάθε κελί M(i,j,k)εξαρτάται από τα κελιά M(i, mid, kleft) και M(mid + 1, j, kright)
(Πηγή [34])

Για την αξιολόγηση της προσέγγισής τους, οι ερευνητές υλοποίησαν τον αλγόριθμο σε μια συστοιχία δέκα υπολογιστών με λειτουργικό σύστημα GNU/Linux, δυο Intel-Xeon CPUs και 1 GB μνήμα RAM στον καθένα. Η ταχύτητα στο χρόνο εκτέλεσης επιταχύνεται έως 16 φορές σε σύγκριση με έναν απλό επεξεργαστή ανάλογα με τον αριθμό των μονάδων επεξεργασίας που χρησιμοποιούνται. Επιπλέον, υλοποίησαν τον αλγόριθμο σε μια συστοιχία δώδεκα SMP κόμβων και πέτυχαν επιτάχυνση στην ταχύτητα 36 φορές χρησιμοποιώντας 48 Alpha EV68 επεξεργαστές .

Εξέτασαν επίσης την ακρίβεια της πρότασης [34] συγκρίνοντας τα αποτελέσματα με τα αποτελέσματα δυο άλλων εργαλείων, του CrystalW και του HMMER. Στον **Error! Reference source not found.**Πίνακας 3.1, φαίνονται τα αποτελέσματα χρησιμοποιώντας ως είσοδο 60 E. Coli SSU rRNA ακολουθίες και 20 E. coli LSU rRNA ακολουθίες. Ο χρόνος υπολογισμού με βάση την προσέγγιση [34] ήταν 1,5 ώρες ενώ ένας απλός Pentium 4 υπολογιστής θα χρειαζόταν μια ολόκληρη μέρα για τον ίδιο υπολογισμό.

	Program		
	ClustalW	HMMER	Parallel CYK
E. coli SSU rRNA	23.9%	66.8%	76.0%
E. coli LSU rRNA	26.8%	65.5%	77.8%

Πίνακας 3.1:Σύγκριση ακρίβειας αποτελεσμάτων με είσοδο 60 E. Coli SSU rRNA ακολουθίες και 20 E. coli LSU rRNA ακολουθίες

(Πηγή [34])

Συμπεραίνουμε λοιπόν ότι η επιτάχυνση στην ταχύτητα υπολογισμού που επιτυγχάνεται εφαρμόζοντας παράλληλη υλοποίηση του CYK αλγορίθμου σε μια συστοιχία κόμβων (cluster nodes) είναι αρκετά σημαντική και ειδικά για μεγάλες RNA ακολουθίες όπου ο υπολογισμός τους θα ήταν απαγορευτικός, αν όχι μη υλοποιήσιμος, με απλές υλοποιήσεις.

3.5 Load Balancing Algorithm in Cluster-based RNA secondary structure Prediction

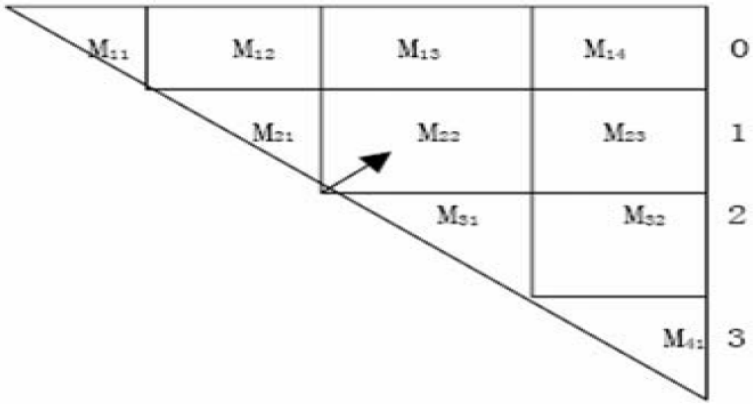
Το 2005 οι Tan, Feng και Sun παρουσίασαν μια παράλληλη υλοποίηση του CYK/inside αλγορίθμου σε μια συστοιχία υπολογιστών [52] και μείωσαν τον χρονική και χωρική πολυπλοκότητα από $O(N^4)$ και $O(N^3)$ σε $O(N^4/P)$ και $O(N^3/P)$ αντίστοιχα, όπου P ο αριθμός των επεξεργαστών.

Όπως έχει ήδη αναφερθεί ο CYK/inside αλγόριθμος μοντελοποιεί μια RNA ακολουθία με βάση ένα μοντέλο συνδιακύμανσης (covariance model - CM), μια SCFG γραμματική με επτά τύπους καταστάσεων, για να προσδιοριστεί η πιθανότητα η αλληλουχία να ανήκει στην RNA οικογένεια. Ο τρισδιάστατος CYK/inside αλγόριθμος δυναμικού προγραμματισμού προσδιορίζει το αποτέλεσμα της πιθανότητας υπολογίζοντας επαναληπτικά έναν τρισδιάστατο πίνακα M(i, j, k) (3D DP matrix). Το κελί M(1,L,1) του πίνακα περιέχει την πιθανότητα του βέλτιστου δέντρου της RNA ακολουθίας.

Η κύρια ιδέα είναι ότι χωρίζουν τον πίνακα σε τμήματα και κατανέμουν τον υπολογισμό των τμημάτων σε επεξεργαστές. Ο παράλληλος αλγόριθμος μπορεί να είναι στατικός ή δυναμικός ανάλογα με το πλήθος των τμημάτων που υπολογίζει κάθε επεξεργαστής.

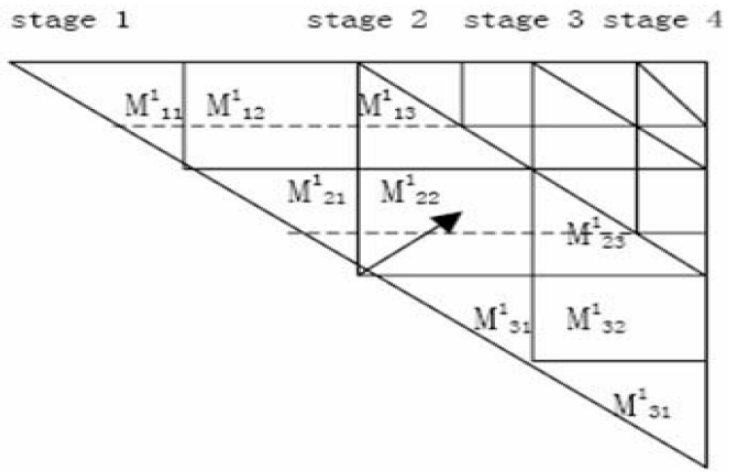
Στην στατική προσέγγιση κάθε επεξεργαστής υπολογίζει ένα σταθερό τμήμα του πίνακα (static partitioning algorithm). Ο αλγόριθμος χωρίζεται σε τόσα στάδια όσα και οι επεξεργαστές και κάθε

στάδιο χωρίζεται σε δυο βήματα. Στο πρώτο βήμα ο επεξεργαστής i υπολογίζει τα κελιά M(i+1,1) του πίνακα και στο δεύτερο βήμα μεταφέρει τα υπολογισμένα κελιά στον επεξεργαστή i-1. Επίσης μεταφέρει και τα αποτελέσματα των κελιών που υπολογίστηκαν στο προηγούμενο στάδιο, δηλαδή στο i+1. Ο συνολικός χρόνος υπολογίζεται σε $O(N^4/P)$ και χωρίζεται στον χρόνο υπολογισμού του πρώτου βήματος και στον χρόνο επικοινωνίας του δεύτερου βήματος, δηλαδή T_{comp} και T_{comm} αντίστοιχα. Ο κυματοειδής υπολογισμός του πίνακα φαίνεται στην Εικόνα 3.8.



Εικόνα 3.8: Ο κυματοειδής υπολογισμός στην στατική προσέγγιση
(Πηγή: [52])

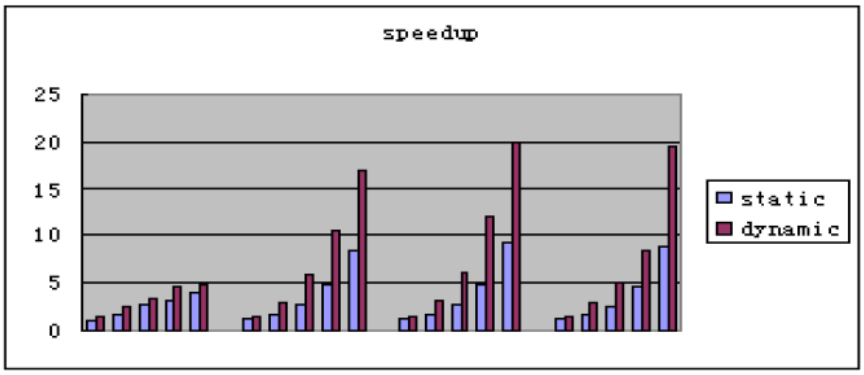
Στη δυναμική προσέγγιση (dynamic partitioning algorithm) ο πίνακας επανακατανέμεται στους επεξεργαστές μεταξύ των σταδίων με σκοπό να χρησιμοποιούνται με πιο αποδοτικό τρόπο. Ο υπολογισμός του πίνακα φαίνεται στην Εικόνα 3.9. Η διακεκομμένη γραμμή αντιπροσωπεύει την επανακατανομή, η οποία πραγματοποιείται όταν ο αριθμός των αδρανών επεξεργαστών υπερβαίνει κάποιο όριο t. Σύμφωνα με τους ερευνητές η ιδανική τιμή του t για την πρόβλεψη της RNA δομής είναι p/2. Ο συνολικός χρόνος ισούται με τον συνολικό χρόνο της στατικής προσέγγισης, δηλαδή ίσος με $O(N^4/P)$, και τον χρόνο της επανακατανομής του πίνακα $T_{repartition}$, ο οποίος ισούται με $O(n^2\log(p)/p)$.



Εικόνα 3.9: Ο κυματοειδής υπολογισμός στην δυναμική προσέγγιση

(Πηγή: [52])

Για την αξιολόγηση της προσέγγισης, οι ερευνητές εφάρμοσαν τον αλγόριθμο σε μια συστοιχία υπολογιστών με 4 AMD επεξεργαστές και 8 GB μνήμη RAM για τέσσερα μήκη RNA ακολουθιών. Τα αποτελέσματα φαίνονται στην Εικόνα 3.10 για την στατική και την δυναμική προσέγγιση.



Εικόνα 3.10: Επιτάχυνση για τέσσερα μήκη αλληλουχιών RNA: 120 (5Secoli), 377 (RnaseP), 1542 (lsuecoli) και 2904 (ssuecoli). Ο αριθμός επεξεργαστών είναι 2, 4, 8, 16 και 32

(Πηγή: [52])

Σύμφωνα με τα αποτελέσματα, η επιτάχυνση στη δυναμική προσέγγιση είναι αρκετά μεγαλύτερη. Στην δυναμική προσέγγιση ο συνολικός χρόνος εκτέλεσης μειώνεται λόγω της βέλτιστης χρήσης των επεξεργαστών παρόλο που προστίθεται ο χρόνος επανακατανομής του πίνακα. Η επιτάχυνση αυξάνεται έως 19 φορές στην δυναμική προσέγγιση για μια συστοιχία 8 υπολογιστών (32 επεξεργαστές) και έως 9 φορές στην στατική προσέγγιση για τον ίδιο αριθμό επεξεργαστών.

3.6 Evolving stochastic context–free grammars for RNA secondary structure prediction

Το 2012 οι Anderson J., Tataru P., Staines J, Hein J and Lyngso R. παρουσίασαν [25] μια πολύ ενδιαφέρουσα τεχνική, υλοποιώντας έναν εξελικτικό αλγόριθμο, για την αυτόματη εύρεση ισχυρότερων SCFG γραμματικών ανάμεσα σε ένα μεγάλο σύνολο SCFG γραμματικών ώστε να πετύχουν καλύτερη πρόβλεψη της δευτεροταγούς δομής του RNA. Επιπλέον, εξέτασαν αυτές τις SCFG γραμματικές, που προέκυψαν, αν είναι διφορούμενες ή όχι, για να ελέγξουν αν η διφορούμενη γραμματική ή ο σχεδιασμός της γραμματικής είναι αυτό που επηρεάζει περισσότερο την απόδοσή της. Με τα αποτελέσματα που προέκυψαν, απέδειξαν ότι για να αυξηθεί η ακρίβεια πρόγνωσης μια γραμματικής δεν απαιτείται η επέκταση της γραμματικής με επιπλέον χαρακτηριστικά (όπως οι ψευδοκόμβοι) αλλά είναι αναγκαία η εξέλιξη της γραμματικής ως προς τον σχεδιασμό της.

Οι SCFG γραμματικές, στις οποίες πραγματοποιείται ο έλεγχος, πρέπει να έχουν μια νέα μορφή που οι παραπάνω ερευνητές ονόμασαν «κανονική μορφή διπλής εκπομπής» (double emission normal form). Θεωρήθηκε ότι αυτή η μορφή θα μπορούσε να σχετίζεται καλύτερα με την δευτεροταγή δομή του RNA σε σχέση με την CNF μορφή. Με βάση αυτή την διπλής εκπομπής κανονική μορφή, οι κανόνες της γραμματικής πρέπει να έχουν την παρακάτω μορφή:

$$T \rightarrow UV$$
$$T \rightarrow .$$
$$T \rightarrow (U)$$

Στην παραπάνω μορφή, τα σύμβολα T, U και V αντιστοιχούν σε μη τερματικά σύμβολα, η τελεία αντιπροσωπεύει μη ζευγαρωμένα νουκλεοτίδια και η παρένθεση αντιπροσωπεύει ζευγαρωμένα νουκλεοτίδια.

Το σύνολο των SCFG γραμματικών για n τερματικά σύμβολα είναι $O(2^{n^3})$ γραμματικές. Ο εξελικτικός αλγόριθμος επιτρέπει την αποτελεσματική εξερεύνηση των γραμματικών αυτών. Ο τρόπος με τον οποίο ο εξελικτικός αλγόριθμος αναζητά τις ισχυρότερες γραμματικές στο σύνολο

των γραμματικών καθορίζεται από το σχεδιασμό του αρχικού πληθυσμού, τη μετάλλαξη, την αναπαραγωγή και τη διαδικασία επιλογής.

Ο αρχικός πληθυσμός είναι ένα μικρό σύνολο SCFG γραμματικών, 16 γραμματικές για την ακρίβεια, της μορφής :

$$S \rightarrow \left\{ \begin{array}{c|c|c|c|c} SS & SB & BS & BB & (S) \\ - & - & - & - & \end{array} \right. .$$
$$B \rightarrow .$$

Επιπλέον, οι πιθανότητες των γραμματικών αυτών προκύπτουν από τον προσδιορισμό μέγιστης πιθανότητας (maximum likelihood - ML). Η μετάλλαξη και η αναπαραγωγή σχεδιάστηκαν για να επιτρέψουν στην SCFG γραμματική να αναπτύξει πρόσθετους κανόνες παραγωγής και μη τερματικά σύμβολα.

Η διαδικασία επιλογής των ισχυρότερων γραμματικών για την καλύτερη πρόβλεψη της δευτεροταγούς RNA δομής καθορίζεται από τις εξής παραμέτρους: την ευαισθησία, την θετική προγνωστική αξία των προβλέψεων και τον λόγο πιθανότητας πραγματικής δομής και προβλεπόμενης δομής. Η ακρίβεια ορίζεται όταν ένα ζεύγος βάσεων είναι αληθώς θετικό (true positive) αν υπάρχει στις προβλεπόμενες και στις πραγματικές δομές, ψευδώς θετικό (false positive) όταν υπάρχει μόνο στην προβλεπόμενη δομή και ψευδώς αρνητικό (false negative) όταν υπάρχει μόνο στην πραγματική δομή. Για την ακρίβεια εισήγαγαν κι άλλες παραμέτρους όπως του μεγαλύτερου προβλεφθέντος μήκους, την ανικανότητα πρόβλεψης δομών και το κόστος πολυπλοκότητας της γραμματικής.

Οι ερευνητές άντλησαν τα δεδομένα τους για την εφαρμογή του εξελικτικού αλγορίθμου και την εξαγωγή αποτελεσμάτων κυρίως από την βάση της RNAStrand αλλά και άλλες βάσεις δεδομένων. Το τελικό σύνολο δεδομένων αποτελούνταν από 369 ακολουθίες νουκλεοτιδίων, ένα μέρος από τις οποίες χρησιμοποιήθηκε για την παραμετροποίηση της SCFG γραμματικής, ένα άλλο μέρος για την αξιολόγηση της απόδοσης πρόβλεψης του εξελικτικού αλγορίθμου και οι υπόλοιπες ακολουθίες

χρησιμοποιήθηκαν για την αξιολόγηση των SCFG γραμματικών που προέκυψαν από τον εξελικτικό αλγόριθμο.

Τα μέτρα για την αξιολόγηση των γραμματικών που προέκυψαν από τον αλγόριθμο ήταν η ευαισθησία, το PPV και το F-score. Η ευαισθησία είναι το κλάσμα των ζευγών βάσεων στις αξιόπιστες δομές που προβλέπονται σωστά. Το PPV (Positive Predictive value) είναι ο αριθμός των προβλεπόμενων ζευγών βάσεων που βρίσκονται στην αξιόπιστη δομή και το F-score υπολογίζεται από την παρακάτω σχέση: 2* αληθώς θετικό /(2* αληθώς θετικό + ψευδώς αρνητικό + ψευδώς θετικό).

Μετά την εφαρμογή του εξελικτικού αλγορίθμου προέκυψαν επτά ισχυρές γραμματικές, τα αποτελέσματα των οποίων φαίνονται στον Πίνακας 3.2. Να σημειώσουμε ότι η γραμματική KH99' είναι η γραμματική του αλγορίθμου KH99[21] σε κανονική μορφή διπλής εκπομπής.

Grammar	Sensitivity	PPV	Ambiguity	Completeness
KH99′	0.496	0.479	No	Yes
GG1	0.503	0.480	Yes	Yes
GG2	0.505	**0.481**	Yes	Yes
GG3	0.474	0.421	Yes	No
GG4	0.474	0.454	Yes	No
GG5	0.469	0.467	Yes	No
GG6	0.486	0.468	Yes	Yes
GG7	**0.526**	0.479	Yes	Yes

Πίνακας 3.2: Τα αποτελέσματα των μέτρων αξιολόγησης για τις επτά γραμματικές που προέκυψαν
(Πηγή: [25])

Από τον Πίνακας 3.2 αξίζει να σημειωθεί πως όλες οι γραμματικές που προέκυψαν ήταν διφορούμενες. Απέδειξαν έτσι ότι οι διφορούμενες γραμματικές μπορεί να είναι και ισχυρές. Επομένως, στον σχεδιασμό μιας SCFG γραμματικής δεν χρειάζεται να εξαλείψουμε την αμφισημία για να είναι αποτελεσματική στην πρόβλεψη.

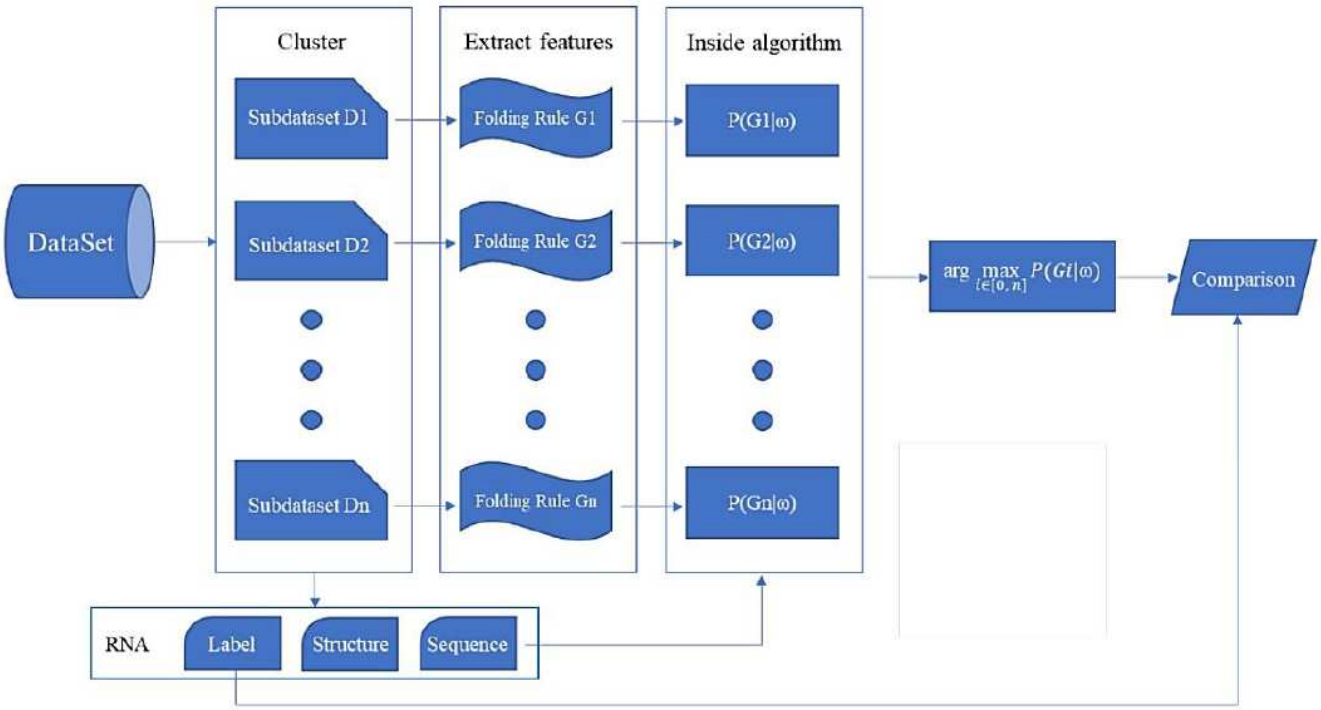
Συμπεραίνουμε επίσης από τον πίνακα ότι κάποιες ισχυρές γραμματικές μπορεί να είναι και ατελείς. Ορίζουμε μια γραμματική ως πλήρη (complete) αν έχει έναν κανόνα παραγωγής για όλες τις πιθανές δομές. Επομένως, η ικανότητα μιας γραμματικής να δημιουργεί όλες τις δομές δεν χρειάζεται να είναι αναγκαστική παράμετρος στον σχεδιασμό της γραμματικής. Τέλος, βρέθηκαν πολλές SCFG

γραμματικές με ισχυρή ακρίβεια πρόγνωσης, εξίσου καλές αλλά και καλύτερες από εκείνες που σχεδιάστηκαν χειροκίνητα.

3.7 Research on folding diversity in statistical learning methods for RNA secondary structure prediction

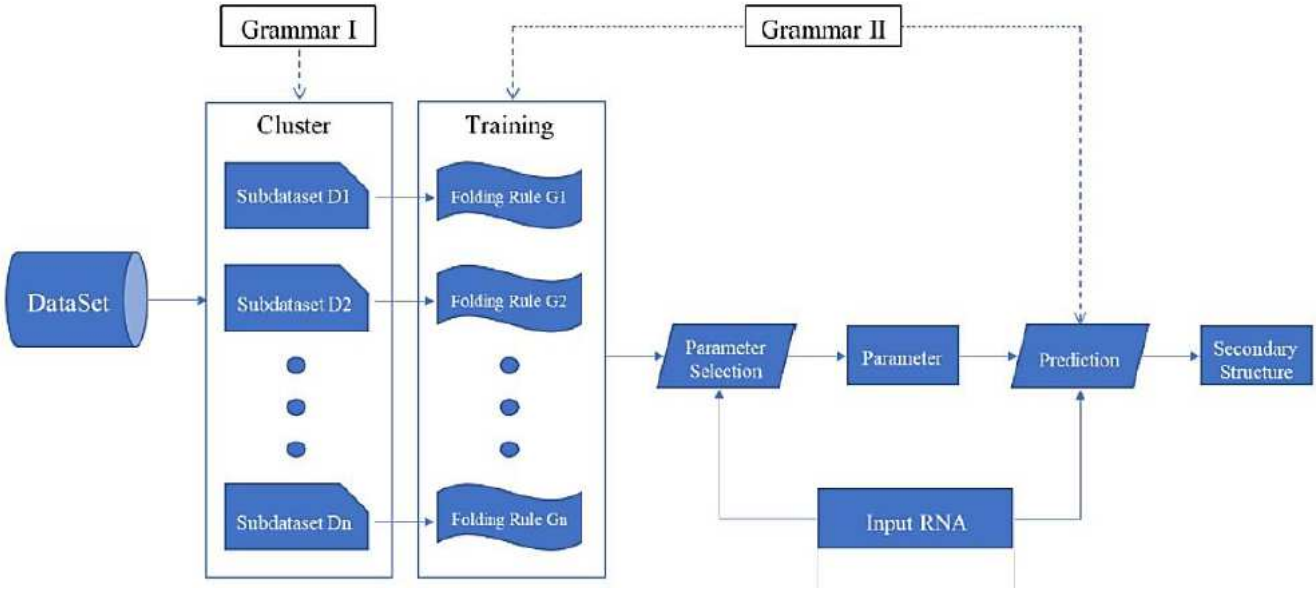
Το 2017, οι Zhu Y., Xie Z., Li Y, Zhu M και Chen Y. σχεδίασαν ένα μοντέλο στατιστικής μάθησης βασισμένο σε κανόνες αναδίπλωσης του RNA για να βελτιώσουν την ακρίβεια πρόβλεψης της δευτεροταγούς δομής του RNA και πρότειναν ένα εργαλείο, το FocusFold [36] για τη διεξαγωγή της πρόβλεψης. Οι κανόνες αναδίπλωσης βασίζονται σε στοχαστικές γραμματικές χωρίς συμφραζόμενα (SCFG). Το μοντέλο στατιστικής εκμάθησης (statistical learning model) εκπαιδεύει τις παραμέτρους για να περιγράψει τους κανόνες αναδίπλωσης και προβλέπει τις άγνωστες δομές RNA μέσω της χρήσης της υπάρχουσας γνωστής δομής του RNA [36].

Με βάση την Εικόνα 3.11, πρότειναν μια μέθοδο που βασίζεται σε SCFG γραμματικές για να ταξινομήσουν τις ακολουθίες RNA με άγνωστες δομές [36]. Πιο συγκεκριμένα, χρησιμοποίησαν ένα σύνολο δεδομένων (Data Set) με γνωστές δομές RNA και το διαίρεσαν σε n υποσύνολα (subdataset cluster) σύμφωνα με τις διαφορές στους κανόνες αναδίπλωσης. Ορίζουν ως κανόνα αναδίπλωσης μια συγκεκριμένη SCFG γραμματική. Για κάθε υποσύνολο εξάγουν έναν κανόνα αναδίπλωσης G_i, στον οποίο εφαρμόζουν τον inside αλγόριθμο για να βρουν την πιθανότητα $P(G_i \mid \omega)$. Η $P(G_i \mid \omega)$ είναι η πιθανότητα η RNA ακολουθία εισόδου να ανήκει στην συγκεκριμένη γραμματική. Στη συνέχεια, βρίσκουν τη μέγιστη πιθανότητα η γραμματική G_i να παράγει την RNA ακολουθία ω.



Εικόνα 3.11: Διάγραμμα ταξινόμησης κανόνων αναδίπλωσης
(Πηγή [36])

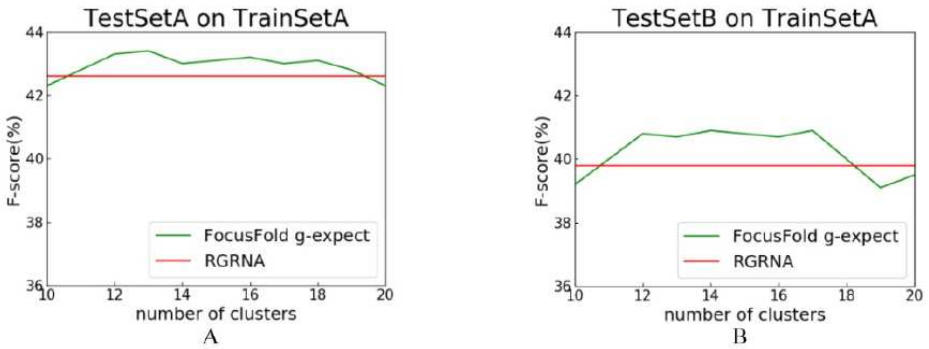
Οι ερευνητές πρότειναν επιπλέον ένα εργαλείο πρόβλεψης της δευτεροταγούς δομής του RNA, το FocusFold, του οποίου η δομή φαίνεται στην Εικόνα 3.12. Το τμήμα συστοιχίας (clustering module) και τμήμα εκπαίδευσης (training module) εξάγουν τους κανόνες αναδίπλωσης, όπως αναφέρθηκε παραπάνω. Το τμήμα επιλογής παραμέτρων (parameter selection module) λαμβάνει ως είσοδο την άγνωστης δομής ακολουθία RNA και τους κανόνες αναδίπλωσης του προηγούμενου τμήματος και διαλέγει τον κανόνα που ταιριάζει καλύτερα στην ακολουθία εισόδου. Η γραμματική με την μεγαλύτερη πιθανότητα που προκύπτει και η ακολουθία εισόδου εκχωρούνται ως είσοδοι στο τμήμα πρόγνωσης (prediction module) για να προκύψει η προβλεπόμενη δευτεροταγής δομή. Κάθε τμήμα του Focusfold μπορεί να υλοποιηθεί ανεξάρτητα με διαφορετικούς αλγορίθμους. Για την πρόβλεψη της δομής χρησιμοποιήθηκε ο αλγόριθμος μέγιστης αναμενόμενης ακρίβειας (maximum expected accuracy - MEA).



Εικόνα 3.12: Δομή του Focusfold

(Πηγή [36])

Για την αξιολόγηση της [36], οι ερευνητές σύγκριναν το Focusfold με ένα άλλο εργαλείο, το RGRNA [53], που προβλέπει την δευτεροταγή δομή με βάση την ελάχιστη ελεύθερη ενέργεια (MFE). Χρησιμοποίησαν ένα σύνολο δεδομένων με γνωστές RNA δομές (Trainset A) από διαφορετικού τύπου και μήκους RNA ακολουθίες. Τα αποτελέσματα με βάση το F-score ως μονάδα μέτρησης φαίνονται παρακάτω στην Εικόνα 3.13 και είναι σαφώς υπέρ του Focusfold, που σημαίνει ότι το Focusfold επιτυγχάνει καλύτερη ακρίβεια πρόβλεψης.



Εικόνα 3.13: Σύγκριση του FocusFold και του RGRNA A. Το TrainSetA αποτελεί τα δεδομένα εκπαίδευσης και τοTestSetA αποτελεί το σύνολο δοκιμών B. Το TrainSetA αποτελεί τα δεδομένα εκπαίδευσης και τοTestSetB αποτελεί το σύνολο δοκιμών

(Πηγή: [36])

Τέλος, η σωστή ταξινόμηση των δεδομένων και ο αριθμός των υποσυνόλων που προκύπτουν χρησιμοποιούνται ως δείκτης αξιολόγησης. Σύμφωνα με τη βιβλιογραφία [36], η επίτευξη καλύτερης πρόβλεψης σχετίζεται με το σύνολο δεδομένων που χρησιμοποιείται για την εκπαίδευση της μεθόδου και όχι με την καθαυτή πρόβλεψη των δεδομένων. Όσο καταλληλότερη είναι η δημιουργία και ο αριθμός των κανόνων αναδίπλωσης τόσο ευστοχότερη είναι η πρόβλεψη της δομής.

3.8 Μέθοδοι δομικής στοίχισης – Stemloc & Consan

Το 2005 ο Holmes Ian [38] και το 2006 οι Dowell R. και Eddy S. [39] παρουσίασαν δυο παρόμοιες προσεγγίσεις βασισμένες στην μέθοδο της ταυτόχρονης στοίχισης και την πρόβλεψης δευτεροταγούς δομής δύο αλληλουχιών RNA, την μέθοδο της δομικής στοίχισης (structural alignment), με την χρήση συζευγμένων (pair) στοχαστικών γραμματικών χωρίς συμφραζόμενα (pairSCFG). Μια pairSCFG είναι μια SCFG γραμματική που έχει επεκταθεί ώστε τα μη τερματικά σύμβολα να εκπέμπουν συσχετισμένα ζεύγη βάσεων για δυο ακολουθίες x και y. Για παράδειγμα, η συμβολική ένδειξη $^{a}_{b}P^{a'}_{b'}$ υποδηλώνει ένα ζεύγος βάσεων a και a’ στην ακολουθία x και ένα ζεύγος βάσεων b και b’ στην ακολουθία y [39]. Και οι δυο προτάσεις επιτυγχάνουν μείωση στον χρόνο εκτέλεσης και στις απαιτήσεις μνήμης για την εκτέλεση των αλγορίθμων δυναμικού προγραμματισμού για pairSCFGs (alignment-and-folding algorithms), περιορίζοντας τα πιθανά ζεύγη βάσεων και τις στοιχίσεις μεταξύ των ακολουθιών. Επιπλέον παρουσίασαν ένα πρόγραμμα ο καθένας, το Stemloc [38] και το ConSan[39] αντίστοιχα, που υλοποιούν τους αλγορίθμους για τον περιορισμό των δεδομένων ανάλυσης (constrained versions).

3.8.1 Accelerated probabilistic inference of RNA structure evolution

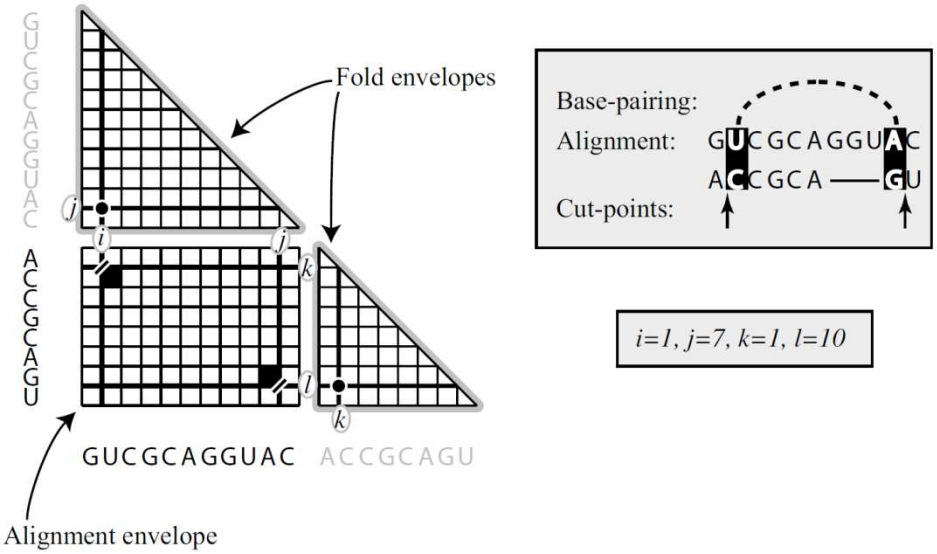
Πιο συγκεκριμένα, ο Holmes παρουσίασε προσαρμοσμένες εκδόσεις των Inside, Outside, CYK και KYC αλγορίθμων που βασίζονται στην προσέγγιση των πλαισίων (envelopes) αναδίπλωσης και στοίχισης που είχε ήδη παρουσιάσει το 2002 [37]. Για δυο ακολουθίες X και Y, η πολυπλοκότητα χρόνου για παράλληλη στοίχιση και σύγκριση δομής είναι $O(X^3Y^3)$ [37]. Η προσέγγιση των πλαισίων αναδίπλωσης και στοίχισης (fold and alignment envelopes) έχει στόχο την μείωση του

χρόνου εκτέλεσης των αλγορίθμων για pairSCFGs και κάνει δυνατή και αποδοτική την εκτέλεσή τους ακόμα και για μεγάλες ακολουθίες RNA. Με την χρήση του πλαισίου αναδίπλωσης (fold envelopes) προ-υπολογίζονται όλα τα πιθανά ζεύγη βάσεων κάθε ακολουθίας και δημιουργείται ένα υποσύνολο με τα 1000 καταλληλότερα ζεύγη. Με την χρήση του πλαισίου στοίχισης (alignment envelopes) προ-υπολογίζονται όλες οι πιθανές στοιχίσεις των κατά ζευγών ακολουθιών και δημιουργείται ένα αντίστοιχο υποσύνολο που περιέχει μόνο τις 100 καλύτερες ευθυγραμμίσεις. Το μέγεθος των συνόλων μπορεί να μεταβληθεί αλλά, μετά από διάφορα πειράματα σε σύνολα RNA ακολουθιών, ο Holmes κατέληξε σε αυτά τα δυο μεγέθη επειδή παρουσιάζουν τις μικρότερες απαιτήσεις σε CPU και μνήμη με τα καλύτερα δυνατά αποτελέσματα στην στοίχιση και πρόβλεψη.

Ο Inside αλγόριθμος υπολογίζει τις πιθανότητες όλων των δέντρων ανάλυσης, μιας συγκεκριμένης pairSCFG γραμματικής, που δημιουργεί τις ακολουθίες X και Y. Ο ψευδοκώδικας του αλγορίθμου είναι διαθέσιμος στην [38]. Ο Outside αλγόριθμος μπορεί να χρησιμοποιηθεί αντιστρόφως του Inside αλγορίθμου για την ανάκτηση των a posteriori/ εκ των υστέρων πιθανοτήτων των δεδομένων ζευγών βάσεων/στηλών. Ο προσαρμοσμένος Inside αλγόριθμος από τον Holmes δημιουργεί τα υποσύνολα αναδίπλωσης με την εξής μέθοδο: επιλέγεται η SCFG γραμματική για την μοντελοποίηση της αναδίπλωσης κάθε RNA ακολουθίας, στη συνέχεια φιλτράρονται ορισμένες δομές RNA με στοχαστική αναδρομή και τέλος δημιουργείται το πλαίσιο από την ένωση των υποακολουθιών των αναδρομών. Για την δημιουργία του πλαισίου στοίχισης ακολουθείται η ίδια μέθοδος με την προαναφερόμενη με την διαφορά ότι χρησιμοποιείται HMM (Hidden Markov Model) για την στοίχιση κάθε ακολουθίας.

Ο pairSCFG CYK αλγόριθμος είναι παρόμοιος με τον Inside αλγόριθμο με την διαφορά ότι δεδομένης της pairSCFG γραμματικής και δυο ακολουθιών εισόδου, υπολογίζει το δέντρο ανάλυσης με την μέγιστη πιθανότητα. Αντιστρόφως, ο KYC αλγόριθμος ανακτά το δέντρο ανάλυσης με την μεγαλύτερη πιθανότητα. Ο προσαρμοσμένος CYK αλγόριθμος δημιουργεί τα υποσύνολα αναδίπλωσης με την εξής μέθοδο: διαλέγει την SCFG γραμματική για την μοντελοποίηση της αναδίπλωσης κάθε RNA ακολουθίας, στη συνέχεια υπολογίζει την μέγιστη πιθανότητα με στοχαστική αναδρομή και τέλος επιλέγει τις υποακολουθίες με την μέγιστη πιθανότητα να είναι μεγαλύτερη από ένα κατώφλι. Αντίστοιχα, χρησιμοποιείται HMM (Hidden Markov Model) για την δημιουργία του πλαισίου στοίχισης.

Στην Εικόνα 3.14 φαίνονται τα πλαίσια αναδίπλωσης (τριγωνικά πλέγματα) και στοίχισης (ορθογώνιο πλέγμα) του προσαρμοσμένου Inside αλγορίθμου.



Εικόνα 3.14: Πλαίσια αναδίπλωσης και στοίχισης - Fold και aligned envelopes
(Πηγή [38])

Οι παραπάνω αλγόριθμοι και μέθοδοι υλοποιούνται στην εφαρμογή που ονομάστηκε Stemloc και είναι γραμμένη σε γλώσσα C++. Η εφαρμογή μπορεί να δεχθεί ως είσοδο είτε πολλαπλές RNA ακολουθίες είτε κατά ζεύγη και προσφέρει την δυνατότητα παραμετροποίησης των περιορισμών, με βάση των οποίων γίνεται η δημιουργία των πλαισίων, με την χρήση Unix εντολών. Για την αξιολόγηση της [38], τα πειράματα εκτελέστηκαν μόνο με ευθυγραμμισμένες ακολουθίες κατά ζεύγη. Η pairSCFG γραμματική που δημιουργήθηκε για το Stemloc είναι μη-διφορούμενη (unambiguous). Ο Holmes συμπέρανε, μετά τα πειράματά του, ότι μια μη-διφορούμενη γραμματική βελτιώνει την ακρίβεια της στοίχισης και της πρόβλεψης στην προσέγγιση [38] που πρότεινε.

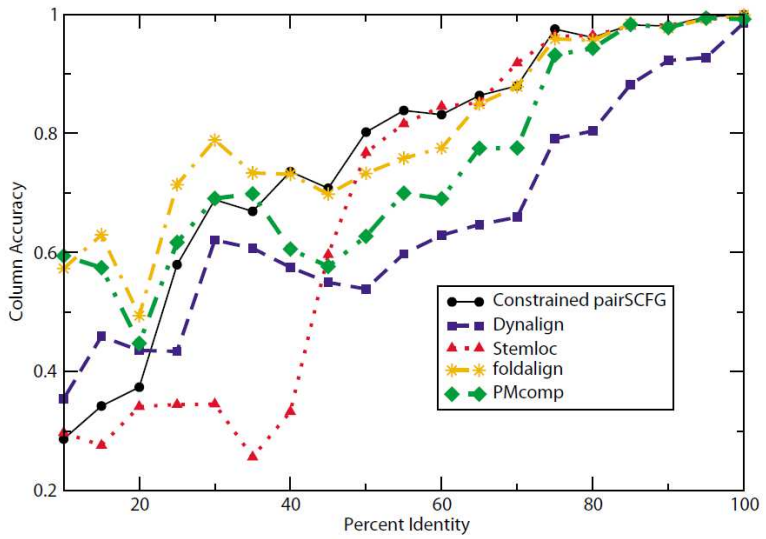
3.8.2 Efficient pairwise RNA structure prediction and alignment using sequence alignment constraints

Οι Dowell R. και Eddy S. [39] παρουσίασαν μια πολύ παρόμοια προσέγγιση με εκείνη του Holmes [38]. Οι ερευνητές της [39] προσάρμοσαν τον CYK αλγόριθμο δομικής στοίχισης βασισμένο σε μια pairSCFG γραμματική για να επιλέξουν ένα ζεύγος θέσεων από τις δύο ακολουθίες εισόδου που

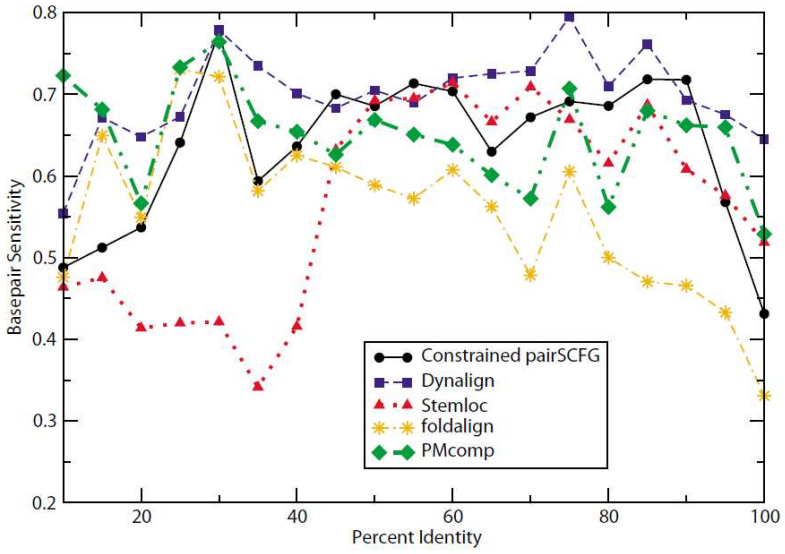
έχουν μεγάλη πιθανότητα να ευθυγραμμιστούν. Το επιλεγμένο αυτό ζεύγος ονομάζεται pin. Τα pins είναι τα αντίστοιχα πλαίσια στοίχισης που εισήγαγε ο Holmes. Δύο διαδοχικά pins ορίζουν τα όρια στοίχισης, όπου εκτελείται η στοίχιση και δίπλωση. Ο ψευδοκώδικας του αλγορίθμου αυτού είναι διαθέσιμος στην στην εργασία των Dowell R. και Eddy S. [39] και το πρόγραμμα Consan αποτελεί την υλοποίηση του αλγορίθμου σε γλώσσα ANSI C.

Και οι δυο προσεγγίσεις είναι βασισμένες σε pairSCFGs. Η γραμματική που χρησιμοποιήθηκε για το Stemloc είναι αρκετά διαφορετική από τη γραμματική που χρησιμοποιήθηκε για το ConSan. Και στην εργασία των Dowell R. και Eddy S. [39] όμως, η pairSCFG γραμματική που δημιουργήθηκε για το ConSan είναι μη-διφορούμενη κατά την δομή και κατά την στοίχιση (structural and alignment unambiguous). Αυτό σημαίνει ότι κάθε στοίχιση και δομή προκύπτουν μόνο από ένα δέντρο ανάλυσης. Επιπλέον, το ConSan λαμβάνει ως είσοδο ευθυγραμμισμένες ακολουθίες κατά ζεύγη ενώ το Stemloc λαμβάνει και πολλαπλές ευθυγραμμισμένες ακολουθίες.

Για την αξιολόγηση της εργασίας των Dowell R. και Eddy S. [39] άντλησαν δεδομένα από την βάση δεδομένων Rfam και χρησιμοποίησαν το PPV score ως μέτρο υπολογισμού της ακρίβειας στην πρόβλεψη. Σύγκριναν επιπλέον την στοίχιση μεταξύ της καταχωρημένης δομικής στοίχισης και της προβλεπόμενης στοίχισης με βάση την εργασία των Dowell R. και Eddy S. [39]. Επίσης σύγκριναν την εν λόγω εργασία [39] με τέσσερις άλλες υλοποιήσεις προσαρμοσμένων αλγορίθμων μεταξύ των οποίων και το Stemloc. Οι εικόνεςΕικόνα *3.15* καιΕικόνα *3.16* δείχνουν την απόδοση των πέντε υλοποιήσεων ως προς την ακρίβεια στοίχισης και την ακρίβεια πρόβλεψης αντίστοιχα.



Εικόνα 3.15: Συγκριτική απόδοση ως προς την ακρίβεια στοίχισης
(Πηγή [39])



Εικόνα 3.16: Συγκριτική απόδοση ως προς την ακρίβεια πρόβλεψης
(Πηγή [39])

Όπως φαίνεται στα παραπάνω αποτελέσματα, το Consan παράγει καλύτερες ευθυγραμμίσεις και δομές από το Stemloc. Σύμφωνα με τους Dowell και Eddy, αυτό συμβαίνει λόγω της ολοκληρωμένης στοίχισης σε αντίθεση με το Consan που η στοίχιση περιορίζεται σε ένα υποσύνολο των pins. Επιπλέον, και πάλι σύμφωνα με τους Dowell και Eddy, η γραμματική του

Stemloc ίσως να είναι διφορούμενη ως προς την στοίχιση, παρόλο που ισχυρίζεται ότι δεν είναι, με αποτέλεσμα να μην υπολογίζει σωστά τις υποβέλτιστες στοιχίσεις με βάση τις πιθανότητές τους.

3.9 Προσεγγίσεις SCFG γραμματικών για πρόβλεψη RNA δευτεροταγούς δομής συμπεριλαμβανομένων ψευδοκόμβων.

Η έρευνα των RNA δομών που περιέχουν ψευδοκόμβους έχουν προσελκύσει το ενδιαφέρον των ερευνητών εδώ και αρκετά χρόνια. Η πειραματική επαλήθευση της δομής αυτών των RNA δομών με φασματοσκοπία ή κρυσταλλογραφία είναι αρκετά ακριβή και χρονοβόρα. Έχουν γίνει αρκετές προτάσεις από πολλούς ερευνητές για την πρόβλεψη της RNA δευτεροταγούς δομής συμπεριλαμβανομένων ψευδοκόμβων βασισμένες στη χρήση γραμματικών. Κατά τον Brown, δεν υπάρχει γραμματική χωρίς συμφραζόμενα που να μπορεί να μοντελοποιήσει ψευδοκόμβους [42]. Στα πλαίσια των SCFG γραμματικών έχουν προταθεί υβριδικές προσεγγίσεις, όπως πολλαπλές στοχαστικές γραμματικές χωρίς συμφραζόμενα (stochastic multiple context free grammar - SMCFG) [27], [30], διασταυρώσεις δύο SCFG γραμματικών [42], παράλληλες γραμματικές που επικοινωνούν μεταξύ τους [43], [45], υλοποιήσεις με μοντέλα συνδιακύμανσης [44] κ.α. ώστε να είναι δυνατή η πρόβλεψη των ψευδοκόμβων.

Το 1995, οι Brown και Wilson [42] πρότειναν ένα μοντέλο διασταύρωσης SCFG γραμματικών που ήταν ικανό να διακρίνει την RNA δομή με ψευδοκόμβους μιας υδατοδιαλυτής βιταμίνη B, της βιοτίνης, ανάμεσα σε RNA ακολουθίες που είναι καταχωρημένες σε βάσεις δεδομένων. Η μια γραμματική θα υπολογίζει τις πιθανότητες για τα ζεύγη βάσεων A-U της RNA ακολουθίας και η άλλη γραμματική θα υπολογίζει τις πιθανότητες για τα ζεύγη βάσεων G-C. Η δομή θα περιλαμβάνει ψευδοκόμβους αν υπάρχει πιθανότητα ζευγών βάσεων και στις δυο γραμματικές παράλληλα. Η υπολογιστική πολυπλοκότητα ήταν $O(n^3)$, όπου n το μέγεθος της ακολουθίας.

Το 2003, οι Cai, Malmberg και Wu πρότειναν ένα μοντέλο παράλληλων γραμματικών που επικοινωνούν μεταξύ τους με απαιτήσεις μνήμης $O(n^4)$ [43]. Υλοποίησαν επίσης έναν αλγόριθμο δυναμικού προγραμματισμού σε γλώσσα C και C++ που λαμβάνει ως είσοδο μια SCFG γραμματική και υπολογίζει τη βέλτιστη δομική στοίχιση μιας ακολουθίας. Το 2006, οι Congzhou He και Liming Cai πρότειναν μια βελτιωμένη υλοποίηση της εργασίας των Cai, Malmberg και Wu [43] που

μειώνει τις απαιτήσεις μνήμης σε $O(n^2)$ [45] ώστε να είναι δυνατή η αναζήτηση ψευδοκόμβων σε μεγάλες RNA ακολουθίες

Το 2006, ο Garcia πρότεινε έναν αλγόριθμο βασισμένο σε μοντέλα συνδιακύμανσης για πρόβλεψη δευτεροταγούς δομής με ψευδοκόμβους, που μειώνει την πολυπλοκότητα σε $O(n^2 \log n)$ [44]. Πραγματοποίησε δυο υλοποιήσεις του αλγορίθμου, μια σειριακή σε GNU/Linux πλατφόρμα με 1024 Mbytes μνήμης και μια παράλληλη σε μια IRIX 6.5 platform με 4 παράλληλους επεξεργαστές και 1024 Mbytes μνήμης.

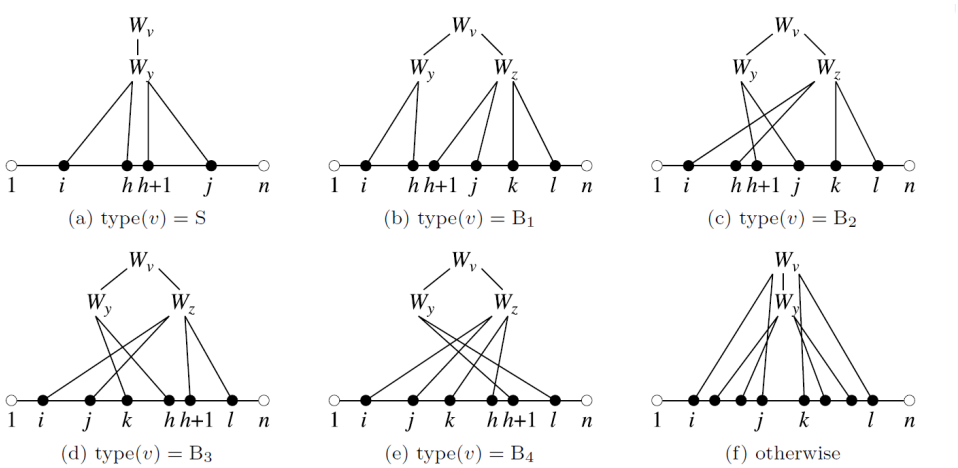
Παρακάτω παρατίθενται πιο αναλυτικά δυο υλοποιήσεις βασισμένες σε πολλαπλές στοχαστικές γραμματικές χωρίς συμφραζόμενα (stochastic multiple context free grammar - SMCFG).

3.9.1 RNA Pseudoknotted structure prediction using stochastic multiple context free grammar

Το 2006, οι Kato Y., Seki H., και Kasami T. πρότειναν μια μέθοδο πρόβλεψης της δευτεροταγούς δομής RNA που περιλαμβάνει και πρόβλεψη για ψευδοκόμβους βασισμένη σε πολλαπλές στοχαστικές γραμματικές χωρίς συμφραζόμενα (stochastic multiple context free grammar - SMCFG) [27]. Οι πολλαπλές γραμματικές χωρίς συμφραζόμενα (MCFG) είναι μια υποκατηγορία γενικευμένων γραμματικών χωρίς συμφραζόμενα (generalized context-free grammars) που εισήγαγε ο Pollard (1984) για να περιγράψει τη σύνταξη των φυσικών γλωσσών. Οι Fujii, Seki, Matsumura και Kasami υλοποίησαν τρεις επεκτάσεις των MCFG γραμματικών ώστε να υποστηρίζουν και πρόγνωση ψευδοκόμβων. Και κατ' επέκταση οι γραμματικές SMCFG εισάγουν την έννοια της στοχαστικότητας στις MCFG γραμματικές.

Στην εργασία των Kato, Yuki, Hiroyuki Seki και Tadao Kasami [27] υλοποιείται ένας αλγόριθμος δυναμικού προγραμματισμού (parsing algorithm) για την εύρεση του πιο πιθανού δέντρου δεδομένης μιας συμβολοσειράς εισόδου και ένας αλγόριθμος εκτίμησης των παραμέτρων πιθανότητας της γραμματικής βασισμένο στον αλγόριθμο inside-outside [23].

Ο αλγόριθμος συντακτικής ανάλυσης (parsing algorithm) για την εύρεση του πιο πιθανού δέντρου είναι βασισμένος στον CYK αλγόριθμο και λειτουργεί ως εξής. Έστω μια συμβολοσειρά $w = a_1 \cdots a_n$ και $\gamma_\upsilon(i,j)$, $\gamma_y(i,j,k,l)$ ο λογάριθμος των μέγιστων πιθανοτήτων ενός υποδέντρου που οδηγεί σε ένα μη τερματικό σύμβολο τερματικό W_υ για μια τελική υποακολουθία $a_i \cdots a_j$ και ενός υποδέντρου που οδηγεί σε ένα μη τερματικό σύμβολο τερματικό W_y για μια πλειάδα (tuple) τελικών υποακολουθιών $(a_i \cdots a_j, a_k \cdots a_l)$ αντίστοιχα. Οι παράμετροι $\gamma_\upsilon(i, i-1)$ και $\gamma_y(i, i-1, j, j-1)$ είναι ο λογάριθμος των μέγιστων πιθανοτήτων για μια κενή ακολουθία ε και ένα ζεύγος ε. Επίσης, οι μεταβλητές $\tau_\upsilon(i, j)$ και $\tau_y(i, j, k, l)$ είναι οι traceback μεταβλητές για την κατασκευή του δέντρου. Ο CYK αλγόριθμος χρησιμοποιεί ένα πλέγμα δυναμικού προγραμματισμού πέντε διαστάσεων για τον υπολογισμό του γ, το οποίο οδηγεί στο λογαριθμικό $\log P(w, \hat{\pi} \mid \theta)$ όπου $\hat{\pi}$ είναι το πιο πιθανό δένδρο και το θ είναι ένα πλήρες σύνολο παραμέτρων πιθανότητας. Για την ανάλυση του βέλτιστου δέντρου υπολογίζονται οι traceback μεταβλητές τ_υ και τ_y. Η διαδικασία υπολογισμού του γ φαίνεται στην Εικόνα 3.17.



Εικόνα 3.17: Διαδικασία υπολογισμού του γ

(Πηγή: [27])

Η πιθανότητα ενός δένδρου ορίζεται ως το προϊόν της πιθανότητας μετάβασης (transition probability) και εκπομπής (emission probability) όλων των κανόνων που χρησιμοποιούνται στην παραγωγή του δέντρου. Για κάθε κανόνα ορίζεται η πιθανότητα μετάβασης $t_\upsilon(y)$ και εκπομπής $e_\upsilon(a_i, a_j)$, όπου υ και y υποδηλώνουν τα μη τερματικά σύμβολα που εμφανίζονται στην αριστερή και την δεξιά πλευρά του εφαρμοζόμενου κανόνα, αντίστοιχα, και a_i και a_j είναι τερματικά σύμβολα που το καθένα υποδηλώνει μία από τις βάσεις a, g, c, u. Για τον εκτίμηση των παραμέτρων

πιθανότητας της γραμματικής, οι ερευνητές της εργασίας [27] υλοποίησαν τον αλγόριθμο inside-outside βασισμένο στον EM (expectation maximization) αλγόριθμο.

Οι ερευνητές για την εφαρμογή της παραπάνω μεθόδου υλοποίησαν τον CYK αλγόριθμο σε γλώσσα ANSI C σε υπολογιστή Intel Pentium D με CPU 2,8 GHz και 2 GB μνήμη RAM. Χρησιμοποίησαν ως δεδομένα τις RNA οικογένειες Corona_pk_3, HDV_ribozyme και Tombus_3_IV από την βάση δεδομένων της Rfam [28]. Οι RNA οικογένειες αυτές περιέχουν και ψευδοκόμβους στην δευτεροταγή δομή τους με βάση τις εγγραφές τους στην Rfam. Η ακρίβεια της πρόγνωσης μετρήθηκε από την αναλογία του αριθμού των σωστών ζευγών βάσεων που προβλέπονται από τον αλγόριθμο προς το συνολικό αριθμό των προβλεπόμενων ζευγών βάσεων (precision metric) και από την αναλογία του αριθμού των σωστών ζευγών βάσεων που προβλέπονται από τον αλγόριθμο προς τον συνολικό αριθμό των ζευγών βάσης που είναι καταχωρημένες στην Rfam (recall metric). Με βάση τα αποτελέσματα που φαίνονται στον *Πίνακας 3.3*, προκύπτει το συμπέρασμα ότι η παραπάνω μέθοδος ήταν αρκετά αποτελεσματική στην ακρίβεια πρόγνωσης της δευτεροταγούς δομής.

Family	Precision [%]			Recall [%]			CPU time [sec]		
	Average	Min	Max	Average	Min	Max	Average	Min	Max
Corona_pk_3	99.4	94.4	100.0	99.4	94.4	100.0	27.8	26.0	30.4
HDV_ribozyme	100.0	100.0	100.0	100.0	100.0	100.0	252.1	219.0	278.4
Tombus_3_IV	100.0	100.0	100.0	100.0	100.0	100.0	244.8	215.2	257.5

Πίνακας 3.3: Αποτελέσματα πρόγνωσης της προσέγγισης [27]

(Πηγή: [27])

Επίσης, στην Εικόνα 3.18 φαίνεται πολύ χαρακτηριστικά ότι η προβλεπόμενη δομή της Corona_pk_3 με βάση την παραπάνω μέθοδο ήταν σχεδόν η ίδια με την δομή στην Rfam. Τα σωστά προβλεπόμενα ζεύγη είναι υπογραμμισμένα.

[Καταχωρημένη δομή στην Rfam]

CUAGUCUUAUACACAAUGGUAAGCCAGUGGUAGUAAAGGUAUAAGAAAUUUGCUAC
UAUGUUA

[[[[[[[[((((((((((( ]]]]]]]])))))))))))

[Προβλεπόμενη δομή με βάση την [27]]

CUAGUCUUAUACACAAUGGUAAGCCAGUGGUAGUAAAGGUAUAAGAAAUUUGCUAC
UAUGUUA

[[[[[[[[((((((((((( ]]]]]]]]))))))))))))

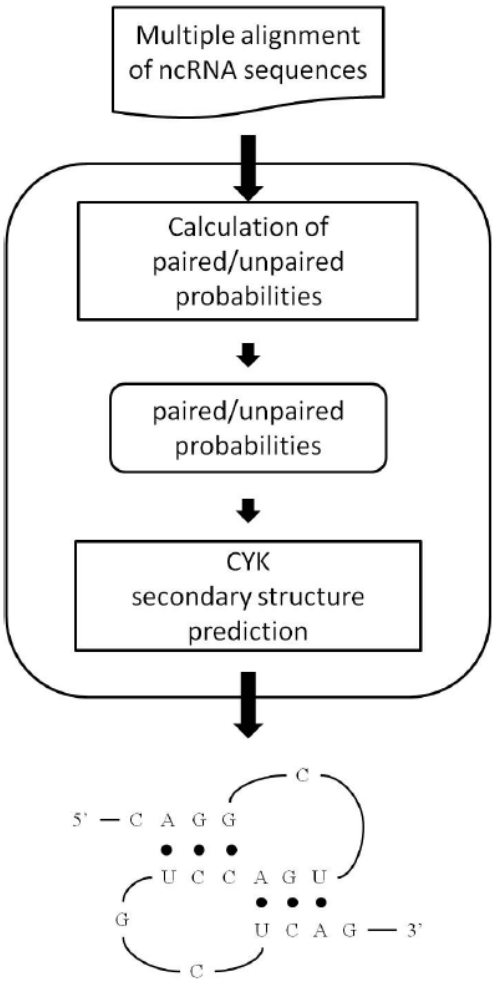
Εικόνα 3.18: Σύγκριση προβλεπόμενης δομής της Corona_pk_3 με την καταχωρημένη της δομή στην RFam *(Πηγή: [27])*

3.9.2 A Grammar-Based Approach to RNA Pseudoknotted Structure Prediction for Aligned Sequences

Το 2009 οι Mizoguchi N., Kato Y. και Seki H. υλοποίησαν [27] σε συνδυασμό με συγκριτικές μεθόδους, ώστε να λαμβάνει ως είσοδο πολλαπλές ακολουθίες RNA και να παράγει την κοινή δευτεροταγή δομή συμπεριλαμβανομένων ψευδοκόμβων [30]. Αυτή ήταν μέχρι τότε η πρώτη μελέτη για πρόβλεψη δευτεροταγούς δομής συμπεριλαμβανομένων ψευδοκόμβων με βάση τη γραμματική σε συνδυασμό με την μέθοδο συγκριτικής ανάλυσης αλληλουχίας. Η ίδια ομάδα ερευνητών είχε υλοποιήσει λίγο νωρίτερα μια μέθοδο πρόβλεψης σε συνδυασμό με την συγκριτική ανάλυση ακολουθιών [29] αλλά η μέθοδος αυτή λάμβανε ως είσοδο στοιχισμένες ακολουθίες κατά ζεύγη. Αυτό είχε ως αποτέλεσμα [29] να μην μπορεί να χρησιμοποιήσει τις εξελικτικές πληροφορίες των ακολουθιών που θα μπορούσαν να αυξήσουν την ακρίβεια της πρόβλεψης.

Η μέθοδος των Mizoguchi, Nobuyoshi, Yuki Kato και Hiroyuki Seki [30] λαμβάνει ως είσοδο πολλαπλές μη κωδικοποιημένες αλληλουχίες RNA και υπολογίζει τις πιθανότητες ζευγών και μη συζευγμένων βάσεων της στήλης εισόδου ώστε να εξάγει την εξελικτική πληροφορία και να την ενσωματώσει στην πρόβλεψη. Στη συνέχεια, ο CYK αλγόριθμος, χρησιμοποιώντας αυτές τις πιθανότητες, προβλέπει την κοινή δευτεροταγή δομή των ακολουθιών με βάση μια συγκεκριμένη

SMCFG (stochastic multiple context free grammar) γραμματική. Μια επισκόπηση της μεθόδου [30] φαίνεται στην Εικόνα 3.19.



Εικόνα 3.19: Επισκόπηση της μεθόδου [30]
(Πηγή [30])

Οι πιθανότητες ζευγών και μη συζευγμένων βάσεων ορίζονται ως η πιθανότητα μετάβασης $t_{\upsilon}(y)$, η πιθανότητα σύζευξης pp(i,j) και η πιθανότητα μη σύζευξης upp(i). Η πιθανότητα σύζευξης pp(i,j) ορίζεται ως η τάση ότι το ζεύγος των στηλών (i,j) σχηματίζει ένα κανονικό ζεύγος βάσεων (GC, AU, GU) στην στοίχιση εισόδου και αντίστοιχα η πιθανότητα μη σύζευξης ισούται με 1 - pp(i,j). Μετά τον υπολογισμό των πιθανοτήτων, ο CYK αλγόριθμος προβλέπει την δομή με τρόπο παρόμοιο με αυτόν της εργασίας [27].

Για την αξιολόγηση της μεθόδου οι ερευνητές άντλησαν τα δεδομένα τους από την βάση δεδομένων Rfam και χρησιμοποίησαν οκτώ RNA οικογένειες που περιέχουν και ψευδοκόμβους. Σύγκριναν τα αποτελέσματά τους με τα αποτελέσματα εφαρμογής δυο άλλων μεθόδων στα ίδια δεδομένα. Η πρώτη μέθοδος αναφοράς ήταν η προηγούμενη υλοποίηση τους με συγκριτική ανάλυση ακολουθίας [29] και η δεύτερη μέθοδος αναφοράς, hxmatch, ήταν μια μέθοδος που συνδύαζε θερμοδυναμική και συγκριτική ανάλυση ακολουθίας [31] και θεωρούνταν μια από τις καλύτερες μεθόδους για την πρόβλεψη δευτεροταγούς δομής με ψευδοκόμβους. Τα αποτελέσματα φαίνονται στον Πίνακας 3.4. Το μέτρο σύγκρισης του πίνακα είναι το F-score σε ποσοστό %.

Family	# of sequences	Length	hxmatch	Pair-SMCFG	Proposed method
Corona_pk3	14	67	89.5	80.1	**97.1**
TMV_UPD-PK3	27	33	87.5	82.3	**88.9**
TMV_UPD-PK2	3	22	**100.0**	87.5	82.4
SBWMV2_UPD-PKb	3	29	**90.9**	78.3	81.9
UPD_PK2	6	22	**83.3**	75.4	66.7
UPSK	6	23	60.0	**67.7**	57.1
PK_IAV	32	48	N/A	29.2	**48.0**
Prion	157	57	**47.6**	17.3	0.0
Average			**69.9**	64.7	65.3

Πίνακας 3.4: Συγκριτική ακρίβεια πρόβλεψης μεταξύ της [29] [30] και [31]
(Πηγή [30])

Σύμφωνα με τα παραπάνω αποτελέσματα, η ακρίβεια πρόβλεψης της [30] είναι καλύτερη από εκείνη της [29] και λίγο χειρότερη από εκείνη της [31]. Η χρονική πολυπλοκότητα της [29] και της [30] είναι $O(n^4)$ ενώ της [31] είναι $O(n^3)$.

4 Υλοποιήσεις Υλικού με βάση SCFG γραμματικές

Σκοπός αυτού του κεφαλαίου είναι να παρουσιάσει ορισμένες υλοποιήσεις σε επίπεδο υλικού που έχουν προταθεί για την χρήση των SCFG γραμματικών στην ανάλυση και πρόβλεψη της δευτεροταγούς δομής του RNA. Οι προσεγγίσεις αυτές είναι παράλληλες προσεγγίσεις σε FPGA πλακέτες είτε αλγορίθμων είτε σειριακών εφαρμογών. Σκοπό έχουν να βελτιώσουν το υπολογιστικό κόστος ενώ παράλληλα να μειώσουν το κόστος διαχείρισης και συντήρησης αντίστοιχων υλοποιήσεων σε συστοιχίες υπολογιστών.

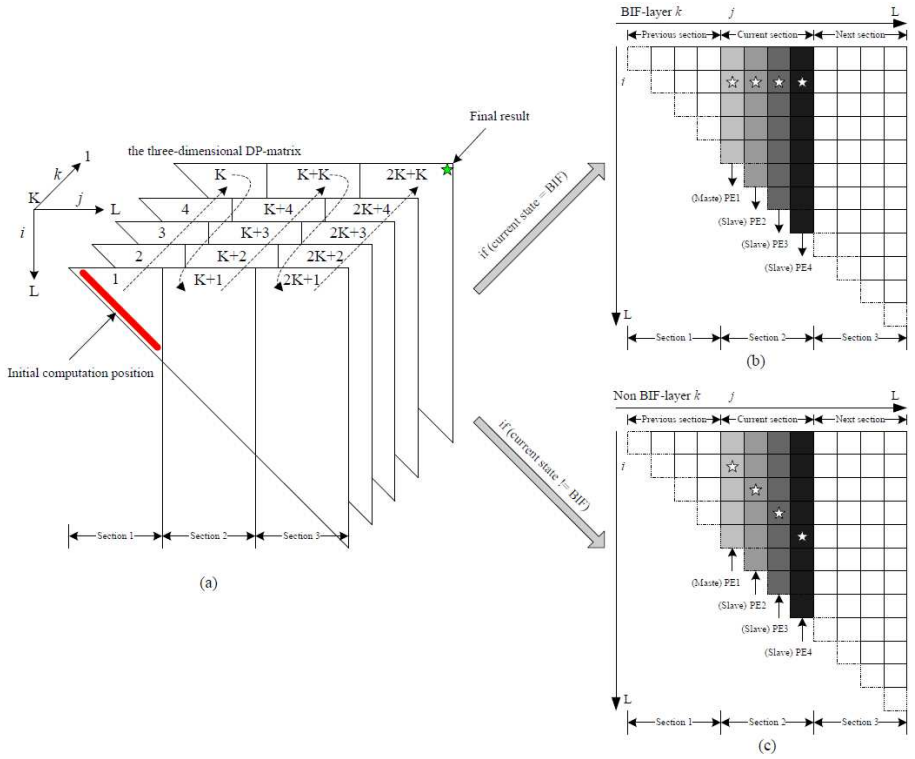
4.1 Fine-grained Parallel Application Specific Computing for RNA Secondary Structure Prediction Using SCFGs on FPGA

Το 2009, οι Dou Y, Xia F, Jiang J παρουσίασαν την πρώτη fine-grained παράλληλη υλοποίηση του CYK/inside αλγορίθμου σε μια FPGA πλακέτα [35] με σκοπό την επιτάχυνση στο χρόνο εκτέλεσης του. Η εν λόγω εργασία [35] είναι μια αρκετά παρόμοια υλοποίηση με την εργασία [33]. Η κύρια διαφορά είναι ότι στην εργασία [33] διαιρεί έναν 2D dynaming programming matrix ενώ στην [35] διαιρεί έναν 3D dynaming programming matrix.

Η υλοποίηση της εργασίας των Dou Y, Xia F και Jiang J [35] πραγματοποιείται σε FPGA πλακέτα με βάση τη συστολική δομή συστοιχίας (systolic array structure) και με βάση τα πειραματικά αποτελέσματα επιτυγχάνεται έως 14 φορές μεγαλύτερη επιτάχυνση υπολογισμού στην FPGA πλακέτα σε σύγκριση με την εφαρμογή Infernal σε έναν Pentium υπολογιστή.

Όπως έχει ήδη αναφερθεί ο CYK/inside αλγόριθμος μοντελοποιεί μια RNA ακολουθία με βάση ένα μοντέλο συνδιακύμανσης (covariance model - CM), μια SCFG γραμματική με επτά τύπους καταστάσεων, για να προσδιοριστεί η πιθανότητα ότι η αλληλουχία ανήκει στην RNA οικογένεια. Ο τρισδιάστατος CYK/inside αλγόριθμος δυναμικού προγραμματισμού προσδιορίζει το αποτέλεσμα της πιθανότητας υπολογίζοντας επαναληπτικά έναν τρισδιάστατο πίνακα M(i, j, k) όπως φαίνεται στην *Εικόνα 4.1*. Το κελί M(1,L,1) του πίνακα περιέχει την πιθανότητα του βέλτιστου δέντρου της RNA ακολουθίας.

Ο παράλληλος CYK/inside αλγόριθμος φαίνεται στην *Εικόνα 4.1*. Αρχικά ο αλγόριθμος χωρίζει τον 3-διάστατο πίνακα, κατά μήκος της διάστασης k, σε μικρότερους 3-διάστατους πίνακες, που ονομάζονται ενότητες (sections) και υπολογίζει την κάθε ενότητα μια προς μια. Αυτό έχει ως στόχο να ελαχιστοποιήσει τις απαιτήσεις αποθήκευσης. Κάθε ενότητα αποτελείται από k επίπεδα (*Εικόνα 4.1*.a). Για την καλύτερη ισορροπία στον φόρτο υπολογισμού (load balancing), κάθε επίπεδο αποτελείται από *N* στήλες (*Εικόνα 4.1*.b και **Error! Reference source not found.**.c) και οι υπολογισμοί κάθε στήλης ανατίθενται σε μια μονάδα επεξεργασίας (processing element - PE). Στην *Εικόνα 4.1*, το κάθε επίπεδο αποτελείται από 4 στήλες. Η συμπλήρωση του πίνακα ξεκινά από τα κελιά που βρίσκονται στη διαγώνιο της πρώτης ενότητας, τα κελιά δηλαδή της κόκκινης γραμμής που φαίνεται στην *Εικόνα 4.1*.α. Οι διακεκομμένες γραμμές της *Εικόνα 4.1*.a υποδεικνύουν την σειρά με την οποία υπολογίζονται τα κελιά του πίνακα. Στην *Εικόνα 4.1*.b φαίνονται τα κελιά του πίνακα που υπολογίζονται παράλληλα σε ένα επίπεδο διακλάδωσης (BIF-layer) και στην *Εικόνα 4.1*.c φαίνονται τα κελιά του πίνακα που υπολογίζονται παράλληλα σε ένα επίπεδο που δεν είναι επίπεδο διακλάδωσης (non BIF-layer). Όταν συμπληρωθούν τα κελιά της πρώτης ενότητας, τότε οι μονάδες επεξεργασίας ξεκινάνε να υπολογίζουν τα κελιά της δεύτερης ενότητας, και ούτω καθεξής.

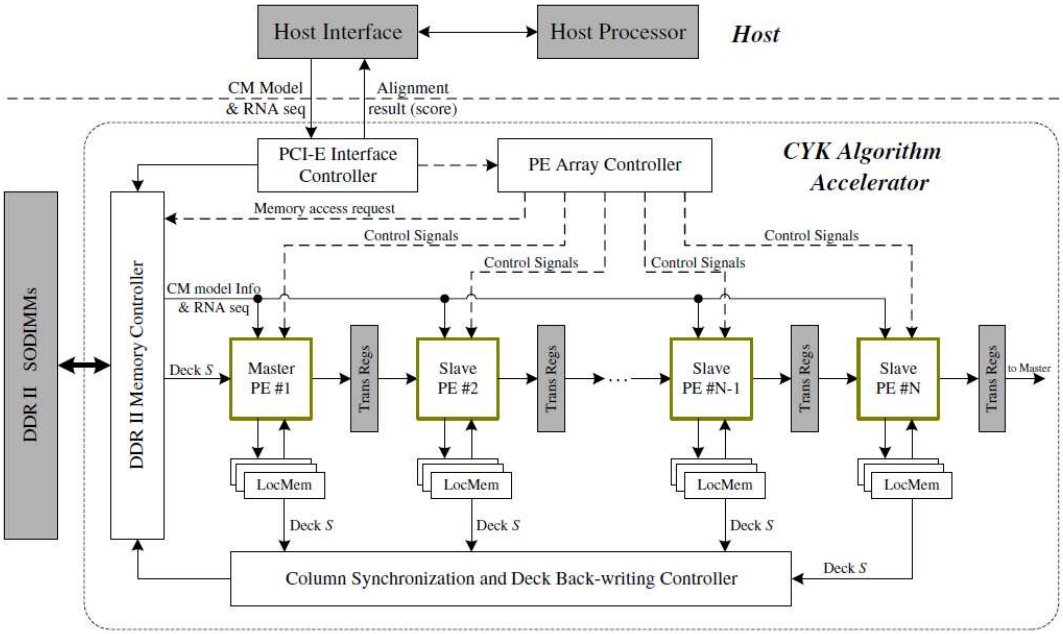


Εικόνα 4.1: Η κατανομή εργασιών και η σειρά υπολογισμού με 4-PE

(Πηγή [35])

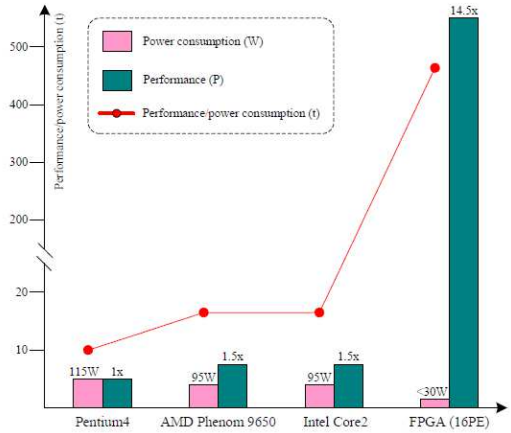
Μια επισκόπηση της προσέγγισης [35] φαίνεται στην *Εικόνα 4.2*. Αποτελείται από έναν επιταχυντή του CYK/inside αλγορίθμου (CYK/inside algorithm accelerator) υλοποιημένο σε μια FPGA πλακέτα Virtex5 XC5VLX330, και έναν υπολογιστή. Ο επιταχυντής λαμβάνει ως είσοδο μια RNA ακολουθία μήκους L και ένα μοντέλο συνδιακύμανσης μήκους k, υπολογίζει τον 3-διάστατο πίνακα και επιστρέφει στον υπολογιστή το κελί M(1,L,1) του πίνακα το οποίο περιέχει την πιθανότητα του βέλτιστου δέντρου της RNA ακολουθίας. Ο επιταχυντής αποτελείται από δυο μνήμες DDRII SODIMMs οι οποίες αποθηκεύουν τον μοντέλο συνδιακύμανσης και ένα μέρος του 3-διάστατου πίνακα, μια PCI-E διεπαφή που μεταφέρει τα δεδομένα από και προς τον υπολογιστή, έναν ελεγκτή (PE Array controller) για την αρχικοποίηση των μονάδων επεξεργασίας και την ανάθεση στηλών στις μονάδες, τις μονάδες επεξεργασίας (κύρια και δευτερεύουσες), έναν ελεγκτή συγχρονισμού (Column Synchronization and Data Back-writing controller) για την εκτέλεση της τρίτης και τέταρτης φάσης του αλγορίθμου, μια τοπική μνήμη για κάθε μονάδα επεξεργασίας για την προσωρινή αποθήκευση δεδομένων (row data cache και data back-writing) και τους καταχωρητές (TransRegs) που χρησιμοποιούνται για την παράδοση επαναχρησιμοποιήσιμων στοιχείων στήλης μεταξύ των PE.

Ο παράλληλος CYK/inside αλγόριθμος χωρίζεται σε πέντε φάσεις: την εισαγωγή δεδομένων, τον υπολογισμό των κελιών, τον συγχρονισμό των δεδομένων, την έξοδο των δεδομένων και την προώθηση του τμήματος. Αρχικά τα δεδομένα φορτώνονται στη μνήμη. Μόνο η κύρια μονάδα επεξεργασίας (master PE) διαβάζει τα δεδομένα από την εξωτερική μνήμη DRAM. Οι υπόλοιπες μονάδες επεξεργασίας (slaves PE) απλώς περιμένουν τα δεδομένα από την τοπική μνήμη της προηγούμενης μονάδας επεξεργασίας. Όταν υπολογιστούν όλα τα κελιά της κάθε στήλης, κάθε μονάδα επεξεργασίας αποθηκεύει τα αποτελέσματα της τοπικής μνήμης στην εξωτερική μνήμη DRAM. Αυτές είναι η τρίτη και τέταρτη φάση του αλγορίθμου, δηλαδή η φάση συγχρονισμού και η φάση εξόδου των δεδομένων. Τέλος, όπως προαναφέρθηκε, μόλις ολοκληρωθεί ο υπολογισμός της πρώτης ενότητας, ξεκινά η προώθηση των μονάδων επεξεργασίας στην επόμενη ενότητα, την πέμπτη δηλαδή φάση του αλγορίθμου.



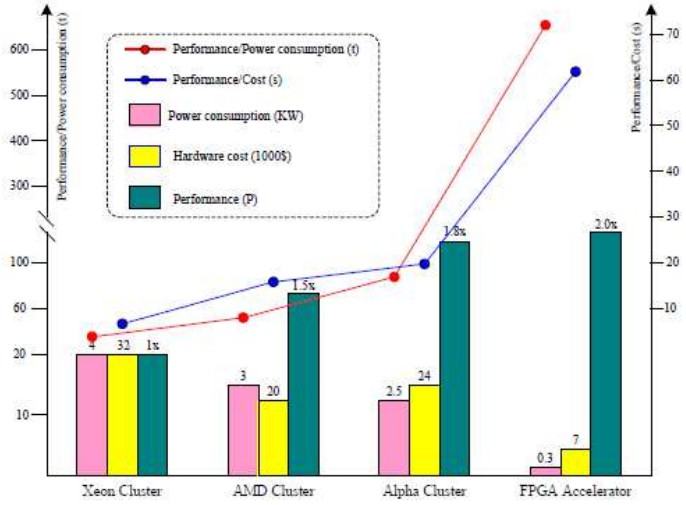
Εικόνα 4.2: Η δομή του επιταχυντή του CYK/inside αλγορίθμου
(Πηγή [35])

Ο παράλληλος CYK/Inside αλγόριθμος υλοποιήθηκε σε FPGA πλακέτα XC5VLX330 και δυο μνήμες DDRII SODIMM των 2 GB. Για την αξιολόγηση της απόδοσης στην εργασία [35], υπολογίστηκε ο χρόνος εκτέλεσης σε σύγκριση με διάφορες υλοποιήσεις λογισμικού μιας CPU. Στην *Εικόνα 4.3*, φαίνεται ότι η επιτάχυνση του αλγορίθμου στην προτεινόμενη υλοποίηση 16 μονάδες επεξεργασίας είναι 14 φορές μεγαλύτερη σε σύγκριση με υλοποιήσεις λογισμικού. Επιπλέον, η κατανάλωση ενέργειας είναι πολύ μικρότερη σε σχέση με τις υπόλοιπες υλοποιήσεις.



Εικόνα 4.3: Δείκτης απόδοσης / κατανάλωσης ενέργειας (FPGA VS. CPU)
(Πηγή [35])

Επιπλέον, η επιτάχυνση στην εργασία [35] συγκρίθηκε με διάφορες υλοποιήσεις σε συστοιχία υπολογιστών με επεξεργαστές Xeon, Alpha και AMD (multi-CPU cluster platforms) σε σχέση με την επιτάχυνση, την κατανάλωση ρεύματος και το κόστος.



Εικόνα 4.4: Απόδοση / κατανάλωση ενέργειας και λόγος απόδοσης / κόστους (FPGA VS. cluster)
(Πηγή [35])

Όπως φαίνεται και στην *Εικόνα 4.4*, το AMD cluster, το Alpha cluster και ο επιταχυντής FPGA [35] επιτυγχάνουν σχεδόν παρόμοια απόδοση σε σχέση με το Xeon cluster. Το κόστος υλικού του FPGA είναι πολύ μικρότερο σε σχέση με το κόστος που απαιτείται για τις υπόλοιπες συστοιχίες υπολογιστών. Επίσης, η κατανάλωση ρεύματος είναι μικρότερη από 0,3 KW σε αντίθεση με τις υπόλοιπες υλοποιήσεις που απαιτούν πολύ μεγαλύτερη κατανάλωση.

4.2 FPQRNA: Hardware-Accelerated QRNA Package For Noncoding RNA Gene Detecting On FPGA

Το 2010 οι Xia F., Dou Y. και Lei G. πρότειναν μια παράλληλη υλοποίηση σε υλικό [33] της QRNA [32] εφαρμογής με σκοπό την επιτάχυνση του χρόνου εκτέλεσης της εφαρμογής. Η υλοποίηση πραγματοποιείται σε FPGA πλακέτα με βάση τη συστολική δομή συστοιχίας (systolic array structure) και με βάση τα πειραματικά αποτελέσματα επιτυγχάνεται έως 18 φορές επιτάχυνση υπολογισμού στην FPGA πλακέτα σε σύγκριση με την εκτέλεση του QRNA σε μια υπολογιστική

πλατφόρμα. Επιπλέον, η κατανάλωση ρεύματος που απαιτείται για την εκτέλεση του FPQRNA είναι σχεδόν η μισή σε σύγκριση με το QRNA.

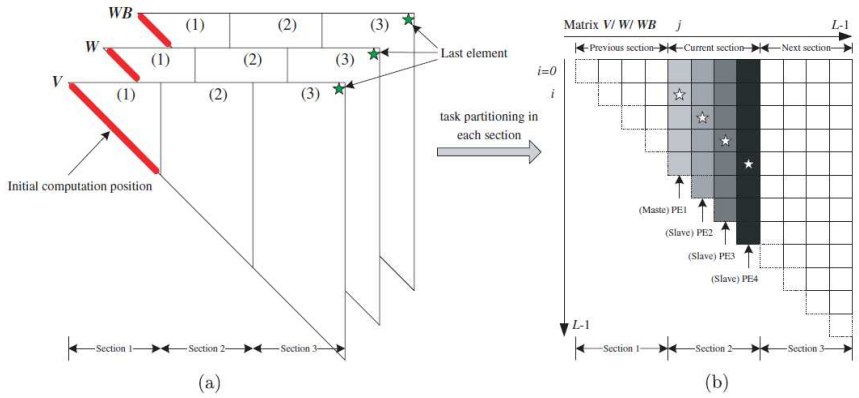
Το QRNA [32] είναι ένα πρόγραμμα που προβλέπει την δομή μη κωδικοποιημένου RNA (non-coding RNA-ncRNA) βασισμένο σε στοιχίσεις ακολουθιών κατά ζεύγη. Ο κώδικάς του είναι γραμμένος σε γλώσσα ANSI C και είναι διαθέσιμο στη διεύθυνση http://www.genetics.wustl.edu. Το QRNA λαμβάνει ως είσοδο στοιχισμένες RNA ακολουθίες κατά ζεύγη και ανιχνεύει τις δευτεροταγείς δομές, χρησιμοποιώντας μοντέλα που βασίζονται σε στοχαστική γραμματική χωρίς συμφραζόμενα (SCFG). Περιέχει τρία μοντέλα που αναλύουν την μετάλλαξη μιας δεδομένης στοίχισης κατά ζεύγη και με βάση το σκορ των μοντέλων εξέλιξης αποφασίζει ποιο από τα τρία περιγράφει καλύτερα την στοίχιση. Τα τρία μοντέλα είναι μια συζευγμένη SCFG γραμματική (pairSCFG) για την μοντελοποίηση της εξέλιξης της δευτεροταγούς RNA δομής (RNA model), ένα pair hidden Μαρκοβιανό μοντέλο (COD model) που περιγράφει την εξέλιξη της ακολουθίας κωδικοποίησης μιας πρωτεΐνης και ένα pair hidden Μαρκοβιανό μοντέλο (OTH model) που μοντελοποιεί μια μη κωδικοποιημένη αλληλουχία. Για το QRNA, η υπολογιστική πολυπλοκότητα για ένα ζεύγος ακολουθιών είναι $O(L^3)$, όπου L το μήκος των ακολουθιών.

Το πιο χρονοβόρο σημείο εκτέλεσης του QRNA είναι ο χρόνος που απαιτείται για τον υπολογισμό του RNA μοντέλου. Έτσι οι ερευνητές θέλησαν να επιταχύνουν αυτό το κομμάτι υπολογισμού της εφαρμογής, υλοποιώντας το FPQRNA. Το QRNA για τον υπολογισμό του RNA μοντέλου χρησιμοποιεί έναν αλγόριθμο δυναμικού προγραμματισμού, σαρώνει τις ακολουθίες εισόδου και συμπληρώνει τρεις δισδιάστατους πίνακες (2D matrix) βαθμολόγησης τον V, τον WB και τον W αντίστοιχα. Η κατάσταση V αντιπροσωπεύει ένα υποσύνολο ακολουθίας στο οποίο τα άκρα είναι σίγουρα ζεύγη βάσεων. Οι καταστάσεις W και WB αντιπροσωπεύουν ένα υποσύνολο στο οποίο τα άκρα είναι είτε ζευγαρωμένα είτε μη ζευγαρωμένα.

Ο FPQRNA αλγόριθμος

Για την επιτάχυνση του υπολογισμού του RNA μοντέλου, ο FPQRNA αλγόριθμος χωρίζει τους τριγωνικούς πίνακες σε μικρά τμήματα, που ονομάζονται ενότητες (sections) και υπολογίζει την κάθε ενότητα μια προς μια από αριστερά προς τα δεξιά, για να ελαχιστοποιήσει τις ανάγκες αποθήκευσης. Αυτό είναι απαραίτητο λόγω της περιορισμένης μνήμης του FPGA αφού για κάθε πίνακα απαιτούνται 16 MB χώρου αποθήκευσης αλλά η χωρητικότητα μνήμης του FPGA είναι

μόλις 4 MB. Κάθε ενότητα περιλαμβάνει τρία επίπεδα, ένα για κάθε έναν πίνακα, που υπολογίζονται παράλληλα. Κάθε επίπεδο χωρίζεται σε n στήλες, όπου n είναι ο αριθμός των μονάδων επεξεργασίας (processing element - PE).

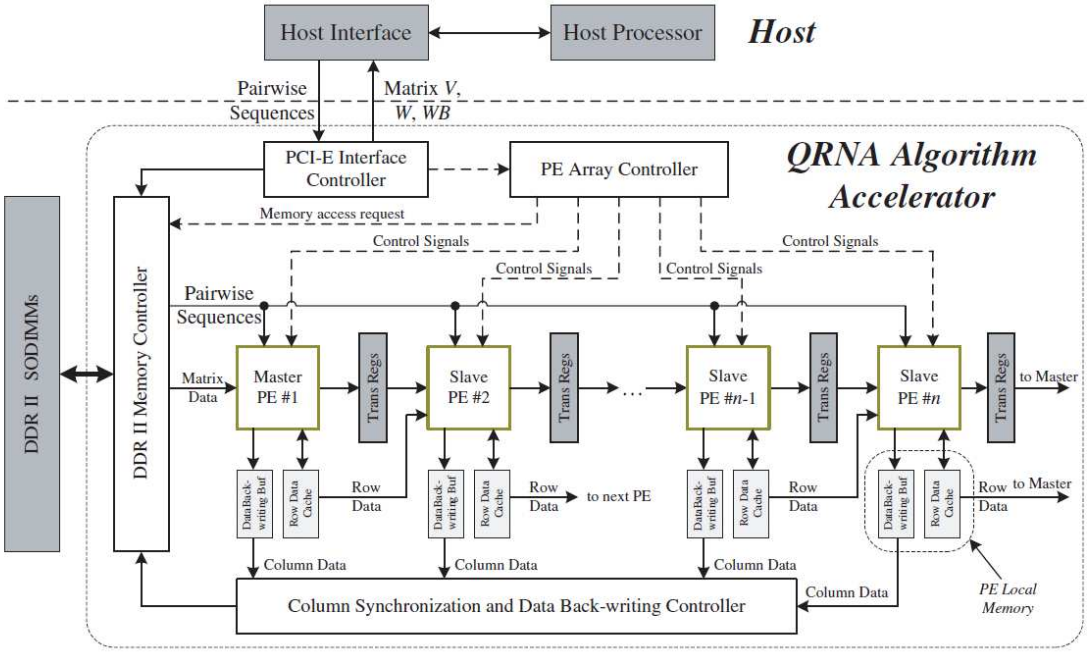


Εικόνα 4.5: Παράδειγμα κατανομής εργασιών και σειρά υπολογισμού για 4 PE
(Πηγή [33])

Ο παράλληλος αλγόριθμος για τον υπολογισμό του RNA μοντέλου χωρίζεται σε πέντε φάσεις: την εισαγωγή δεδομένων, τον υπολογισμό των κελιών, τον συγχρονισμό των δεδομένων, την έξοδο των δεδομένων και την προώθηση του τμήματος. Αρχικά τα δεδομένα φορτώνονται στη μνήμη. Μόνο η κύρια μονάδα επεξεργασίας (master PE) διαβάζει τα δεδομένα από την εξωτερική μνήμη DRAM. Οι υπόλοιπες μονάδες επεξεργασίας (slaves PE) απλώς περιμένουν τα δεδομένα από την τοπική μνήμη της προηγούμενης μονάδας επεξεργασίας. Η συμπλήρωση του πίνακα ξεκινά με τον παράλληλο υπολογισμό των κελιών (V (i, j),W(i, j) και WB(i, j)) που βρίσκονται στη διαγώνιο της πρώτης ενότητας, όπως φαίνεται στην *Εικόνα 4.5*.α και**Error! Reference source not found.**.β. Όταν συμπληρωθούν τα κελιά της πρώτης ενότητας, τότε οι μονάδες επεξεργασίας ξεκινάνε να υπολογίζουν τα κελιά της δεύτερης ενότητας, και με τον ίδιο τρόπο της τρίτης ενότητας. Όταν υπολογιστούν όλα τα κελιά της κάθε στήλης, κάθε μονάδα επεξεργασίας αποθηκεύει τα αποτελέσματα της τοπικής μνήμης στην εξωτερική μνήμη DRAM. Αυτές είναι η τρίτη και τέταρτη φάση του αλγορίθμου, δηλαδή η φάση συγχρονισμού και η φάση εξόδου των δεδομένων. Τέλος, μόλις ολοκληρωθεί ο υπολογισμός της πρώτης ενότητας, ξεκινά η προώθηση των μονάδων επεξεργασίας στην επόμενη ενότητα.

Μια επισκόπηση της προσέγγισης [33] φαίνεται στην *Εικόνα 4.6*. Αποτελείται από έναν επιταχυντή του QRNA αλγορίθμου (QRNA algorithm accelerator) υλοποιημένο σε μια FPGA πλακέτα Virtex5

XC5VLX330 και έναν υπολογιστή. Ο επιταχυντής λαμβάνει ως είσοδο μια διπλή RNA στοίχιση αλληλουχίας μήκους L, υπολογίζει το μοντέλο RNA χρησιμοποιώντας τον SCFG/inside αλγόριθμο και επιστρέφει τους τρεις πίνακες V, W και WB στον υπολογιστή για να προβλέψει σε ποιο από τα τρία μοντέλα ανήκουν οι ευθυγραμμισμένες ακολουθίες.



Εικόνα 4.6: Αρχιτεκτονική του επιταχυντή του QRNA αλγορίθμου
(Πηγή [33])

Όπως φαίνεται στην *Εικόνα 4.6*, ο επιταχυντής αποτελείται από δυο μνήμες DDRII SODIMMs, μια PCI-E διεπαφή που μεταφέρει τα δεδομένα (ευθυγραμμισμένες ακολουθίες προς υπολογισμό και τα τελικά αποτελέσματα) από και προς τον υπολογιστή, έναν ελεγκτή (PE Array controller) για την αρχικοποίηση των μονάδων επεξεργασίας και την ανάθεση στηλών στις μονάδες, τις μονάδες επεξεργασίας (κύρια και δευτερεύουσες), έναν ελεγκτή συγχρονισμού (Column Synchronization and Data Back-writing controller) για την εκτέλεση της τρίτης και τέταρτης φάσης του αλγορίθμου, μια τοπική μνήμη για κάθε μονάδα επεξεργασίας για την προσωρινή αποθήκευση δεδομένων (row data cache και data back-writing) και τους καταχωρητές (TransRegs) που χρησιμοποιούνται για την παράδοση επαναχρησιμοποιήσιμων στοιχείων στήλης μεταξύ των PE.

Για την αξιολόγηση της απόδοσης της [33], η παραπάνω αρχιτεκτονική υλοποιήθηκε σε μια μεγάλης κλίμακας FPGA πλακέτα Virtex5 XC5VLX330 και δυο μνήμες DDRII SODIMM των 2

GB και υπολογίστηκε ο χρόνος εκτέλεσης και η εγκυρότητα των αποτελεσμάτων. Για σύγκριση των αποτελεσμάτων, μετρήθηκε επίσης ο χρόνος εκτέλεσης της QRNA εφαρμογής στις πλατφόρμες AMD 9650, Intel Q9400 Quad CPU και Intel Core2 Duo E7400.

Όσον αφορά την εγκυρότητα των αποτελεσμάτων, η υλοποίηση στην εργασία [33] έδωσε ακριβώς τα ίδια αποτελέσματα με την εκτέλεση της QRNA εφαρμογής σε υλοποίηση λογισμικού.

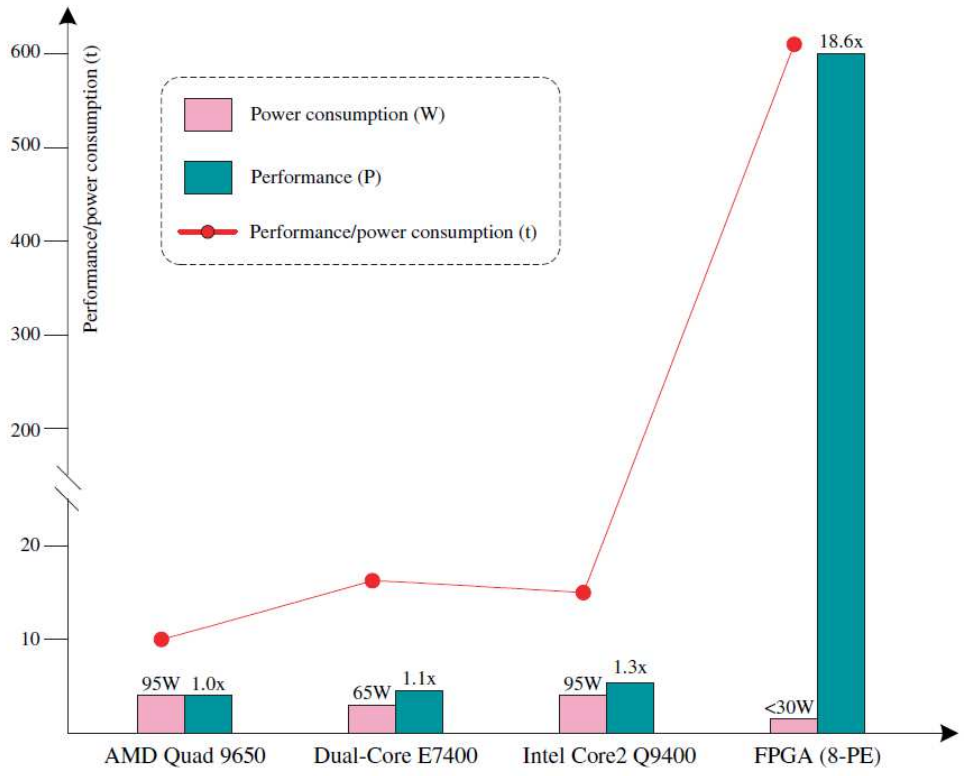
Όσον αφορά τον χρόνο υπολογισμού, σύμφωνα με τα αποτελέσματα που φαίνονται στον Πίνακας 4.1, η επίδοση στην εργασία [33] είναι έως 18 φορές μεγαλύτερη, ανάλογα με το μήκος των ακολουθιών, σε σύγκριση με τις υλοποιήσεις της QRNA σε λογισμικό.

Test Platforms	$L = 284$		$L = 505$		$L = 765$		$L = 996$	
	Time	Speedup	Time	Sp	Time	Sp	Time	Sp
AMD[a]	11266	1	76015	1	293516	1	715781	1
Intel[b]	10625	1.06	68890	1.10	262359	1.12	640875	1.12
Intel[c]	8763	1.29	56453	1.35	210891	1.39	545969	1.31
FPGA[d]	1397	8.06	6687	11.37	20276	14.48	38378	18.65

Note: hardware environment
[a]AMD Phenom 9650 Quad CPU, 2.3 GHz, 3.0 GB memory;
[b]Intel Core2 Duo E7400 2.80 GHz CPU, 2.0 GB memory;
[c]Intel Core2 Quad Q9400 2.66 GHz CPU, 3.0 GB memory;
[d]FPGA accelerator (XC5VLX330) with 8-PE, 160 MHz, 4.0 GB memory.

Πίνακας 4.1: χρόνοι εκτέλεσης μεταξύ της [33] και τριών υλοποιήσεων λογισμικού
(Πηγή [33])

Τέλος, σύμφωνα με την *Εικόνα 4.7*, η κατανάλωση ρεύματος που απαιτείται [33] είναι μικρότερη από τη μισή σε σύγκριση με τις υλοποιήσεις λογισμικού.



Εικόνα 4.7: Ο λόγος απόδοσης / κατανάλωσης ενέργειας σε σύγκριση με τους μικροεπεξεργαστές γενικής χρήσης (Πηγή [33])

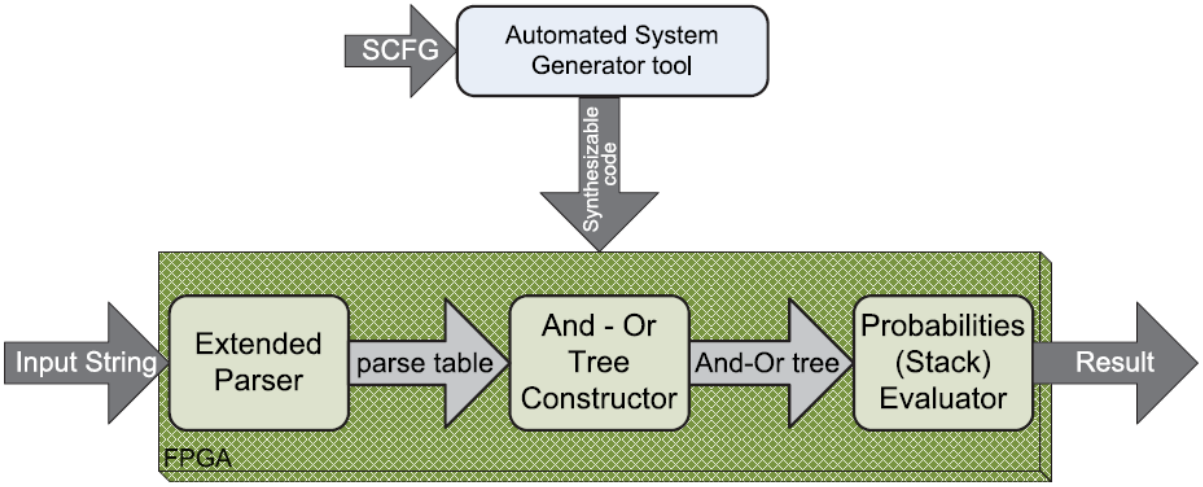
4.3 Parallel Hardware Stochastic Context-Free Parsers

Το 2016, η ομάδα των Παυλάτου Χ., Δημόπουλου Α. και Παπακωνσταντίνου Γ. παρουσίασαν την πρώτη διεθνώς παράλληλη υλοποίηση ενός SCFG αναλυτή (SCFG parser) σε επίπεδο υλικού με βάση τον αλγόριθμο Earley [18]. Η προσέγγιση αυτή είναι μια βελτίωση των δημοσιευμένων εργασιών [17] και [13] ώστε να υποστηρίζει στοχαστικές γραμματικές χωρίς συμφραζόμενα (stochastic context-free grammars - SCFG). Η προσέγγιση της εργασίας [17] επεκτάθηκε ώστε να δημιουργεί ένα AND-or δέντρο ανάλυσης, το οποίο είναι αναγκαίο για την αξιολόγηση των πιθανοτήτων.

Πιο συγκεκριμένα, υλοποίησαν μια πλατφόρμα σε υλικό, η οποία λαμβάνει ως είσοδο μια SCFG γραμματική και παράγει τον αναλυτή για την συγκεκριμένη γραμματική σε HDL (Hardware Design Language) κώδικα σε γλώσσα Verilog. Έπειτα ο κώδικας αυτός φορτώνεται σε FPGA πλακέτα όπου ο αναλυτής, λαμβάνοντας μια συμβολοσειρά εισόδου, δημιουργεί το AND-or δέντρο

ανάλυσης, ώστε να διαχειρίζεται επιπλέον και διφορούμενες (ambiguous) γραμματικές, και αξιολογεί τις inner και forward πιθανότητες της συμβολοσειράς.

Μια **επισκόπηση** της προσέγγισης [18] φαίνεται στην *Εικόνα 4.8*.



Εικόνα 4.8: Επισκόπηση της προσέγγισης [18]
(Πηγή: [18])

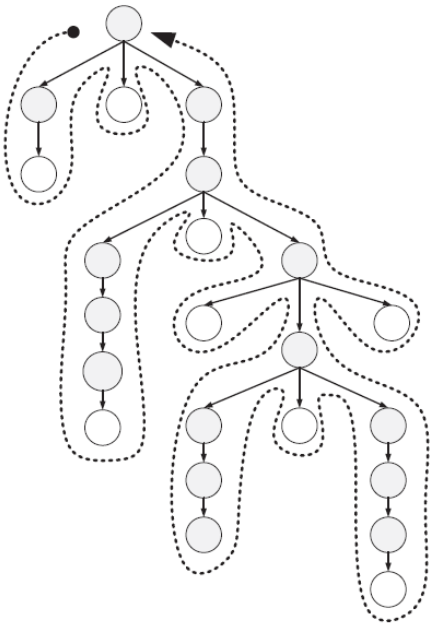
Όπως φαίνεται και στην *Εικόνα 4.8*, η προτεινόμενη υλοποίηση αποτελείται από τρεις υπομονάδες, τον αναλυτή ο οποίος δημιουργεί τον πίνακα ανάλυσης (Extended Parser), τον κατασκευαστή του AND-or δέντρου ανάλυσης (AND-or Tree Constructor) και τον αξιολογητή πιθανοτήτων (Probabilities Evaluator).

Η πλατφόρμα λαμβάνει ως είσοδο μια SCFG γραμματική σε BNF μορφή [66], δεδομένου ότι η υλοποίηση του αναλυτή βασίζεται στον αλγόριθμο Earley. Ο αναλυτής χρησιμοποιείται για την συντακτική ανάλυση, αναλύει δηλαδή τους συντακτικούς κανόνες της SCFG γραμματικής και στη συνέχεια δημιουργεί το δέντρο ανάλυσης. Σημειώνεται ότι στην περίπτωση μιας διφορούμενης γραμματικής (ambiguous grammar) δημιουργεί όλα τα δέντρα ανάλυσης.

Ο κατασκευαστής του AND-or δέντρου ανάλυσης ξεκινάει να διαβάζει τον πίνακα ανάλυσης από το πάνω δεξιά κελί και μόλις βρει τον αρχικό κανόνα της γραμματικής τον αποθηκεύει στην ρίζα του δέντρου. Οι κανόνες που δημιουργούν τον αρχικό κανόνα αποθηκεύονται ως κλαδιά δέντρου.

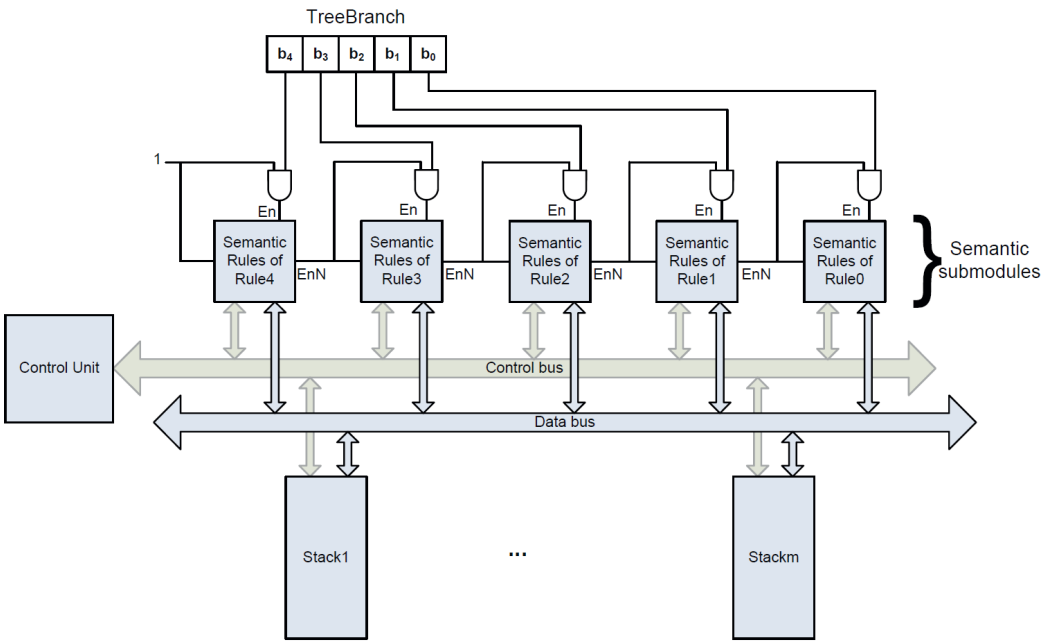
Στην περίπτωση μιας διφορούμενης γραμματικής δημιουργούνται όλα τα δέντρα. Η κατασκευή ενός πιθανού επόμενου δέντρου ξεκινά μόλις ολοκληρωθεί το προηγούμενο δέντρο. Το δέντρο που προκύπτει μεταφέρεται στην επόμενη υπομονάδα, στον σημασιολογικό αξιολογητή (Semantic Evaluator), για να «διακοσμηθεί» με τα χαρακτηριστικά της γραμματικής.

Ο αξιολογητής πιθανοτήτων (Probabilities Evaluator) πραγματοποιεί την σημασιολογική ανάλυση της SCFG γραμματικής. Αποτελείται από μονάδες υλικού (hardware modules), οι οποίες αξιολογούν τα κληρονομούμενα (inherited) και τα συνθετικά (synthesized) κατηγορήματα. Η υλοποίηση των μονάδων είναι μια υλοποίηση στοίβας (stack-based approach) που χρησιμοποιεί push και pop εντολές για την αξιολόγηση των κατηγορημάτων (attributes) της γραμματικής. Όλοι οι σημασιολογικοί κανόνες της κατηγορικής (attribute grammar -AG) γραμματικής μετατρέπονται σε push και pop εντολές. Ο σημασιολογικός αξιολογητής διαβάζει το δέντρο από την κορυφή μέχρι το κάτω μέρος και από αριστερά προς τα δεξιά, όπως φαίνεται στην *Εικόνα 4.9*.



Εικόνα 4.9: Διάσχιση του δέντρου ανάλυσης
(Πηγή:[17])

Η αρχιτεκτονική του σημασιολογικού αξιολογητή, που είναι υλοποιημένος σε γλώσσα Verilog, φαίνεται στην Εικόνα 4.10.



Εικόνα 4.10: Αρχιτεκτονική σημασιολογικού αξιολογητή
(Πηγή: [17])

Όπως αναφέρθηκε, ο αξιολογητής διασχίζει το δέντρο από αριστερά προς τα δεξιά. Κάθε κλαδί είναι σε δυαδική μορφή μήκους p, όπου p είναι ο αριθμός των συντακτικών κανόνων. Όταν τα κατηγορήματα του κλαδιού πρέπει να αξιολογηθούν, η δυαδική τιμή του κλαδιού αποθηκεύεται σε μια παράμετρο που ονομάζεται TreeBranch, όπως φαίνεται στην Εικόνα 4.10. Επιπλέον, υλοποιείται ένα σήμα En (Enable) το οποίο ενεργοποιεί την αντίστοιχη μονάδα (module) που είναι υπεύθυνη για την αξιολόγηση του εκάστοτε κανόνα. Η κάθε μονάδα, όταν ολοκληρώσει την αξιολόγηση, ενεργοποιεί το σήμα En. Το σήμα περνάει από μια λογική πύλη AND μαζί με το επόμενο bit της παραμέτρου TreeBranch και ενεργοποιείται η επόμενη σημασιολογική μονάδα (semantic submodule). Κάθε μονάδα επικοινωνεί μέσω ενός διαύλου ελέγχου (Control bus) και ενός διαύλου δεδομένων (data bus) με τις στοίβες που χρειάζεται. Η μονάδα ελέγχου (Control Unit) σε κάθε κύκλο ρολογιού ενεργοποιεί μια σημασιολογική υπομονάδα η οποία με τη σειρά της στέλνει τα κατάλληλα σήματα στις αντίστοιχες στοίβες ώστε να εκτελεστούν οι push και pop εντολές.

Τα πλεονεκτήματα της προσέγγισης

- Επιτρέπεται η εισαγωγή γραμματικής σε οποιαδήποτε μορφή, αφού η υλοποίηση βασίζεται στον αλγόριθμο Earley.

- Υποστηρίζει τις διφορούμενες γραμματικές με πιο αποτελεσματικό τρόπο με την χρήση ενός AND-or δέντρου σε αντίθεση με διάφορες άλλες υλοποιήσεις που χρειάζεται να κατασκευάσουν όλα τα δέντρα ανάλυσης.
- Όλες οι υπομονάδες φορτώνονται στην ίδια FPGA πλακέτα. Η υλοποίηση σε FPGA έχει το πλεονέκτημα ότι η κατανάλωση ρεύματος είναι πολύ μικρή. Επίσης, καθώς είναι δυνατός ο επαναπρογραμματισμός ενός FPGA, επιτρέπει την εφαρμογή αλγορίθμων με διαφορετικές υπολογιστικές απαιτήσεις (computing requirements) στην ίδια πλατφόρμα υλικού.

Για την αξιολόγηση της προτεινόμενης υλοποίησης στην εργασία [18] εκτελέστηκαν πειράματα μεταξύ της υλοποίησης [18] και υλοποιήσεων λογισμικού (Stolcke software). Όλες οι υλοποιήσεις στην εργασία [18] εκτελέστηκαν σε FPGA πλακέτα XilinxVirtex-6. Στα πειράματα χρησιμοποιήθηκαν διφορούμενες γραμματικές και συμβολοσειρές εισόδου με διαφορετικά μήκη. Η επιτάχυνση στην εργασία [18] είναι μία έως δύο τάξεις μεγέθους μεγαλύτερη, σε σύγκριση με τις υλοποιήσεις λογισμικού, ανάλογα με την εφαρμογή, το μήκος της συμβολοσειράς και τον αριθμό των παραγόμενων δέντρων.

Να σημειώσουμε εδώ ότι η υλοποίηση στην εργασία [18] επιβαρύνεται με το κόστος της κατασκευής του AND-or δέντρου. Για το σκοπό αυτό η ομάδα των ερευνητών παρουσίασαν μια βελτιωμένη υλοποίηση της [18] που αξιολογεί απευθείας τις πιθανότητες από τον πίνακα ανάλυσης χωρίς να δημιουργεί το δέντρο ανάλυσης. Η υλοποίηση αυτή παρουσιάζεται στην εργασία [20].

5 Συμπεράσματα

Η πρόβλεψη της δευτεροταγούς RNA δομής είναι ένα σημαντικό πρόβλημα που απασχολεί μέχρι σήμερα την επιστήμη της βιοπληροφορικής. Οι στοχαστικές γραμματικές χωρίς συμφραζόμενα (SCFG) χρησιμοποιήθηκαν για την πρόβλεψη δευτεροταγούς RNA δομής στις αρχές της δεκαετίας του '90 από τον Sakakibara (1994) [40] και τους Eddy και Durbin (1994) [41]. Και οι δυο αυτές πρώτες προσεγγίσεις βασίστηκαν σε SCFG γραμματικές για την εύρεση μιας κοινής δομής ενός συνόλου ομόλογων RNA ακολουθιών. Από τότε μέχρι σήμερα έχουν προταθεί πολλές υλοποιήσεις αλγορίθμων που βασίζονται σε μια ή σε πολλαπλές RNA ακολουθίες. Οι υλοποιήσεις αυτές μπορεί να είναι σειριακές ή παράλληλες, σε υλικό ή λογισμικό επίπεδο.

5.1 Προσεγγίσεις με βάση των αριθμών των ακολουθιών εισόδου

Όπως προαναφέρθηκε, ένας πρώτος διαχωρισμός των προσεγγίσεων γίνεται με βάση τις ακολουθίες εισόδου. Υπάρχουν δυο τύποι πρόβλεψης: η μέθοδος πρόβλεψης μιας ακολουθίας και η μέθοδος πρόβλεψης πολλαπλών ακολουθιών.

Οι μέθοδοι πρόβλεψης μιας ακολουθίας με βάση τις SCFG γραμματικές χρησιμοποιούν διάφορες παραμέτρους για την πρόβλεψη της δευτεροταγούς δομής κάθε αλληλουχίας RNA. Συνήθως, οι υλοποιήσεις λαμβάνουν ως είσοδο μια RNA ακολουθία και μια γραμματική και με βάση τον εκάστοτε αλγόριθμο που χρησιμοποιούν ελέγχουν αν η ακολουθία ανήκει στην γραμματική και/ή σε μια RNA οικογένεια [17], [18], [27], [34], [35], [36].

Οι μέθοδοι πρόβλεψης με βάση τις SCFG γραμματικές επεκτάθηκαν για να εξάγουν μια κοινή δευτεροταγή δομή με βάση πολλαπλές ομόλογες αλληλουχίες RNA. Ομόλογες RNA ακολουθίες είναι αυτές που έχουν κοινή εξελικτική προέλευση. Οι ομόλογες ακολουθίες πρέπει πρώτα να στοιχηθούν (alignment) έτσι ώστε να εξεταστεί η κοινή τους προέλευση και η διατηρημένη δομή μιας RNA οικογένειας. Το αποτέλεσμα της στοίχισης είναι τα ομόλογα τμήματα σε κάθε ακολουθία να εμφανίζονται στην ίδια στήλη της στοίχισης. Η στοίχιση των αλληλουχιών μπορεί να υλοποιηθεί είτε κατά ζεύγη, όπως στις εργασίες [37], [38], [39], είτε πολλαπλές ακολουθίες ταυτόχρονα, όπως στις εργασίες [26], [30], [40], [41]. Η δεύτερη είναι μια επιτυχημένη προσέγγιση για την πρόβλεψη της δομής αλλά προϋποθέτει αξιόπιστη στοίχιση. Για να είναι αξιόπιστη η στοίχιση απαιτείται

μεγάλος αριθμός RNA ακολουθιών και τόσο διαφορετικών μεταξύ τους ώστε η συνδιακύμανση να είναι ικανοποιητική. Υπάρχουν τριών ειδών προσεγγίσεις ως προς την πρόβλεψη RNA κοινής δομής πολλαπλών RNA ακολουθιών.

Πρώτη προσέγγιση: Πρώτα στοίχιση, μετά αναδίπλωση

Πολλά προγράμματα για τη συγκριτική ανάλυση του RNA με βάση SCFG γραμματικές προϋποθέτουν στοιχισμένες RNA ακολουθίες για την πρόβλεψη της κοινής τους δομής, όπως αναφέρεται στη βιβλιογραφία [21], [22], [24], [29], [32]. Το πλεονέκτημα της προσέγγισης αυτής είναι η ταχύτητα. Μπορεί να εφαρμοστεί σε σχεδόν οποιοδήποτε αριθμό ακολουθιών και παίρνει περίπου τον ίδιο χρόνο υπολογισμού με την πρόβλεψη της δομής για μία μόνο ακολουθία του ίδιου μήκους με το μήκος στοίχισης [47]. Το μειονέκτημα είναι ότι η ποιότητα της πρόβλεψης δομής εξαρτάται από την ποιότητα της στοίχισης καθώς οι μη σωστά στοιχισμένες βάσεις μπορούν να αλλοιώσουν το αποτέλεσμα της συνδιακύμανσης των ακολουθιών.

Δεύτερη προσέγγιση: Ταυτόχρονη στοίχιση και αναδίπλωση

Η ιδέα της χρήσης αλγόριθμου δυναμικού προγραμματισμού για ταυτόχρονη αναδίπλωση και στοίχιση ενός συνόλου ακολουθιών εισήχθη το 1985 από τον Sankoff [48]. Αρκετά προγράμματα, βασίζονται στην μέθοδο της ταυτόχρονης στοίχισης και πρόβλεψης της δευτεροταγούς δομής αλληλουχιών RNA. Όμως, για δυο ακολουθίες X και Y, η πολυπλοκότητα χρόνου για παράλληλη στοίχιση και σύγκριση δομής είναι $O(X^3Y^3)$. Το υπολογιστικό κόστος καθιστά την ταυτόχρονη στοίχιση και πρόβλεψη περισσοτέρων ακολουθιών μη υλοποιήσιμη. Έτσι, ο Holmes [37], [38] και οι Dowell & Eddy [39] υλοποίησαν, ανεξάρτητα, προσαρμοσμένες εκδόσεις των alignment-and-folding αλγορίθμων με βάση SCFG γραμματικές περιορίζοντας τα πιθανά ζεύγη βάσεων και τις ευθυγραμμίσεις μεταξύ των ακολουθιών με σκοπό να επιταχύνουν την εκτέλεση και να μειώσουν τις απαιτήσεις μνήμης. Το Stemloc [38] μπορεί να φτάσει στην πολυπλοκότητα χρόνου και μνήμης $O(n^2)$ και το ConSan [39] μπορεί να φτάσει στον χρόνο $O(n^3)$ και την πολυπλοκότητα μνήμης $O(n^2)$ για δύο αλληλουχίες μήκους n.

Τρίτη προσέγγιση: Πρώτα αναδίπλωση, μετά στοίχιση

Στην τρίτη προσέγγιση προσδιορίζεται ένα σύνολο χαμηλών δευτεροταγών δομών ελεύθερης ενέργειας για κάθε μία από τις πολλαπλές αλληλουχίες και στη συνέχεια αναλύονται οι προβλεπόμενες δομές για να βρεθεί η χαμηλότερη δομή ελεύθερης ενέργειας κοινή σε όλες τις

αλληλουχίες. Δεν είναι γνωστές υλοποιήσεις με βάση SCFG γραμματικές. Η προσέγγιση αυτή έχει το πλεονέκτημα ότι είναι ταχύτερη από την πρώτη προσέγγιση. Έχει το μειονέκτημα ότι εντοπίζει την κοινή δομή, η οποία δεν προβλέπει ακριβώς ποια ζεύγη είναι ομόλογα. Απαιτείται επιπλέον επεξεργασία για την εύρεση των ομόλογων ζευγών [47].

Επιπλέον, οι SCFG γραμματικές μπορούν να συνδυαστούν με μοντέλα μοριακής εξέλιξης για να παράγουν μια κοινή δομή με μεγαλύτερη ακρίβεια σε σχέση με την πρόβλεψη μιας μεμονωμένης ακολουθίας. Η πρώτη προσέγγιση αναπτύχθηκε από τους Knudsen και Hein, οι οποίοι υλοποίησαν το Pfold [21, 22]. Το Pfold συνδυάζει μια SCFG γραμματική και ένα εξελικτικό μοντέλο για να παράγει ένα εξελικτικό δέντρο και μια κοινή δευτεροταγή δομή από ένα στοιχισμένο σύνολο αλληλουχιών RNA. Το Pfold ξεχωρίζει, σε σχέση με άλλες προσεγγίσεις που λαμβάνουν ως είσοδο στοιχισμένες RNA ακολουθίες, στο γεγονός ότι δίνει μια πιθανότητα σε κάθε νουκλεοτίδιο να είναι οποιοδήποτε άλλο νουκλεοτίδιο. Έτσι ο Pfold γίνεται πιο ισχυρός στις περιπτώσεις λανθασμένης στοίχισης όπως αναφέρθηκε προηγουμένως.

5.2 Προσεγγίσεις με βάση την υλοποίηση

Όπως αναφέρθηκε σε προηγούμενα κεφάλαια, ένας πρώτος περιορισμός των SCFG γραμματικών είναι ότι δεν μπορούν να μοντελοποιήσουν RNA ψευδοκόμβους. Παραλλαγές των SCFG γραμματικών [27], [30], [42] - [45] έχουν προταθεί για την πρόβλεψη δομών RNA με ψευδοκόμβους.

Ο σημαντικότερος περιορισμός των SCFG είναι η υπολογιστική τους πολυπλοκότητα. Ο χρόνος εκτέλεσης των αλγορίθμων για την πρόβλεψη δομής με βάση SCFG γραμματικές υπολογίζεται σε $O(L^3)$ ή χειρότερος, όπου L το μήκος της ακολουθίας. Για την βελτίωση του υπολογιστικού κόστους έχουν προταθεί παράλληλες υλοποιήσεις των αλγορίθμων. Επομένως, ένας άλλος διαχωρισμός των προσεγγίσεων γίνεται με βάση την υλοποίησή τους, η οποία μπορεί να είναι σειριακή, όπως αναφέρεται στη βιβλιογραφία [22], [27], [30], [32], [38], [39], ή παράλληλη [18], [33] - [35], [52], σε υλικό [13], [17], [33], [35] ή λογισμικό [22], [26], [27], [32], [34], [37] – [39], [42] – [45]. Οι αλγόριθμοι δυναμικού προγραμματισμού για την πρόβλεψη της δευτεροταγούς δομής του RNA βασίζονται στον κυματοειδή υπολογισμό (wavefront computation), που σημαίνει ότι κάθε στοιχείο

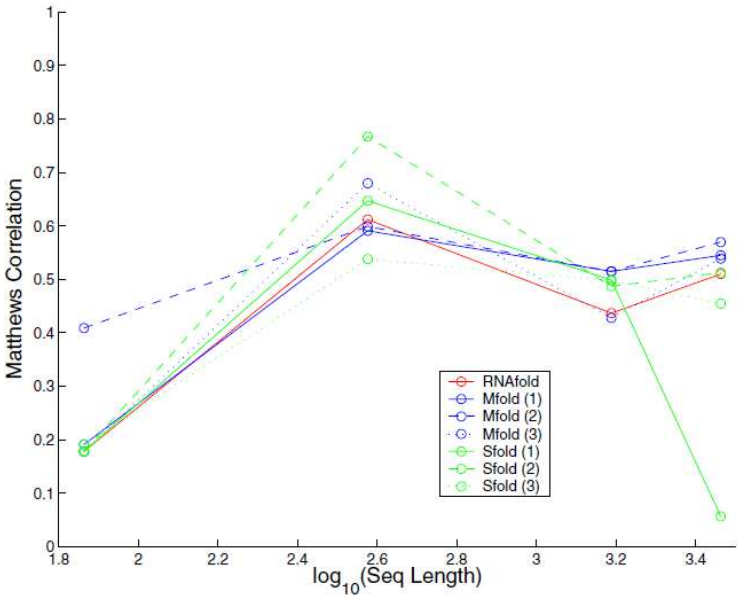
ενός δισδιάστατου ή τρισδιάστατου πίνακα εξαρτάται από τον υπολογισμό προηγούμενων στοιχείων. Επομένως, οι υπολογισμοί μπορεί να πραγματοποιηθούν με παράλληλο τρόπο με σκοπό την επιτάχυνση στην εκτέλεση. Διάφορες παράλληλες μέθοδοι έχουν προταθεί χρησιμοποιώντας παράλληλη multicore αρχιτεκτονική, παραδείγματος χάριν η coarse-grained υλοποίηση ή fine-grained υλοποίησεις σε FPGA [18], [33], [35].

Σε επίπεδο λογισμικού, η εργασία [24] παραλληλοποιεί το Pfold, υλοποιώντας παράλληλα threads για τον υπολογισμό του φυλογενετικού δέντρου και την εκτέλεση του inside–outside αλγορίθμου. Η εργασία [34] προτείνει μια παράλληλη προσέγγιση του CYK/inside αλγορίθμου, μεταξύ στρωμάτων (layers) 3-διάστατου πίνακα, για την στοίχιση μιας αλληλουχίας RNA χρησιμοποιώντας SCFG γραμματική και την coarse-grained υλοποίηση του αλγορίθμου σε μια συστοιχία υπολογιστών (PC cluster) και μια συστοιχία SMPs. Ωστόσο, η χρήση συστοιχιών υπολογιστών έχει πολύ υψηλό κόστος διαχείρισης και συντήρησης. Η εργασία [52] προτείνει επίσης μια παράλληλη προσέγγιση του CYK/inside αλγορίθμου και την υλοποίηση του αλγορίθμου σε μια συστοιχία υπολογιστών (PC cluster).

Παράλληλες υλοποιήσεις σε υλικό επίπεδο αναφέρονται στη βιβλιογραφία [18], [33], [35]. Η εργασία [18] προτείνει μια παράλληλη προσέγγιση του αλγορίθμου Earley και την υλοποίησή του σε FPGA πλακέτα. Η εργασία [33] παραλληλοποιεί σε FPGA πλακέτα την QRNA εφαρμογή λογισμικού, με σκοπό την επιτάχυνση του χρόνου εκτέλεσης της εφαρμογής. Η εργασία [35] παρουσιάζει την πρώτη fine-grained παράλληλη υλοποίηση του CYK/inside αλγορίθμου σε μια FPGA πλακέτα. Η εργασία [35] είναι μια αρκετά παρόμοια υλοποίηση με την [33]. Η κύρια διαφορά είναι ότι η εργασία [33] διαιρεί έναν 2D dynaming programming matrix ενώ η εργασία [35] διαιρεί έναν 3D dynaming programming matrix. Σε fine-grained αρχιτεκτονικές, τα κελιά στην ίδια διαγώνιο υπολογίζονται παράλληλα. Για να είναι αποτελεσματική μια τέτοια υλοποίηση, χρειάζεται οι επεξεργαστές να επικοινωνούν γρήγορα μεταξύ τους, όπως συμβαίνει σε περιπτώσεις συστολικής δομή συστοιχίας, για παράδειγμα στην υλοποίηση της εργασίας[33].

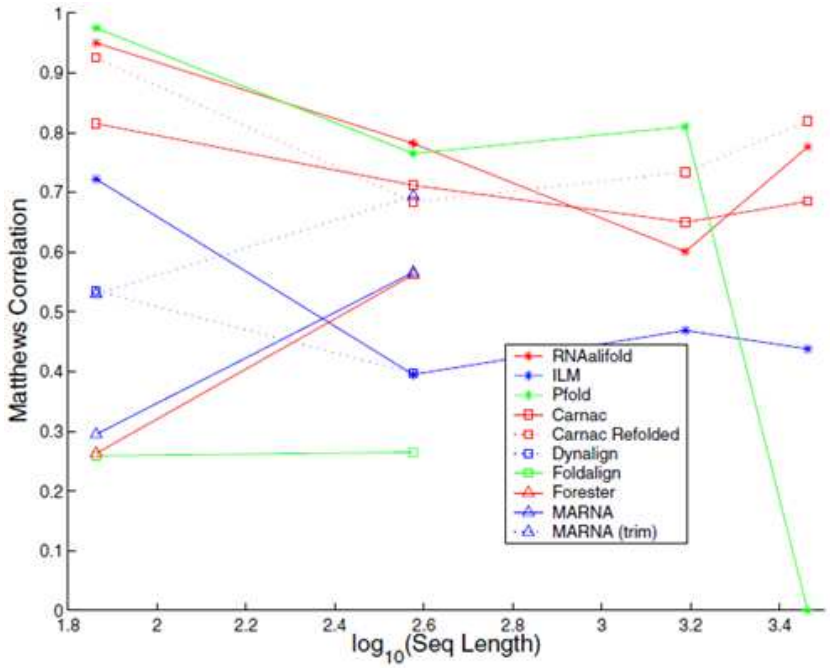
5.3 Μια αξιολόγηση των συγκριτικών προσεγγίσεων πρόβλεψης της δομής του RNA

Το 2004, οι Gardner και Robert Giegerich μέτρησαν την ακρίβεια διάφορων μεθόδων πρόβλεψης δευτεροταγούς RNA δομής που είχαν προταθεί μέχρι εκείνη την περίοδο, σε ένα σύνολο 4 RNA οικογενειών [46]. Οι μέθοδοι αυτοί ήταν μέθοδοι μονής ακολουθίας και πολλαπλών ακολουθιών με πληροφορίες στοίχισης, ταυτόχρονη στοίχιση και αναδίπλωση και/ή αναδίπλωση και μετά στοίχιση. Ανάμεσα στις μεθόδους ήταν και το Pfold που αναλύθηκε σε προηγούμενο κεφάλαιο. Σαν μονάδα μέτρησης της ακρίβειας χρησιμοποίησαν τον συντελεστή συσχέτισης Matthew, που παίρνει τιμές από -1 για ανακριβείς προβλέψεις έως 1 για πολύ ακριβείς προβλέψεις. Με βάση τα αποτελέσματα, κατέληξαν στο συμπέρασμα ότι οι μέθοδοι συγκριτικής ανάλυσης, όπως το Pfold, οδηγούν σε καλύτερη ακρίβεια πρόβλεψης σε σύγκριση με μεθόδους μονής ακολουθίας. Το συμπέρασμα αυτό αντικατοπτρίζεται και στις παρακάτω εικόνες, όπου η ακρίβεια πρόβλεψης σε μεθόδους μονής ακολουθίας δεν ξεπερνάει το 0,8 σε αντίθεση με την ακρίβεια πρόβλεψης σε μεθόδους πολλαπλών ακολουθιών, η οποία ήταν 0.9 για τις περισσότερες προβλέψεις. Επιπλέον φαίνεται στην Εικόνα 5.2, ότι οι μέθοδοι πολλαπλών ακολουθιών με ταυτόχρονη στοίχιση και αναδίπλωση, όπως το Foldalign και Dynalign, δεν μπορούν να προβλέψουν την δομή για μεγάλου μήκους RNA ακολουθίες. Επομένως, η εφαρμογή αυτών των μεθόδων είχε περιθώρια βελτίωσης, όπως για παράδειγμα τα Stemloc και ConSan που προτάθηκαν τα επόμενα χρόνια.



Εικόνα 5.1: Συσχέτιση μεθόδων μονής ακολουθίας

(Πηγή [46])



Εικόνα 5.2: Συσχέτιση συγκριτικών μεθόδων σε δεδομένα υψηλής ομοιότητας
(Πηγή: [46])

5.4 Μια συγκριτική αξιολόγηση των παράλληλων υλοποιήσεων του CYK/inside αλγορίθμου

Όπως αναφέρθηκε σε προηγούμενα κεφάλαια, οι εργασίες [34] και η [52] παρουσίασαν παράλληλες προσεγγίσεις του CYK/inside αλγορίθμου σε επίπεδο λογισμικού. Με βάση τα πειραματικά αποτελέσματα, η εργασία [34] πέτυχε αύξηση της επιτάχυνσης έως 19 φορές σε μια συστοιχία υπολογιστών με 32 AMD επεξεργαστές συνολικά. Η εργασία [52] πέτυχε αύξηση της επιτάχυνσης έως 16 φορές σε μια συστοιχία υπολογιστών με 20 Xeon επεξεργαστές και έως 36 φορές σε μια συστοιχία υπολογιστών με 48 EV68 επεξεργαστές συνολικά. Συγκρίνοντας τις δυο προσεγγίσεις στις ίδιου μήκους RNA ακολουθίες (SSU rRNA2 μήκους 1545 και LSU rRNA μήκους 2904) και με τον ίδιο αριθμό επεξεργαστών, φαίνεται ότι η εργασία [34] επιτυγχάνει καλύτερο χρόνο εκτέλεσης με 16 επεξεργαστές για την μεγαλύτερη ακολουθία ενώ σχεδόν ίδιο χρόνο εκτέλεσης για την SSU rRNA2 ακολουθία.

1542	2	169.734	169.562	0.172	1.55
	4	85.172	84.609	0.563	3.09
	8	42.707	42.273	0.434	6.16
	16	22.051	21.113	0.938	11.94
	32	13.223	10.543	2.68	19.91
2904	2	2774.266	2771.047	3.219	1.54
	4	1442.59	1379.574	63.016	2.96
	8	859.289	702.219	157.07	4.97
	16	512.27	347.523	164.747	8.34
	32	218.258	77.207	141.051	19.57

Εικόνα 5.3: Χρόνος εκτέλεσης σε δευτερόλεπτα στη δυναμική προσέγγιση

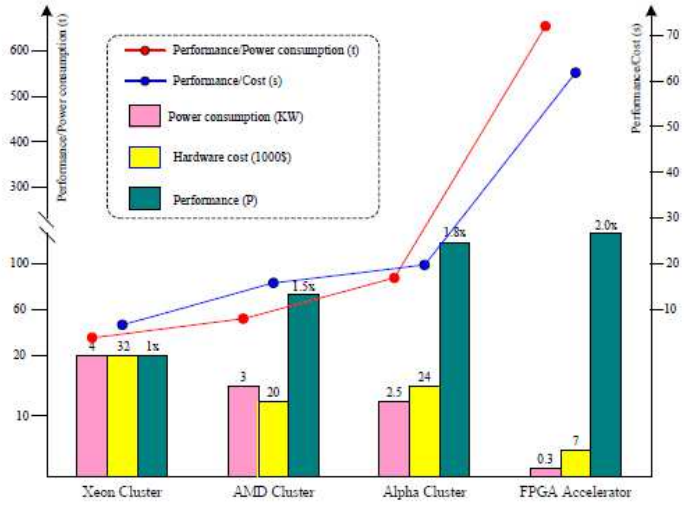
(Πηγή [52])

	Number of processors					
P	1	2	4	8	16	20
SSU rRNA 2	1119	746 (1.5)	352 (3.2)	182 (6.1)	95 (11.8)	79 (14.1)
LSU rRNA	15 768	9793 (1.6)	4778 (3.3)	2530 (6.2)	457 (12.6)	1058 (14.9)

Εικόνα 5.4: Χρόνος εκτέλεσης σε δευτερόλεπτα στην προσέγγιση 34 με Xeon επεξεργαστές

(Πηγή [34])

Επιπλέον, ο CYK/inside αλγόριθμος παραλληλοποιείται στην προσέγγιση [35] σε μια FPGA πλακέτα, όπου σύμφωνα με τα πειραματικά αποτελέσματα της εργασία [35] η υλοποίηση σε υλικό παρουσιάζει σχεδόν παρόμοια απόδοση με αυτή σε λογισμικό επίπεδο (συστοιχία AMD επεξεργαστών) αλλά το κόστος υλικού και η κατανάλωση ρεύματος στην FPGA πλακέτα είναι πολύ μικρότερα.



Εικόνα 5.5: Απόδοση / κατανάλωση ενέργειας και λόγος απόδοσης / κόστους (FPGA VS. cluster)
(Πηγή [35])

5.5 Επίλογος

Στο πλαίσιο αυτής της εργασίας μελετήθηκαν περίπου σαράντα αναφορές που προσεγγίζουν το πρόβλημα της πρόγνωσης της δευτεροταγούς δομής του RNA με βάση τις στοχαστικές γραμματικές χωρίς συμφραζόμενα. Από την ανασκόπηση αυτών των προσεγγίσεων φαίνεται ότι το παραπάνω πρόβλημα υφίσταται μέχρι σήμερα ως πεδίο έρευνας παρόλο που έχει σημειωθεί μεγάλη πρόοδος. Οι ερευνητές τείνουν να προτείνουν παράλληλες υλοποιήσεις με σκοπό να πετύχουν γρήγορη ανάλυση και πρόγνωση. Οι περισσότερες παράλληλες υλοποιήσεις που μελετήθηκαν στη βιβλιογραφία αφορούσαν υλοποιήσεις λογισμικού. Παρ' όλα αυτά, υπάρχουν, και μάλιστα πρόσφατες, παράλληλες υλοποιήσεις σε FPGA πλακέτες που μπορούν να πετύχουν παρόμοια απόδοση σε σύγκριση με την απόδοση σε λογισμικό επίπεδο με επιπλέον όφελος το μικρότερο κόστος υλικού και κατανάλωση ρεύματος σε FPGA πλακέτες. Τα διαθέσιμα δεδομένα και το μήκος των προς εξέταση RNA ακολουθιών καθορίζουν σε μεγάλο βαθμό την απόδοση της υλοποίησης.

Επιπλέον, οι ερευνητές τείνουν να ενσωματώνουν, όπου είναι δυνατό και εξελικτικές πληροφορίες πολλαπλών RNA ακολουθιών στις προτεινόμενες προσεγγίσεις ώστε να πετύχουν προβλέψεις με μεγαλύτερη ακρίβεια, όπως στις δημοσιεύσεις [21], [30], [32], [60], [68]. Τέλος, μια άλλη προσέγγιση με μεγάλη δυναμική η οποία χρησιμοποιείται αρκετά από τους ερευνητές για την βελτίωση των προβλέψεων είναι η εισαγωγή προϋπάρχουσας γνώσης, δηλαδή η βελτίωση των

γραμματικών μέσα από την ανάλυση/στατιστική εκμάθηση ήδη γνωστών δομών, όπως στις δημοσιεύσεις [25], [27], [36].

Βιβλιογραφία

1. https://www.rnasociety.org/about/what-is-rna/
2. https://el.wikipedia.org/wiki/RNA
3. Durbin, R., Eddy, S. R., Krogh, A., & Mitchison, G. Biological sequence analysis: probabilistic models of proteins and nucleic acids. Cambridge university press, 1998.
4. Hofacker, Ivo L., Peter F. Stadler, and Peter F. Stadler. "RNA secondary structures." Reviews in Cell Biology and Molecular Medicine (2006).
5. Lorsch, Jon. Laboratory methods in enzymology: RNA. Vol. 530. Academic Press, 2013.
6. https://el.wikipedia.org/wiki/Κεντρικό_δόγμα_της_μοριακής_βιολογίας
7. Μπάγκος, Π., 2015. Βιοπληροφορική. [ηλεκτρ. βιβλ.] Αθήνα:Σύνδεσμος Ελληνικών Ακαδημαϊκών Βιβλιοθηκών. Διαθέσιμο στο: http://hdl.handle.net/11419/5016
8. Zuker, Michael. "[20] Computer prediction of RNA structure." Methods in enzymology. Vol. 180. Academic Press, 1989. 262-288.
9. Fürtig, B., Richter, C., Wöhnert, J., & Schwalbe, H. "NMR spectroscopy of RNA." Chembiochem 4.10 (2003): 936-962.
10. Ge, Ping, and Shaojie Zhang. "Computational analysis of RNA structures with chemical probing data." Methods 79 (2015): 60-66.
11. Zuker, Michael, and Patrick Stiegler. "Optimal computer folding of large RNA sequences using thermodynamics and auxiliary information." Nucleic acids research 9.1 (1981): 133-148.
12. Chiang, Y. T., and King-Sun Fu. "Parallel parsing algorithms and VLSI implementations for syntactic pattern recognition." IEEE Transactions on Pattern Analysis & Machine Intelligence 3 (1984): 302-314.
13. C. Pavlatos, A. C. Dimopoulos, A. Koulouris, T. Andronikos, I. Panagopoulos and G. Papakonstantinou, Efficient reconfigurable embedded parsers, Comput. Lang. Syst. Struct. 35 (2009) 196-215.
14. Moscola, James, Ron K. Cytron, and Young H. Cho. "Hardware-accelerated RNA secondary-structure alignment." ACM Transactions on Reconfigurable Technology and Systems (TRETS) 3.3 (2010): 14.
15. Xia, F., Dou, Y., Zhou, D., & Li, X. "Fine-grained parallel RNA secondary structure prediction using SCFGs on FPGA." Parallel Computing 36.9 (2010): 516-530.

16. Earley, Jay. "An efficient context-free parsing algorithm." Communications of the ACM 13.2 (1970): 94-102.
17. Dimopoulos, Alexandros C., Christos Pavlatos, and George Papakonstantinou. "A platform for the automatic generation of attributes evaluation hardware systems." Computer Languages, Systems & Structures 36.2 (2010): 203-222.
18. Pavlatos, Christos, Alexandros C. Dimopoulos, and George Papakonstantinou. "Parallel hardware stochastic context-free parsers." International Journal of Pattern Recognition and Artificial Intelligence 30.04 (2016): 1650008.
19. Lorenz, R., Wolfinger, M. T., Tanzer, A., & Hofacker, I. L. "Predicting RNA secondary structures from sequence and probing data." Methods 103 (2016): 86-98.
20. Dimopoulos, Alexandros C., Christos Pavlatos, and George Papakonstantinou. "Hardware Inexact Grammar Parser." International Journal of Pattern Recognition and Artificial Intelligence 31.11 (2017): 1759025.
21. Knudsen, Bjarne, and Jotun Hein. "RNA secondary structure prediction using stochastic context-free grammars and evolutionary history." Bioinformatics (Oxford, England) 15.6 (1999): 446-454.).
22. Knudsen, Bjarne, and Jotun Hein. "Pfold: RNA secondary structure prediction using stochastic context-free grammars." Nucleic acids research 31.13 (2003): 3423-3428.
23. Lari, Karim, and Steve J. Young. "The estimation of stochastic context-free grammars using the inside-outside algorithm." Computer speech & language 4.1 (1990): 35-56.
24. Sükösd, Z., Knudsen, B., Værum, M., Kjems, J., & Andersen, E. S. "Multithreaded comparative RNA secondary structure prediction using stochastic context-free grammars." BMC bioinformatics 12.1 (2011): 103.
25. Anderson, J. W., Tataru, P., Staines, J., Hein, J., & Lyngsø, R. "Evolving stochastic context-free grammars for RNA secondary structure prediction." BMC bioinformatics 13.1 (2012): 78.
26. Grate, Leslie. "Automatic RNA secondary structure determination with stochastic context-free grammars." ISMB. 1995.
27. Kato, Yuki, Hiroyuki Seki, and Tadao Kasami. "RNA pseudoknotted structure prediction using stochastic multiple context-free grammar." IPSJ Digital Courier 2 (2006): 655-664.
28. Griffiths-Jones, S., Bateman, A., Marshall, M., Khanna, A., & Eddy, S. R. "Rfam: an RNA family database." Nucleic acids research 31.1 (2003): 439-441.

29. Mizoguchi, Nobuyoshi, Yuki Kato, and Hiroyuki Seki. "Pairwise RNA pseudoknotted structure prediction based on stochastic grammar." 20th Intl. Conf. Genome Inform.(GIW2009) poster, pp. P018–1–P018–2. 2009.
30. Mizoguchi, Nobuyoshi, Yuki Kato, and Hiroyuki Seki. "A grammar-based approach to RNA pseudoknotted structure prediction for aligned sequences." Computational Advances in Bio and Medical Sciences (ICCABS), 2011 IEEE 1st International Conference on. IEEE, 2011.
31. Witwer, Christina, Ivo L. Hofacker, and Peter F. Stadler. "Prediction of consensus RNA secondary structures including pseudoknots." IEEE/ACM Transactions on Computational Biology and Bioinformatics (TCBB) 1.2 (2004): 66-77.
32. Rivas, Elena, and Sean R. Eddy. "Noncoding RNA gene detection using comparative sequence analysis." BMC bioinformatics 2.1 (2001): 8.
33. Xia, Fei, Yong Dou, and Guo-Qing Lei. "FPQRNA: Hardware-Accelerated QRNA Package for Noncoding RNA Gene Detecting on FPGA." Journal of Bioinformatics and Computational Biology 8.04 (2010): 743-761.
34. Liu, Tong, and Bertil Schmidt. "Parallel RNA secondary structure prediction using stochastic context-free grammars." Concurrency and Computation: Practice and Experience 17.14 (2005): 1669-1685.
35. Dou, Yong, Fei Xia, and Jingfei Jiang. "Fine-grained parallel application specific computing for rna secondary structure prediction using scfgs on fpga." Proceedings of the 2009 international conference on Compilers, architecture, and synthesis for embedded systems. ACM, 2009.
36. Zhu, Y., Xie, Z., Li, Y., Zhu, M., & Chen, Y. P. P. "Research on folding diversity in statistical learning methods for RNA secondary structure prediction." International Journal of Biological Sciences 14.6 (2018): 622-632.
37. Holmes, Ian, and Gerald M. Rubin. "Pairwise RNA structure comparison with stochastic context-free grammars." Biocomputing 2002. 2001. 163-174.
38. Holmes, Ian. "Accelerated probabilistic inference of RNA structure evolution." BMC bioinformatics 6.1 (2005): 73.
39. Dowell, Robin D., and Sean R. Eddy. "Efficient pairwise RNA structure prediction and alignment using sequence alignment constraints." BMC bioinformatics 7.1 (2006): 400.

40. Sakakibara, Y., Brown, M., Hughey, R., Mian, I. S., Sjölander, K., Underwood, R. C., & Haussler, D. "Stochastic context-free grammars for tRNA modeling." Nucleic acids research 22.23 (1994): 5112-5120.
41. Eddy, Sean R., and Richard Durbin. "RNA sequence analysis using covariance models." Nucleic acids research 22.11 (1994): 2079-2088.
42. Brown, Michael, and Charles Wilson. "RNA pseudoknot modeling using intersections of stochastic context free grammars with applications to database search." Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing. 1996.
43. Cai, Liming, Russell L. Malmberg, and Yunzhou Wu. "Stochastic modeling of RNA pseudoknotted structures: a grammatical approach." Bioinformatics 19.suppl_1 (2003): i66-i73.
44. García, Rafael. "Prediction of RNA pseudoknotted secondary structure using Stochastic Context Free Grammars (SCFG)." CLEI Electronic Journal 9.2 (2006).
45. Song, Y., Liu, C., Malmberg, R. L., He, C., & Cai, L. "Memory efficient alignment between RNA sequences and stochastic grammar models of pseudoknots." International journal of bioinformatics research and applications 2.3 (2006): 289-304.
46. Gardner, Paul P., and Robert Giegerich. "A comprehensive comparison of comparative RNA structure prediction approaches." BMC bioinformatics 5.1 (2004): 140.
47. Mathews, David H., Walter N. Moss, and Douglas H. Turner. "Folding and finding RNA secondary structure." Cold Spring Harbor perspectives in biology (2010): a003665.
48. Sankoff, David. "Simultaneous solution of the RNA folding, alignment and protosequence problems." SIAM journal on applied mathematics 45.5 (1985): 810-825.
49. Sethupathy, Praveen, Benoit Corda, and Artemis G. Hatzigeorgiou. "TarBase: A comprehensive database of experimentally supported animal microRNA targets." Rna 12.2 (2006): 192-197.
50. https://en.wikipedia.org/wiki/X-ray_crystallography
51. Zuker, Michael, and David Sankoff. "RNA secondary structures and their prediction." Bulletin of mathematical biology 46.4 (1984): 591-621.
52. Tan, Guangming, Shengzhong Feng, and Ninghui Sun. "load balancing algorithm in cluster-based RNA secondary structure prediction." Parallel and Distributed Computing, 2005. ISPDC 2005. The 4th International Symposium on. IEEE, 2005.

53. Li, J., Xu, C., Liang, H., Cong, W., Wang, Y., Luan, K., & Liu, Y. "RGRNA: prediction of RNA secondary structure based on replacement and growth of stems." Computer methods in biomechanics and biomedical engineering 20.12 (2017): 1261-1272.
54. https://en.wikipedia.org/wiki/Formal_grammar
55. Chomsky, Noam. "Three models for the description of language." IRE Transactions on information theory 2.3 (1956): 113-124.
56. https://en.wikipedia.org/wiki/Nuclear_magnetic_resonance_spectroscopy
57. Habashy, Shahira M. "Stochastic context-free grammars method aided calculation of the local folding potential of target RNA."
58. Brown, Michael PS. "Small subunit ribosomal RNA modeling using stochastic context-free grammars." ISMB. 2000.
59. Younger, Daniel H. "Recognition and parsing of context-free languages in time n3." Information and control 10.2 (1967): 189-208.
60. Pedersen, J. S., Bejerano, G., Siepel, A., Rosenbloom, K., Lindblad-Toh, K., Lander, E. S.,. & Haussler, D. "Identification and classification of conserved RNA secondary structures in the human genome." PLoS computational biology 2.4 (2006): e33.
61. Moscola, James, Ron K. Cytron, and Young H. Cho. "Hardware-accelerated RNA secondary-structure alignment." ACM Transactions on Reconfigurable Technology and Systems (TRETS) 3.3 (2010): 14.
62. Griffiths-Jones, S., Grocock, R. J., Van Dongen, S., Bateman, A., & Enright, A. J. "miRBase: microRNA sequences, targets and gene nomenclature." Nucleic acids research 34.suppl_1 (2006): D140-D144.
63. Hsu, S. D., Lin, F. M., Wu, W. Y., Liang, C., Huang, W. C., Chan, W. L., ... & Chien, C. H "miRTarBase: a database curates experimentally validated microRNA–target interactions." Nucleic acids research 39.suppl_1 (2010): D163-D169.
64. Zuker, Michael. "Mfold web server for nucleic acid folding and hybridization prediction." Nucleic acids research 31.13 (2003): 3406-3415.
65. Hofacker, Ivo L. "Vienna RNA secondary structure server." Nucleic acids research 31.13 (2003): 3429-3431.
66. Backus, John W. "The syntax and semantics of the proposed international algebraic language of the Zurich ACM-GAMM conference." Proceedings of the International Conference on Information Processing, 1959 (1959).

67. https://en.wikipedia.org/wiki/Phylogenetic_tree
68. Anderson, J. W., Haas, P. A., Mathieson, L. A., Volynkin, V., Lyngso, R., Tataru, P., & Hein, J. "Oxfold: kinetic folding of RNA using stochastic context-free grammars and evolutionary information." Bioinformatics 29.6 (2013): 704-710.
69. https://en.wikipedia.org/wiki/Chomsky_normal_form