ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ
ΣΧΟΛΗ ΗΛΕΚΤΡΟΛΟΓΩΝ ΜΗΧΑΝΙΚΩΝ ΚΑΙ ΜΗΧΑΝΙΚΩΝ ΥΠΟΛΟΓΙΣΤΩΝ

ΤΟΜΕΑΣ ΣΥΣΤΗΜΑΤΩΝ ΜΕΤΑΔΟΣΗΣ ΠΛΗΡΟΦΟΡΙΑΣ ΚΑΙ ΤΕΧΝΟΛΟΓΙΑΣ ΥΛΙΚΩΝ

Υπολογιστική βελτιστοποίηση της θεραπείας Exon Skipping για τη Μυϊκή Δυστροφία Duchenne με μεθόδους μηχανικής μάθησης

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

της

Τετιάνα Πριντζί

Επιβλέπων: Δημήτριος-Διονύσιος Κουτσούρης
Καθηγητής Ε.Μ.Π.

Αθήνα, Δεκέμβριος 2020

ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ
ΣΧΟΛΗ ΗΛΕΚΤΡΟΛΟΓΩΝ ΜΗΧΑΝΙΚΩΝ ΚΑΙ ΜΗΧΑΝΙΚΩΝ ΥΠΟΛΟΓΙΣΤΩΝ
ΤΟΜΕΑΣ ΣΥΣΤΗΜΑΤΩΝ ΜΕΤΑΔΟΣΗΣ ΠΛΗΡΟΦΟΡΙΑΣ ΚΑΙ ΤΕΧΝΟΛΟΓΙΑΣ ΥΛΙΚΩΝ

Υπολογιστική βελτιστοποίηση της θεραπείας Exon Skipping για τη Μυϊκή Δυστροφία Duchenne με μεθόδους μηχανικής μάθησης

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

της

Τετιάνα Πριντζί

Επιβλέπων: Δημήτριος-Διονύσιος Κουτσούρης
Καθηγητής Ε.Μ.Π.

Εγκρίθηκε από την τριμελή εξεταστική επιτροπή την 9η Δεκεμβρίου 2020.

.....	.....	.....
Δημήτριος-Διονύσιος Κουτσούρης Καθηγητής Ε.Μ.Π.	Γιώργος Ματσόπουλος Καθηγητής Ε.Μ.Π.	Παναγιώτης Τσανάκας Καθηγητής Ε.Μ.Π.

Αθήνα, Δεκέμβριος 2020

.....
ΤΕΤΙΑΝΑ ΠΡΙΝΤΙΪ
Διπλωματούχος Ηλεκτρολόγος Μηχανικός και Μηχανικός Υπολογιστών Ε.Μ.Π.

Περίληψη

Ο σκοπός αυτής της διπλωματικής εργασίας είναι να διερευνηθεί η βέλτιστη εφαρμογή της θεραπευτικής μεθόδου exon skipping για τη μυασθένεια Duchenne (DMD). Με χρήση μεθόδων μηχανικής μάθησης, διερευνάται ποιά είναι τα κατάλληλα σημεία πρόσδεσης των ολιγονουκλεοτιδίων στο εξώνιο, που πρέπει να αποκοπεί κατά το μάτισμα, έτσι ώστε να αποκατασταθεί το πλαίσιο ανάγνωσης του mRNA και να παραχθεί λειτουργική παραλλαγή της δυστροφίνης.

Μελετήσαμε δύο σύνολα δεδομένων που αντιστοιχούν σε Phosphorodiamidate Morpholino Oligomer (PMO) και 2' O Methyl Phosphorothioate (2OMePS) ολιγονουκλεοτίδια. Εφαρμόσαμε διαφορετικές προσεγγίσεις επιλογής χαρακτηριστικών για την εκπαίδευση μοντέλων μηχανικής μάθησης, που με τη σειρά τους αντιστοιχούσαν σε πολλαπλούς αλγορίθμους παλινδρόμησης και ταξινόμησης. Κάναμε περαιτέρω πειράματα στο 2OMePS σύνολο δεδομένων λόγω έλλειψης συνοχής στα χαρακτηριστικά του.

Τα αποτελέσματα των πειραμάτων μας παρουσιάζουν συνέπεια σε σχέση με την υπάρχουσα βιβλιογραφία. Το πιο σημαντικό χαρακτηριστικό που συμβάλλει στο exon skipping, επιβεβαιώθηκε να είναι η ενέργεια πρόσδεσης του ολιγονουκλεοτιδίου στην εξωνική αλληλουχία. Η διπλωματική αυτή προτείνει το παραπάνω χαρακτηριστικό να συνδυάζεται με το ποσοστό των GC βάσεων στην αλληλουχία στόχο, για μέγιστη αποτελεσματικότητα. Καθώς τα παραπάνω φαίνεται να είναι κοινά και στις δύο ομάδες ολιγονουκλεοτιδίων, παρατηρήθηκαν κάποια ακόμη σημαντικά χαρακτηριστικά για το 2OMePS σύνολο δεδομένων, τα οποία ενθαρρύνεται να μελετήθούν περισσότερο στο μέλλον.

Λέξεις κλειδιά:
exon skipping, DMD, δυστροφίνη, ολιγονουκλεοτίδια, PMO, 2OMePS, μηχανική μάθηση

Abstract

The aim of this thesis is to study how to achieve optimal performance of the exon skipping therapy for Duchenne muscular dystrophy (DMD). Machine learning techniques are used in order to detect the optimal oligonucleotide binding sites on the targeted exon. Exon skipping during mRNA splicing is intended to restore the reading frame and eventually induce the production of a semifunctional protein that can replace dystrophine.

We used two datasets for our studies, each representing a different oligo chemistry kind: Phosphorodiamidate Morpholino Oligomer (PMO) and 2' O Methyl Phosphorothioate (2OMePS). We performed different feature selection approaches in order to train our models, which were based on multiple regression and classification algorithms. Additional experiments were performed on 2OMePS dataset, as it lacks feature coherence.

The results of our studies are consistent with the existing bibliography. The most important feature for exon skipping was confirmed to be the binding energy of the oligo to the target. This thesis suggests the percentage of GC bases in the target sequence to be combined with the above feature for optimal results. Although these results are similar for both datasets, there were additional features that seemed to induce exon skipping for 2OMePS dataset. Further studies are encouraged to be performed in this area.

Keywords:
exon skipping, DMD, dystrophine, oligonucleotides, PMO, 2OMePS, machine learning

Ευχαριστίες

Αρχικά θα ήθελα να ευχαριστήσω τους επιβλέποντές μου από το Ε.Κ.Ε.Φ.Ε. Δημόκριτος, τον Δρ. Γιώργο Παλιούρα και την Δρ. Αναστασία Κριθαρά, καθώς και τον Τάσο Νεντίδη, που μου έδωσαν την ευκαιρία να συνεργαστώ μαζί τους για την εκτέλεση αυτής της διπλωματικής. Τους ευχαριστώ πολύ για την καθοδήγησή τους κατά τη διάρκεια αυτής της μακριάς πορείας, αλλά και για την αμέριστη υπομονή τους και τη συνεχή παρότρυνση που μου προσφέρανε. Ακόμη, θα ήθελα να εκφράσω πόσο ευγνώμων είμαι για την προθυμία τους να με καθοδηγήσουν εκ αποστάσεως κατά την παραμονή μου στο εξωτερικό αλλά και εν μέσω των δύσκολων συνθηκών της πανδημίας που βιώνουμε σήμερα.

Θέλω να ευχαριστήσω ακόμη τον καθηγητή μου απο το Ε.Μ.Π. Δημήτριο-Διονύσιο Κουτσούρη και την Δρ. Ουρανία Πετροπούλου που μου έδωσαν την ευελιξία της συνεργασίας με το Ε.Κ.Ε.Φ.Ε. Δημόκριτος και αποτέλεσαν τον σύνδεσμό μου με τη σχολή.

Η πορεία μου μέσα στη Σχολή ΗΜΜΥ και η διαδικασία εκτέλεσης αυτής της διπλωματικής συνοδεύτηκε από πολλές αλλαγές στη ζωή μου. Θα ήθελα να ευχαριστήσω την οικογένειά μου που με βοηθάει με τη στήριξή της σε όλες τις επιλογές μου. Αφιερώνω αυτή τη διπλωματική στη μητέρα μου, η οποία μου έδωσε τη δυνατότητα και τα εφόδια να σπουδάσω και με την υπερπροσπάθειά της με ενέπνευσε να αγωνίζομαι για τους στόχους μου παρά τις δυσκολίες και τα εμπόδια.

Τέλος, θέλω να ευχαριστήσω όλους τους φίλους μου που αποτέλεσαν το δίχτυ ασφαλείας μου όλα αυτά τα χρόνια. Η δεύτερη οικογένειά μου, το σπίτι μου στην Ελλάδα, είναι τα άτομα με τα οποία πορευτήκαμε μαζί και δεν με άφησαν ούτε στιγμή χωρίς υποστήριξη. Ο καθένας με τον δικό του τρόπο, με βοήθησε να εκτελέσω αυτό το έργο και είμαι βαθιά ευγνώμων. Ιδιαίτερα ευχαριστώ τον Μάνο που με στήριξε τους τελευταίους μήνες αυτής της προσπάθειας και πίστεψε σε μένα στις πιο δύσκολες στιγμές.

Περιεχόμενα

Κατάλογος Σχημάτων

Κατάλογος Πινάκων

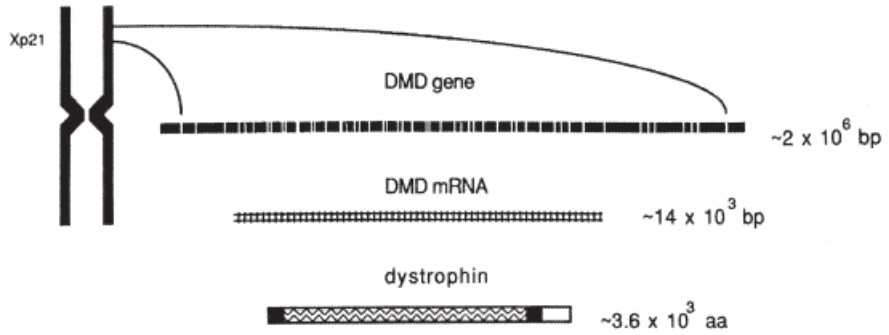
Κεφάλαιο 1

Εισαγωγή

1.1 Exon Skipping θεραπεία στη Μυϊκή Δυστροφία Duchenne

Η δυστροφίνη (Σχήμα 1.1) είναι πρωτεΐνη που βρίσκεται σε σκελετικούς και καρδιακούς μύες και κωδικοποιείται από το μεγαλύτερο γονίδιο του ανθρώπινου οργανισμού. Οι μεταλλάξεις στο γονίδιο της δυστροφίνης μπορούν να επιφέρουν την παραγωγή δυσλειτουργικής πρωτεΐνης ή ακόμη και τον τερματισμό της παραγωγής της. Οι γενετικές ασθένειες που προκύπτουν, ονομάζονται μυϊκές δυστροφίες και ο φαινότυπός τους επηρεάζεται άμεσα από την αλλαγή που έχει υποστεί η δυστροφίνη ή στη χειρότερη περίπτωση, την απουσία της. Επειδή οι μυϊκές δυστροφίες είναι Χ-φυλοσύνδετες ασθένειες, παρουσιάζονται μόνο στους άντρες, καθώς οι γυναίκες μπορούν να είναι φορείς.

Η πιο σοβαρή μυοπάθεια είναι η μυϊκή δυστροφία Duchenne (DMD), που συναντάται σε παιδική ηλικία σε 1:3500 άντρες και οδηγεί, με ταχείς ρυθμούς, σε απώλεια βάδισης και τον θάνατο. Σε αυτή την περίπτωση η μετάλλαξη είναι καταστροφική, αφού αποτρέπει τη σύνθεση της δυστροφίνης και οι μύες υπερτροφούν με αποτέλεσμα να χάσουν την κινητικότητά τους. Η ασθένεια αυτή δεν θεραπεύεται ολοκληρωτικά. Υπάρχουν όμως τρόποι επέμβασης στην έκφραση της γονιδιακής πληροφορίας που μπορούν να αποκαταστήσουν την



Σχήμα 1.1: Η τοποθεσία του γονιδίου της δυστροφίνης στο χρωμόσωμα Xp21, το γονίδιο, το μεταφρασμένο mRNA και η παραγόμενη πρωτεΐνη. Το σχήμα προέρχεται από τους *Darras et al.*[1]

παραγωγή της δυστροφίνης, αν και όχι στην κανονική μορφή της. Η θεραπευτική αυτή μέθοδος οδηγεί σε φαινότυπο με λιγότερο σοβαρές επιπλοκές στον ασθενή (μυϊκή δυστροφία Becker) και σημαντική βελτίωση της ποιότητας ζωής του. Όλα τα παραπάνω παρουσιάζονται πιο αναλυτικά στο *Dystrophinopathies* των *Darras et al.*[1]

Ο τελικός φαινότυπος είναι άμεσο αποτέλεσμα της γονιδιακής έκφρασης. Πιο αναλυτικά, στους ευκαρυωτικούς οργανισμούς, κατά τη διάρκεια της μεταγραφής ενός γονιδίου δημιουργείται το λεγόμενο πρόδρομο mRNA, το οποίο περιέχει αλληλουχίες που φέρουν γονιδιακή πληροφορία (εξώνια) και άλλες που παρεμβαίνουν ανάμεσα τους και δεν έχουν αυτήν την ιδιότητα (εσώνια). Στη συνέχεια, το πρόδρομο mRNA υπόκειται σε διαδικασία ωρίμανσης, κατά την οποία γίνεται αποκοπή των εσωνίων και συρραφή των εξωνίων μεταξύ τους. Η διαδικασία αυτή ονομάζεται μάτισμα και αποτελεί το κλειδί για τη θεραπευτική μέθοδο exon skipping, με την οποία ασχολείται αυτή η εργασία.

Αυτό που επιδιώκεται με τη θεραπευτική προσέγγιση exon skipping είναι η σωστή ανάγνωση της γονιδιακής πληροφορίας κατά τη μετάφραση του ώριμου mRNA μέσω παρεμβολής στη διαδικασία του ματίσματος, όπου δίνεται η δυνατότητα να αποκοπεί, μαζί με τα εσώνια, το εξώνιο που έχει προσβληθεί από τη μετάλλαξη [18]. Η

πρωτεΐνη που παράγεται δεν είναι πλήρως λειτουργική, εφόσον της λείπουν ορισμένα αμινοξέα, αλλά γίνεται αποδεκτή από το κύτταρο και επιτελεί τον ρόλο της δυστροφίνης.

Το εργαλείο που χρησιμοποιείται από τους βιολόγους για την επιτυχή παράληψη ενός εξωνίου, είναι οι μικρές αλληλουχίες βάσεων νουκλεϊκού όξεος, που ονομάζονται ολιγονουκλεοτίδια (antisense oligonucleotides ή AONs), και μπορούν να εισχωρίσουν από την κυτταρική μεμβράνη και να προσδεθούν σε κατάλληλα σημεία του εξωνίου που έχει υποστεί μετάλλαξη [12]. Για να έχουμε λειτουργικά και ευέλικτα ολιγονουκλεοτίδια, επιβάλλεται περιορισμός στο μήκος τους που αυτόματα μας οδηγεί στο ερώτημα ποια είναι τα σημεία του εξωνίου που πρέπει να καλυφθούν από το ολιγονουκλεοτίδιο έτσι ώστε να έχουμε μέγιστη πιθανότητα παράληψής του, από το μηχανισμό του ματίσματος, κατά τη συρραφή του ώριμου mRNA.

1.2 Αντικείμενο διπλωματικής

Η διπλωματική αυτή μελετά τα χαρακτηριστικά των εξωνίων που πιστεύεται πως σχετίζονται με την ενίσχυση και την καταστολή του ματίσματος, έτσι ώστε να αποφανθεί σε ποιες ακριβώς αλληλουχίες θα πρέπει να είναι συμπληρωματικά τα ολιγονουκλεοτίδια για τη μέγιστη επιτυχία του exon skipping. Προς το παρόν, δύο είναι τα είδη ολιγονουκλετιδίων που έχουν δημιουργηθεί και μελετηθεί από τους ειδικούς και η διαφορά τους έγκειται στη χημική τους σύνθεση. Πρόκειται για τα Phosphorodiamidate Morpholino Oligomer (PMO) και τα 2' O Methyl Phosphorothioate (2OMePS) και παρακάτω γίνετα μελέτη και των δύο.

Η επιλογή των χαρακτηριστικών γίνεται με μεθόδους μηχανικής μάθησης με επίτηρηση (supervised machine learning) που αφορούν αλγορίθμους παλινδρόμησης και ταξινόμησης.

1.3 Συνεισφορά

Στην εργασία αυτή επιχειρείται η εφαρμογή διαφόρων μεθόδων επιλογής σημαντικών χαρακτηριστικών, έτσι ώστε να αποφευχθούν συμπεράσματα που είναι εξαρτημένα από τα εργαλεία επιλογής και να φανούν αυτά που είναι ισχυρά και παρουσιάζουν συνέπεια.

Η ποικιλία μεθόδων που επιτελούνται σε αυτή τη διπλωματική αφορά τα χαρακτηριστικά του συνόλου δεδομένων, τον τρόπο επιλογής των χαρακτηριστικών και τους αλγορίθμους μηχανικής μάθησης για την εκπαίδευση του μοντέλου εξιολόγησης.

Έπειτα από την αξιολόγηση των αποτελεσμάτων της πειραματικής διαδικασίας, φιλοδοξείται η πρόταση ενός μοντέλου που μπορεί να χρησιμοποιηθεί για την πρόβλεψη της αποδοτικότητας των ολιγονουκλεοτιδίων λαμβάνοντας ως είσοδο δεδομένων τα χαρακτηριστικά με τη μέγιστη συμβολή στην επιτυχία του exon skipping.

1.4 Οργάνωση Κειμένου

Το κείμενο διαρθρώνεται στα εξής κεφάλαια:

Στο **Κεφάλαιο 2** παρουσιάζεται το βιολογικό υπόβαθρο που θεωρείται απαραίτητο για την κατανόηση του σκοπού αυτής της εργασίας.

Στο **Κεφάλαιο 3** γίνεται σύντομη επεξήγηση των υπολογιστικών εργαλείων που χρησιμοποιούνται στο πειραματκό μέρος της διπλωματικής.

Στο **Κεφάλαιο 4** αναφέρονται τα εργαλεία εξαγωγής χαρακτηριστικών και οι σχετικές έρευνες γύρω από τη θεραπεία της μυοπάθειας Duchenne με exon skipping.

Στο Κεφάλαιο 5 γίνεται παρουσίαση της μεθοδολογίας που ακολουθείται στο υπολογιστικό κομμάτι της διπλωματικής ώστε να επιλεγούν τα βέλτιστα χαρακτηριστικά για την αποδοτικότητα του exon skipping.

Στο Κεφάλαιο 6 αναλύονται τα αποτελέσματα των πειραμάτων που πραγματοποιήθηκαν ώστε να αναπτυχθούν τα τελικά συμπεράσματα .

Κεφάλαιο 2

Βιολογικό Υπόβαθρο

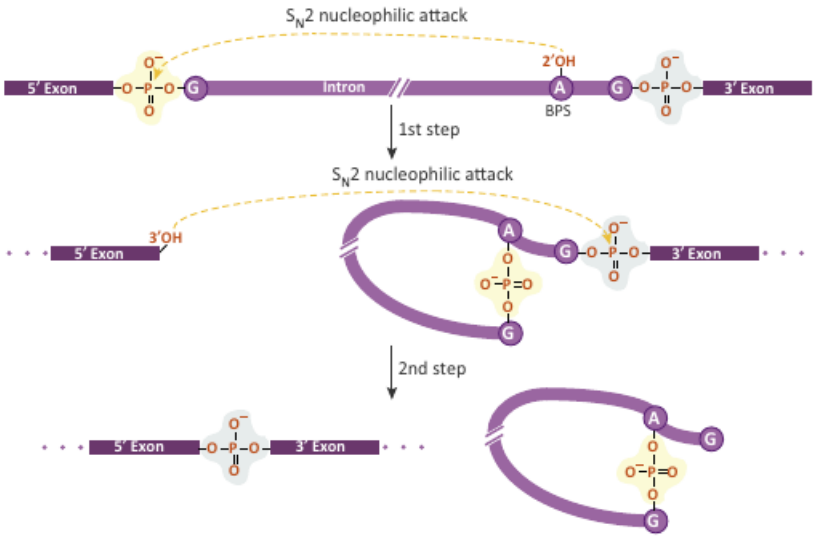
2.1 Μάτισμα (Splicing)

Κατά τη δημιουργία του ώριμου mRNA γίνεται αποκοπή και απομάκρυνση των εσωνίων από το πρόδρομο mRNA. Για να επιτευχθεί κάτι τέτοιο, θα πρέπει να εντοπιστούν τα σημεία που καθορίζουν το τέλος ενός εσωνίου και την αρχή του διπλανού εξωνίου. Χάρη στην ασάφεια που κυριαρχεί γύρω από αυτόν τον ορισμό, δημιούργουνται εναλλακτικά μοτίβα αποκοπής εσωνίων, με αποτέλεσμα να διευρύνεται σημαντικά η ποικολομορφία στη μεταγραφή του γονιδίου ενός πολυκυτταρικού οργανισμού.

Κατά τη διάρκεια της ανάπτυξης και της διαφοροποίησης ενός κυττάρου στον ανθρώπινο οργανισμό γίνεται εκτεταμένη χρήση της γονιδιακής ρύθμισης, μιας και πάνω από το 70% του ανθρώπινου γονιδιόματος υπόκειται σε εναλλακτικό μάτισμα [11]. Εάν όμως κατα τη διαδικασία του ματίσματος του mRNA συμβούν απρόβλεπτες αλλαγές και λάθη, το αποτέλεσμα μπορεί να επιφέρει σοβαρές γενετικές ασθένειες, όπως είναι η μυϊκή δυστροφία Duchenne.

Για τον ορισμό των εξωνίων χρησιμοποιούνται τα κανονικά σήματα ματίσματος (Σχήμα 2.1) που βρίσκονται μέσα στα εσώνια:

- 5′ θέση ματίσματος (5’ ss) ή θέση δότη (donor site)
- θέση διακλάδωσης (branch site)



Σχήμα 2.1: Το μάτισμα του RNA πραγματοποιείται σε δύο βήματα. Στο πρώτο βήμα, το σημείο διακλάδωσης (branch point site ή BPS) ενώνεται με τη φωσφορική ομάδα μεταξύ του 5′ άκρου του εξωνίου και του 3′ άκρου του εσωνίου. Στο δεύτερο βήμα, το 5′ άκρο του εξωνίου συνδέεται με το 3′ άκρο του επόμενου εξωνίου αποκόβοντας το εσώνιο. Το σχήμα προέρχεται από τους *Papasaikas et al.*[2]

- 3′ θέση ματίσματος (3’ ss) ή θέση αποδέκτη (acceptor site)

Τα σήματα αυτά, όμως, δεν επαρκούν για τον ορισμό των εξωνίων με ακρίβεια [9, 10]. Κατά το μήκος των εσωνίων συναντιούνται πολλές αλληλουχίες που μοιάζουν με εξώνια λόγω του μήκους τους και της ύπαρξης των θέσεων ματίσματος στα άκρα τους, τα λεγόμενα ψευδοεξώνια, αλλά δεν υπόκεινται ποτέ σε μάτισμα επειδή δεν αναγνωρίζονται ως εξώνια από τον μηχανισμό ματίσματος. Έτσι, η παρουσία των κανονικών θέσεων ματίσματος είναι αναγκαία αλλά όχι επαρκής συνθήκη για να οριστεί μια αλληλουχία ως εξώνιο.

Για αυτό, υπάρχουν βοηθητικές εξωνικές αλληλουχίες, που λέγονται ενισχυτές (ESEs) και καταστολείς(ESSs) ματίσματος και είναι πολύ σημαντικοί παράγοντες για το εναλλακτικό μάτισμα, αν και είναι παρούσες και στα εξώνια κανονικής έκφρασης [11]. Οι ενισχυτές ματίσματος αναγνωρίζονται από τις SR (serine/arginine-rich) πρω-

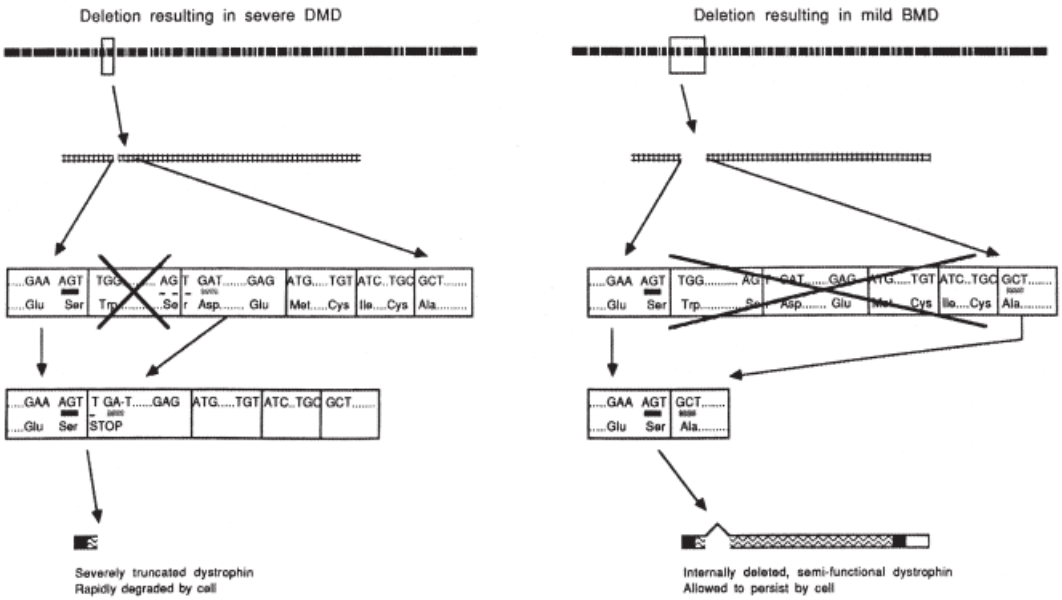
τεΐνες, οι οποίες προτρέπουν τον μηχανισμό ματίσματος να προσδεθεί στις γειτονικές θέσεις ματίσματος ή και αναιρούν τη δράση των γειτονικών καταστολέων ματίσματος. Οι καταστολείς ματίσματος (ESSs) είναι σημεία πρόσδεσης πρωτεϊνών (hnRNP) που συμβάλλουν στο exon skipping κατά το μάτισμα. Θα πρέπει να σημειωθεί πως υπάρχουν και αντίστοιχες βοηθητικές αλληλουχίες που βρίσκονται στα εσώνια (ISEs, ISSs).

Όπως έχει αναφερθεί πιο πάνω, οι SR πρωτεΐνες αλληλεπιδρούν με πρωτεΐνες του μηχανισμού ματίσματος (spliceosome) και ειδικές αλληλουχίες του πρόδρομου mRNA, παίζοντας έτσι καθοριστικό ρόλο στη διαδικασία της αποκοπής των εσωνίων και του εναλλακτικού ματίσματος. Από τις SR πρωτεΐνες που έχουν παρατηρηθεί [4], οι παρακάτω εξετάζονται σε αυτήν την εργασία ως παράγοντες συμβολής στο exon skipping.

- SF2/ASF (Splicing Factor 2/Alternative Splicing Factor)
- SC35
- SRp40
- SRp55
- BRCA1
- Tra2β
- 9G8

2.2 Exon Skipping

Για να κατανοήσουμε τη δράση του exon skipping θα πρέπει να δούμε πως συντίθεται η δυστροφίνη. Μία πρωτεΐνη αποτελείται από αμινοξέα, τα οποία κωδικοποιούνται από τριπλέτες ριβονουκλεϊκού οξέος, τα κωδικόνια. Αν κατα τη μετάλλαξη του γονιδίου της



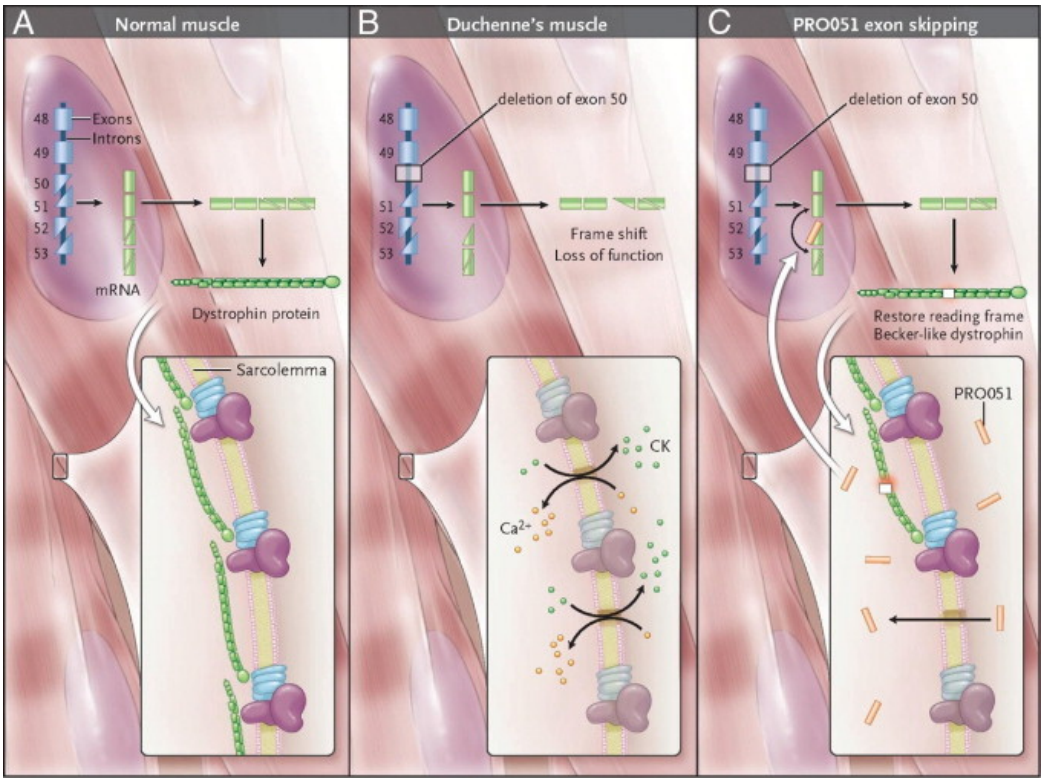
Σχήμα 2.2: Αριστερά βλέπουμε πως μία μικρή απώλεια βάσεων από το γονίδιο της δυστροφίνης προκαλεί την αποτορπή της παραγωγής της και επομένως την εφάνιση της μυασθένειας Duchenne (DMD). Αντιθέτως, στη δεξιά εικόνα, βλέπουμε απώλεια μεγαλύτερου μήκους να επιφέρει πολύ πιο ομαλό φαινότυπο που αποτελεί τη μυασθένεια Becker (BMD). Στην πρώτη περίπτωση, το πλαίσιο ανάγνωσης των βάσεων ανά τριάδες χαλάει και αμέσως μετά την διαγραφή των βάσεων δημιουργείται καινούργιο κωδικόνιο (TGA) που μεταφράζεται σε τερματισμό της μετάφρασης. Αντίστοιχα, στη δεύτερη περίπτωση, παρά τη διαγραφή περισσότερων βάσεων, τυχαίνει να διατηρείται η σωστή ανάγνωση των βάσεων που απομένουν και να παράγεται ημιλειτουργική πρωτεΐνη, η οποία δεν απορρίπτεται από το κύτταρο. Το σχήμα προέρχεται από τους *Darras et al.*[1]

πρωτεΐνης διασπαστεί το πλαίσιο ανάγνωσης των βάσεων, με την προσθήκη ή την αφαίρεση βάσης από κάποια εξωνική τριπλέτα, για παράδειγμα, η γονιαδική πληροφορία αλλάζει ριζικά, κωδικοποιώντας λανθασμένα αμινοξέα ή ακόμη και δημιουργώντας αλληλουχίες τερματισμού της μετάφρασης (Σχήμα 2.2).

Ο σκοπός του exon skipping είναι η ανάκτηση του σωστού πλαισίου ανάγνωσης κατά τη διάρκεια του ματίσματος του πρόδρομου mRNA [3]. Αυτό επιτυγχάνεται με την αποτροπή της συρραφής του εξωνίου με τη λανθασμένη πληροφορία, παραλείποντας το στην αλληλουχία του ώριμου mRNA (Σχήμα 2.3). Έτσι, κατά τη διάρκεια της μετάφρασης, οι τριπλέτες των κωδικονίων θα είναι ορθές, παρόλο που ορίσμενες θα λείπουν. Ως αποτέλεσμα αυτής της διαδικασίας, θα προκύψει λειτουργική πρωτεΐνη με πολύ πιο ήπιο φαινότυπο μυασθένειας (BMD).

2.3 Μηχανισμός Ματίσματος (Spliceosome)

Σε αυτό το σημείο θα πρέπει να σταθούμε στο ερώτημα, γιατί τα κύτταρα διαθέτουν πολύτιμες ώρες και πόρους για τη μεταγραφή και τελικά την αποκοπή των εσωνίων. Η απάντηση μπορεί να βρεθεί στην προσπάθεια να κατανοηθεί η προέλευση των εσωνίων και η καθοριστική τους εξάρτηση από τον μηχανισμό ματίσματος. Είναι πιθανό, τα εσώνια να προέρχονται από κινητά γενετικά στοιχεία, με δυνατότητα συναρμογής του εαυτού τους, και συνεπώς εύκολης εξάπλωσης σε όλο το γονιδίωμα. Τα στοιχεία αυτά ονομάζονται εσώνια Ομάδας II και δεν συναντιούνται σε ευκαρυωτικά πυρηνικά γονιδιώματα. Η βασική διαφορά μεταξύ των εσωνίων και των εσωνίων Ομάδας II βρίσκεται στην ικανότητα των τελευταίων να τοποθετούν με ακρίβεια στο χώρο αντιδραστικές χημικές ομάδες, χάρη στην περίπλοκη τριτοταγή δομή τους. Αντιθέτως, η αποκοπή των εσωνίων του πρόδρομου mRNA εξαρτάται από τον μηχανισμό ματίσματος, η μοριακή δομή του οποίου διευκολύνει τις αντι-



Σχήμα 2.3: Παράδειγμα δράσης φαρμακευτικής ουσίας με ολιγονουκλεοτίδια στη μυασθένεια Duchenne. **A:**Μάτισμα της δυστροφίνης σε έναν υγιή μυ. **B:** Ασθενής με DMD λόγω διαγραφής του εξωνίου 55. Τα υπόλοιπα εξώνια υπόκεινται σε μάτισμα, αλλά με λανθασμένο πλαίσιο ανάγνωσης το οποίο κάνει αδύνατη την παραγωγή της δυστροφίνης. **C:** Η δράση του ολιγονουλεοτιδιακού φαρμάκου PRO051 που στοχεύει στο εξώνιο 51. Με την παράληψη του εξωνίου 51, ενώνονται μεταξύ τους τα εξώνια 49 και 52 και αποκαθίστεται το σωστό πλαίσιο ανάγνωσης. Ως αποτέλεσμα, παράγεται η ημιλειτουργική δυστροφίνη τύπου Becker. Το σχήμα προέρχεται από τους *Hoffman et al.*[3]

δράσεις μεταφοράς φωσφορυλίου καλύπτοντας τα σημεία ματίσματος με τμήματα RNA του ιδίου, και ανακατασκευάζει το καταλυτικό κέντρο των εσωνίων Ομάδας II.

Καθοριστικό ρόλο στη λειτουργεία του μηχανισμού ματίσματος έχουν και οι πρωτεΐνες του, οι οποίες ρυθμίζουν τη ριβοζύμη (καταλυτικό RNA). Συγκεκριμένα, οι πρωτεΐνες αυτές λειτουργούν σαν κοιλότητα σταθεροποίησης για την καταλυτική τοποθεσία και διασφαλίζουν πως αυτή παραμένει ανοιχτή και προσβάσιμη. Η ιδιότητα αυτή είναι πολύ σημαντική επειδή επιτρέπει την εξυπηρέτηση αλληλουχιών με διαφορετικά μήκη, αλλά και ενδείκνυται για μάτισμα αμέσως μετά την ολοκλήρωση της μεταγραφής του πρόδρομου mRNA. Φαίνεται πως για κάποια εσώνια το μάτισμα μπορεί να ξεκινήσει πριν ολοκληρωθεί η μεταγραφή και οι δύο μηχανισμοί είναι τόσο στενά συνδεδεμένοι που μπορεί να χρησιμοποιούν ακόμη και κοινούς πόρους. Έτσι, η προσβασιμότητα συγκεκριμένων νουκλεοτιδιακών αλληλουχιών αποτελεί σημαντικό υποψήφιο παράγοντα για την επιτυχία της μεθόδου exon skipping. Για πιο αναλυτική περιγραφή των παραπάνω, ο αναγνώστης παραπέμπεται στη δουλειά των *Papasaikas et al.*[2].

Κεφάλαιο 3

Υπολογιστικό Υπόβαθρο

Για την εφαρμογή μεθόδων μηχανικής μάθησης σε ένα σύνολο δεδομένων ακολουθούνται ορισμένα βήματα ώστε να προκύψει ένα αξιόπιστο αποτέλεσμα [22, 23].

- **Αναπαράσταση δεδομένων (Data representation):** Αρχικά, θα πρέπει να μελετήσουμε τα δεδομένα με τα οποία πρόκειται να δουλέψουμε. Αυτό επιτυγχάνεται με τον οπτικό έλεγχο των ιδίων ή κάποιων στατιστικών μεγεθών. Η οπτικοποίηση δεδομένων με γραφήματα είναι επίσης αρκετά βοηθητική όταν προσπαθούμε να αποκτήσουμε διαίσθηση σχετικά με τα δεδομένα μας.

- **Προ-επεξεργασία δεδομένων (Data pre-processing):** Σε αυτό το βήμα γίνονται όλες οι απαραίτητες ενέργειες για τη διόρθωση και την περαιτέρω διαμόρφωση των δεδομένων εισόδου. Για παράδειγμα, σε ορισμένες περιπτώσεις μπορεί κάποιες τιμές να λείπουν (missing values) από το σύνολό μας. Υπάρχουν διάφορες τεχνικές για την αντιμετώπιση αυτού του προβλήματος. Ένα άλλο σημαντικό ζήτημα είναι η επιλογή των χαρακτηριστικών (feature selection). Πολλές φορές χρειάζεται να μειώσουμε των αριθμό των χαρακτηριστικών για λόγους μείωσης υπολογιστικών πόρων. Θα πρέπει να ελέγχεται, επίσης, εαν τα χαρακτηριστικά παρουσιάζουν μεγάλη

συσχέτιση μεταξύ τους μιας και αυτό μπορεί να βλάψει την αποτελεσματικότητα του μοντέλου μας.

- **Εκπαίδευση του μοντέλου (Model training):** Έχοντας καταλήξει με τη μορφή του συνόλου δεδομένων θα πρέπει να το χρησιμοποιήσουμε για να εκπαιδεύσουμε να μοντέλο μηχανικής μάθησης. Επιλέγουμε έναν αλγόριθμο και εισάγουμε το μεγαλύτερο ποσοστό των δεδομένων, αφήνοντας το υπόλοιπο για τον έλεγχο, που είναι το επόμενο βήμα.
- **Έλεγχος και αξιολόγηση του μοντέλου (Model testing and evaluation):** Είναι αναγκαίο να δοκιμάσουμε πως λειτουργεί το μοντέλο που εκπαιδεύσαμε με καινούργια δεδομένα, δηλαδή αυτά που δεν χρησιμοποιήθηκαν στο προηγούμενο βήμα. Για το σκοπό αυτό μπορούν να χρησμοποιηθούν διάφορες μέθοδοι αξιολόγησης σε συνδυασμό με τα αποτελέσματα του μοντέλου. Θα πρέπει να σημειωθεί πως υπάρχει κίνδυνος το μοντέλο μας να προσαρμοστεί σε υπερβολικό βαθμό στις ιδιαιτερότητες των δεδομένων με τα οποία εκπαιδεύεται και να μην είναι κατάλληλο για γενική χρήση με καινούργια δεδομένα εισόδου. Το πρόβλημα αυτό ονομάζεται overfitting.

3.1 Μέθοδοι προσέγγισης επιλογής χαρακτηριστικών

Σε αυτό το κεφάλαιο γίνεται πιο αναλυτική αναφορά στην επιλογή των χαρακτηριστικών μέσα στα πλαίσια προ-επεξεργασίας ενός συνόλου δεδομένων. Μεγάλο κομμάτι αυτής της διπλωματικής ασχολείται με τη διερεύνηση γύρω από αυτό το ζήτημα.

Ένας, αρκετά συνηθισμένος, τρόπος επιλογής χαρακτηριστικών ονομάζεται filter method. Κατά την εκτέλεση της μεθόδου αυτής γίνεται στατιστικός έλεγχος της συσχέτισης των χαρακτηριστικών

με τη μεταβλητή πρόβλεψης p-value και επιλέγονται για την εκπαίδευση μόνο τα χαρακτηριστικά που ικανοποιούν κάποια ικανοποιητική οριακή τιμή.

Μία πολύ διαφορετική κατηγορία μεθόδων επιλογής χαρακτηριστικών είναι οι wrapper methods, όπου γίνεται εκπαίδευση με κάποιον αλγόριθμο μηχανικής μάθησης σε υποσύνολα χαρακτηριστικών και επανεκτιμάται σε κάθε βήμα πρόσθεσης ή αφαίρεσης χαρακτηριστικού η απόδοση του αλγορίθμου. Για τη διπλωματική αυτή επιλέγονται δύο προσεγγίσεις της παραπάνω κατηγορίας:

- **Forward selection**: Εφαρμόζεται αλγόριθμος μηχανικής μάθησης για κάθε ένα χαρακτηριστικό από το σύνολο των δεδομένων και επιλέγεται αρχικά αυτό με τη μεγαλύτερη απόδοση. Στη συνέχεια, προστίθενται ένα ένα όλα τα χαρακτηριστικό σε αυτό που είχε επιλεγεί αρχικά και επιλέγεται αυτό το οποίο έχει τη μεγαλύτερη απόδοση σε συνδυασμό με το χαρακτηριστικό που είχε επιλεγεί πρώτο. Η διαδικασία αυτή συνεχίζεται έως ότου να εξαντληθούν τα χαρακτηριστικά που επιλέγονται και προστίθενται σε αυτά με την υψηλότερη απόδοση.

- **Backward elimination**: Αρχικά εφαρμόζεται αλγόριθμος μηχανικής μάθησης για όλο το σύνολο των χαρακτηριστικών και εκτιμάται η απόδοσή του. Στη συνέχεια, αφαιρείται ένα χαρακτηριστικό και εκτιμάται η αλλαγή στην απόδοση. Η διαδικασία αυτή επαναλαμβάνεται για όλα τα χαρακτηριστικά έτσι ώστε τελικά να αφαιρεθεί αυτό χωρίς το οποίο το σύνολο των δεδομένων ήταν πιο αποτελεσματικό. Η διαδικασία αυτή συνεχίζεται έως ότου να εξαντληθούν όλα τα χαρακτηριστικά, αφού έχουν αφαιρεθεί σταδιακά με κριτήριο τη μέγιστη αύξηση στην απόδοση.

Στο πειραματικό μέρος παρουσιάζονται αποτελέσματα εφαρμογής και των δύο μεθόδων με διάφορους συνδυασμούς αλγορίθμων μηχανικής μάθησης. Κάθως η μία μέθοδος λειτουργεί με αντίστροφη

λογική σχετικά με την άλλη, επιχειρείται να βρεθεί ποιά από τις δύο θα είναι πιο αποτελεσματική για τα δεδομένα μας και αν παίζει ρόλο ο αλγόριθμος με τον οποίο συνδυάζεται.

3.2 Αλγόριθμοι Μηχανικής Μάθησης

Οι αλγόριθμοι επιβλεπόμενης μηχανικής μάθησης που έχουν επιλεγεί χωρίζονται σε αλγορίθμους παλινδρόμησης (regression) και ταξινόμησης (classification). Οι αλγόριθμοι παλινδρόμησης χτίζουν ένα μοντέλο που βασίζεται στα χαρακτηριστικά εισόδου, ώστε να προβλέψουν τη συνεχή μεταβλητή εξόδου. Στην περίπτωση των αλγορίθμων ταξινόμησης όμως, η ανεξάρτηση μεταβλητή εξόδου είναι διακριτή και κατηγοριοποιείται με βάση τα χαρακτηριστικά εισόδου.

Αφού δημιουργηθεί ένα μοντέλο, όπως έχει προαναφερθεί, θα πρέπει να εκτιμηθεί πόσο καλό είναι. Σε αυτή την εργασία, η αξιολόγηση για τους αλγορίθμους παλινδρόμησης γίνεται με τον υπολογισμό του συντελεστή προσδιορισμού R^2 και για τους αλγορίθμους ταξινόμησης με την τιμή της ακρίβειας (accuracy).

$$\boldsymbol{R^2 = 1 - \frac{\sum_i (y_i - y')^2}{\sum_i (y_i - \bar{y})^2}}$$

όπου y_i παρατηρούμενη τιμή, y' προβλεπόμενη τιμή και $\bar{y}$ η μέση τιμή των παρατηρήσεων

$$\boldsymbol{accuracy = \frac{TP+TN}{P+N}}$$

*όπου **TP** οι αληθώς θετικές παρατηρήσεις, **TN** οι αληθώς αρνητικές παρατηρήσεις, **P** όλες οι θετικές παρατηρήσεις και **N** όλες οι αρνητικές παρατηρήσεις*

3.2.1 Αλγόριθμοι Παλινδρόμησης

Σε αυτήν την υποενότητα παρουσιάζονται συνοπτικά κάποιοι από τους αλγορίθμους παλινδρόμησης, εφόσον έχουν επιλεγεί να χρησιμοποιηθούν στο πειραματικό μέρος. Τα αποτελέσματά τους παρουσιάζονται στο Κεφάλαιο 6.

Γραμμική Παλνδρόμηση(Linear Regression)

Στη γραμμική παλινδρόμηση το μοντέλο που χτίζεται, περιγράφει μία γραμμική σχέση μεταξύ των ανεξάρτητων μεταβλητών εισόδου ($\boldsymbol{x_{ij}}$) έτσι ώστε να παράξουν την τιμή της εξαρτημένης μεταβλητής εξόδου ($\boldsymbol{y_i}$).

Στην πραγματικότητα, δηλαδή, αναζητείται η ευθεία που που περιγράφει με βέλτιστο τρόπο τις τιμές της μεταβλητής εξόδου, ελαχιστοποιώντας την πιθανότητα λάθους:

$$\boldsymbol{y_i} = \boldsymbol{\alpha} + \boldsymbol{\beta_1 x_{i1}} + ... + \boldsymbol{\beta_N x_{iN}}, \text{ } \textit{όπου } \boldsymbol{N} \textit{ το σύνολο των χαρακτηριστικών}$$

Για να οριστούν οι παράμετροι $\boldsymbol{\alpha}$ και $\boldsymbol{\beta_{1..N}}$, εφαρμόζεται η μέθοδος των ελαχίστων τετραγώνων, η οποία βρίσκει το μικρότερο άθροισμα των τετραγώνων της διαφοράς μεταξύ των πραγματικών τιμών του $\boldsymbol{y}$ και των τιμών που προκύπτουν από τη μοντελοποίηση.

Παλινδρόμηση k-Πλησιέστερων Γειτόνων (k-Nearest Neighbors Regression)

Η μέθοδος των k-Πλησιέστερων Γειτόνων (k-NN) βασίζεται στις τιμές των k πλησιέστερων δειγμάτων του συνόλου εκπαίδευσης. Πιο συγκεκριμένα, εαν εφαρμοστεί παλινδρόμηση, η τιμή της εξαρτημένης μεταβλητής υπολογίζεται από τον μέσο όρο των k- γειτόνων της, όπου k ακέραιος αριθμός ορισμένος από τον χρήστη.

3.2.2 Αλγόριθμοι Ταξινόμησης

Παρομοίως, εδώ παρουσιάζονται κάποιοι από τους αλγορίθμους ταξινόμησης, εφόσον έχουν επιλεγεί να χρησιμοποιηθούν στο πειραματικό μέρος του Κεφαλαίου 6.

Λογιστική Παλινδρόμηση Logistic Regression

Παρά το όνομά της, η λογιστική παλινδρόμηση είναι αλγόριθμος ταξινόμησης, αφού η εξαρτημένη μεταβλητή $\boldsymbol{y_i}$ μπορεί να πάρει μόνο δύο τιμές (μηδέν ή ένα). Η πιθανότητα αυτή εκφράζεται μέσω τις σιγμοειδούς συνάρτησης $\boldsymbol{f(x)}$:

$$\boldsymbol{f(x)} = \frac{\mathbf{1}}{\mathbf{1}+e^{-\boldsymbol{\alpha x}}}, \text{ όπου } \boldsymbol{\alpha} \text{ η παράμετρος κλήσης}$$

Ταξινόμηση k-Πλησιέστερων Γειτόνων (k-Nearest Neighbors Classification)

Η ταξινόμηση στην περίπτωση του αλγορίθμου των k πλησιέστερων γειτόνων, βασίζεται στην ίδια λογική με την παλινδρόμησή του. Εδώ όμως, κάθε γείτονας ανήκει σε μία κλάση και έτσι το σημείο προς ταξινόμηση θα ανατεθεί στην κλάση της πλειοψηφίας των γειτόνων του.

Κεφάλαιο 4

Σχετικές Εργασίες

4.1 Μελέτες Χαρακτηριστικών

Διάφορες έρευνες έχουν πραγματοποιηθεί στην προσπάθεια ορισμού κάποιων κατευθυντήριων γραμμών για τον σχεδιασμό των ολιγονουκλεοτιδίων. Για να εντοπιστούν τα χαρακτηριστικά που συμβάλλουν στην επιτυχία του exon skipping, έχει γίνει η μελέτη τους για διαφορετικά ολιγονουκλεοτίδια και έχει εκτιμηθεί in vivo η αποδοτικότητά τους.

Οι *AartsmaRus et al.* στο [12], μετά από μελέτη, εντοπίζουν τέσσερις παραμέτρους, με βάση τις οποίες το 79% των 2OMePS ολιγονουκλεοτιδίων ταξινομούνται σωστά σε αποδοτικά και μη. Ο διαχωρισμός των σημαντικών παραμέτρων βασίστηκε στην έντονη διαφορά των τιμών τους στα ολιγονουκλεοτίδια με ποσοστό επιτυχίας άνω των 5% και σε αυτά με χαμηλότερη απόδοση και ο αλγόριθμος που χρησιμοποιήθηκε ήταν η ταξινόμηση με Principal Component Analysis. Τελικά, προτείνεται να σχεδιάζονται ολιγονουκλεοτίδια πλούσια σε εξαμερείς RESCUE-ESE αλληλουχίες, όπως και σημεία πρόσδεσης της SR πρωτεΐνης SC35. Αντιθέτως, τα σήμεία πρόσδεσης της SR πρωτεΐνης Tra2$\boldsymbol{\beta}$ ενθαρρύνεται να μη συμπεριλαμβάνονται, αφού έχει παρατηρηθεί η έντονη παρουσία τους στα μη αποδοτικά ολιγονουκλεοτίδια. Η τελευταία σημαντική παράμετρος είναι η υψηλή ενέργεια πρόσδεσης του ολιγονουκλεοτιδίου

στην αλληλουχία-στόχο. Θα πρέπει να σημειωθεί πως εξετάστηκαν συνολικά 54 παράμετροι και όλες αφορούσαν αλληλουχίες εντός των εξωνίων.

Σε επόμενη έρευνά τους οι *Aartsma-Rus et al.* [13] συγκρίνουν την αποτελεσματικότητα των ολιγονουκλεοτιδίων που στοχεύουν σε εξωνικές αλληλουχίες και αυτών που προορίζονται για τα σημεία ματίσματος (splice sites). Παρά το ίδιο μήκος τους, οι ολιγονουκλεοτιδικές αλληλουχίες, αυτών των δυο κατηγοριών, διαφέρουν στις υπόλοιπες θερμοδυναμικές παραμέτρους τους. Αυτά που στοχεύουν σε εξωνικές αλληλουχίες έχουν υψηλότερη ενέργεια πρόσδεσης, υψηλότερη θερμοκρασία τήξης, πιο πολλές αλληλουχίες γουανίνης-κυτοσίνης (GCs) και λιγότερες αδενίνης (As). Οι ευνοούμενες θερμοδυναμικές ιδιότητες των εξωνίων, σε σχέση με αυτές των εσωνίων, καθιστούν τις αλληλουχίες πρόσδεσης των ολιγονουκλεοτιδίων εσωτερικά των εξωνίων πιο κατάλληλες και αποτελεσματικές.

Οι *Dwi Pramono et al.* [16] παρουσίασαν 61% αποτελεσματικά ολιγονουκλεοτίδια (exon skipping σε >25% του συνόλου των mRNA) σχεδιάζοντάς τα με παρουσία των ESEs και λαμβάνοντας υπόψιν τους την εύκολη πρόσβαση στην αλληλουχία-στόχο κατά τη διάρκεια της μεταγραφής, καθώς και το μήκος της. Αναλύοντας τα ολιγονουκλεοτίδια μετέπειτα των πειραμάτων, διαπίστωσαν πως η αποτελεσματικότητα τους σχετίζεται άμεσα με τη χαμηλότερη αθροιστική τους θέση (ACP).

Δύο ακόμα μεγάλες έρευνες έχουν πραγματοποιηθεί με σκοπό να εξετάσουν την αποτελεσματικότητα του exon skipping με χρήση των 2OMePS ολιγονουκλεοτιδίων στο γονίδιο της δυστροφίνης. Οι *Wilton et al.* [14] σχεδίασαν ολιγονουκλεοτίδια για το κάθε ένα από τα εξώνια της δυστροφίνης, μιας και το εξώνιο που πρέπει να παραβλεφθεί εξαρτάται κάθε φορά από τη φύση και την τοποθεσία της μετάλλαξης. Στη συνέχεια τα εξώνια ταξινομήθηκαν σε 4 κατηγορίες, σύμφωνα με τα αποτελέσματα του exon skipping. Στις τρεις πρώτες κατηγορίες μπήκαν τα εξώνια με υψηλή (>30%), μεσαία (<30% και >10%) και χαμηλή (<10%) αποδοτικότητα, ενώ

στην τέταρτη μπήκαν όσα χρειαζόντουσαν πάνω από ένα ολιγονουκλεοτίδιο για να παραβλεφθούν ή/και μετακινούσαν μαζί τους και τα γειτονικά εξώνιά τους. Οι *Harding et al.* [15], με τη σειρά τους, έκαναν παρόμοια έρευνα για να αποδείξουν πως το μήκος των ολιγονουκλεοτιδίων είναι σημαντικός παράγοντας που σχετίζεται με την αποτελεσματικότητα του exon skipping. Υποστήριξαν πως όσο μεγαλύτερες αλληλουχίες έχουμε, τόσο μεγαλύτερα τα ποσοστά επιτυχίας. Αυτό έρχεται σε συμφωνία με τις παραπάνω έρευνες, αν λάβουμε υπόψιν πως μεγαλύτερο μήκος σημαίνει περισσότερες αλληλουχίες γουανίνης-κυτοσίνης (GCs) και μεγαλύτερη ενάργεια πρόσδεσης στην αλληλουχία-στόχο.

Παρόμοιες μελέτες έχουν γίνει και για τα PMO ολιγονουκλεοτίδια από τους *Popplewell et al.* [17] Έπειτα από σχεδιασμό των ολιγονουκλεοτιδικών αλληλουχιών (μήκους 25 και 30 βάσεων) για ορισμένα εξώνια, μελετήθηκαν οι ιδιότητες των εξωνίων αυτών και παραγματοποιήθηκε *in vivo* έλεγχος αποδοτικότητας. Στη συνέχεια, έγινε *in silico* άναλυση των PMOs για να εντοπιστούν τα πιο σημαντικά χαρακτηριστικά αυτών που ήταν αποτελεσματικά στο exon skipping. Η στατιστική ανάλυση έδειξε πως ο πιο σημαντικός παράγοντας είναι η ενέργεια πρόσδεσης του ολιγονουκλεοτιδίου στο εξώνιο, μιας και αυτός ο σταθερός δεσμός φαίνεται να προκαλεί αλλαγές στη δευτεροταγή δομή του πρόδρομου mRNA και να αποτρέπει τις SR πρωτεΐνες από το μάτισμα. Άλλος ένας παράγοντας, σχετικός με τη δευτεροταγή δομή του πρόδρομου mRNA είναι τα 'ανοιχτά' σημεία του, δηλαδή αυτά στα οποία έχουν εύκολη πρόσβαση οι SR πρωτεΐνες, και επομένως αν καλυφθούν από τα ολιγονουκλεοτίδια θα αποτραπεί και πάλι το μάτισμα. Όπως και σε προηγούμενες μελέτες, το μήκος των ολιγονουκλεοτιδίων και τα σημεία πρόσδεσης των SR πρωτεϊνών (SF2/ASF, SC35) έδειξαν να επηρεάζουν σημαντικά τη διαδικασία.

Σε μια προσπάθεια γενίκευσης των συμπερασμάτων από τις παραπάνω έρευνες και αποβολής των στοιχείων που προκύπτουν λόγω μοναδικών ιδιοτήτων της κάθε ερευνητικής διαδικασίας, οι *Echigoya*

et al. ενοποίησαν τα χαρακτηριστικά τους και τα εμπλούτισαν με επιπλέον παραλλαγές που μπορεί να προσθέτουν πληροφορία για την αποτελεσματικότητα του exon skipping.

Για παράδειγμα, δημιουργήθηκαν επιπλέον μετρήσεις για την ενέργεια πρόσδεσης του ολιγονουκλεοτιδίου στην αλληλουχία στόχο. Αντί για την περιοχή ολόκληρου του εξωνίου, λαμβάνεται υπόψην η ευρύτερη περιοχή αλληλουχίας στο κέντρο της οποίας βρίσκεται ο στόχος πρόσδεσης, ακόμα και αν χρειάζεται να συμπεριληφθούν κομμάτια εσωνίων. Όσο για την αναδίπλωση του πρόδρομου mRNA, η πρόβλεψη γίνεται για αλληλουχίες μικρότερου μήκους, ώστε να επιτευχθεί μεγαλύτερη ακρίβεια.

Επιπλέον, προστέθηκαν καινούργια χαρακτηριστικά:

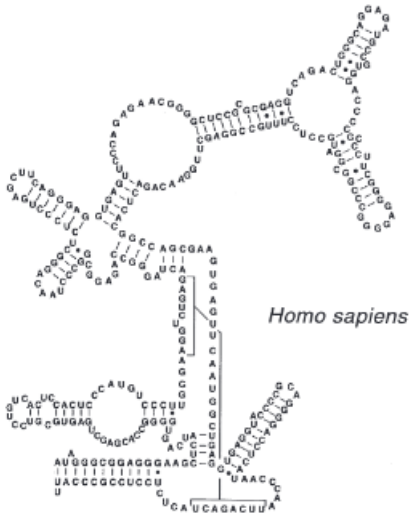
- Οι μετρήσεις του neighborhood inference (NI) [19] που υποδηλώνουν την τάση πρόσδεσης παραγόντων ματίσματος στην εκάστοτε ακολουθία

- Η κατηγοριοποίηση των εξωνίων της δυστροφίνης σε σχέση με τα χαρακτηριστικά ματίσματος που διαθέτουν, όπως πραγματοποιήθηκε από τους *Malueka et al.* [20]

- Η διαφορά στο περιεχόμενο των GC βάσεων μεταξύ των εξωνίων και των γειτονικών τους εσωνίων, μιας και φαίνεται να είναι σημαντικός παράγοντας για τον ορισμό των εξωνίων

- Η μέση τιμή ποσοστού προβλεπόμενων δομών για τις οποίες κάθε στοχευμένη βάση είναι ξεδιπλωμένη (L1) και τη μέση συμμετοχή της κάθε βάσης κατά τη διάρκεια της εικονικής μεταγραφής (L3) σύμφωνα με τους [21]

Τα μόνα χαρακτηριστικά που τελικά φάνηκε να επηρεάζουν σημαντικά την αποδοτικότητα του exon skipping και στα δύο σύνολα δεδομένων (PMO & 2OMePS) είναι η ενέργεια πρόσδεσης του ολιγονουκλεοτιδίου στο RNA και η απόσταση (σε βάσεις) της αλληλουχίας στόχου από τη θέση αποδέκτη (3′ ss).



Σχήμα 4.1: Δευτεροταγής δομή ριβονουκλεϊκου οξέος στον ανθρώπινο οργανισμό. Οι Watson-Crick δεσμοί αναπαριστώνται με ευθείες γραμμές, ενώ τα τμήματα μη συμπληρωματικών βάσεων σχηματίζουν βρόγχους. Το σχήμα προέρχεται από τους *Pace et al.*[5]

4.2 Εργαλεία Εξαγωγής Χαρακτηριστικών

Τα χαρακτηριστικά που έχουν μελετηθεί σε έρευνες βελτιστοποίησης της θεραπείας exon skipping μπορούν να ανακτηθούν με χρήση πολλών υπολογιστικών εργαλείων που αφορούν τη δευτεροταγή δομή ριβονουκλεϊκού οξέος, τις θερμοδυναμικές ιδιότητές του αλλά και εντοπιστμό ορισμένων ακολουθιών ειδικής σημασίας. Στη συνέχεια παρουσιάζονται όλα τα εργαλεία.

4.2.1 Πρόβλεψη και ανάλυση δευτεροταγούς δομής RNA

Η δευτεροταγής δομή μίας αλληλουχίας νουκλεϊκού ή ριβονουκλεϊκού οξέος προκύπτει από τις αναδιπλώσεις που κάνει στο χώρο όταν σχηματίζονται Watson-Crick και GU ζεύγη βάσεων, ενώ παρεμβάλλονται αλληλουχίες που δημιουργούν βρόγχους, όπως φαίνεται στο Σχήμα 4.1.

Το RNAstructure [6] είναι λογισμικό πρόβλεψης της δευτεροταγούς δομής του RNA, μονής ή διπλής έλικας, που βασίζεται σε κανόνες θερμοδυναμικής. Μπορεί να χρησιμοποιηθεί, επίσης, με αρκετά μεγάλη ακρίβεια για πρόβλεψη της δομής των ολιγονουκλεοτιδίων. Ο υπολογισμός χρησιμοποιεί διάφορες εκδοχές αλγορίθμων που βασίζονται στην ποσοτικοποίηση της αλλαγής της ελεύθερης ενέργειας [7] κατά την αναδίπλωση της αλληλουχίας βάσεων υπο εξέταση.

Άλλο ένα εργαλείο, με πολλές υπολογιστικές δυνατότητες, είναι το πακέτο λογισμικού ViennaRNA [8]. Ένας από τους αλγορίθμους που εφαρμόζει (cofoldRNA) έχει χρησιμοποιηθεί για την πρόβλεψη της ενέργειας πρόσδεσης του ολιγονουκλεοτιδίου στην αλληλουχία στόχο και της ελεύθερης ενέργειας αλληλουχίας μονής έλικας. Όπως και ο RNAstructure, ετσι και ο cofoldRNA βασίζεται στις ενεργειακές παραμέτρους τύπου Turner [7].

4.2.2 Εξωνικές Θέσεις Ματίσματος

Για τον εντοπισμό των εξωνικών αλληλουχιών που συμβάλλουν στον μηχανισμό του ματίσματος (ESEs), εφαρμόζονται τρία υπολογιστικά εργαλεία: RESCUE-ESE [9], ESEfinder [10] και Human Splicing Finder [11].

Για τη δημιουργία του ESEfinder, πραγματοποιήθηκαν *in vivo* πειράματα με τυχαίες ολιγονουκλεοτιδικές αλληλουχίες από τις οποίες, τελικά, επιλέχθηκαν αυτές που αντιστοιχούν σε ESE μοτίβα μίας υποομάδας SR πρωτεϊνών (SF2/ASF, SC35, SRp40, SRp55) και ταξινομήθηκαν σε πίνακες που αναπαριστούν τη συχνότητα εμφάνισής τους (Σχήμα 4.2) Με το λογισμικό του ESEfinder μπορούν να εντοπιστούν τα πιθανά μοτίβα των ESEs (μήκους 6-8 βάσεων νουκλεϊκού οξέος) σε μία αλληλουχία βάσεων εισόδου και να δωθούν οι τιμές της συχνότητάς τους. Θα πρέπει να σημειωθεί όμως, πως δεν υπάρχει απόλυτη συσχέτιση της υψηλής συχνότητας των αλλη-

Σχήμα 4.2: ESE μοτίβα που αντιστοιχούν σε ορισμένες SR πρωτεΐνες, σύμφωνα με το ESEfinder [10]. Το μέγεθος των γραμμάτων αντιστοιχεί στη συχνότητα που έχει παρατηρηθεί για κάθε βάση στη συγκεκριμένη θέση. Τα πορτοκαλί γράμματα αναπαριστούν τις βάσεις που ξεπέρασαν το κατώφλι που δίνεται για κάθε πρωτεΐνη.

λουχιών με την παρουσία των ESEs (ή το αντίστροφο) και έτσι η μέθοδος αυτή δεν χαρακτηρίζεται πάντα από μεγάλη ακρίβεια.

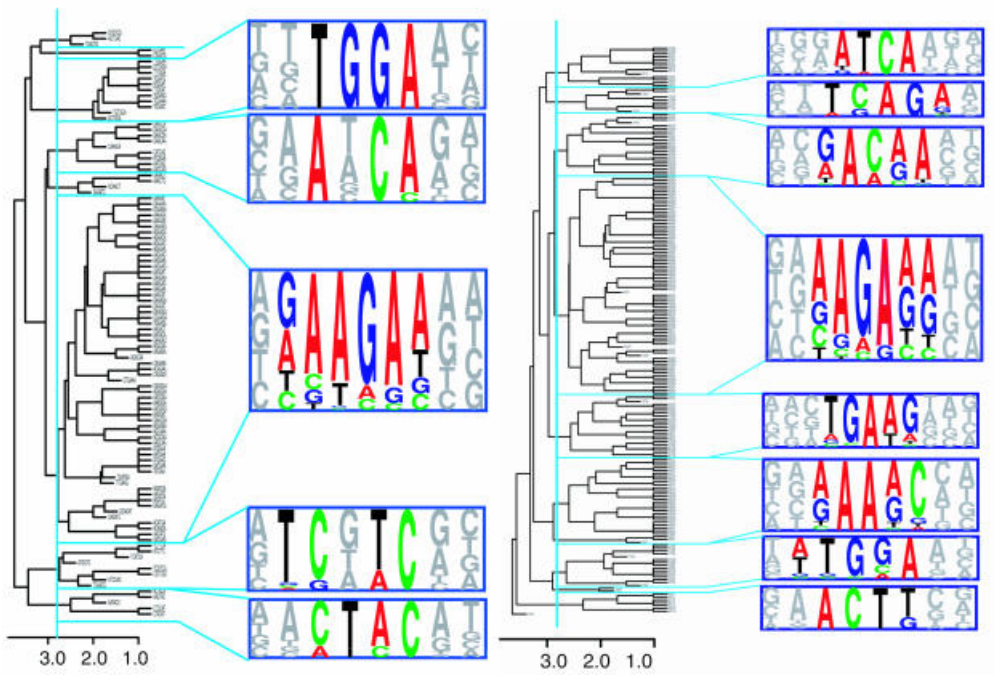
Μία πολύ διαφορετική προσέγγιση έχει ληφθεί για τον εντοπισμό τωνESEs από το λογισμικό του RESCUE-ESE. Οι αλληλουχίες που επιλέχθηκαν ως υποψήφιες έχουν μήκος έξι βάσεων (εξαμερείς) και ικανοποιούν δύο κριτήρια:

1. Η συχνότητα εμφάνισης τους στα εξώνια είναι σημαντικά υψηλότερη από ότι στα εσώνια

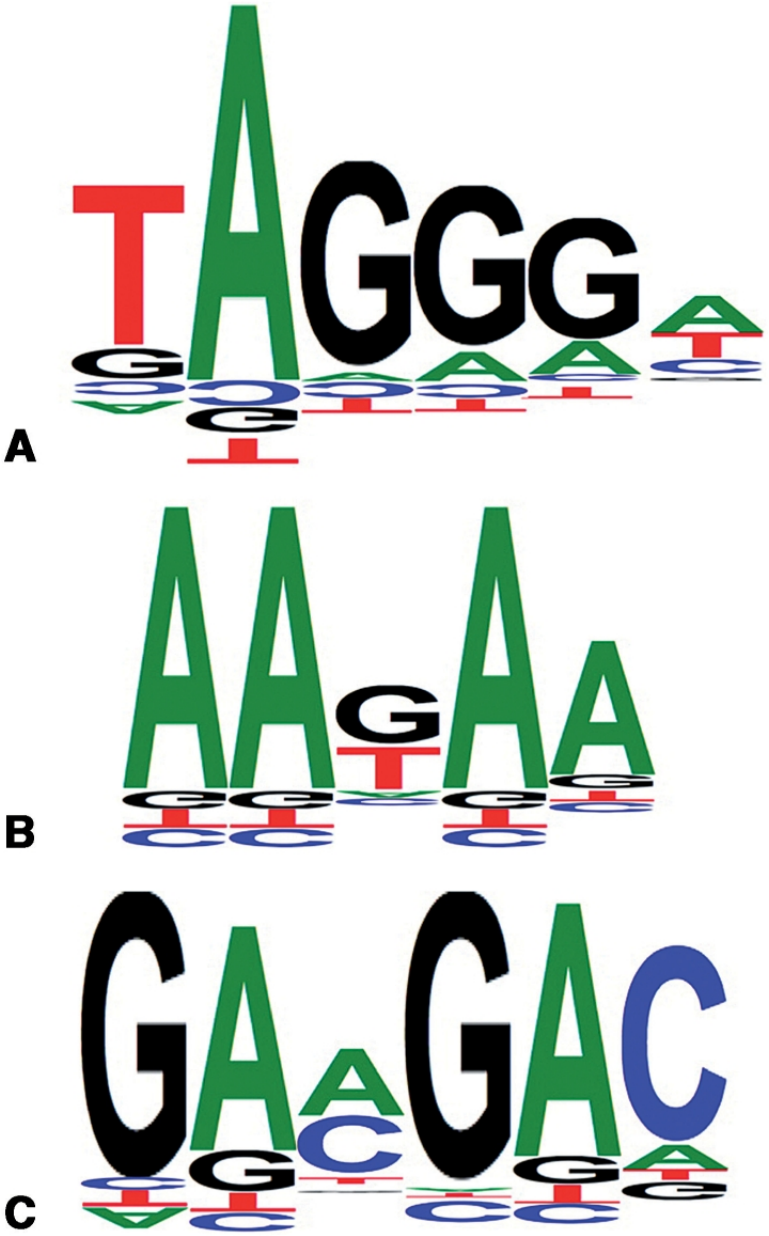
2. Συναντιούνται πολύ πιο συχνά στα εξώνια με ασθενείς θέσεις ματίσματος από οτι στα εξώνια με ισχυρές θέσεις ματίσματος

Οι εξαμερείς ακολουθίες που επιλέχθηκαν (Σχήμα 4.3), κατηγοριοποιήθηκαν σε δέκα ομάδες. Δύο ομάδες αφορούσαν τη 5′ θέση ματίσματος, άλλες πέντε την 3′ θέση ματίσματος και οι υπόλοιπες τρείς αφορούσαν και τις δύο θέσεις ματίσματος. Στη συνέχεια έγιναν *in vivo* πειράματα και εντοπίστηκε επιθυμητή δραστηριότητα με αλληλουχίες όλων των κατηγοριών. Επιπρόσθετα, εισήχθησαν σημειακές μεταλλάξεις και παρατηρήθηκε μεταβολή στη διαδικασία του ματίσματος στη συντριπτική πλειοψηφία των πειραμάτων.

Τέλος, έχουμε το Human Splicing Finder (HSF), ένα πιο γενικό εργαλείο που μπορεί να αναγνωρίσει όλες τις αλληλουχίες που συμβάλλουν στο μάτισμα: τις κανονικές θέσεις ματίσματος (5'ss, branch point, 3'ss), τους ενισχυτές και καταστολείς ματίσματος των εξωνίων (ESEs, ESSs) και τους ενισχυτές και καταστολείς ματίσματος των εσωνίων (ISEs, ISSs). Ο HSF χρησιμοποιεί τους αλγορίθμους των ESEfinder και RESCUE-ESE μεταξύ άλλων και επιχειρεί όχι μόνο την αναγνώριση των παραπάνω αλληλουχιών, αλλά και την παρουσίαση της επίδρασης των μεταλλάξεων πάνω τους. Πέρα από τη συγκέντρωση των θέσεων ματίσματος από πολλούς αλγορίθμους σε ένα κοινό υπολογιστικό εργαλείο, έχει γίνει επίσης αναγνώριση καινούργιων μοτίβων για τις αλληλουχίες ενισχυτών ματίσματος που αντιστοιχούν στις SR πρωτεΐνες 9G8, Tra2$\boldsymbol{\beta}$ και για



Σχήμα 4.3: ESE μοτίβα σύμφωνα με το RESCUE-ESE [9]. Όλες οι υποποψήφιες αλληλουχίες, τελικά, τεαξινομούνται σε 5 μοτίβα της 5′ θέσης ματίσματος και 8 μοτίβα της 3′ θέσης ματίσματος. Το μέγεθος των γραμμάτων αντιστοιχεί στη συχνότητα που έχει παρατηρηθεί για κάθε βάση στη συγκεκριμένη θέση.



(α′) Μοτίβα αλληλουχιών για τις πρωτεΐνες (A) hnRNP A1, (B) Tra2b και (C) 9G8

(β′) Μοτίβο αλληλουχιών για τη θέση διακλάδωσης (splicing branch point)

Σχήμα 4.4: Human Splicing Finder μοτίβα για κανονικές και βοηθητικές θέσεις ματίσματος. Τα ανάποδα γράμματα υποδεικνύουν τις βάσεις που δεν πέρασαν το κατώφλι που έχει οριστεί από τους συγγραφείς.

την αλληλουχία καταστολέα ματίσματος στην οποία προσδένεται η ριβονουκλεοπρωτεΐνη hnRNPA1 (Σχήμα 4.4).

Κεφάλαιο 5

Μεθοδολογία

Σε αυτή τη διπλωματική επιχειρείται ο εντοπισμός των πιο σημαντικών χαρακτηριστικών που συμβάλλουν στο exon skipping με εφαρμογή διάφορων προσεγγίσεων σε σχέση με την επιλογή τους, την ταξινόμησή τους και την εκτίμηση της συμβολής τους στην αποδοτικότητα.

5.1 Εργαλεία υλοποίησης

Ο κώδικας της εργασίας έχει γραφτεί σε Python και είναι βασισμένος στη βιβλιοθήκη μηχανικής μάθησης ελεύθερου λογισμικού scikit-learn. Άλλες σημαντικές βιβλιοθήκες που χησιμοποιήθηκαν είναι οι βιβλιοθήκες ανοιχτού λογισμικού NumPy και pandas. Με τη χρήση της βιβλιοθήκης NumPy επιτυγχάνεται η εκτέλεση πολύπλοκων μαθηματικών πράξεων σε σύνολα δεδομένων τα οποία έχουν την μορφή πινάκων. Ο χειρισμός, η προσπέλαση, καθώς και η αποθήκευση των δεδομένων αυτών επιτελείται με τη χρήση της βιβλιοθήκης pandas, η οποία προσφέρει ενέργειες για τη διαχείριση μεγάλων συνόλων δομών δεδομένων. Για την παράλληλη εκτέλεση του κώδικα σε περιπτώσεις πολύ μεγάλου όγκου δεδομένων, έγινε χρήση του πακέτου multiprocessing της Python.

5.2 Ταξινόμηση χαρακτηριστικών

Για κάθε πείραμα που πραγματοποιείται, χρησιμοποιούνται δύο wrapper μέθοδοι επιλογής χαρακτηριστικών, forward selection και backward elimination. Σε κάθε στάδιο πρόσθεσης ή αφαίρεσης χαρακτηριστικού εφαρμόζεται η μέθοδος του k-fold cross validation:

- Το σύνολο δεδομένων χωρίζεται σε ορισμένο αριθμό υποσυνόλων (k), από τα οποία το ένα επιτελεί το σύνολο ελέγχου (test set) και τα υπόλοιπα σχηματίζουν το σύνολο εκπαίδευσης (training set)
- Ένας αλγόριθμος μηχανικής μάθησης εφαρμόζεται στο training set για τη δημιουργία ενός μοντέλου
- Γίνεται πρόβλεψη για το test set με χρήση του εκπαιδευμένου μοντέλου και το αποτέλεσμα αποθηκεύεται σε λίστα
- Η διαδικασία επαναλαμβάνεται έως ότου κάθε υποσύνολο να έχει χρησιμοποιηθεί ως test set

Έτσι, για να αξιολογηθεί η συνεισφορά ενός χαρακτηριστικού σε κάθε στάδιο της εκάστοτε wrapper μεθόδου, γίνεται εκτίμηση των προβλέψεων που έχουν αποθηκευτεί σε κάθε βήμα της παραπάνω μεθόδου σε σχέση με τις πραγματικές τιμές της εξαρτημένης μεταβλητής του συνόλου δεδομένων, δηλαδή της αποτελεσματικότητας του exon skipping. Ο τρόπος εκτίμησης εξαρτάται από τον αλγόριθμο μηχανικής μάθησης που έχει επιλεγεί. Σε αυτή τη διπλωματική, η μέθοδος αξιολόγησης για τους αλγορίθμους παλινδρόμησης είναι το μέτρο $\boldsymbol{R^2}$, καθως για τους αλγορίθμους ταξινόμησης χρησιμοποιείται η ακρίβεια (accuracy).

Με τον παραπάνω τρόπο, το πρόγραμμα μας δίνει στην έξοδό του τα χαρακτηριστικά ταξινομημένα σύμφωνα με τη συνεισφορά τους, καθώς και τις τιμές εκτίμησης αποτελεσματικότητας σε κάθε στάδιο.

5.3 Επιλογή χαρακτηριστικών

Κάποια από τα χαρακτηριστικά του συνόλου δεδομένων ανήκουν σε ίδιες υποκατηγορίες. Για παράδειγμα, υπάρχουν πολλαπλές μετρήσεις της ενέργειας πρόσδεσης του ολιγονουκλεοτιδίου στο εξώνιο, αφού διερευνάται ποια ακριβώς περιοχή γύρω από την αλληλουχία στόχο είναι σημαντική σε σχέση με αυτό το χαρακτηριστικό και οι τιμές εξάγωνται από δύο υπολογιστικά εργαλεία.

Μία προσέγγιση είναι να χρησιμοποιηθούν όλα τα χαρακτηριστικά του συνόλου δεδομένων και να ταξινομηθούν με τρόπο που περιγράφεται πιο πάνω. Ενώ, εναλλακτικά, δοκιμάζεται να επιλεγεί ένα χαρακτηριστικό από κάθε υποκατηγορία, μιας και θα έχουν μεγάλη συσχέτιση μεταξύ τους, και να ταξινομηθούν τα χαρακτηριστικά που ανήκουν σε διαφορετικές υποκατηγορίες.

Για την επιλογή ενός χαρακτηριστικού από κάθε υποκατηγορία, συνήθως εφαρμόζεται κάποιο μέτρο που το καθιστά βέλτιστο σε σχέση με τα υπόλοιπα. Για παράδειγμα, στο [18] υπολογίζεται το p-value κάθε χαρακτηριστικού μίας υποκατηγορίας σε σχέση με την αποδοτικότητα του exon skipping και επιλέγεται αυτό με τη μικρότερη τιμή. Σε αυτή τη διπλωματική, εφαρμόστηκε διαφορετική προσέγγιση:

- Σχηματισμός όλων των δυνατών συνδυασμών με ένα χαρακτηριστικό από κάθε υποκατηγορία
- Ταξινόμηση των χαρακτηριστικών σε κάθε ομάδα που έχει προκύψει από τους συνδυασμούς
- Επιλογή της ομάδας με τη μέγιστη τιμή εκτίμησης αποδοτικότητας ($\boldsymbol{R^2}$ ή **accuracy**) για την εξαγωγή του τελικού συμπεράσματος

Κεφάλαιο 6

Αποτελέσματα Πειραμάτων

Όλα τα δεδομένα που χρησιμοποιούνται σε αυτήν την εργασία προέρχονται από τη δουλειά των *Echigoya et al.* [18], η οποία έχει περιγραφεί στο προηγούμενο κεφάλαιο. Έχουμε εμπνευστεί, επίσης, από τη μεθοδολογία τους, στην οποία γίνεται επιλογή χαρακτηριστικών από κάθε υποκατηγορία με p-value (filter method) και στη συνέχεια χτίζεται ένα μοντέλο με Linear Regression για την αξιολόγηση της αποδοτικότητας του exon skipping.

Αποφασίσαμε να χρησιμοποιήσουμε διαφορετική μέθοδο επιλογής χαρακτηριστικών, αλλά και να δοκιμάσουμε όλα τα χαρακτηριστικά ταυτόχρονα. Όσο για το μοντέλο εκπαίδευσης και αξιολόγησης δοκιμάσαμε διάφορους αλγορίθμους παλινδρόμησης και ταξινόμησης που παρουσιάζονται παρακάτω. Η μόνη μέθοδος που δεν έδωσε ικανοποιητικά αποτελέσματα και δεν παρουσιάζεται παρακάτω, είναι η παλινδρόμηση με SVM (Support Vector Machine). Μία εις βάθους ενασχόληση με αυτόν τον αλγόριθμο ίσως να αποβεί πιο αποτελεσματική, αλλά δεν πραγμοτοποιείται στην παρούσα εργασία.

6.1 PMO σύνολο δεδομένων

Για το σύνολο δεδομένων που αφορά τα PMO ολιγονουκλεοτίδια εφαρμόζονται αλγόριθμοι παλινδόμησης και ταξινόμησης με τιμές της

αποδοτικότητας του exon skipping σε μορφή συνεχούς και διακριτής μεταβλητής αντίστοιχα. Το μέγεθος του συνόλου αυτού δεν είναι πολύ μεγάλο καθώς έχει 66 δείγματα ολιγονουκλεοτιδίων. Οι στήλες με τα χαρακτηριστικά είναι 56 και παρουσιάζονται στο Πίνακα 6.1.

	Χαρακτηριστικά
1	Targeted exon
2	Distance from acceptor (position of last base relative to acceptor)
3	ACP
4	Distance from Donor (position of 1st base relative to donor)
5	Length
6	Exon Length
7	Length of exon when blocked by oligo
8	Exon Malueka Category
9	niscore
10	niscore_per_base
11	ACC_FL
12	ACC_LAST8
13	ACC_LAST15
14	ACC_BEST8
15	ACC_AVE
16	%GC oligo
17	GCs (number of)
18	%GC 5' intron 200 bases upstream
19	%GC fold increase target over 5' intron
20	%GC increase target over 5' intron
21	Decrease in GC% due to blocking by oligo
22	Decrease in ratio of exon to intron %GC due to oligo blocking
23	%GC exon

24	Total GCs in exon
25	Total GCs in target
26	# exon GCs blocked by oligo
27	%GC of exon when blocked by oligo
28	Exon v intron %GC
29	Exon v intron %GC after blocking by oligo
30	dG (TargetAsExon, RNAstructure)
31	dG (50BaseFlanksAroundTarget, RNA structure)
32	dG (100BaseFlanks, RNAstructure)
33	dG (200BaseFlanks, RNAstructure)
34	dG (ExonStartTo10BaseDownFromOligo, RNAstructure)
35	RNAeval targetoligo
36	dG TargetAsExon oligotarget (RNAcofold)
37	dG 50bp flanks oligotarget (RNAcofold)
38	dG 100bp flanks oligotarget (RNAcofold)
39	dG 150bp flanks oligotarget (RNAcofold)
40	Free E oligo dimer (RNAcofold)
41	Free E target monomer (RNAcofold)
42	Free E oligo monomer (RNAcofold)
43	% open (Mfold)
44	Ends in open loops
45	%overlap with hybrid peak
46	#rescue ESE sites
47	% overlap with rescue ESE site
48	%overlap with PESE
49	% overlap with PESS
50	SF2/ASF ESEfinder value over threshold
51	BRCA1 ESEfinder value over threshold
52	SC35 ESEfinder value over threshold

53	SRp40 ESEfinder value over threshold
54	SRp55 ESEfinder value over threshold
55	Tra2B
56	9G8

Πίνακας 6.1: Χαρακτηριστικά του συνόλου δεδομένων για τα PMO ολιγονουκλεοτίδια. Η ευθεία γραμμή χωρίζει τα χαρακτηριστικά σε υπο-ομάδες.

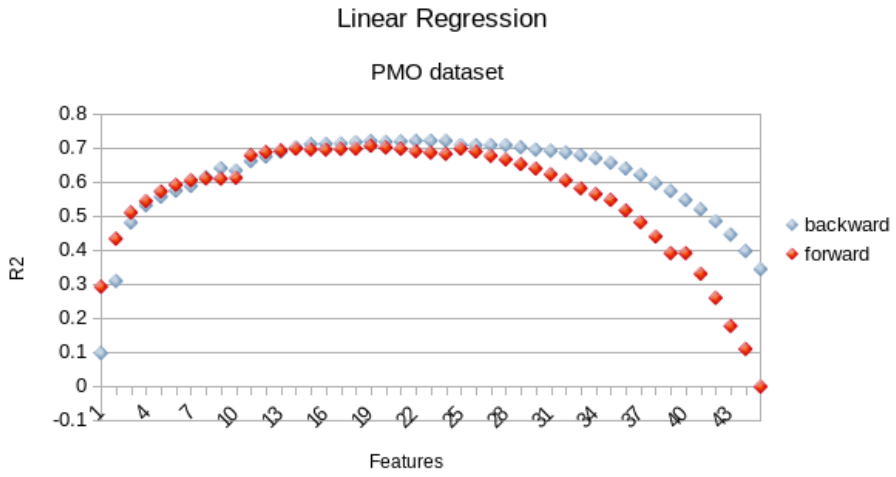
6.1.1 Linear Regression

Linear Regression Features	R^2
dG (ExonStartTo10BaseDownFromOligo, RNAstructure)	0.290032
%GC exon	0.430665
dG (TargetAsExon, RNAstructure)	0.507843
Length	0.541393
Distance from Donor (position of 1st base relative to donor)	0.569763
9G8	0.589094
ACC_FL	0.602376

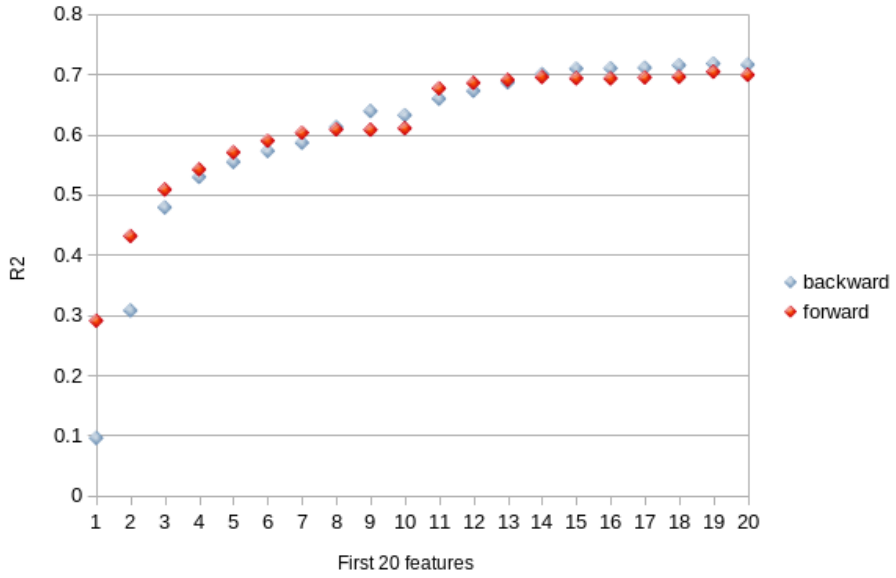
Πίνακας 6.2: Forward selection με Linear Regression για όλα τα χαρακτηριστικά του PMO συνόλου δεδομένων

Linear Regression Features	R^2
Total GCs in target	0.095226
%GC oligo	0.306866
%GC increase target over 5' intron	0.47827
dG TargetAsExon oligo::target (RNAcofold)	0.528919
SF2/ASF ESEfinder value over threshold	0.553865
Length	0.571985
ACC_AVE	0.585668
'Exon v intron %GC	0.611655

Πίνακας 6.3: Backward elimination με Linear Regression για όλα τα χαρακτηριστικά του PMO συνόλου δεδομένων



(α′) Γράφημα για το σύνολο των χαρακτηριστικών



(β′) Γράφημα για τα πρώτα είκοσι χαρακτηριστικά

Σχήμα 6.1: Ταξινόμηση όλων των χαρακτηριστκών του PMO συνόλου δεδομένων με Linear Regression

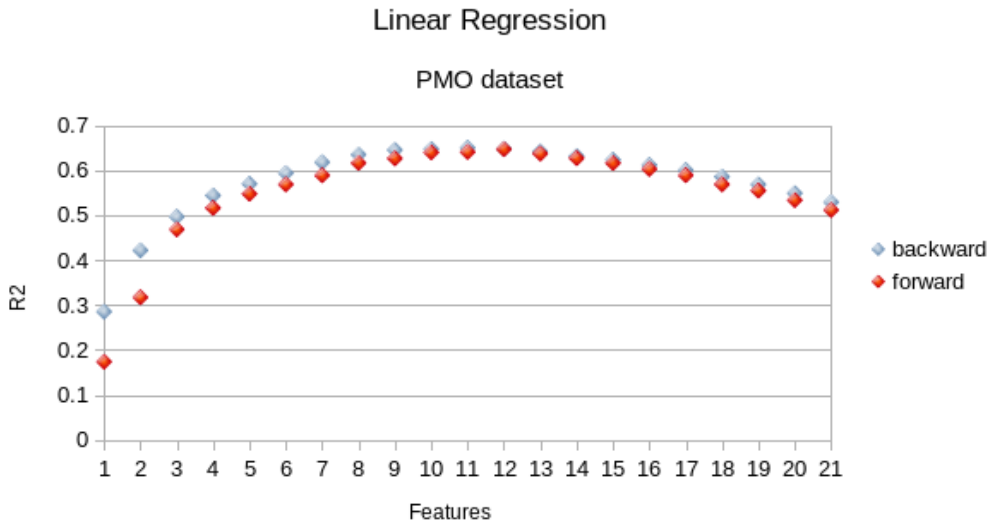


Figure 6.2: Ταξινόμηση συνδυασμού χαρακτηριστκών του PMO συνόλου δεδομένων με Linear Regression

Linear Regression Features	R^2
dG (200BaseFlanks RNAstructure)	0.173818
%GC of exon when blocked by oligo	0.316967
ACP	0.467998
% overlap with PESS	0.515761

Πίνακας 6.4: Forward selection με Linear Regression για συνδυασμό χαρακτηριστικών από υποκατηγορίες του PMO συνόλου δεδομένων

Linear Regression Features	R^2
Length	0.285156
Decrease in GC% due to blocking by oligo	0.421471
RNAeval target-oligo	0.496118
ACP	0.543538

Πίνακας 6.5: Backward elimination με Linear Regression για συνδυασμό χαρακτηριστικών από υποκατηγορίες του PMO συνόλου δεδομένων

Η μέθοδος ταξινόμησης χαρακτηριστικών με χρήση όλου του συνόλου δεδομένων και του αλγόριθμου γραμμικής παλινδρόμησης μας

δίνει μέγιστη τιμή του $\boldsymbol{R^2}$ 0.70 και 0.71 με εφαρμογή forward selection και backward elimination αντίστοιχα (Σχήμα 6.1). Δηλαδή, η αποδοτικότητα είναι σχεδόν πανομοιότυπη. Τα χαρακτηριστικά που καταλαμβάνουν θέσεις κορυφής στην ταξινόμηση και έχουν και τη μεγαλύτερη επίδραση στην αποδοτικότητα είναι διαφορετικά, ανήκουν όμως σε δύο ίδιες υποκατηγορίες: ενέργεια πρόσδεσης dG ολιγονουκλεοτιδίου και αριθμός GC βάσεων(Πίνακες 6.2 και 6.3).

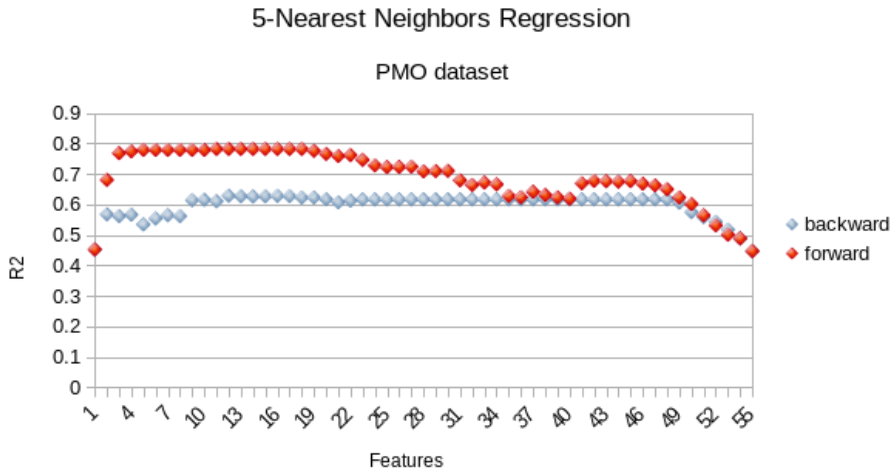
Όταν εφαρμόζεται η μέθοδος που χωρίζει τα χαρακτηριστικά σε υποκατηγορίες, δεν επιτυγχάνεται τόσο υψηλή αποδοτικότητα μιας και η μέγιστη τιμή του $\boldsymbol{R^2}$ είναι το 0.64 (Σχήμα 6.2). Βλέπουμε και πάλι στην πρώτη τριάδα των ταξηνομημένων χαρακτηριστικών την ενέργεια πρόσδεσης dG του ολιγονουκλεοτιδίου και το ποσοστό των GC βάσεων (Πίνακες 6.4 και 6.5).

Παρατηρήσεις:

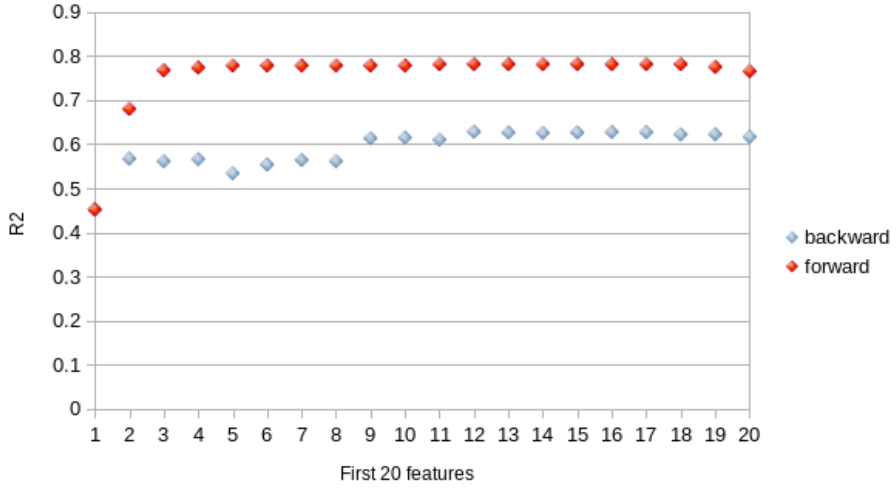
- Η ταξινόμηση όλων των χαρακτηριστικών εκτιμάται πιο αποτελεσματική από αυτήν με επιλεγμένα χαρακτηριστικά από κάθε υποκατηγορία.
- Τα αποτελέσματα από τις δύο προσεγγίσεις ταξινόμησης, forward selection και backward elimination, είναι παρόμοια σε όλες τις περιπτώσεις εφαρμογής.
- Βέλτιστα χαρακτηριστικά: *dG (ExonStartTo10BaseDownFromOligo, RNAstructure)*, *%GC oligo*, *%GC exon*, *dG (TargetAsExon, RNAstructure)* και *Length*

6.1.2 k-Nearest Neighbors Regression

Άλλη μία μεθοδος παλινδρόμησης που εφαρμόζεται είναι ο αλγόριθμος των k-Nearest Neighbors, όπου έχουν επιλεγεί 5 πλησιέστεροι γείτονες (k=5). Έγιναν δοκιμές με μεγαλύτερο και μικρότερο αριθμό γειτόνων για να οδηγηθούμε στο συμπέρασμα πως αυτός



(α΄) Γράφημα για το σύνολο των χαρακτηριστικών



(β΄) Γράφημα για τα πρώτα είκοσι χαρακτηριστικά

Σχήμα 6.3: Ταξινόμηση όλων των χαρακτηριστκών του PMO συνόλου δεδομένων με k-Nearest Neighbors Regression.

5-NN Regression Features	$\boldsymbol{R^2}$
dG (TargetAsExon, RNAstructure)	0.451757
%GC of exon when blocked by oligo	0.680028
dG (ExonStartTo10BaseDownFromOligo, RNAstructure)	0.767498
Length	0.773552

Πίνακας 6.6: Forward selection με 5-NN Regression για όλα τα χαρακτηριστικά του PMO συνόλου δεδομένων

5-NN Regression Features	$\boldsymbol{R^2}$
dG (TargetAsExon, RNAstructure)	0.451757
dG (ExonStartTo10BaseDownFromOligo, RNAstructure)	0.567132
dG TargetAsExon oligo::target (RNAcofold)	0.561213
Length of exon when blocked by oligo	0.565900

Πίνακας 6.7: Backward elimination με 5-NN Regression για όλα τα χαρακτηριστικά του PMO συνόλου δεδομένων

5-NN Regression Features	$\boldsymbol{R^2}$
dG (TargetAsExon, RNAstructure)	0.451757
%GC of exon when blocked by oligo	0.680028
Length	0.709566
ACP	0.710695

Πίνακας 6.8: Forward selection με 5-NN Regression για συνδυασμό χαρακτηριστικών από υποκατηγορίες του PMO συνόλου δεδομένων

5-NN Regression Features	$\boldsymbol{R^2}$
dG (TargetAsExon, RNAstructure)	0.451757
%GC of exon when blocked by oligo	0.680028
Length	0.709566
Free E oligo monomer (RNAcofold)	0.698087

Πίνακας 6.9: Backward elimination με 5-NN Regression για συνδυασμό χαρακτηριστικών από υποκατηγορίες του PMO συνόλου δεδομένων

ήταν ο βέλτιστος. Η ταξινόμηση όλων των χαρακτηριστικών με forward selection επιτυγχάνει τη μέγιστη απόδοση ($\boldsymbol{R^2 = 0.78}$), η

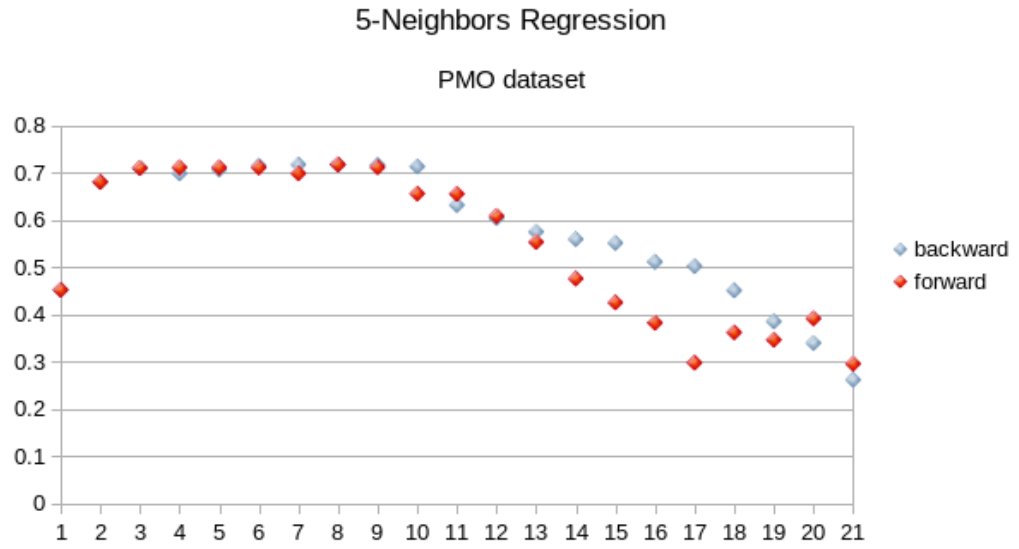


Figure 6.4: Ταξινόμηση συνδυασμού χαρακτηριστκών του PMO συνόλου δεδομένων με k-Nearest Neighbors Regression.

οποία οφείλεται κυρίως στα τρία πρώτα χαρακτηριστικά.

Πρώτα επιλέγεται το *dG (TargetAsExon, RNAstructure)*, έπειτα το βέλτιστο χαρακτηριστικό που το συμπληρώνει είναι το *%GC of exon when blocked by oligo* και το τελικό χαρακτηριστικό που έχει σημαντική επίδραση στην τιμή του $\boldsymbol{R^2}$ είναι το *dG (ExonStartTo10BaseDownFromOligo, RNAstructure)*(Πίνακας 6.6).

Όταν όμως επιλέγεται μόνο ένα χαρακτηριστικό από κάθε υποκατηγορία, βλέπουμε πως η μέγιστη επίδοση είναι $\boldsymbol{R^2 = 0.71}$ με τα δύο πρώτα χαρακτηριστικά ακριβώς ίδια με αυτά που έχουν περιγραφεί για την ταξινόμηση όλων των χαρακτηριστικών (Σχήμα 6.8). Αφού δεν μπορεί να ακολουθήσει άλλη μία τιμή ενέργειας πρόσδεσης, επιλέγεται το μήκος του ολιγονουκλεοτιδίου, που έχει λιγότερο σημαντική συνεισφορά.

Αξίζει να σημειωθεί, πως η προσέγγιση ταξινόμησης με backward elimination ξεχωρίζει τρία χαρακτηριστικά που ανήκουν στην ίδια υποκατηγορία και αφορούν την ενέργεια πρόσδεσης και τελικά δεν καταφέρνει να φτάσει την αποδοτικότητα στις τιμές του forward

selection (Σχήμα 6.3) αλλά μόλις στο 0.62. Όταν η προσέγγιση αυτή εφαρμόζεται για τα χαρακτηριστικά επιλεγμένα από υποκατηγορίες, καταφέρνει λίγο καλύτερα αποτελέσματα (Πίνακας 6.9).

Παρατηρήσεις:

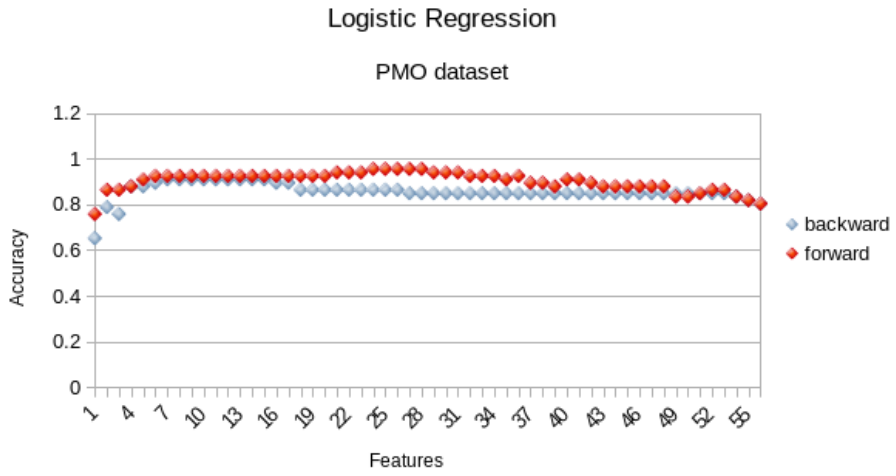
- Η ταξινόμηση όλων των χαρακτηριστικών εκτιμάται αρκετά πιο αποτελεσματική από αυτήν με επιλεγμένα χαρακτηριστικά από κάθε υποκατηγορία στην περίπτωση της ταξινόμησης με forward selection.
- Τα αποτελέσματα από τις δύο προσεγγίσεις ταξινόμησης, forward selection και backward elimination, είναι παρόμοια όταν τα χαρακτηριστικά έχουν επιλεγεί από υποκατηγορίες, αλλά η προσέγγιση με forward selection δίνει αρκετά καλύτερα αποτελέσματα στην ταξινόμηση όλων των χαρακτηριστικών.
- Βέλτιστα χαρακτηριστικά *dG (TargetAsExon, RNAstructure)* και *%GC of exon when blocked by oligo.*

6.1.3 Logistic Regression

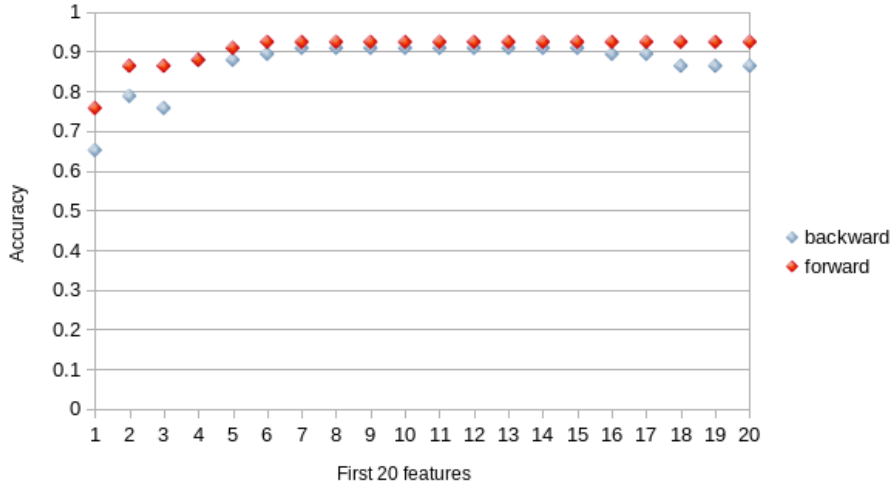
Logistic Regression Features	**Accuracy**
dG (TargetAsExon, RNAstructure)	0.757575
%GC of exon when blocked by oligo	0.863636
ACP	0.863636
Length	0.878787
Distance from acceptor (position of last base relative to acceptor)	0.909090
# exon GCs blocked by oligo	0.924242

Πίνακας 6.10: Forward selection με Logistic Regression Classification για όλα τα χαρακτηριστικά του PMO συνόλου δεδομένων

Όταν εφαρμόζεται αλγόριθμος ταξινόμησης στο PMO σύνολο δεδομένων, και συγκεκριμένα ο Logistic Regression, επιτυγχάνε-

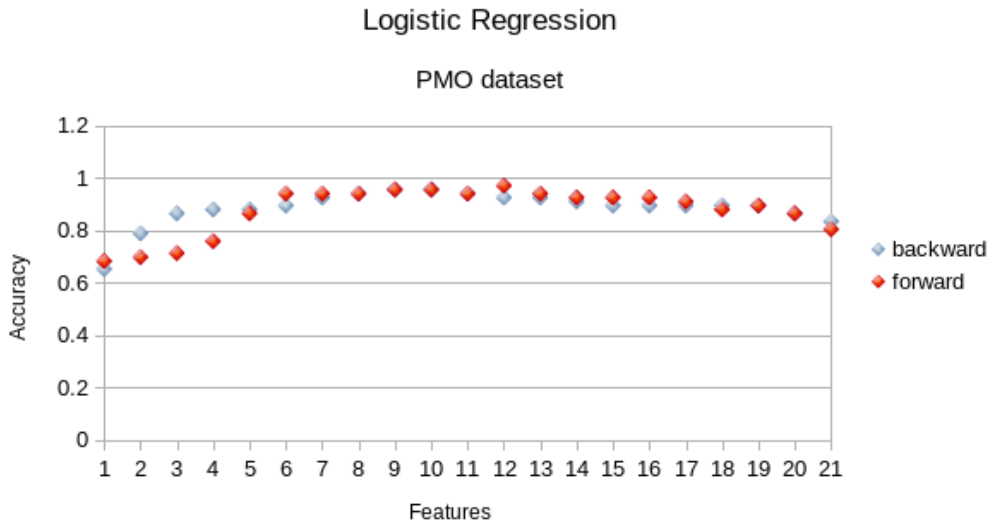


(α΄) Γράφημα για το σύνολο των χαρακτηριστικών



(β΄) Γράφημα για τα πρώτα είκοσι χαρακτηριστικά

Σχήμα 6.5: Ταξινόμηση όλων των χαρακτηριστκών του PMO συνόλου δεδομένων με Logistic Regression Classification



Σχήμα 6.6: Ταξινόμηση συνδυασμού χαρακτηριστκών του PMO συνόλου δεδομένων με Logistic Regression

Logistic Regression Features	**Accuracy**
Length	0.651515
%GC of exon when blocked by oligo	0.787878
Distance from Donor (position of 1st base relative to donor)	0.863636

Πίνακας 6.11: Backward elimination με Logistic Regression για συνδυασμό χαρακτηριστικών από υποκατηγορίες του PMO συνόλου δεδομένων

ται μέγιστη τιμή του accuracy με forward selection σε συγκεκριμένο υποσύνολο συνδυασμού χαρακτηριστικών. Όπως φαίνεται στο Σχήμα 6.6, η εκτίμηση για την αποδοτικότητα του μοντέλου φτάνει στο 0.96, ενώ στην περίπτωση χρήσης του συνόλου των χαρακτηριστικών το αποτέλεσμα είναι ελαφρώς χειρότερο (Σχήμα 6.5).

Μας ενδιαφέρει όμως να εντοπίσουμε το βέλτιστο αποτέλεσμα με μικρό αριθμό χαρακτηριστικών και σε αυτήν την περίπτωση ξεχωρίζει η ενέργεια πρόσδεσης dG του ολιγονουκλεοτιδίου στο εξώνιο και το ποσοστό των GC βάσεων που κρύβονται από το ολιγονουκλεοτίδιο, σύμφωνα με τον Πίνακα 6.10.

Παρατηρήσεις:

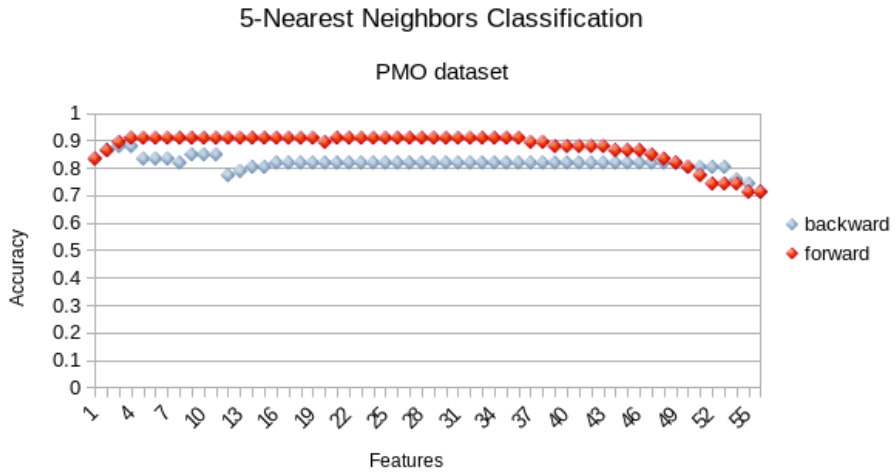
- Βέλτιστη αποδοτικότητα επιτυγχάνεται με forward selection σε υποσύνολο χαρακτηριστικών από διαφορετικές υποκατηγορίες
- Επειδή θέλουμε να επιλέξουμε μικρό αριθμό χαρακτηριστικών, επιλέγουμε την αντίστοιχη μέθοδο με ολόκληρο το σύνολο των δεδομένων
- Στην ταξινόμηση όλων των χαρακτηριστικών παρατηρούμε πως δεν επαναλαμβάνονται στις πρώτες θέσεις τα χαρακτηριστικά από ίδιες υποκατηγορίες
- Βέλτιστα χαρακτηριστικά: *dG (TargetAsExon, RNAstructure)* και *%GC of exon when blocked by oligo*

6.1.4 k-Nearest Neighbors Classification

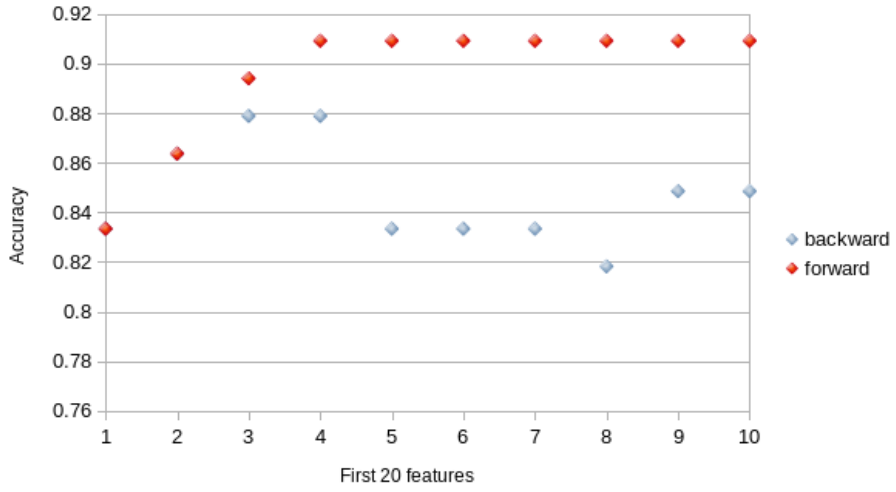
5NN Classification Features	**Accuracy**
Length of exon when blocked by oligo	0.833333
dG TargetAsExon oligo::target (RNAcofold)	0.863636
SC35 ESEfinder value over threshold	0.893939
Exon Length	0.909090

Πίνακας 6.12: Forward selection με 5-NN Classification για όλα τα χαρακτηριστικά του PMO συνόλου δεδομένων

Στην περίπτωση της ταξινόμησης με τον αλγόριθμο των πλησιέστερων γειτόνων (όπου επιλέγονται επίσης 5 γείτονες για λόγους συνοχής), και πάλι βλέπουμε την μέθοδο επιλογής χαρακτηριστικών forward selection να φτάνει υψηλότερα αποτελέσματα από τη μέθοδο backward elimination. Η τιμή του accuracy φτάνει στο 0.9 με μόλις τέσσερα χαρακτηριστικά και για πρώτη φορά φαίνεται να έχει



(α′) Γράφημα για το σύνολο των χαρακτηριστικών

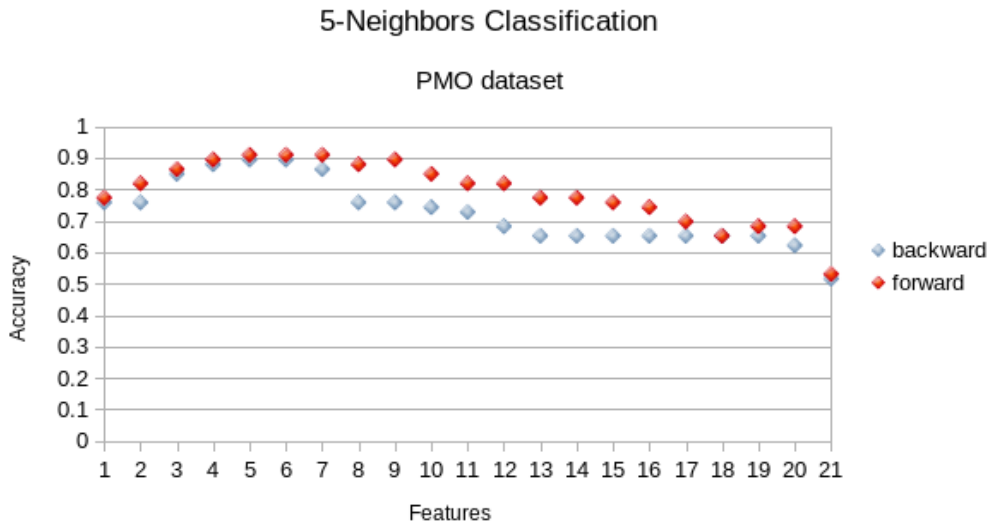


(β′) Γράφημα για τα πρώτα είκοσι χαρακτηριστικά

Σχήμα 6.7: Ταξινόμηση όλων των χαρακτηριστκών του PMO συνόλου δεδομένων με 5-Nearest Neighbors Classification.

5NN Classification Features	Accuracy
Length of exon when blocked by oligo	0.833333
dG TargetAsExon oligo::target (RNAcofold)	0.863636
dG 150bp flanks oligo::target (RNAcofold)	0.878787

Πίνακας 6.13: Backward elimination με 5-NN Classification για όλα τα χαρακτηριστικά του PMO συνόλου δεδομένων



Σχήμα 6.8: Ταξινόμηση συνδυασμού χαρακτηριστκών του PMO συνόλου δεδομένων με 5-NN Classification

σημαντική επίδραση η θέση πρόσδεσης της SR πρωτεΐνης SC35.

Όσο για την ταξινόμηση με τα χαρακτηριστικά επιλεγμένα από υποκατηγορίες, παρατηρούμε επίσης αρκετά καλά αποτελέματα (Σχήμα 6.8) με τη μέθοδο forward selection να προηγείται ελαφρώς της μεθόδου backward elimination. Φαίνεται να έχουμε πολύ καλή τιμή ακρίβειας με συνδυασμό ενέργειας πρόσδεσης και ποσοστού GC βάσεων (Πίνακες 6.14 και 6.15).

Παρατηρήσεις:

- Η μέθοδος προσέγγισης επιλογής χαρακτηριστικών forward

5NN ClassificationFeatures	**Accuracy**
dG TargetAsExon oligo::target (RNAcofold)	0.772727
%GC of exon when blocked by oligo	0.818181
Length	0.863636

Πίνακας 6.14: Forward selection με 5-NN Classification για συνδυασμό χαρακτηριστικών από υποκατηγορίες του PMO συνόλου δεδομένων

5NN ClassificationFeatures	**Accuracy**
dG (TargetAsExon RNAstructure)	0.757575
Tra2B	0.757575
%GC increase target over 5' intron	0.848484

Πίνακας 6.15: Backward elimination με 5-NN Classification για συνδυασμό χαρακτηριστικών από υποκατηγορίες του PMO συνόλου δεδομένων

selection δίνει τα καλύτερα αποτελέσματα

- Βέλτιστα χαρακτηριστικά: *Length of exon when blocked by oligo*, *dG TargetAsExon oligo::target (RNAcofold)*, *%GC of exon when blocked by oligo* και *SC35 ESEfinder value over threshold*

6.2 2OMePS σύνολο δεδομένων

Το σύνολο δεδομένων 2OMePS δεν είναι ομοιόμορφο, μιας και αποτελείται από υποσύνολα διάφορων ερευνών (Πίνακας 6.16). Έτσι, αναμένεται τα αποτελέσματα να μην έχουν την ίδια συνέπεια με το PMO σύνολο δεδομένων. Το πλεονέκτημα που έχουμε σε αυτήν την περίπτωση, όμως, είναι το μεγαλύτερο μέγεθος του συνόλου δεδομένων που ανέρχεται σε 292 δείγματα. Επιπλέον, η αποτελεσματικότητα του exon skipping έχει καταγραφεί με διαφορετικούς τρόπους από τις επιμέρους έρευνες και η ομοιογενής μορφή που μπορούμε να της προσδόσουμε την καθιστά διακριτή μεταβλητή και έτσι

κατάλληλη για τους αλγορίθμους ταξινόμησης και όχι παλινδρόμησης.

	Χαρακτηριστικά	**Υποσύνολο**
1	Targeted exon	AR, AR2009, H, W, DP
2	Distance from acceptor (position of last base relative to acceptor)	AR, AR2009, H, W, DP
3	ACP	AR, AR2009, H, W, DP
4	Distance from Donor (position of 1st base relative to donor)	AR, AR2009, H, W, DP
5	Length	AR, AR2009, H, W, DP
6	Length Cat.	AR, AR2009, H, W, DP
7	Exon Length	AR, AR2009, H, W, DP
8	Length of exon when blocked by oligo	AR, AR2009, H, W, DP
9	Exon Malueka Category	AR, AR2009, H, W, DP
10	L1 (WeeEtal)	AR, W
11	L3 (WeeEtal)	AR, W
12	niscore	AR, AR2009, H, W, DP
13	niscore_per_base	AR, AR2009, H, W, DP
14	ACC_FL	AR, AR2009, H, W, DP
15	ACC_LAST8	AR, AR2009, H, W, DP

16	ACC_LAST15	AR, AR2009, H, W, DP
17	ACC_BEST8	AR, AR2009, H, W, DP
18	ACC_AVE	AR, AR2009, H, W, DP
19	%GC oligo	AR, AR2009, H
20	GCs (number of)	AR, AR2009, H
21	%GC 5' intron 200 bases upstream	AR, AR2009, H, W, DP
22	%GC fold increase target over 5' intron	AR, AR2009, H
23	%GC increase target over 5' intron	AR, AR2009, H
24	Decrease in GC% due to blocking by oligo	AR, AR2009, H
25	Decrease in ratio of exon to intron %GC due to oligo blocking	AR, AR2009, H
26	%GC exon	AR, AR2009, H, W, DP
27	Total GCs in exon	AR, AR2009, H, W, DP
28	Total GCs in target	AR, AR2009, H
29	# exon GCs blocked by oligo	AR, AR2009, H
30	%GC of exon when blocked by oligo	AR, AR2009, H
31	Exon v intron %GC	AR, AR2009, H, W, DP

32	Exon v intron %GC after blocking by oligo	AR, AR2009, H
33	dG (TargetAsExon, RNAstructure)	AR, AR2009, H, W, DP
34	dG (50BaseFlanksAroundTarget, RNA structure)	AR, AR2009, H, W, DP
35	dG (100BaseFlanks, RNAstructure)	AR, AR2009, H, W, DP
36	dG (200BaseFlanks, RNAstructure)	AR, AR2009, H, W, DP
37	dG (ExonStartTo10BaseDownFromOligo, RNAstructure)	AR, AR2009, H, W, DP
38	RNAeval targetoligo	AR, AR2009, H, W, DP
39	dG TargetAsExon oligo::target (RNAcofold)	AR, AR2009, H, W, DP
40	dG 50bp flanks oligo::target (RNAcofold)	AR, AR2009, H, W, DP
41	dG 100bp flanks oligo::target (RNAcofold)	AR, AR2009, H, W, DP
42	dG 150bp flanks oligo::target (RNAcofold)	AR, AR2009, H, W, DP
43	Free E oligo dimer (RNAcofold)	AR, AR2009, H, W, DP
44	Free E target monomer (RNAcofold)	AR, AR2009, H, W, DP
45	Free E oligo monomer (RNAcofold)	AR, AR2009, H, W, DP
46	3' SS	AR
47	5' SS	AR
48	% open (Mfold)	AR
49	#rescue ESE sites	AR, AR2009

50	SF2/ASF ESEfinder value over threshold	AR, AR2009, H, W
51	SC35 ESEfinder value over threshold	AR, AR2009, H, W
52	SRp40 ESEfinder value over threshold	AR, AR2009, H, W
53	SRp55 ESEfinder value over threshold	AR, AR2009, W
54	Tra2B	AR, AR2009
55	9G8	AR, AR2009

Πίνακας 6.16: Χαρακτηριστικά του συνόλου δεδομένων για τα 2OMePS ολιγονουκλεοτίδια. Η ευθεία γραμμή χωρίζει τα χαρακτηριστικά σε υπο-ομάδες. Για κάθε χαρακτηριστικό υποδεικνύεται σε ποια 2OMePS υποσύνολα αυτό ανήκει. **AR**: AartsmaRus, **AR2009**: AartsmaRus2009, **H**: Harding, **W**: Wilton, **DP**: DwiPramono

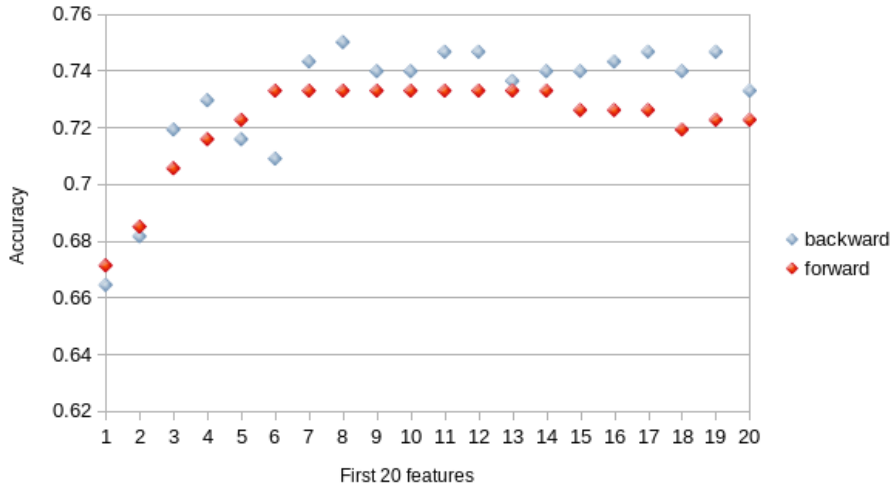
6.2.1 Logistic Regression

Logistic Regression Features	**Accuracy**
dG (ExonStartTo10BaseDownFromOligo, RNAstructure)	0.671232
Targeted exon	0.684931
L1 (WeeEtal)	0.705479
Exon Length	0.715753

Πίνακας 6.17: Forward selection με Logistic Regression για όλα τα χαρακτηριστικά του 2OMePS συνόλου δεδομένων

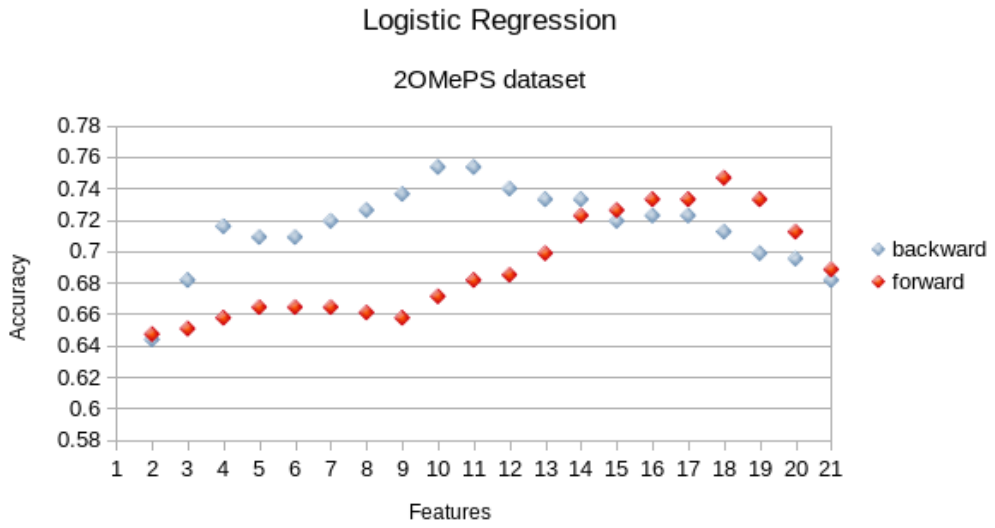
Με την εφαρμογή του αλγορίθμου Logistic Regression παρατηρείται μέγιστη αποδοτικότητα με την ταξινόμηση backward elimination στα χαρακτηριστικά που έχουν επιλεγεί από υποκατηγορίες (accuracy 0.75). Στην περίπτωση αυτή, η διαφορά με την προσέγγιση ταξινόμησης forward selection είναι σημαντική και μπορούμε να τη δούμε στο Σχήμα 6.10. Τα χαρακτηριστικά που ξεχωριζούν

(α΄) Γράφημα για το σύνολο των χαρακτηριστικών



(β΄) Γράφημα για τα πρώτα είκοσι χαρακτηριστικά

Σχήμα 6.9: Ταξινόμηση όλων των χαρακτηριστκών του 2OMePS συνόλου δεδομένων με Logistic Regression



Σχήμα 6.10: Ταξινόμηση συνδυασμού χαρακτηριστκών του 2OMePS συνόλου δεδομένων με Logistic Regression

Logistic Regression Features	**Accuracy**
#rescue ESE sites	0.643835
Tra2B	0.681506
dG (TargetAsExon, RNAstructure)	0.715753

Πίνακας 6.18: Backward elimination με Logistic Regression για όλα τα χαρακτηριστικά του 2OMePS συνόλου δεδομένων

Logistic Regression Features	**Accuracy**
dG (TargetAsExon, RNAstructure)	0.664383
Targeted exon	0.681506
#exon GCs blocked by oligo	0.719178
5' SS	0.729452

Πίνακας 6.19: Backward elimination με Logistic Regression για συνδυασμό χαρακτηριστικών από υποκατηγορίες του 2OMePS συνόλου δεδομένων

είναι η ενέργεια πρόσδεσης dG στο εξώνιο, το εξώνιο στο οποίο στοχεύει το ολιγονουκλεοτίδιο και ο αριθμός των GC βάσεων που κρύβονται κατά την πρόσδεση (Πίνακας 6.26).

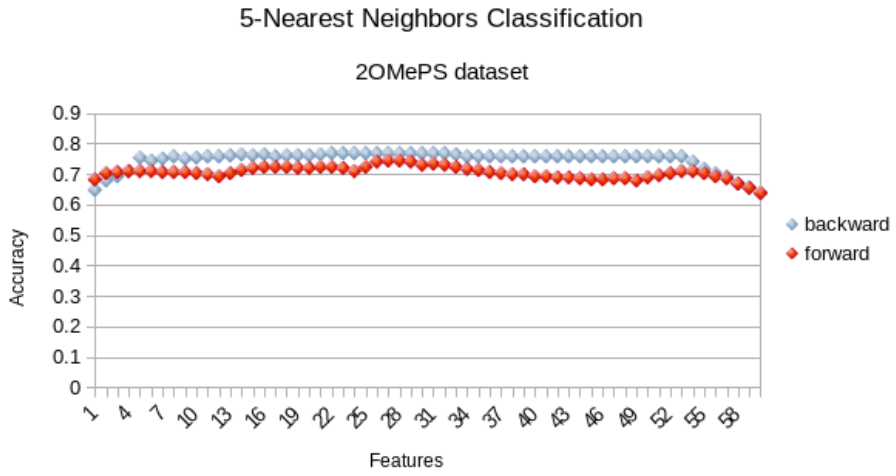
Όταν χρησιμοποιούνται όλα τα χαρακτηριστικά, η μέγιστη τιμή ακρίβειας είναι λιγο χαμηλότερη και οι δύο μέθοδοι ταξινόμησης δεν παρουσιάζουν μεγάλες διαφορές (Σχήμα 6.9). Έχουν όμως διαφορετκή εκτίμηση για τη διαλογή των καλύτερων χαρακτηριστικών. Η ταξινόμηση με backward elimination ξεχωρίζει τις θέσεις ενισχυτών ματίσματος και τις θέσεις πρόσδεσης της SR πρωτεΐνης *Tra2β*, καθώς και την ενέργεια πρόσδεσης του ολιγονουκλεοτιδίου στο εξώνιο (Πίνακας 6.18). Η ταξινόμηση με forward selection δίνει αποτελέσματα που έχουν περισσότερες ομοιότητες με αυτά που είδαμε πιο πάνω. Δηλαδή, ξεχωρίζει η ενέργεια πρόσδεσης dG και το στοχευμένο εξώνιο (Πίνακας 6.17)

Παρατηρήσεις:

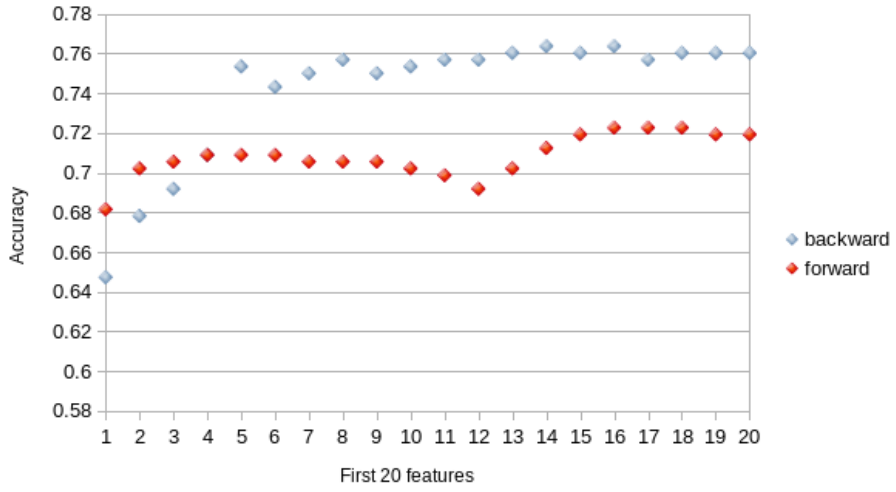
- Η βέλτιστη ταξινόμηση επιτυγχάνεται με εφαρμογή backward elimination μεθόδου σε συνδυασμό επιλεγμένων χαρακτηριστικών.
- Τα αποτελέσματα από τις δύο προσεγγίσεις ταξινόμησης, forward selection και backward elimination, είναι παρόμοια όταν εφαρμόζονται σε σύνολα δεδομένων με όλα τα χαρακτηριστικά.
- Βέλτιστα χαρακτηριστικά *dG (TargetAsExon, RNAstructure)* και *dG (ExonStartTo10BaseDownFromOligo, RNAstructure)*

6.2.2 k-Nearest Neighbors Classification

Στην περίπτωση της ταξινόμησης με τους πλησιέστερους γείτονες, η μέγιστη τιμή ακρίβειας επιτυγχάνεται όταν το σύνολο δεδομένων περιλαμβάνει όλα τα χαρακτηριστικά. Παρ΄όλα αυτά, έχουμε καλύτερα αποτελέσματα με λίγα χαρακτηριστικά όταν αυτά ανήκουν σε ξεχωριστές υποκατηγορίες (Σχήματα 6.11 και 6.12). Είναι αξιοσημείωτο πως και με τους δύο τρόπους επιλογής χαρακτηριστικών, επιλέγονται τρία ίδια που ανήκουν σε ξεχωριστές υποκατηγορίες και

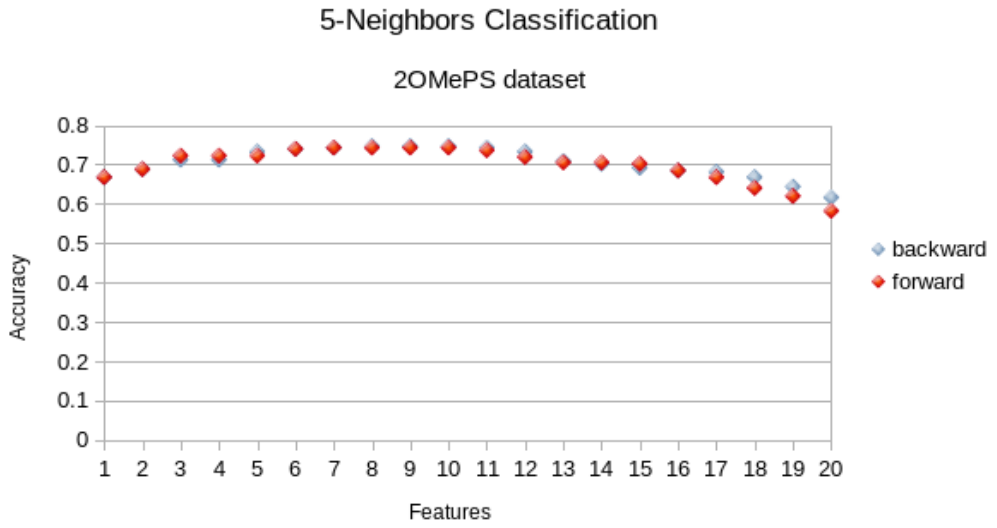


(α΄) Γράφημα για το σύνολο των χαρακτηριστικών



(β΄) Γράφημα για τα πρώτα είκοσι χαρακτηριστικά

Σχήμα 6.11: Ταξινόμηση όλων των χαρακτηριστκών του 2OMePS συνόλου δεδομένων με 5-Nearest Neighbors Classification



Σχήμα 6.12: Ταξινόμηση συνδυασμού χαρακτηριστκών του 2OMePS συνόλου δεδομένων με 5-Nearest Neighbors Classification

φτάνουν την ακρίβεια στην τιμή 0.72, όπως φαίνεται στους Πίνακες 6.27 και 6.28. Εδώ βλέπουμε να επιλέγονται για πρώτη φορά η 3′ θέση ματίσματος και το μήκος του ολιγονουκλεοτιδίου ως πιο σημαντικά χαρακτηριστικά και να συμπληρώνονται με το μήκος και την ενέργεια πρόσδεσης dG.

Στην περίπτωση ταξινόμησης όλων των χαρκτηριστικών, επιλέγονται πρώτα αυτά που έχουμε δει σε προηγούμενα αποτελέσματα και αφορούν την ενέργεια πρόσδεσης dG, τις GC βάσεις και το στοχευμένο εξώνιο (Πίνακες 6.20 και 6.21).

5-NN Classification Features	**Accuracy**
dG 150bp flanks oligo::target (RNAcofold)	0.681506
Exon v intron %GC	0.702054
L3 (WeeEtal)	0.705479

Πίνακας 6.20: Forward selection με 5-NN Classification για όλα τα χαρακτηριστικά του 2OMePS συνόλου δεδομένων

5-NN Classification Features	Accuracy
Targeted exon	0.647260
Length of exon when blocked by oligo	0.678082
RNAeval target-oligo	0.691780
% open (Mfold)	0.708904
%GC increase target over 5' intron	0.753424

Πίνακας 6.21: Backward elimination με 5-NN Classification για όλα τα χαρακτηριστικά του 2OMePS συνόλου δεδομένων

5-NN Classification Features	Accuracy
3' SS	0.667808
Length	0.688356
dG 50bp flanks oligo::target (RNAcofold)	0.722602

Πίνακας 6.22: Forward selection για συνδυασμό χαρακτηριστικών από υποκατηγορίες του 2OMePS συνόλου δεδομένων

5-NN Classification Features	Accuracy
3' SS	0.667808
Length	0.688356
dG (50BaseFlanksAroundTarget RNA structure)	0.712328

Πίνακας 6.23: Backward elimination για συνδυασμό χαρακτηριστικών από υποκατηγορίες του 2OMePS συνόλου δεδομένων

Παρατηρήσεις:

- Βέλτιστη απόδοση επιτυγχάνεται με ταξινόμηση χαρακτηριστικών που έχουν επιλεγεί από ξεχωριστές υποκατηγορίες, με παρόμοιες τιμές και ίδια χαρακτηριστικά στις δύο προσεγγίσεις (forward selection & backward elimination).
- Στην ταξινόμηση με όλα τα χαρακτηριστικά οι δύο μέθοδοι, forward selection και backward elimination έχουν παρόμοια απόδοση.
- Παρατηρείται για πρώτη φορά ως πολύ σημαντικό χαρακτηρι-

Υποσύνολο	**Μέγεθος**
AartsmaRus	113
AartsmaRus2009	41
Harding	33
Wilton	78
DwiPramono	23

Πίνακας 6.24: Τα υποσύνολα του 2OMePS συνόλου δεδομένων και τα μεγέθη τους.

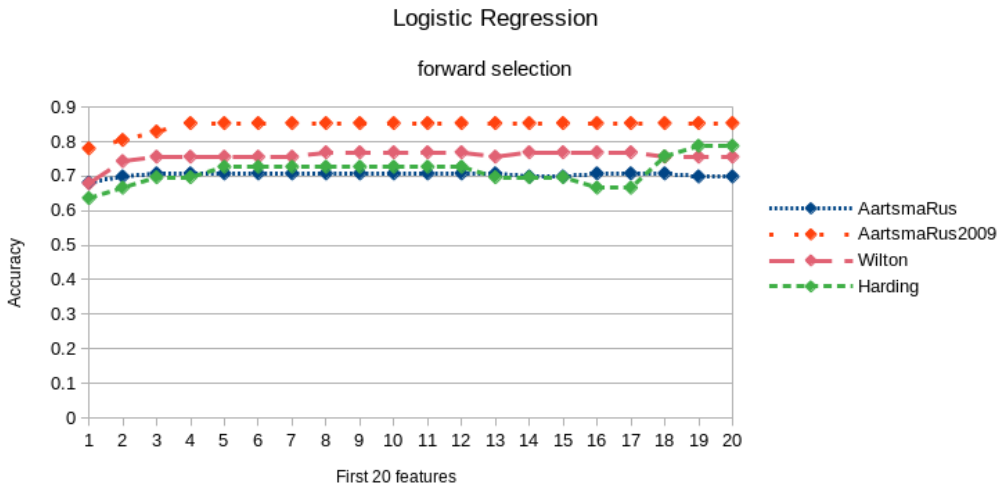
στικό η 3′ θέση ματίσματος.

6.3 Υποσύνολα του 2OMePS συνόλου δεδομένων

Όπως έχει αναφερθεί πιο πάνω, το 2OMePS συνόλο δεδομένων αποτελείται από υποσύνολα που έχουν μελετηθεί από διαφορετικούς ερευνητές και υπο διαφορετικές συνθήκες. Έτσι, δεν παρουσιάζεται ομοιογένεια σε όλα τα χαρακτηριστικά που περιλαμβάνονται στη δική μας μελέτη. Για να καλυφθούν οι τιμές που λείπουν από τα δεδομένα μας (missing values) έχουμε επιλέξει να τις αντικαταστήσουμε με μηδενικά. Η μεθοδολογία αυτή είναι μία γρήγορη αλλά όχι απαραίτητα αξιόπιστη λύση.

Παρατηρήθηκε παραπάνω πως η συμβολή κάποιων χαρακτηριστικών, όπως είναι ορισμένες SR πρωτείνες, είναι πολύ σημαντική στην αποδοτικότητα του exon skipping. Κάτι τέτοιο δεν παρουσιάστηκε στα αποτελέσματα που βγάλαμε για το PMO σύνολο δεδομένων. Επομένως, προκύπτει το ερώτημα εάν αυτή η διαφοροποίηση οφείλεται στη διαφορετική χημεία των ολιγονουκλεοτιδίων ή είναι αποτέλεσμα του χειρισμού που έχουμε κάνει για μη ομοιογενές σύνολο δεδομένων.

Παρακάτω παρουσιάζεται η αποδοτικότητα των 2OMePS υποσυνόλων ξεχωριστά για τις διάφορες μεθόδους και αλγορίθμους



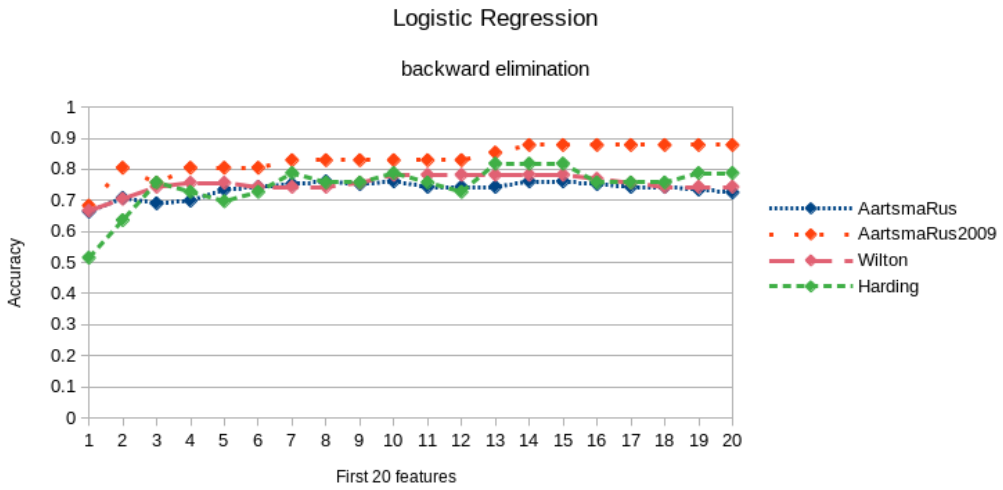
Σχήμα 6.13: Forward selection ταξινόμηση όλων των χαρακτηριστκών των 2OMePS υποσυνόλων δεδομένων με Logistic Regression

ταξινόμησης με χρήση όλων των χαρακτηριστικών του κάθε υποσυνόλου. Εξαιρείται το υποσύνολο των *Dwi Pramono et al.* [16] μιας και λόγω πολύ μικρού μεγέθους δεν παρουσιάζει εύστοχα αποτελέσματα.

6.3.1 Logistic Regression

Με την εφαρμογή του Logistic Regression σε κάθε υποσύνολο ξεχωριστά παρατηρείται βέλτιστη απόδοση στο σύνολο *AartsmaRus2009* και αν συγκρίνουμε τις μεθόδους ταξινόμησης θα δούμε πως αυτή με forward selection είναι η πιο αποτελεσματική (Πίνακες 6.25 και 6.26). Τα δύο μόλις χαρακτηριστικά που επιτυγχάνουν accuracy 0.8 είναι η μείωση του ποσοστού των GC βάσεων λόγω της πρόσδεσης του ολιγονουκλεοτιδίου και η απόσταση από τη θέση δότη, δηλαδή την 5′ θέση ματίσματος.

Για το υποσύνολο *AartsmaRus* η αποδοτικότητα είναι λίγο πιο χαμηλή και επιτυγχάνεται με τον αριθμό των θέσεων ενισχυτών ματίσματος και την πρωτείνη Tra**2**$\boldsymbol{\beta}$ εφαρμόζοντας και πάλι τη μέθοδο forward selection.



Σχήμα 6.14: Backward elimination ταξινόμηση όλων των χαρακτηριστκών των 2OMePS υποσυνόλων δεδομένων με Logistic Regression

Για τα υποσύνολα *Wilton* και *Harding* το μήκος του εξωνίου-με την ενέργεια πρόσδεσης dG και αντίστοιχα η απόσταση από τη θέση του δότη και το μικρό μήκος του ολιγονουκλεοτιδίου φαίνεται να αποτελούν τα χαρακτηριστικά με τη μεγαλύτερη αποδοτικότητα.

Παρατηρήσεις:

- Βέλτιστη απόδοση επιτυγχάνει η ταξινόμηση χαρακτηριστικών με forward selection σε όλα τα υποσύνολα, αν και με όχι πολύ μεγάλη διαφορά σε σχέση με την backward elimination.
- Όλα τα υποσύνολα έχουν σχετικά διαφορετικά χαρακτηριστικά που συμβάλλουν με βέλτιστο τρόπο στην επιτυχία του exon skipping.
- Φαίνεται πως το σύνολο δεδομένων *AartsmaRus2009* έχει σταθερά την υψηλότερη τιμή ακρίβειας σε σχέση με τα υπόλοιπα 2OMePS σύνολα.

Logistic Regression Features	Accuracy
AartsmaRus	
ACC_LAST8	0.681416
niscore_per_base	0.699115
L3 (WeeEtal)	0.707965
AartsmaRus2009	
Decrease in GC% due to blocking by oligo	0.780488
Distance from Donor (position of 1st base relative to donor)	0.804878
Length	0.829268
dG (ExonStartTo10BaseDownFromOligo, RNAstructure)	0.853659
Wilton	
Exon Length	0.679487
dG (100BaseFlanks, RNAstructure)	0.743589
ACC_LAST8	0.756410
Harding	
Distance from Donor (position of 1st base relative to donor)	0.636363
short oligo	0.666666
ACP	0.696969

Πίνακας 6.25: Forward selection με Logistic Regression για όλα τα χαρακτηριστικά των υποσυνόλων του 2OMePS συνόλου δεδομένων

6.3.2 k-Nearest Neighbors Classification

Με την εφαρμογή του αλγορίθμου ταξινόμησης με 5-Nearest Neighbors έχουμε και πάλι βέλτιστη απόδοση στο σύνολο *AartsmaRus2009* και η ακρίβεια φτάνει στο 0.85 όταν εφαμόζεται η μέθοδος forward selection και τις πρώτες θέσεις καταλαμβάνουν τα χαρακτηριστικά που αφορούν την απόσταση από τη 5′ θέση ματίσματος και το εξώνιο που στοχοποιείται.

Ακολουθεί το υποσύνολο *Wilton*, το οποίο οφείλει την αποδοτικότητά του στην ενέργεια πρόσδεσης dG και στο στοχευμένο εξώνιο. Ενώ στο υποσύνολο *Harding* φαίνεται να είναι πιο σημαντικά τα χαρακτηριστικά προσβασιμότητας (accessibility) των βάσεων και μήκους των ολιγονουκλεοτιδίων.

Για το τέλος έχουμε το υποσύνολο *AartsmaRus* όπου πρωτα-

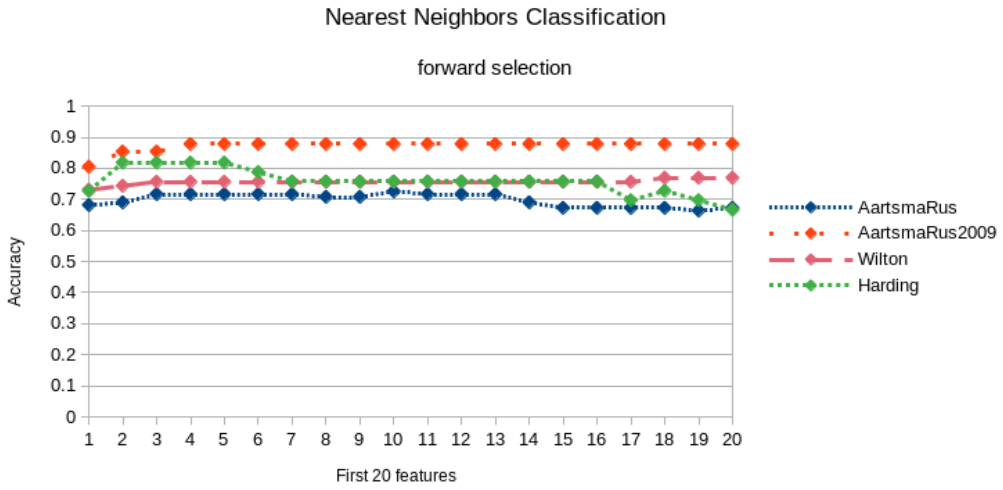
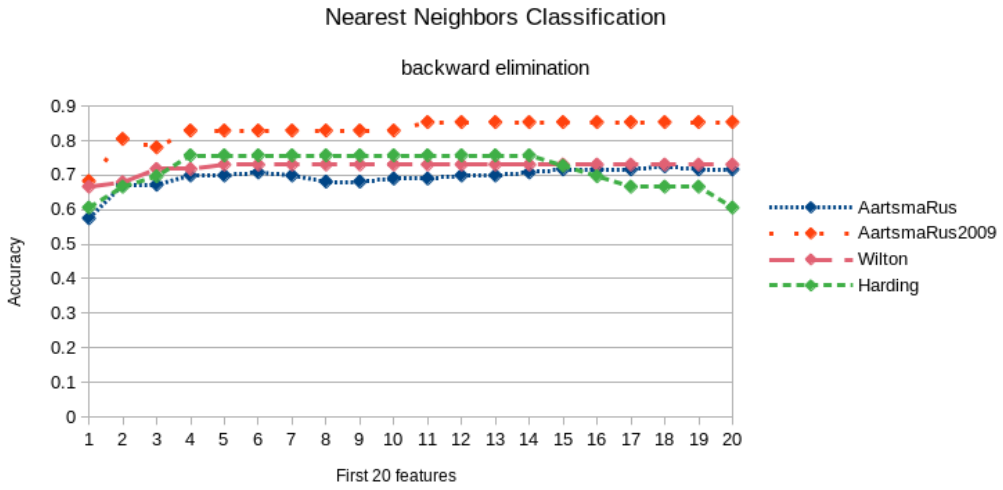


Figure 6.15: Forward selection ταξινόμηση όλων των χαρακτηριστκών των 2OMePS υποσυνόλων δεδομένων με 5-NN Classification



Σχήμα 6.16: Backward elimination ταξινόμηση όλων των χαρακτηριστκών των 2OMePS υποσυνόλων δεδομένων με 5-NN Classification

Logistic Regression Features	Accuracy
AartsmaRus	
#rescue ESE sites	0.663717
Tra2B	0.707965
AartsmaRus2009	
# exon GCs blocked by oligo	0.682926
Exon Length	0.804878
Wilton	
Exon Malueka Category_C	0.666666
Length of exon when blocked by oligo	0.705128
dG (50BaseFlanksAroundTarget, RNA structure)	0.743589
dG 50bp flanks oligo::target (RNAcofold)	0.756410
Harding	
SRp40 ESEfinder value over threshold	0.515151
dG 150bp flanks oligo::target (RNAcofold)	0.636363
dG 50bp flanks oligo::target (RNAcofold)	0.757575

Πίνακας 6.26: Backward elimination με Logistic Regression για όλα τα χαρακτηριστικά των υποσυνόλων του 2OMePS συνόλου δεδομένων

γωνιστεί ο αριθμός των GC βάσεων και πάλι η προσβασιμότητα, αλλά των 8 τελευταίων βάσεων από την 3′ άκρη της στοχευμένης αλυσίδας. Όλα τα παραπάνω παρουσιάζονται γραφικά στα Σχήματα 6.15 και 6.16 και πιο αναλυτικά στους Πίνακες 6.27 και 6.28.

Παρατηρήσεις:

- Βέλτιστη απόδοση επιτυγχάνει η ταξινόμηση χαρακτηριστικών με forward selection σε όλα τα υποσύνολα, αν και με όχι πολύ μεγάλη διαφορά σε σχέση με την backward elimination.
- Όλα τα υποσύνολα έχουν σχετικά διαφορετικά χαρακτηριστικά που συμβάλλουν με βέλτιστο τρόπο στην επιτυχία του exon skipping.
- Φαίνεται πως το σύνολο δεδομένων *AartsmaRus2009* έχει σταθερά την υψηλότερη τιμή ακρίβειας σε σχέση με τα υπόλοιπα 2OMePS σύνολα.

5-NN Classification Features	**Accuracy**
AartsmaRus	
Total GCs in target	0.681416
ACC_LAST8	0.690265
SF2/ASF ESEfinder value over threshold	0.716814
AartsmaRus2009	
Distance from Donor (position of 1st base relative to donor)	0.804878
Targeted exon	0.853658
Wilton	
dG TargetAsExon oligo::target (RNAcofold)	0.730769
Targeted exon	0.743589
dG (50BaseFlanksAroundTarget, RNA structure)	0.756410
Harding	
ACC_AVE	0.727272
Length	0.818181

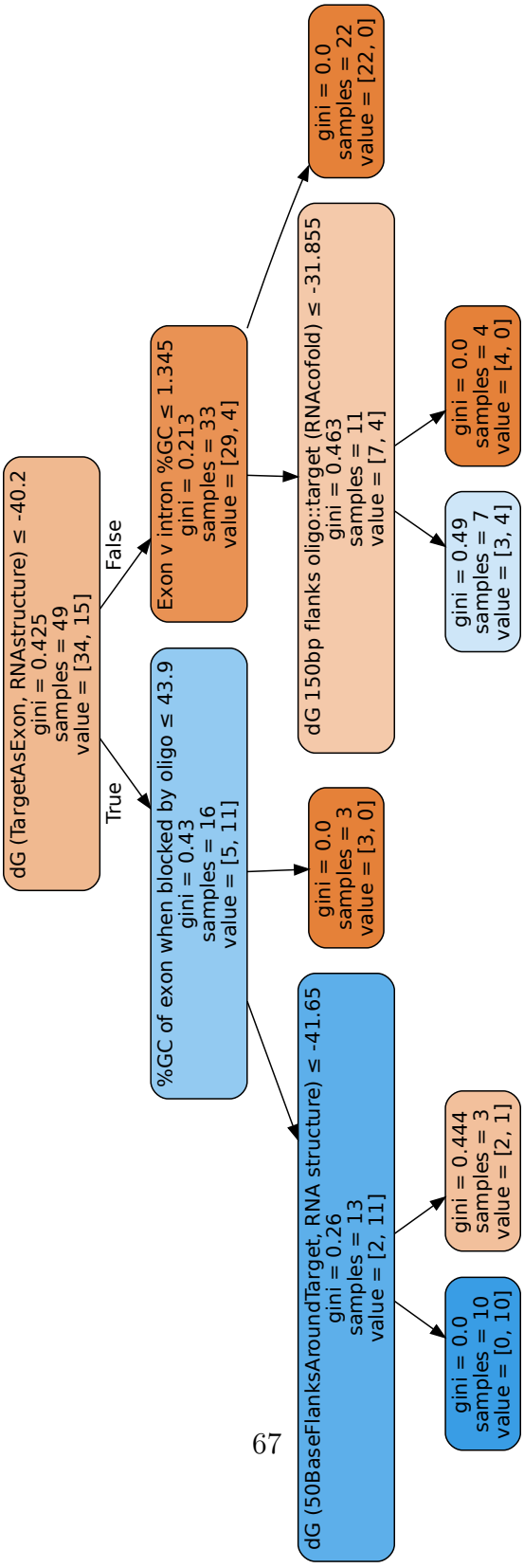
Πίνακας 6.27: Forward selection με 5-NN Classification για όλα τα χαρακτηριστικά των υποσυνόλων του 2OMePS συνόλου δεδομένων

6.4 Οπτικοποίηση των Αποτελεσμάτων

Αφού έχουμε συλλέξει τα βέλτιστα χαρακτηριστικά για την επιτυχία του exon skipping με όλες τις παραπάνω μεθόδους παλινδρόμισης και ταξινόμησης, σε αυτό το κεφάλαιο επιχειρείται η οπτικοποίηση της ταξινόμησής τους. Το δέντρο απόφασης είναι ένα πολύ βοηθητικό εργαλείο για να αποκτήσουμε διαίσθηση σχετικά με το πως ταξινομούνται τα χαρακτηριστικά και ποια η βαρύτητά τους.

Για το σκοπό αυτό, για το PMO σύνολο δεδομένων επιλέγονται όλα τα χαρακτηριστικά που σχετίζονται με την ενέργεια πρόσδεσης dG και την ποσότητα των GC βάσεων, μιας και αυτά είδαμε να δίνουν τα καλύτερα αποτελέσματα στα προηγούμενα πειράματα. Περιορίζουμε επίσης τον αριθμό διακλαδώσεων του δέντρου στις τρεις, για να μην έχουμε βαθύ και δυσνόητο γράφημα. Το δέντρο απόφασης έχει ακρίβεια 0.89 παρουσιάζεται στο Σχήμα 6.17 ενώ στο

Σχήμα 6.17: Δέντρο απόφασης για το PMO σύνολο δεδομένων

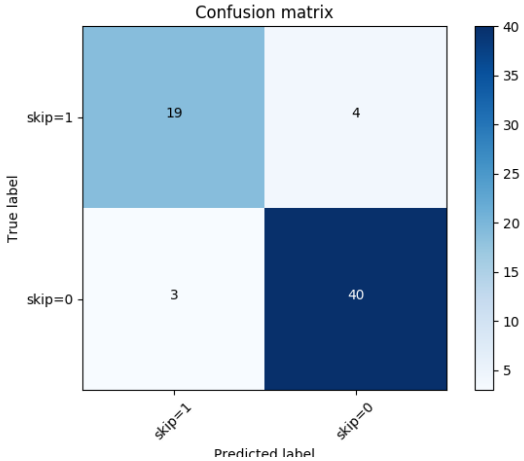


5-NN Classification Features	**Accuracy**
AartsmaRus	
5' SS	0.575221
# exon GCs blocked by oligo	0.672566
AartsmaRus2009	
9G8	0.682926
dG 100bp flanks oligo::target (RNAcofold)	0.804878
Wilton	
Length of exon when blocked by oligo	0.666666
Distance from acceptor (position of last base relative to acceptor)	0.679487
Exon Length	0.717948
Harding	
Distance from acceptor (position of last base relative to acceptor)	0.606060
Length of exon when blocked by oligo	0.666666
Targeted exon	0.696969
dG 150bp flanks oligo::target (RNAcofold)	0.757575

Πίνακας 6.28: Backward elimination με 5-NN Classification για όλα τα χαρακτηριστικά των υποσυνόλων του 2OMePS συνόλου δεδομένων

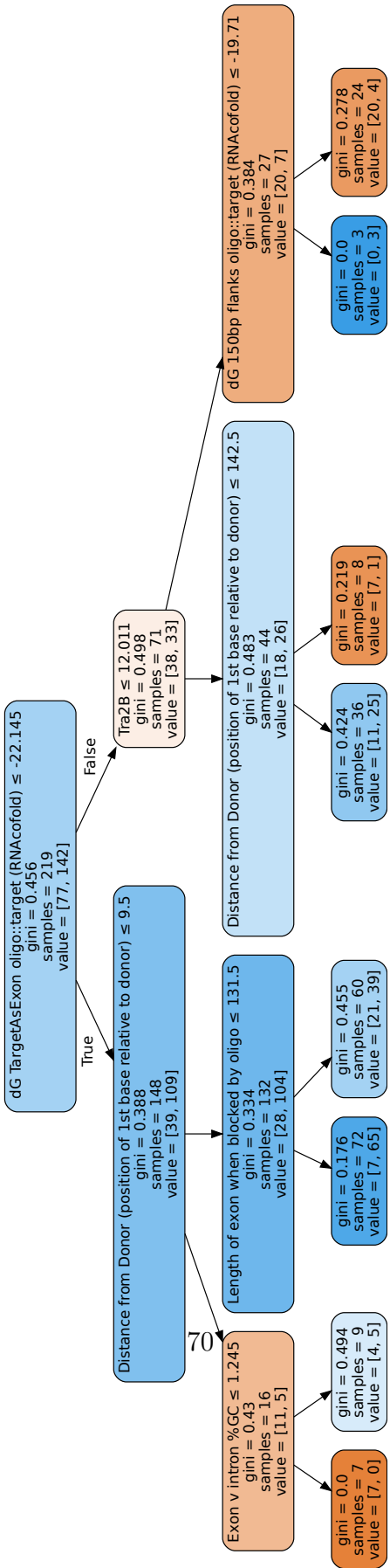
Σχήμα 6.18 βλέπουμε το confusion matrix της ταξινόμησης.

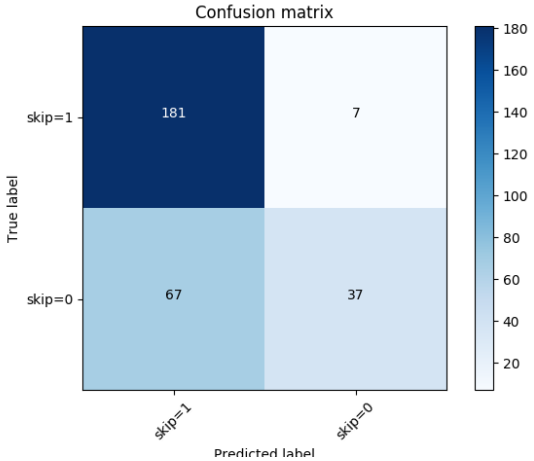
Για το 2OMePS σύνολο δεδομένων επιλέγουμε και πάλι το υποσύνολο χαρακτηριστικών της ενέργειας πρόσδεσης dG και κάποια ακόμη χαρακτηριστικά που είδαμε να ξεχωρίζουν σε όλα τα πειράματα που που αφορούσαν τη συγκεκριμένη χημεία (Σχήμα 6.19). Η ακρίβεια είναι αρκετα χαμηλότερη, φτάνει το 0.75, και αν δούμε το confusion matrix στο Σχήμα 6.20 θα παρατηρήσουμε πως υπάρχει σχεδόν διπλάσιος αριθμός περιπτώσεων όπου η πρόβλεψη είναι θετική ενώ στην πραγματικότητα δεν έχουμε exon skipping.



Σχήμα 6.18: Confusion matrix για το δέντρο απόφασης του PMO συνόλου δεδομένων

Σχήμα 6.19: Δέντρο απόφασης για το 2OMePS σύνολο δεδομένων





Σχήμα 6.20: Confusion matrix για το δέντρο απόφασης του 2OMePS συνόλου δεδομένων

Κεφάλαιο 7

Συμπεράσματα

Μελετώντας όλα τα πειραματικά δεδομένα, παρουσιάζεται παρακάτω προσπάθεια ανάλυσης και εξαγωγής συμπερασμάτων ως προς διάφορους υπολογιστικούς παράγοντες.

- **Επιλογή χαρακτηριστικών:**
  Η προσπάθειά μας να χωρίσουμε τα χαρακτηριστικά σε υποκατηγορίες, με το λογικό κριτήριο της παρόμοιας σημασίας τους, και να δημιουργήσουμε όλους τους δυνατούς συνδυασμούς μεταξύ τους, δεν απέδωσε εντυπωσιακά αποτελέσματα. Παρατηρήσαμε πως το μοντέλο μας έκανε καλύτερες προβλέψεις όταν για την εκπαίδευσή του χρησιμοποιούσαμε όλα τα χαρακτηριστικά. Αυτό συνεπάγεται πως υπήρχαν περιπτώσεις όπου στην κορυφή της ταξινόμησης είχαμε χαρακτηριστικά από την ίδια υποκατηγορία και αυτό όχι μόνο δεν έβλαπτε το μοντέλο, αλλά αντιθέτως το ενίσχυε. Θα είχε ενδιαφέρον, σε μελλοντική έρευνα να γίνει μαθηματική μελέτη της συσχέτισης όλων των χαρακτηριστικών μεταξύ τους και να επιχειρηθεί ξανά η επιλογή τους, με λιγότερο προφανή κριτήρια αυτή τη φορά.

- **Ταξινόμηση χαρακτηριστικών:**
  Οι δύο μέθοδοι ταξινόμησης χαρακτηριστικών, forward selection και backward elimination, παρουσίασαν παρόμοια αποτελέσματα σε όλο τον όγκο των πειραμάτων. Ακόμη και όταν

μία μεθοδος υπερίσχυε της άλλης, η διαφορά στην τιμή αξιολόγησης του προκύπτοντος μοντέλου δεν ήταν πολύ μεγάλη. Αξίζει να σημειωθεί όμως, πως συχνά οι δύο μέθοδοι πρότειναν διαφορετικά χαρακτηριστικά ως βέλτιστα.

- **Αλγόριθμοι μηχανικής μάθησης:**
  Ανάμεσα στους αλγορίθμους που χρησιμοποιήθηκαν σε αυτή τη διπλωματική, δεν παρατηρήθηκε αξιοσημείωτη διαφορά στην αποδοτικότητα, εκτός από τον Linear Regression που ήταν αναμενόμενο να βρεθεί χαμηλότερα. Το γεγονός αυτό είναι θετικό μιας και δείχνει συνέπεια και αξιοπιστία των πειραματικών αποτελεσμάτων. Σε επόμενα πειράματα, θα μπορούσε να γίνει χρήση περισσότερων αλγορίθμων. Έχει προταθεί και παραπάνω η χρήση του αλγορίθμου Support Vector Machines (SVM).

Όπως αναφέρεται στην αρχή αυτής της εργασίας, ο σκοπός της είναι να προσδιοριστούν τα χαρακτηριστικά που συμβάλλουν με βέλτιστο τρόπο στην επιτυχία της θεραπευτικής μεδόθου exon skipping. Παρά τα ποικίλα πειραματικά αποτελέσματα του Κεφαλαίου 6, είναι εύκολο να παρατηρηθεί πως η ενέργεια πρόσδεσης του ολιγονουκλεοτιδίου dG και η ποσότητα των GC βάσεων είναι οι δύο κατηγορίες χαρακτηριστικών που πρωταγωνιστούν στις πρώτες θέσεις ιεραρχίας. Το μήκος του ολιγονουκλεοτιδίου, καθώς και του εξωνίου, φαίνεται να έχουν επίσης σημαντικό ρόλο σε αυτή τη διαδικασία. Αυτό μπορεί να σχετίζεται με τα δύο προηγούμενα χαρακτηριστικά. Μεγαλύτερο μήκος μας δίνει περισσότερους δεσμούς μεταξύ των συμπληρωματικών βάσεων, αλλά και πιθανότατα περισσότερες GC βάσεις.

Τα παραπάνω ισχύουν σχεδόν σε απόλυτο βαθμό για το σύνολο δεδομένων των PMO ολιγονουκλεοτιδίων. Στην περίπτωση των 2OMePS ολιγονουκλεοτιδίων μπορούμε να δούμε και άλλα χαρακτηριστικά να συμβάλλουν σε μεγάλο βαθμό στην αποδοτικότητα του exon skipping. Τα χαρακτηριστικά αυτά είναι η προσβασιμότη-

τα των βάσεων κατά τη διαδικασία του ματίσματος, η απόσταση του σημείου πρόσδεσης του ολιγονουκλεοτιδίου από την 5′ θέση ματίσματος (ορισμένεςφορές και από την 3′), η παρουσία των θέσεων ματίσματος (ESEs) και το σημείο πρόσδεσης της SR πρωτεΐνης Tra2$\boldsymbol{\beta}$. Δεν αποκλείεται τα χαρακτηριστικά αυτά να μην είναι ίδια για κάθε εξώνιο. Δηλαδή, τα ολιγονουκλεοτίδια και το σημείο πρόσδεσής τους να πρέπει να αντιστοιχούν σε κάθε συγκεκριμένο εξώνιο που πρέπει να παραβλεφθεί.

Τα παραπάνω αποτελέσματα έρχονται σε συμφωνία με τις προηγούμενες έρευνες που έχουν διεξαχθεί και τις οποίες έχουμε αναφέρει αναλυτικά στο Κεφάλαιο 4. Η καινοτομία που προκύπτει από αυτή τη διπλωματική είναι η έντονη σημασία του ποσοστού των GC βάσεων και στις δύο χημείες ολιγονουκλεοτιδίων καθώς και το αναμφισβήτητο γεγονός πως ο συνδυασμός αυτού του χαρακτηριστικού με την ενέργεια πρόσδεσης δίνει ένα πολύ καλό αποτέλεσμα.

Βιβλιογραφία

[1] Basil T. Darras, Caroline C. Menache-Starobinski, Veronica Hinton, and Louis M. Kunkel *Dystrophinopathies*, 2015 doi:10.1016/B978-0-12-417044-5.00030-5

[2] Panagiotis Papasaikas, Juan Valcárcel *The Spliceosome:The Ultimate RNA Chaperone and Sculptor*, Trends in Biochemical Sciences, January 2016, Vol. 41, No. 1

[3] Hoffman EP, Bronson A, Levin AA, et al. *Restoring dystrophin expression in duchenne muscular dystrophy muscle progress in exon skipping and stop codon read through.* Am J Pathol. 2011;179(1):12-22. doi:10.1016/j.ajpath.2011.03.050

[4] Graveley BR. *Sorting out the complexity of SR protein functions.* RNA. 2000;6(9):1197-1211. doi:10.1017/s1355838200000960

[5] Pace NR, Thomas BC, Woese CR *Probing RNA structure, function, and history by comparative analysis.* The RNA World Cold Spring Harbor Laboratory PressGesteland RF, Cech TR, Atkins JF, 2 1999, 113-141

[6] Reuter JS, Mathews DH. *RNAstructure: software for RNA secondary structure prediction and analysis.* BMC Bioinformatics. 2010;11:129. Published 2010 Mar 15. doi:10.1186/1471-2105-11-129

[7] Mathews DH, Turner DH. *Prediction of RNA secondary structure by free energy minimization.* Curr Opin Struct Biol. 2006;16(3):270-278. doi:10.1016/j.sbi.2006.05.010

[8] Lorenz R, Bernhart SH, Höner Zu Siederdissen C, et al. *ViennaRNA Package 2.0.* Algorithms Mol Biol. 2011;6:26. Published 2011 Nov 24. doi:10.1186/1748-7188-6-26

[9] Fairbrother WG, Yeo GW, Yeh R, et al. *RESCUE-ESE identifies candidate exonic splicing enhancers in vertebrate exons.* Nucleic Acids Res. 2004;32(Web Server issue):W187-W190. doi:10.1093/nar/gkh393

[10] Cartegni L, Wang J, Zhu Z, Zhang MQ, Krainer AR. *ESEfinder: A web resource to identify exonic splicing enhancers.* Nucleic Acids Res. 2003;31(13):3568-3571. doi:10.1093/nar/gkg616

[11] Desmet FO, Hamroun D, Lalande M, Collod-Béroud G, Claustres M, Béroud C. *Human Splicing Finder: an online bioinformatics tool to predict splicing signals.* Nucleic Acids Res. 2009;37(9):e67. doi:10.1093/nar/gkp215

[12] Aartsma-Rus A, van Vliet L, Hirschi M, et al. *Guidelines for antisense oligonucleotide design and insight into splice-modulating mechanisms.* Mol Ther. 2009;17(3):548-553. doi:10.1038/mt.2008.205

[13] Aartsma-Rus A, Houlleberghs H, van Deutekom JC, van Ommen GJ, 't Hoen PA. *Exonic sequences provide better targets for antisense oligonucleotides than splice site sequences in the modulation of Duchenne muscular dystrophy splicing.* Oligonucleotides. 2010;20(2):69-77. doi:10.1089/oli.2009.0215

[14] Wilton SD, Fall AM, Harding PL, McClorey G, Coleman C, Fletcher S. *Antisense oligonucleotide-induced exon skipping across the human dystrophin gene transcript.* Mol Ther. 2007;15(7):1288-1296. doi:10.1038/sj.mt.6300095

[15] Harding PL, Fall AM, Honeyman K, Fletcher S, Wilton SD. *The influence of antisense oligonucleotide length on dystrophin exon skipping.* Mol Ther. 2007;15(1):157-166. doi:10.1038/sj.mt.6300006

[16] Pramono ZA, Wee KB, Wang JL, et al. *A prospective study in the rational design of efficient antisense oligonucleotides for exon skipping in the DMD gene.* Hum Gene Ther. 2012;23(7):781-790. doi:10.1089/hum.2011.205

[17] Popplewell LJ, Trollet C, Dickson G, Graham IR. *Design of phosphorodiamidate morpholino oligomers (PMOs) for the induction of exon skipping of the human DMD gene.* Mol Ther. 2009;17(3):554-561. doi:10.1038/mt.2008.287

[18] Echigoya Y, Mouly V, Garcia L, Yokota T, Duddy W. *In silico screening based on predictive algorithms as a design tool for exon skipping oligonucleotides in Duchenne muscular dystrophy.* PLoS One. 2015;10(3):e0120058. Published 2015 Mar 27. doi:10.1371/journal.pone.0120058

[19] Stadler MB, Shomron N, Yeo GW, Schneider A, Xiao X, Burge CB. *Inference of splicing regulatory activities by sequence neighborhood analysis.* PLoS Genet. 2006;2(11):e191. doi:10.1371/journal.pgen.0020191

[20] Malueka RG, Takaoka Y, Yagi M, et al. *Categorization of 77 dystrophin exons into 5 groups by a decision tree using indexes of splicing regulatory factors as decision markers.* BMC Genet. 2012;13:23. Published 2012 Mar 31. doi:10.1186/1471-2156-13-23

[21] Wee KB, Pramono ZAD, Wang JL, MacDorman KF, Lai PS, et al. *Dynamics of Co-Transcriptional Pre-mRNA Folding Influences the Induction of Dystrophin Exon Skipping by Antisense Oligonucleotides.* PLOS ONE 3(3): e1844. https://doi.org/10.1371/journal.pone.0001844

[22] Sergios Theodoridis, Konstantinos Koutroumbas *2008. Pattern Recognition, Fourth Edition (4th. ed.).* Academic Press, Inc., USA.

[23] Duda, R. O., Hart, P. E.,, Stork, D. G. *(2001). Pattern Classification.* New York: Wiley. ISBN: 978-0-471-05669-0

[24] Vasiliki Triantafyllidou *Computational methods for optimization of therapeutic approaches for Duchenne Muscular Dystrophy*