ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ

ΣΧΟΛΗ ΕΦΑΡΜΟΣΜΕΝΩΝ ΜΑΘΗΜΑΤΙΚΩΝ ΚΑΙ ΦΥΣΙΚΩΝ ΕΠΙΣΤΗΜΩΝ

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

«Υπολογιστικές μέθοδοι για βελτιστοποίηση γενετικών θεραπειών για τη μυϊκή δυστροφία Duchenne»

Τριμελής εξεταστική επιτροπή
Επίκουρος Καθηγητής Κωνσταντίνος Κουσουρής (Επιβλέπων)
Τομέας Φυσικής.

Αναπληρωτής Καθηγητής Αλέξανδρος Γεωργακήλας
Τομέας φυσικής

Καθηγητής Γεώργιος Τσιπολίτης
Τομέας Φυσικής

Αθήνα, 2019

Περίληψη

Η Μυϊκή Δυστροφία Duchenne, μια θανατηφόρος γενετική ασθένεια που πλήττει 1 στα 5000 νεογέννητα αγόρια παγκοσμίως, οφείλεται σε μεταλλάξεις στο μεγαλύτερο ανθρώπινο γονίδιο, το γονίδιο της δυστροφίνης. Τα συμπτώματα εμφανίζονται από την ηλικία των 3 ετών και το μέσο προσδόκιμο ζωής υπολογίζεται στα 25 χρόνια χωρίς να υπάρχει γνωστή θεραπεία μέχρι σήμερα.

Πολύ ελπιδοφόρα θεωρείται η μέθοδος του exon skipping με CRISPR από τις γενετικές θεραπείες που έχουν δοκιμαστεί. Κατά το exon skipping γίνεται παράκαμψη ενός ή περισσότερων εξονίων από το μηχανισμό ματίσματος ώστε να αποκατασταθεί το πλαίσιο ανάγνωσης και να παραχθεί μια μικρότερη σε μήκος πρωτεΐνη αλλά λειτουργική. Αυτό επιτυγχάνεται με τη χρήση της πρωτεΐνης Cas9 που έχουν τα βακτήρια σε συνδυασμό με ένα sgRNA όπου οδηγεί την πρωτεΐνη στο επιθυμητό σημείο του γονιδίου ώστε η πρωτεΐνη να κόψει το DNA. Αφού κοπεί το η αλυσίδα στο επιθυμητό σημείο, για να μην καταστραφεί η αλυσίδα, τα δύο κομμάτια της αλυσίδας ξανακολλάνε μεταξύ τους δημιουργώντας συνήθως μια μετάλλαξη η οποία μπορεί να οδηγήσει σε μια νέα αλληλουχία που δεν κωδικοποιεί το εξόνιο.

Στη παρούσα εργασία γίνεται προσπάθεια μελέτης και πρόβλεψης των χαρακτηριστικών των sgRNA που τους επιτρέπουν να είναι αποτελεσματικά για τη θεραπεία της Μυϊκής Δυστροφίας Duchenne. Από πειραματικά δεδομένα σε ανθρώπινα κύτταρα, γίνεται προσπάθεια εκπαίδευσης ενός αλγορίθμου μηχανικής εκμάθησης για τη δημιουργία ενός μοντέλου πρόβλεψης.

Παρατηρήθηκε ότι η δομή της αλυσίδας στο σημείο που στοχεύουμε και το εαν η θέση του σημείου είναι στην κωδικοποιούμενη περιοχή (εξόνια) αποτελούν τα κυριότερα χαρακτηριστικά των αλυσίδων που πέτυχαν το επιθυμητό αποτέλεσμα στα κύτταρα.

Η κατανόηση του τρόπου εισόδου στα κύτταρα, ένα μεγαλύτερο σετ δεδομένων ή και μοντελοποίηση σε άλλους οργανισμούς θα μπορούσαν να είναι επόμενοι στόχοι για βελτίωση του μοντέλων που δημιουργήθηκαν.

Ευχαριστίες

Αρχικά θα ήθελα να ευχαριστήσω τον κ. Παλιούρα για την ευκαιρία να εκπονήσω την διπλωματική μου εργασία στο εργαστήριο τους και για την βοήθεια και κατανόηση σε όλη τη διάρκεια. Επίσης τον κ. Νεντίδη για την συνεργασία όλη αυτή την περίοδο. Θα ήθελα να ευχαριστήσω τον κ Κουσουρή που δέχτηκε να είναι επιβλέπων της εργασίας αλλά κυρίως για την στήριξη στην επιλογή του θέματος. Τέλος να ευχαριστήσω τους κ. Γεωργακήλα και κ. Τσιπολίτη που δέχτηκαν να συμμετάσχουν στην τριμελή εξεταστική επιτροπή.

Τριμελής εξεταστική επιτροπή
Επίκουρος Καθηγητής Κωνσταντίνος Κουσουρής (Επιβλέπων)
Τομέας Φυσικής.

Αναπληρωτής Καθηγητής Αλέξανδρος Γεωργακήλας
Τομέας φυσικής

Καθηγητής Γεώργιος Τσιπολίτης
Τομέας Φυσικής

Περιεχόμενα

ΕΥΡΕΤΗΡΙΟ ΟΡΩΝ

MD: Muscular Dystrophy

DMD: Duchenne Muscular Dystrophy

BMD: Becker Muscular Dystrophy

CRISPR: Clustered regularly Interspaced Palindromic Repeats

sgRNA: Single guide RNA

PAM: Protospacer Adjacent Motif

XGBoost: Extreme Gradient Boosting

SMOTE: Synthetic Minority Over-sampling Technique

ΕΥΡΕΤΗΡΙΟ ΕΙΚΟΝΩΝ

1 Ορισμός Προβλήματος Και Δομή Κειμένου

1.1 Ορισμός Προβλήματος

Σκοπός της παρούσας εργασίας είναι η πρόβλεψη μέσω αλγορίθμων Μηχανικής Μάθησης αποτελεσματικών Single Guide RNA (sgRNA) για τη γονιδιακή θεραπεία της Μυϊκής Δυστροφίας Duchenne (DMD) με CRISPR cas9. Τα sgRNAs χρησιμοποιούνται για οδηγήσουν την πρωτεΐνη cas9 στο τμήμα του DNA που μας ενδιαφέρει να επιτευχθεί exon-skipping ώστε να αυξηθεί η παραγωγή δυστροφίνης σε ασθενείς της μυϊκής δυστροφίας Duchenne. Μέσω εκπαίδευσης αλγορίθμων μηχανικής μάθησης με δημοσιευμένα δεδομένα αναζητήσαμε να βρούμε ποια χαρακτηριστικά ενός sgRNA είναι περισσότερο σημαντικά για την αποτελεσματική τους χρήση στη θεραπεία της DMD.

1.2 Δομή Κειμένου

Στο 2ο κεφάλαιο της εργασίας γίνεται παρουσίαση της πρωτεΐνης της δυστροφίνης, του γονιδίου που είναι υπεύθυνο για τη παραγωγή της και της ασθένειας της Μυϊκής Δυστροφίας Duchenne και παρουσιάζονται οι θεραπείες που δοκιμάζονται με έμφαση στο exon skipping με CRISPR. Τέλος παρουσιάζονται τα sgRNAs και η σχετική βιβλιογραφία για την χρήση τους στην DMD καθώς και η σχετική βιβλιογραφία για τα εργαλεία πρόβλεψης για sgRNAs που υπάρχουν διαθέσιμα. Στο 3ο κεφάλαιο παρουσιάζεται ο τρόπος συλλογής και οργάνωσης των δεδομένων, γίνεται μια εισαγωγή στη μηχανική εκμάθηση με έμφαση στους αλγόριθμο XGB που υλοποιήσαμε σε python. Στο 4ο κεφάλαιο παρουσιάζονται αναλυτικά τα αποτελέσματα εκπαίδευσης των διάφορων αλγορίθμων και η μεταξύ τους σύγκριση για τα διαθέσιμα δεδομένα. Στο 5ο κεφάλαιο γίνεται ανάλυση των αποτελεσμάτων και προτείνονται τρόποι βελτίωσης των μοντέλων.

2 DMD KAI CRISPR

2.1 Μυϊκή Δυστροφία Duchene (DMD)

Οι Μυϊκές Δυστροφίες (MD) είναι γενετικές παθήσεις που χαρακτηρίζονται από προοδευτική μυϊκή αδυναμία και φθορά κυρίως των σκελετικών μυών και σε κάποιες περιπτώσεις και του καρδιακού μυ.

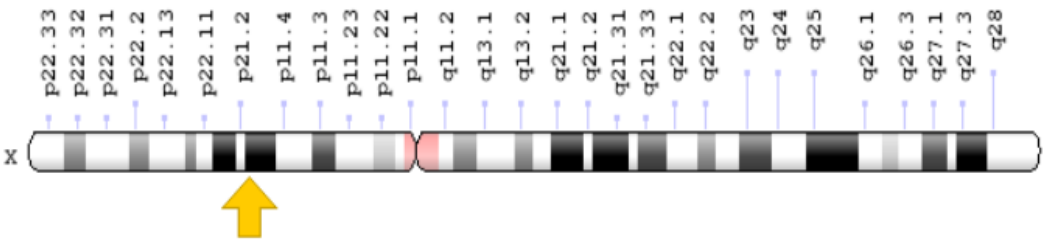
Σύμφωνα με το NINDS (National Instute of Neurological Disorders and Stroke) του NIH υπάρχουν εννέα (9) κατηγορίες Μυϊκής Δυστροφίας:

- Duchenne
- Beckers
- Οσφυοπυελική
- Συγγενής
- Μυοτονική
- Προσωποωμοπλατοβραχιόνιος
- Περιφερική
- Οφθαλμοφαρυγγική
- Emery – Dreifuss [1]

Η μυϊκή δυστροφία Duchenne(DMD) έχει πάρει το όνομά της από τον νευρολόγο Guilaume Duchenne που τη περιέγραψε στο βιβλίο του[2]. Υπήρχαν αναφορές για την ασθένεια ήδη από το Edward Meryon (1852) και από τον John Little (1853) [3].

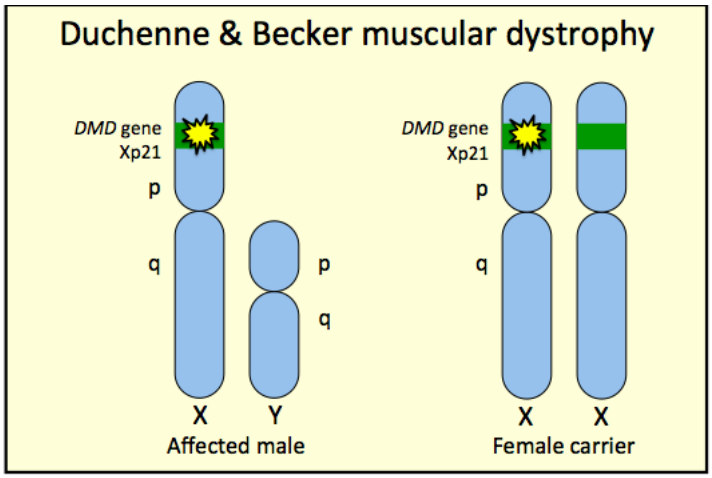
Μέχρι το 1968 θεωρούταν εγκεφαλική ασθένεια, όμως το 1871 αποδείχθηκε η μυϊκή φύση της ασθένειας και ο Duchenne πρότεινε κάποιες μεθόδους διάγνωσης[5].

Στις αρχές της δεκαετίας του 1980 με τη πρόοδο της κυτταρογενιτικής έγινε χαρτογράφηση του γονιδίου υπεύθυνο για την ασθένεια στη ζώνη 21 του χρωμοσώματος Χ. Το γονίδιο DMD έχει 14kb κωδικοποιούμενη περιοχή, 79 εξώνια και εκτείνεται 2.2 mega-βάσεις στο κοντό βραχίωνα του Χ χρωμοσώματος(1% του χρωμοσώματος Χ) (εικόνα 1).



Εικόνα 1: γονίδιο DMD στο κοντό βραχίονα του Χρωμοσώματος Χ

Λόγω της θέσης του γονιδίου στο DNA η εμφάνιση της ασθένειας είναι πιο συχνή σε αγόρια(1 στις 5000 γεννήσεις σε σύγκριση με τα κορίτσια(1 στις 50εκ. Γεννήσεις) (εικόνα 2) [4,5].



Εικόνα 2: Αρσενικά θηλυκα χρωμοσώματα X

2.2 Το Γονίδιο Της δυστροφίνης

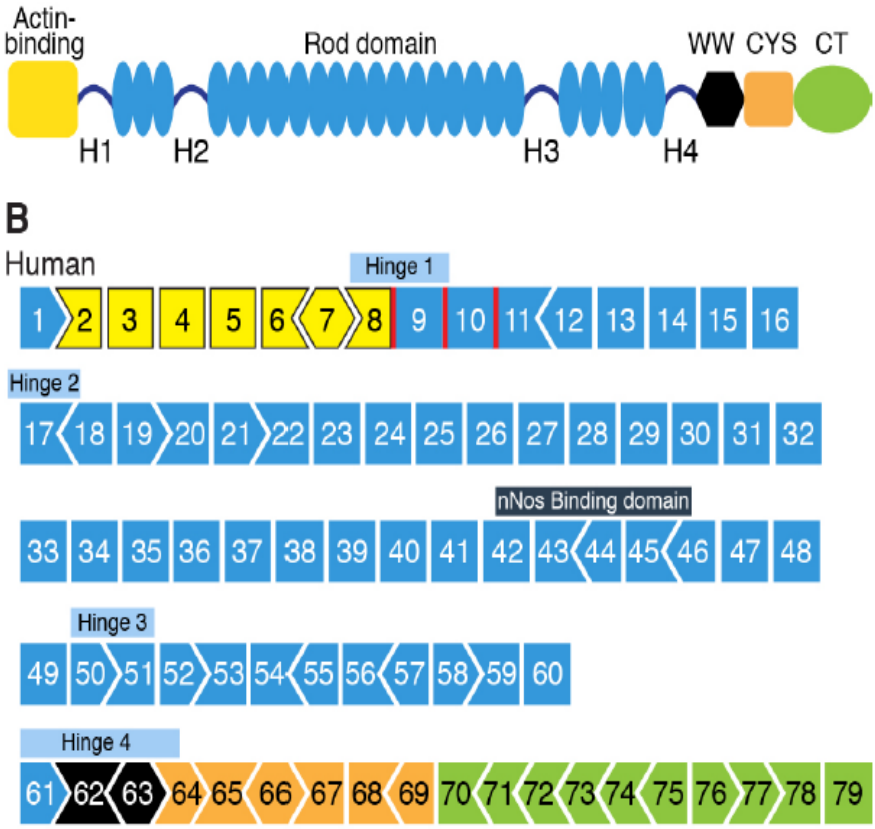
Τα ιντρόνια του γονιδίου είναι μεγάλα, με κάποια κοντά στα 5΄ 3ο του γονιδίου και μεταξύ των εξονίων 44 και 45 που είναι σχεδόν 200 kb. Η γονιδιωματική (εξονική και ιντρονική) αλληλουχία του γονιδίου DMD μπορεί να ληφθεί από το National Center for Biotechnology Information (NCBI) / Genbank site, αλλά και από τις ιστοσελίδες του Leiden Muscular Dystrophy. Το γονιδιακό προϊόν DMD της δυστροφίνης αναγνωρίζεται σε Western blots (κηλίδες) πρωτεϊνών των ανθρώπινων σκελετικών μυών χρησιμοποιώντας αντισώματα αντιδυστροφίνης.

Με τη χρήση της ανοσοκυτταροχημείας, η δυστροφίνη υπήρξε εντοπισμένη στην κυτταροπλασμική πλευρά της πλασμικής μεμβράνης των μυϊκών ινών. Η δυστροφίνη έχει τέσσερις λειτουργικές περιοχές:

- την αμινο-τερματική ή ακτίνας περιοχή (amino-terminal domain)
- την περιοχή της ράβδου (rod domain)
- την πλούσια σε κυστεΐνη περιοχή (cysteine-rich domain)
- την καρβοξυ-τερματική περιοχή (carboxy-terminal domain)

Η αμινο-τερματική περιοχή (εξόνια 1-8) δεσμευεται με ακτίνα μέσω τριών ακτίνων υψηλής συγγένειας θέσεων δέσμευσης. Η περιοχή της ράβδου (εξόνια 9-63) είναι μεγάλη, αποτελούμενη από ομόλογες επαναλήψεις τύπου «σπεκτρίνης» σχηματίζοντας ανα-ελικοειδή δομή. Οι επαναλήψεις διακόπτονται από δύο μη ελικοειδή περιοχές γνωστές ως «αρθρώσεις», το οποίο πιστεύεται

ότι προσδίδουν ευελιξία στην περιοχή της ράβδου κατά τη διάρκεια της μυϊκής σύσπασης. Ο τομέας πλούσιος σε κυστεΐνη (εξόνια 64-69) είναι μια περιοχή του γονιδίου DMD κοντά στο καρβοξύ άκρο που φαίνεται να είναι η περιοχή δέσμευσης της δυστρογλυκάνης. Ο καρβοξυ-τερματικός τομέας (εξόνια 70-79) δεσμεύεται με ένα σύμπλοκο πρωτεϊνών που συνδέει το κυτταροσκελετό με πρωτεΐνες μεμβράνης που με τη σειρά τους δεσμεύονται με πρωτεΐνες στην εξωκυτταρική μήτρα.



Εικ. 3 Η δομή του γονιδίου της δυστροφίνης από τον Erik Olson

Η δυστροφίνη είναι μέρος ενός μεγάλου, στενά συνδεδεμένου συμπλόκου γλυκοπρωτεΐνης που περιέχει άλλες πρωτεΐνες, τις λεγόμενες πρωτεΐνες που σχετίζονται με τη δυστροφίνη (DAP). Το γονίδιο DMD μπορεί να παράγει αρκετές ισομορφές κυττάρου εξειδικευμένου τύπου δυστροφίνης διαφορετικών μοριακών βαρών, το καθένα από τα οποία οδηγείται από ένα διακριτό υποκινητή. Αυτοί οι υποκινητές οδηγούν τη μεταγραφή του γονιδίου δυστροφίνης από το δικό τους πρώτο εξόνιο και δημιουργούν ισομορφές δυστροφίνης διαφόρων μοριακών βαρών.

Οι εγκεφαλικές, μυϊκές και Purkinje κυτταρικές δυστροφίνες προβλέπεται να έχουν μοριακό βάρος 427 kilodaltons (kDa), ενώ του νευρικού / γενικού και Schwann κυττάρου οι δυστροφίνες κωδικοποιούνται από καρβοξυ-τερματικές μεταγραφές και έχουν προβλέψιμα μοριακά βάρη 71 και 116 kDa. χαρακτηρίζονται ως B / Dp427, M / Dp427, P / Dp427, G / Dp71 και S / Dp116 δυστροφίνες, αντίστοιχα. Οι δυστροφίνες του αμφιβληστροειδούς (R / Dp260) και του εγκεφάλου / νεφρού (BK / Dp140) έχουν επίσης περιγραφεί. Άλλες

παραλλαγές δυστροφίνης (δηλαδή εναλλακτικά συραμμένες μεταγραφές) προκύπτουν από εναλλακτικό σύραμμα των εξονίων 71, 74, 78 και 79 και εκφράζονται σε ιστούς μυών, εγκεφάλου και καρδιάς, εναλλακτική σύνδεση συμβαίνει στις περιοχές που εμπλέκονται στη σύνδεση της δυστροφίνης με το σύμπλεγμα DAP και, ως εκ τούτου, μπορεί να ρυθμίσει τη σύνδεσή τους με διάφορους ιστούς.

Η μυϊκή δυστροφίνη εκφράζεται σε σκελετικούς (όλοι οι τύποι μυϊκών ινών), ομαλούς και καρδιακούς μύες, καθώς και στο εξωτερικό πλεγματοειδές στρώμα του αμφιβληστροειδή. Η έκφραση αυτής και πιθανώς άλλων ισομορφών ρυθμίζεται αναπτυξιακά. Η δυστροφίνη εντοπίζεται πρώτα στον ανθρώπινο εμβρυϊκό μυ στις 9 εβδομάδες κύησης και η έκφρασή της αυξάνεται καθώς οι μυοβλάστες διαφοροποιούνται σε πολυπυρηνικούς μυοσωλήνες. Η υψηλότερη έκφραση της δυστροφίνης εγκεφάλου βρίσκεται στο νεοσύκλατο και τον ιππόκαμπο. Η κυτταρική δυστροφίνη της Purkinje περιλαμβάνει την περισσότερη ή σχεδόν όλη την παρεγκεφαλιδική δυστροφίνη.

Το Dp71 εκφράζεται σε γλοιακά κύτταρα, σπλάχνα και ώριμο καρδιακό και εμβρυϊκό σκελετικό μυ, αλλά όχι σε ώριμο σκελετικό μυ. Ενώ η μυϊκή δυστροφίνη ανιχνεύεται στη μεμβράνη πλάσματος των σκελετικών μυϊκών ινών, φαίνεται να είναι ιδιαίτερα άφθονη σε νευρομυϊκές και μυοτεννικές διασταυρώσεις. Οι διαφορετικές ισομορφές πρέπει να αλληλοεπιδρούν με διαφορετικές πρωτεΐνες στους διάφορους ιστούς στους οποίους εκφράζονται. Στην DMD, οι πιο σημαντικές από αυτές τις ισομορφές είναι οι τρεις πλήρους μήκους, μεγάλου μοριακού βάρους (427 kDa) δυστροφίνες, που υπάρχουν στον σκελετικό και λείο μυ (μυϊκός τύπος, Dp427), στον εγκεφαλικό φλοιό και τον ιππόκαμπο(τύπου εγκεφάλου, Dp427) και στα Purkinje κύτταρα(τύπος Purkinje, Dp427).

Από τις μεταλλάξεις DMD / BMD που εντοπίστηκαν μέχρι τώρα, οι περισσότερες είναι διαγραφές που ανιχνεύθηκαν σε περίπου 50-65% των αρρένων με DMD και 65-70% των αρσενικών με BMD. Μερικές επαναλήψεις του γονιδίου έχουν επίσης αναφερθεί σε ένα μικρό ποσοστό ασθενών με DMD και BMD (περίπου 5-10%). Οι διαγραφές ή οι επικαλύψεις περιλαμβάνουν 1 ή περισσότερα εξόνια του γονιδίου DMD. Στο DMD και BMD, μερικές διαγραφές και αντιγραφές συγκεντρώνονται σε ένα ανασυνδυασμό δύο καυτών σημείων, ένα εγγύς στο 5' άκρο του γονιδίου στην περιοχή των εξονίων 3-7 (σχεδόν 30%) και ένα ακόμη περιφερικό, που περιλαμβάνει τα εξόνια 44-53 (σχεδόν 70%), με επιπλέον διαγραφές στο υπόλοιπο του γονιδίου. Τα καυτά σημεία φαίνεται να εμφανίζονται στις περιοχές του γονιδίου όπου τα εσόνια είναι μακρά, για παράδειγμα το εσόνιο 44 μεταξύ των εξονίων 44 και 45 είναι εξαιρετικά μεγάλο, τα σημεία διακοπής του γονιδιώματος του 5' hotspot (καυτό σημείο) βρίσκονται εντός των εσονίων 2 και 7. Στο υπόλοιπο 25-35% των ασθενών με DMD χωρίς ανιχνεύσιμες διαγραφές ή επαναλήψεις και 20-30% των ανδρών με BMD,οι μοριακές βλάβες αντιπροσωπεύουν απλές αλλαγές βάσης, μικρές διαγραφές ή εισαγωγές ή σφαλμάτων συρραφής. Οι ανόητες μεταλλάξεις συμβαίνουν συχνότερα στην DMD από ότι στην BMD, σε έκταση 20-25% των περιπτώσεων DMD σε σύγκριση με λιγότερο από το 5% των περιπτώσεων BMD.

Σε DMD και BMD, ένα σημαντικό τμήμα αλλαγών αλληλουχίας είναι μεταλλάξεις θέσης συρραφής και μικρές μεταλλάξεις εισαγωγής / εξάλειψης (in/ del). Σε καμία από τις παθήσεις DMD και BMD δεν υπάρχουν μεταλλάξεις τύπου missense μιας κοινής αιτίας. Ατυπικές μεταλλάξεις (βαθιά εσονικά, εκείνα που σπάνια εμφανίζονται στις μη μεταφρασμένες περιοχές 5' και 3') αντιπροσωπεύουν λιγότερο από το 1% όλων των μεταλλάξεων του γονιδίου DMD. Δημοσιευμένες μελέτες απέτυχαν να αποκαλύψουν οποιαδήποτε εμφανή συσχέτιση μεταξύ του μεγέθους των διαγραφών του γονιδίου της δυστροφίνης και τη σοβαρότητα και την εξέλιξη του φαινοτύπου DMD / BMD. Η μοριακή βάση της μυϊκής δυστροφίας του Duchenne έναντι της Becker φαίνεται να σχετίζεται με την διάσπαση ή συντήρηση του μεταφραστικού πλαισίου ανάγνωσης αμινοξέων με τη διαγραφή των μεταλλάξεων. Με άλλα λόγια, κατά τη διάρκεια της σύνθεσης του ώριμου mRNA, τα άκρα σύνδεσης των εξονίων (μετά από σύνδεση των εσονίων) πρέπει να είναι σε φάση, προκειμένου να διατηρηθεί το σωστό μεταφραστικό πλαίσιο ανάγνωσης ανοικτό.

Μια διαγραφή που αντιπαραθέτει εξόνια που μετατοπίζουν το μεταφραστικό πλαίσιο ανάγνωσης (εκτός του πλαισίου διαγραφής) συνήθως οδηγεί σε ασταθές mRNA και τελικά περικόπτει σοβαρά ένα μόριο δυστροφίνης, η οποία υποβαθμίζεται γρήγορα στο κύτταρο και οδηγεί σε ένα πιο σοβαρό φαινότυπο DMD. Το φαινόμενο αυτό είναι γνωστό ως διαμεσολάβηση που προκαλείται από σφάλματα και έχει ως στόχο να εξαλείψει την παθολογική αλληλόμορφο για να αποφευχθούν κυρίαρχα αρνητικά αποτελέσματα. Άλλες μεταλλάξεις διαταράσσοντας το πλαίσιο ανάγνωσης περιλαμβάνουν διακοπή μεταλλάξεων και μερικών μεταλλάξεων συρραφής καθώς και επαναλήψεις. Μια διαγραφή η οποία αντιπαραθέτει εξόνια που διατηρούν το μεταφραστικό πλαίσιο ανάγνωσης (διαγραφή πλαισίου) θα οδηγήσει σε μια εσωτερικά διαγραμμένη αλλά ημιεστιακή πρωτεΐνη δυστροφίνης (με ανέπαφο αμινο και καρβοξυ-άκρα) τα οποία μπορούν να παραμείνουν σε κάποια ποιότητα στο κύτταρο και έτσι να οδηγήσει σε έναν πιο ήπιο φαινότυπο της BMD.

Άλλες μεταλλάξεις εντός πλαισίου περιλαμβάνουν επαναλήψεις, μερικές μεταλλάξεις συναρμολόγησης και τις περισσότερες αλλαγές μη μονόπλευρης βάσης. Ωστόσο, εξαιρέσεις από την "υπόθεση πλαισίου ανάγνωσης" εμφανίζεται σε περίπου 8% των περιπτώσεων. Όταν συγκρίνονται οι παραλείψεις εντός πλαισίου σε ασθενείς με BMD, κατά κανόνα ορισμένες περιοχές της πρωτεΐνης της δυστροφίνης φαίνονται πιο κρίσιμες από άλλες διαγραφές στις περιοχές που κωδικοποιούν τις αμινο-τερματικές ή καρβοξυ-τερματικές περιοχές της δυστροφίνης έχουν ως αποτέλεσμα πιο αυστηρούς φαινότυπους από τις μεταλλάξεις που επηρεάζουν την περιοχή της ράβδου.

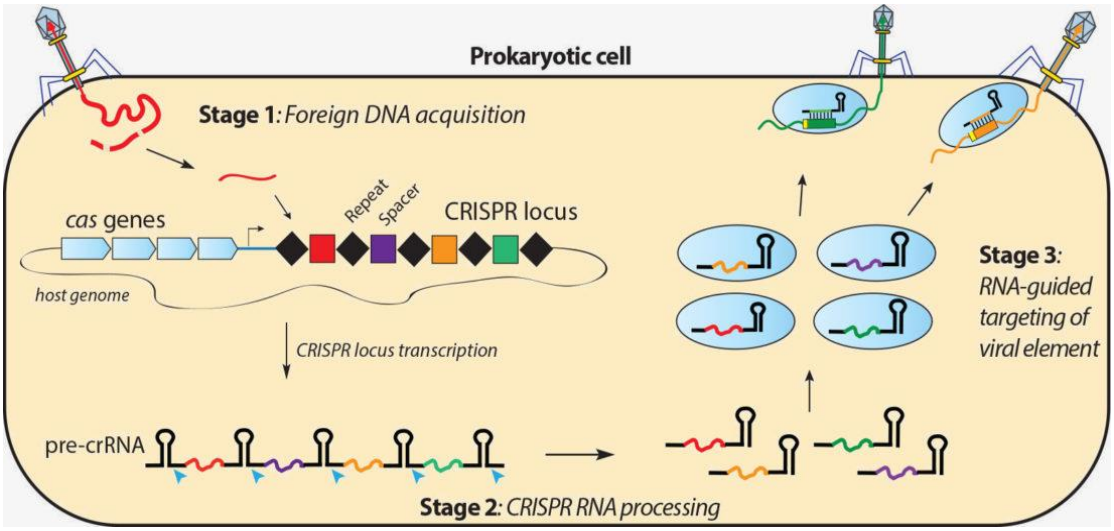
Εντούτοις, υπάρχουν εξαιρέσεις από αυτόν τον κανόνα και συζητούνται παρακάτω. Εξαιρέσεις συμβαίνουν συχνότερα με διαγραφές εξονίων 3 έως 7 ή εξόνιο 45, οι εξαιρέσεις αναλύονται επίσης παρακάτω. Αν και αυτές οι μεταλλάξεις είναι εκτός πλαισίου μπορεί να οδηγήσουν σε BMD, DMD ή σε ένα ενδιάμεσο φαινότυπο. Γεγονότα που αφορούν την παράκαμψη των εξονίων ή την ενεργοποίηση νέων κρυπτικών μεταφραστικών τοποθεσιών εκκίνησης μπορούν να δημιουργήσουν καταστάσεις στις οποίες προφανώς εξέρχονται των απαγορευμένων πλαισίων και συμπεριφέρονται ως διαγραφές πλαισίου ως προς το πλαίσιο ή αντίστροφα.

Εναλλακτικά, το οι γραμματικοί μετασχηματισμοί (πχ. μετατόπιση ή επανατοποθέτηση ριβοσωματικών πλαισίων) μπορεί να βοηθήσουν στην αποκατάσταση του πλαισίου ανάγνωσης του mRNA σε ασθενείς BMD με εκτός πλαισίου 3 έως 7 διαγραφές εξονίων. Επιπλέον, μεταβολές στη σοβαρότητα του φαινοτύπου μεταξύ ασθενών με παρόμοιες μεταλλάξεις (π.χ. διαγραφή του εξονίου 45) έχουν αναφερθεί, υποστηρίζοντας την συνεισφορά άλλων παραγόντων στην φαινοτυπική έκφραση αυτών των μεταλλάξεων. Επιπλέον, πολύ μεγάλες διαγραφές ή διαγραφές πρωτεΐνης σε δεσμευτικές περιοχές (π.χ. πλούσια σε κυστεΐνη περιοχή), ακόμη και εάν είναι εντός πλαισίου, μπορεί να οδηγήσουν σε σοβαρό φαινότυπο.

2.3 CRISPR Στα Βακτήρια

Ιστορικά, η εκτίμησή μας για τα μικροβιακά ανοσοποιητικά συστήματα περιοριζόταν στους έμφυτους μηχανισμούς άμυνας (π.χ., τροποποίηση περιορισμού και μεταγωγή υποδοχέα), αλλά ένα νουκλεϊκό οξύ-βασισμένο βασίζεται στο προσαρμοστικό ανοσοποιητικό σύστημα. Τα βακτήρια και τα αρχαία αποκτούν ανθεκτικότητα σε ιούς και αμφιβληστροειδή πλασμίδια με ενσωμάτωση μικρών θραυσμάτων ξένου νουκλεϊκού οξύ στο χρωμόσωμα ξενιστή στο ένα άκρο ενός επαναλαμβανόμενου στοιχείου γνωστού ως CRISPR (**C**lustered **R**egularly **I**nterspaced **S**hort **P**alindromic **R**epeat). .

Η CRISPR μεσολαβούμενη προσαρμοστική ανοσία προχωρά σε τρία διαφορετικά στάδια. Μόλις τα βακτήρια μολύνονται, ξένες αλληλουχίες DNA που δεν έχουν αποκτηθεί πριν, συλλαμβάνονται και ενσωματώνονται στο γενετικό τόπο CRISPR ως νέοι αποστάτες. Ο τόπος CRISPR μεταγράφεται και επεξεργάζεται για να παράγει ώριμα RNAs CRISPR, το καθένα από τα οποία κωδικοποιεί μια μοναδική διαχωριστική αλληλουχία (CRISPR RNA crRNA biogenesis). Κάθε crRNA συνδέεται με τις πρωτεΐνες του τελεστή Cas που χρησιμοποιεί crRNA ως οδηγούς για την σίγαση ξένων γενετικών στοιχείων που ταιριάζουν με την ακολουθία crRNA (παρεμβολή) (Εικ.1). Με απλά λόγια αυτός ο μηχανισμός διασπά το ξένο DNA οδηγώντας το στην αποσυνθεσή του.

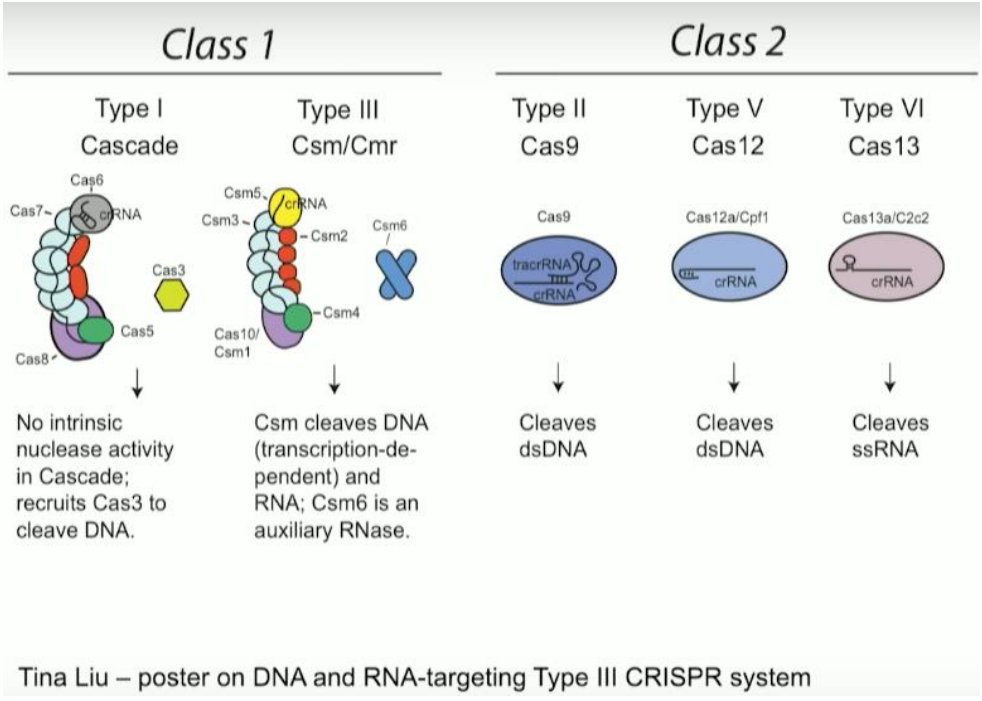


Εικ.4 Στάδιο 1: Κατά την μόλυνση, νέες ξένες αλληλουχίες DNA συλλαμβάνονται και ενσωματώνονται στον τόπο CRISPR (CRISPR Locus) του ξενιστή ως νέοι αποστάτες (spacers). **Στάδιο 2**: Ο τόπος CRISPR μεταγράφεται και επεξεργάζεται για τη δημιουργία ώριμων RNA CRISPR, το καθένα από τα οποία κωδικοποιεί μια μοναδική αλληλουχία αποστάτη. **Στάδιο 3**: Κάθε crRNA συνδέεται με πρωτεΐνες Cas που χρησιμοποιούν crRNAs ως οδηγούς για την σιωπή ξένων γενετικών στοιχείων που ταιριάζουν με την αλληλουχία crRNA. [20]

Αν και αυτά τα τρία βασικά στάδια φαίνονται κοινά σε όλα τα συστήματα CRISPR, CRISPR loci και οι πρωτεΐνες σε κάθε στάδιο είναι πολύ διαφορετικές μεταξύ των διαφόρων τύπων βακτηριδίων [6]. Αν και αυτά τα τρία βασικά στάδια φαίνονται κοινά σε όλα τα συστήματα CRISPR, CRISPR loci και οι πρωτεΐνες που μεσολαβούν σε κάθε στάδιο της προσαρμοστικής ανοσίας είναι αξιοσημείωτα διαφορετικές.[6]

Όλα τα CRISPR έχουν κοινή αρχιτεκτονική. Μία ακολουθία αδενίνης και θυμίνης (AT) ονομάζεται ηγέτης και πλευρίζει συχνά τους CRISPR τόπους. Οι αλληλουχίες οδηγοί περιέχουν στοιχεία προαγωγέα και δεσμευτικές θέσεις για ρυθμιστικές πρωτεΐνες κρίσιμες για την έκφραση του crRNA και για την απόκτηση νέας ακολουθίας. Κάθε CRISPR τόπος αποτελείται από μια σειρά από βραχείες αλληλουχίες επανάληψης, τυπικού μήκος 20-50 ζευγών βάσης, οι οποίες διαχωρίζονται με μοναδικές διαχωριστικές αλληλουχίες παρόμοιου μήκους.

Οι επαναλαμβανόμενες ακολουθίες μέσα σε ένα τόπο CRISPR διατηρούνται, αλλά οι επαναλήψεις σε διαφορετικούς τόπους CRISPR μπορεί να διαφέρουν και στη σειρά και στο μήκος. (Σχήμα 1) Το τελευταίο σύστημα ταξινόμησης για τα συστήματα CRISPR-Cas, το οποίο λαμβάνει υπόψη το ρεπερτόριο των Cas γονιδίων και την ομοιότητα αλληλουχίας μεταξύ των πρωτεϊνών Cas και την αρχιτεκτονική του τόπου, περιλαμβάνει δύο κατηγορίες που επί του παρόντος υποδιαιρούνται σε έξι τύπους και 19 υποτύπους. Τα συστήματα κλάσης 1 (συμπεριλαμβανομένων των τύπων I, III και IV) υπάρχουν στα βακτήρια και τα αρχαία, και όλα περιλαμβάνουν πολλαπλές πρωτεΐνες Cas για να σχηματίσουν σύμπλοκα παρακολούθησης. Τα συστήματα κλάσης 2 CRISPR-Cas (τύποι II, V και VI), περιορίζονται σχεδόν εξ ολοκλήρου σε βακτήρια, το σύμπλεγμα τελεστών είναι που αντιπροσωπεύεται από μία μεμονωμένη πολυπεπτιδική πρωτεΐνη. [7][8]



Εικ.5 Ταξινόμηση των CRISPR Cas συστημάτων

2.4 CRISPR Ένα Εργαλείο Για Τροποποίηση Αλληλουχιών DNA

Αυτές οι λειτουργίες ήταν συναρπαστικές και έχει γίνει έρευνα για την αξιοποίηση αυτών των πρωτεϊνών με σκοπό την επεξεργασία γονιδίων. Αρχικά, η έρευνα επικεντρώθηκε στα συστήματα κλάσης 1 αλλά ήταν δύσκολο να χρησιμοποιηθούν ως τεχνολογία για την επεξεργασία γονιδίων επειδή θα απαιτούσε πολλαπλούς συνδυασμούς πρωτεϊνών που πρέπει να γίνουν στα κύτταρα και να συναρμολογηθούν. Αυτό το σύστημα πολλών "συστατικών" θα είναι πολύπλοκο για να δουλέψει σε έναν ετερόλογο κυτταρικό τύπο.

Από την άλλη πλευρά, τα συστήματα κλάσης 2 ήταν απλούστερα και απλοποιήθηκαν περαιτέρω με την ικανότητα συνδυασμού των δύο φυσικών RNAs (CRISPR RNA και tracer RNA) σε μία ενιαία μορφή οδηγού. Ως εκ τούτου είναι ευκολότερο να αξιοποιηθούν ως ένα εργαλείο. Αυτή η ταχέως αναπτυσσόμενη τεχνολογία περιλαμβάνει κυρίως τους ακόλουθους τύπους γενετικών διαταραχών: το knockout γονιδίου (KO), το knockin γονιδίου (KI) για επεξεργασία γονιδιώματος , και αναστολή ή ενεργοποίηση έκφρασης γονιδίου (CRISPRi / a) (Πίνακας 1)

Genetic Manipulation	Application	Cas9	gRNA	Additional Considerations
Knockout	Permanently disrupt gene function in a particular cell type or organism without a specific preferred mutation	Cas9 (or Cas9 nickase)	Single (or dual) gRNA targeting 5′ exon or essential protein domains	High-fidelity Cas enzymes increase specificity. Dual-nickase approach increases specificity but is less efficient. Each putative knockout allele must be experimentally verified.
Edit	Generate a specific user-defined sequence change in a particular gene, such as generating a point mutation or inserting a tag	Cas9 (or Cas9 nickase); Base editor	Single (or dual) gRNA targeting the region where the edit should be made	HDR requires a repair template and displays reduced efficiency compared to NHEJ knockout. Base editors can make a limited set of mutations.
Repress or Interfere (CRISPRi)	Reduce expression of a particular gene(s) without permanently modifying the genome	dCas9-repressor (such as dCas9-KRAB) or dCas9	gRNA(s) targeting promoter elements of target gene	dCas9-KRAB is more effective than dCas9 alone for mammalian cell lines.
Activate (CRISPRa)	Increase expression of an endogenous gene(s) without permanently modifying the genome	dCas9-activator (such as dCas9-VP64)	gRNA(s) targeting promoter elements of target gene	Many different activators exist, including the multi-plasmid SAM system.

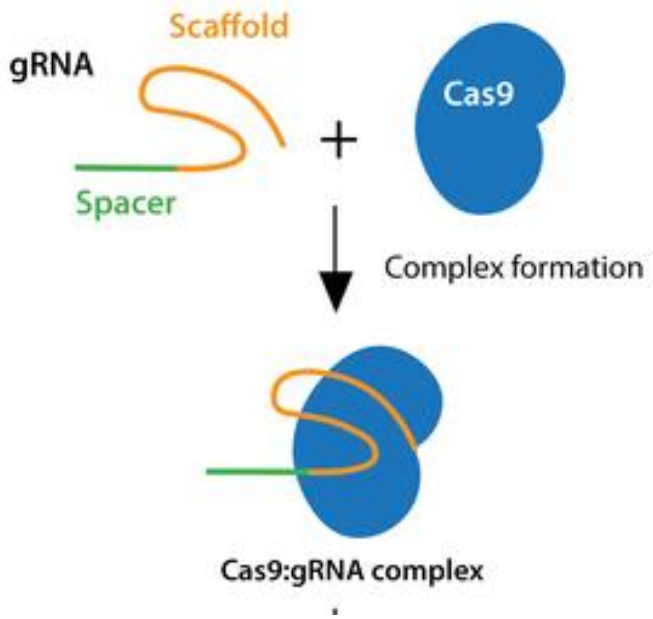
Πινακας 1 Διαφορες εφαρμογές του CRISPR για την επεξεργασία γονιδίων

Τα επεξεργασμένα συστήματα CRISPR περιέχουν δύο στοιχεία: ένα οδηγό RNA (gRNA) και μία CRISPR συνδεδεμένη ενδονουκλεάση (πρωτεΐνη Cas). Το gRNA είναι ένα σύντομο συνθετικό RNA που αποτελείται από ακολουθία ικριωμάτων που είναι απαραίτητες για την δέσμευση Cas και ένα καθορισμένο από το χρήστη αποστάτη 20 νουκλεοτιδίων που ορίζει τον γονιδιωματικό στόχο που πρέπει να τροποποιηθεί. Έτσι, μπορεί κανείς να αλλάξει τον γονιδιωματικό στόχο του Cas με απλή αλλαγή της αλληλουχίας του στόχου που υπάρχει στο gRNA (Εικόνα 3)

Το σύστημα CRISPR Cas9 που χρησιμοποιείται σήμερα για την επεξεργασία γονιδιώματος είναι ένα CRISPR τύπου II Cas system προσαρμοσμένο από τον Streptococcus pyogenes, αλλά εξετάστηκαν και άλλες εναλλακτικές λύσεις

επίσης. Το Streptococcus pyogenes Cas9 (**SpCas9**) θα χρησιμοποιηθεί ως παράδειγμα αναφοράς.

Μόλις εκφραστεί, η πρωτεΐνη Cas9 και το gRNA σχηματίζουν ένα σύμπλεγμα ριβονουκλεοπρωτεΐνης μεταξύ του ικριώματος gRNA και των επιφανειακά εκτεθειμένων θετικά φορτισμένων αυλακώσεων Cas9. Το Cas9 υφίσταται μια μεταβολή της διαμόρφωσης κατά τη δέσμευση του gRNA που μετατοπίζει το μόριο από μια ανενεργή, μη δεσμευτική διαμόρφωση στο DNA σε μια ενεργή διαμόρφωση δέσμευσης DNA. Είναι σημαντικό ότι, η διαχωριστική περιοχή του gRNA παραμένει ελεύθερη να αλληλοεπιδρά με το στοχευμένο DNA.



Εικ. 6 Δομή συμπλόκου Cas0-sgRNA

Προκειμένου οι πρωτεΐνες Cas να κόψουν το ξένο DNA, η αλληλουχία στόχος πρέπει να είναι δίπλα σε μια συγκεκριμένη αλληλουχία που ονομάζεται PAM (Protospacer Adjacent Motif) η οποία είναι μια μικρή αλληλουχία νουκλεϊκού οξέος που δεν περιέχεται στο βακτηριακό DNA. Με αυτόν τον τρόπο

προστατεύονται από τη διάσπασή του δικού τους DNA. Διαφορετικοί οργανισμοί έχουν διαφορετικές ακολουθίες PAM. (Πίνακας 2)

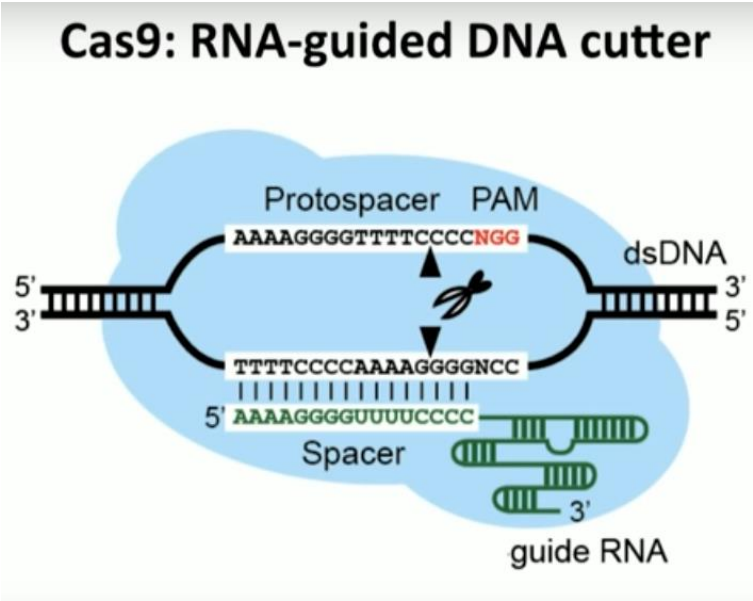
Species/Variant of Cas9	PAM Sequence
Streptococcus pyogenes (SP); SpCas9	3' NGG
SpCas9 D1135E variant	3' NGG (reduced NAG binding)
SpCas9 VRER variant	3' NGCG
SpCas9 EQR variant	3' NGAG
SpCas9 VQR variant	3' NGAN or NGNG
Staphylococcus aureus (SA); SaCas9	3' NNGRRT or NNGRR(N)
Acidaminococcus sp. (AsCpf1) and Lachnospiraceae bacterium (LbCpf1)	5' TTTV
AsCpf1 RR variant	5' TYCV
LbCpf1 RR variant	5' TYCV
AsCpf1 RVR variant	5' TATV
Neisseria meningitidis (NM)	3' NNNNGATT
Streptococcus thermophilus (ST)	3' NNAGAAW
Treponema denticola (TD)	3' NAAAAC
Additional Cas9s from various species	PAM sequence may not be characterized

Πινακας 2 PAM αλληλουχίες για διάφορες πρωτεΐνες Cas

Λόγω των περιορισμών που προκλήθηκαν από τις ακολουθίες PAM ορισμένοι ερευνητές καθιέρωσαν την εφικτότητα της κατασκευής ενός ευρέος φάσματος εφαρμογών των Cas9s με τροποποιημένες και βελτιωμένες ιδιότητες PAM. (Πίνακας 2) Το Cas9 θα συνδέσει μόνο μια δεδομένη θέση εάν η αλληλουχία spacer του gRNA μοιράζεται επαρκή ομολογία με το στοχευμένο DNA . Μόλις το σύμπλεγμα Cas9-gRNA δεσμεύσει έναν υποθετικό στόχο DNA, η αλληλουχία του σπόρου (8-10 βάσεις στο 3' άκρο του gRNA) θα αρχίσει να συγκολλάται στο στοχευμένο DNA. Εάν οι σπόροι και οι στοχευμένες αλληλουχίες DNA ταιριάζουν, το gRNA θα συνεχίσει να επανασυνδέεται με το στοχευμένο DNA σε μία κατεύθυνση 3' έως 5'.

Ο λόγος για τον οποίο οι αναντιστοιχίες μεταξύ της αλληλουχίας στόχου και της αλληλουχίας 3' σπόρου καταργούν τελείως τη διάσπαση του στόχου, ενώ οι αναντιστοιχίες προς το άκρο 5' που είναι απομακρυσμένες από το PAM συχνά επιτρέπουν την διάσπαση του στόχου, μπορεί να εξηγηθεί από το γεγονός ότι το ιικό DNA εξελίσσεται και το DNA που αποκτάται στο αποστάτες μπορεί να μην ταιριάζει ακριβώς με ένα άλλο εισβάλλον DNA.

Η νουκλεάση Cas9 έχει δύο λειτουργικές περιοχές ενδονουκλεάσης: RuvC και HNH ("μοριακό ψαλίδι'). Το Cas9 υφίσταται μια δεύτερη αλλαγή διαμόρφωσης κατά τη σύνδεση των θέσεων αυτών με στόχο τα πεδία νουκλεάσης να αποκοπούν από τούς αντίθετους κλώνους του στοχευμένου DNA. Το τελικό αποτέλεσμα του Cas9-η μεσολαβούμενη διάσπαση DNA είναι ένα διάλειμμα διπλού έλικα (DSB) εντός του στοχευμένου DNA (3 νουκλεοτίδια ανοδικά της αλληλουχίας PAM).



Εικ.7 Απεικόνιση του κοψίματος της αλυσίδας με CRISPR

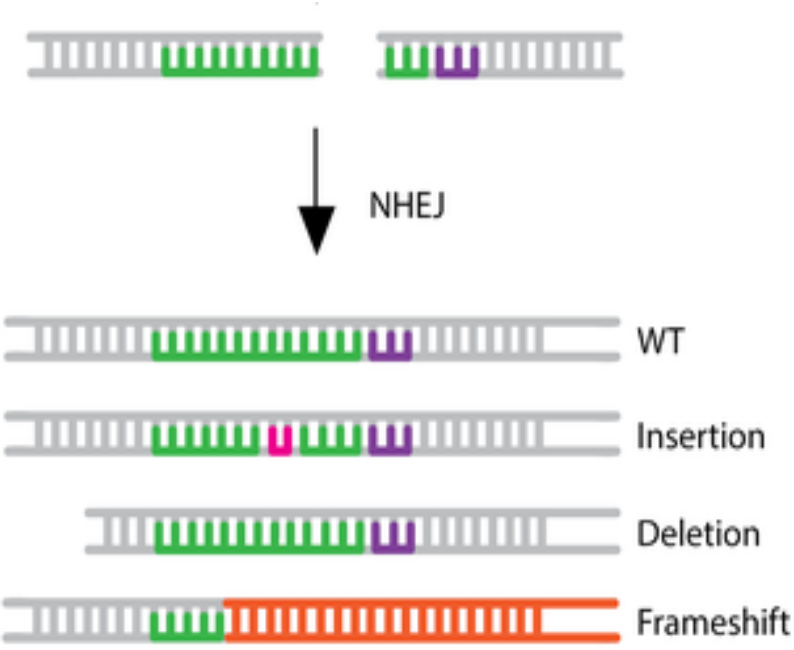
Αυτό το κόψιμο μπορεί να επισκευαστεί είτε με Non-homologous End Joining (NHEJ) ή με Homology Directed Repair (HDR)(HDR).

Non-homologous End Joining (NHEJ)

Τις περισσότερες φορές, αυτό το DSB επισκευάζεται με μη ομόλογους και τελικούς μηχανισμούς ένωσης (NHEJ), γεγονός που οδηγεί σε τυχαίες παρεμβολές και / ή διαγραφές (indels) στη θέση της διάσπασης. Στο πλαίσιο της θεραπείας με DMD, αυτός ο μηχανισμός επιδιόρθωσης μπορεί γενικά να χρησιμοποιηθεί σε τρεις διαφορετικές στρατηγικές.

Πρώτον, εάν χρησιμοποιείται ένα μοναδικό gRNA για στόχευση διάσπασης σε ή κοντά σε ένα πρόωρο σήμα τερματισμού, σε μεταλλαγμένο ή μεταλλαγμένο εξόνιο DMD, σχηματισμό ινδελίου από NHEJ μπορεί να εξαλείψει το σήμα λήξης ή / και να επαναφέρει το πλαίσιο ανάγνωσης πίσω στην κανονική διαμόρφωση (Ανασυγκρότηση NHEJ).

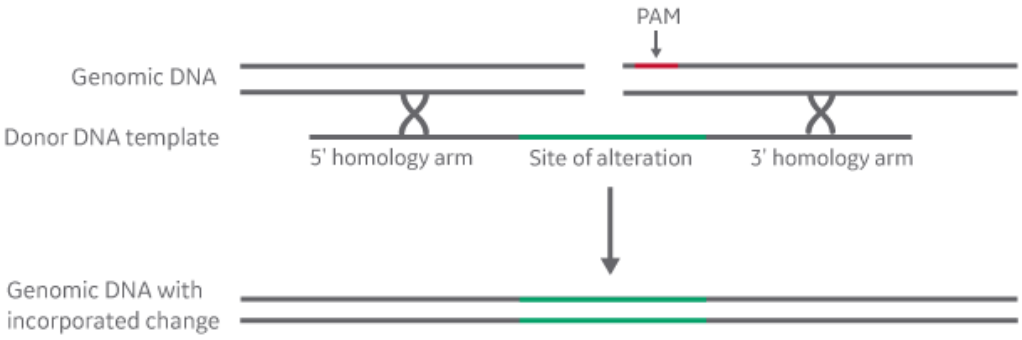
Δεύτερον, εάν ένα απλό gRNA χρησιμοποιείται για στόχευση διάσπασης σε ή κοντά σε ακολουθίες σύνδεσης σε εξόνια ή εσόνια DMD, οι συμβολές μπορούν να διαταράξουν αυτές τις περιοχές και να επιτρέψει να συμβεί η παράκαμψη ενός εξονίου εκτός πλαισίου (κλασική παραβίαση εξονίου). Τελικά, εάν χρησιμοποιούνται τουλάχιστον δύο gRNAs για να στοχεύσουν τη διάσπαση σε ξεχωριστά εξόνια ή εσόνια, μπορούν να επιτευχθούν διαγραφές ενός ή περισσοτέρων εξονίων για την αποκατάσταση του πλαισίου ανάγνωσης DMD (άμεση παράλειψη εξονίου). Ιδιαίτερο ενδιαφέρον για την επανεξέταση αυτή θα είναι η δεύτερη και η τρίτη προσέγγιση που αναφέρονται. Όπως θα δούμε, αυτές οι στρατηγικές, που χρησιμοποιούνται μόνες ή σε συνδυασμό, έχουν χρησιμοποιηθεί για την ανάπτυξη διάφορων πιθανών θεραπειών DMD τόσο σε vitro όσο και σε νίνο.



Εικ.8 Απεικόνιση της ανάπλασης του DNA μετά από το κόψημο της αλυσίδας με τη χρήση NHEJ

Homology Directed Repair (HDR)

Εκτός από το NHEJ, τα κύτταρα είναι σε θέση να χρησιμοποιούν έναν πιο ακριβή μηχανισμό επισκευής γνωστού ως Homologous Directed Repair (HDR). Αυτός ο μηχανισμός επισκευής μπορεί να αξιοποιηθεί για την εισαγωγή ειδικών νουκλεοτιδικών τροποποιήσεων στο γονιδιωματικό DNA. Ένα πρότυπο επισκευής DNA, με υψηλό βαθμό ομολογίας προς την ακολουθία άμεσα ανοδικά και καθοδικά της επιδιωκόμενης θέσης επεξεργασίας, εισάγεται στο κύτταρο μαζί με την κατάλληλη gRNA και Cas9 νουκλεάση. Στην παρουσία αυτού του κατάλληλου προτύπου, ο λιγότερο ευαίσθητος σε σφάλματα μηχανισμός HDR μπορεί να κάνει πιστά τις επιθυμητές αλλαγές στην θέση DSB επαγόμενη από Cas9 μέσω ανασυνδυασμού. Κατά το σχεδιασμό της επισκευής του πρότυπου, βεβαιωθείτε ότι είτε η ακολουθία στόχου δεν ακολουθείται αμέσως από την ακολουθία PAM ή ότι η ακολουθία PAM είτε αποκλείεται είτε μεταλλάσσεται. Αυτό είναι για να αποφευχθεί η αποικοδόμηση του προτύπου επισκευής από το ίδιο σύστημα CRISPR Cas9.



Εικ.9 Απεικόνιση της ανάπλασης του DNA μετά από το κόψημο της αλυσίδας με τη χρήση HDR

Off-Target Effects

Μια σημαντική εκτίμηση κατά τη χρήση του CRISPR Cas9 ως εργαλείου επεξεργασίας γονιδιώματος είναι η έκταση που η εκτός στόχου διάσπαση λαμβάνει χώρα. Το γεγονός εκτός στόχου μπορεί να οδηγήσει σε μεταλλάξεις του InDel σε χώρους που δεν προορίζονταν αρχικά και επομένως, να θέτει σε κίνδυνο τα φαινοτυπικά αποτελέσματα που ελήφθησαν. Αρκετές μελέτες έχουν αξιολογήσει αυτή την ιδιαιτερότητα του συστήματος CRISPR Cas9 και έχουν δείξει ότι, γενικά, οι αναντιστοιχίες προς το άκρο 5 'της περιοχής στόχευσης των 20 ζευγών βάσεων του gRNA είναι ανεκτές.

Ωστόσο είναι δύσκολο να προβλεφθεί πώς αυτές οι αναντιστοιχίες επηρεάζουν τα εκτός στόχου αποτελέσματα του CRISPR system Cas9. Έχουν αναφερθεί περιστατικά όπου υπάρχει αναντιστοιχία στο 5' άκρο της περιοχής στόχευσης του gRNA όπου δεν ήταν ανεκτά και άλλες περιπτώσεις όπου δεν συμφωνούν στο 3' άκρο της περιοχής στόχευσης . Συνολικά, τα αποτελέσματα εκτός στόχου εισάγονται από το CRISPRT σύστημα Cas9 είναι μεταβλητά σε συχνότητα και δύσκολο να προβλεφθούν. [9]

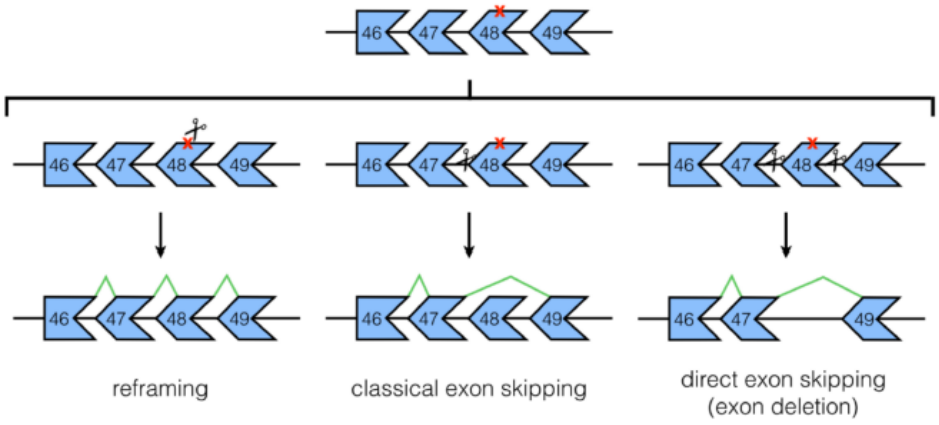
2.4 Χρηση CRISPR Για Τη Θεραπεία DMD

Για την θεραπεία DMD με βάση τα προαναφερθέντα, είναι τώρα εύκολο να κατανοήσουμε πώς το σύστημα CRISPR / Cas9 system μπορεί να προσαρμοστεί για χρήση στην ακριβή επεξεργασία γονιδιωμάτων. Τα gRNAs μπορεί να σχεδιάστηκαν για να επάγουν στοχευμένες DSBs σε οποιαδήποτε επιλεγμένη ακολουθία DNA κατά τη χορήγηση με Cas9, υπό την προϋπόθεση ότι η αλληλουχία στόχος είναι κοντά σε μια θέση PAM που αναγνωρίζεται από το χρησιμοποιούμενο ένζυμο Cas9.

Τον περισσότερο χρόνο, αυτό το DSB επισκευάζεται με μηχανισμούς μη ομόλογου άκρου σύνδεσης (NHEJ), οι οποίοι οδηγούν σε τυχαία συμβολές στο σημείο της διάσπασης. Στο πλαίσιο της θεραπείας με DMD, αυτός η ο μηχανισμός επισκευής μπορεί γενικά να χρησιμοποιηθεί σε τρεις διαφορετικές στρατηγικές (σχήμα). .

Πρώτον, εάν ένα μόνο gRNA χρησιμοποιείται για στόχευση διάσπασης σε ή κοντά σε ένα πρόωρο σήμα τερματισμού σε ένα μεταλλαγμένο ή μεταλλαγμένο DMD εξόνιο, ο σχηματισμός ινδελίου από το NHEJ μπορεί να εξαλείψει το σήμα τερματισμού και / ή να επαναφέρει το πλαίσιο ανάγνωσης πίσω στην κανονική διαμόρφωση (αναμόρφωση του NHEJ). Δεύτερον, αν χρησιμοποιείται ένα gRNA για διάσπαση στόχου σε ή κοντά σε αλληλουχίες συρραφής σε εξόνια ή εσόνια DMD, οι ινδελές μπορούν να διαταράξουν αυτές τις θέσεις και επιτρέπουν να συμβεί η παραβίαση ενός εξονίου εκτός πλαισίου (κλασική παράλειψη εξονίου). Και τέλος, αν τουλάχιστον δύο gRNAs χρησιμοποιούνται για στόχευση, διάσπαση στόχου σε ξεχωριστά εξόνια ή εσόνια, διαγραφές ενός ή περισσότερων εξονίων μπορούν να επιτευχθούν για την αποκατάσταση του πλαισίου ανάγνωσης DMD (άμεση παράλειψη εξονίου).

Ιδιαίτερου ενδιαφέροντος θα ήταν η δεύτερη και η τρίτη προσέγγιση που αναφέρθηκαν. Αυτές οι στρατηγικές, που χρησιμοποιούνται μόνες ή σε συνδυασμό μεταξύ τους, έχουν χρησιμοποιηθεί για την ανάπτυξη διαφόρων δυνατοτήτων DMD τόσο in vitro όσο και in vivo. Τα DSB που δημιουργούνται από το CRISPR / Cas9 μπορούν επίσης να επισκευαστούν μέσω άλλου μηχανισμού: HDR. Στο HDR, αντί να ενώσει τυχαία τα δύο DSBs μαζί, ένα πρότυπο με άκρα ομόλογα προς κάθε άκρο του DSB χρησιμοποιείται για την επακριβή επιδιόρθωση της βλάβης ώστε να έχει την ίδια ακολουθία πριν από το DSB. Ωστόσο, καθώς συμβαίνει αυτό σε πολύ χαμηλότερες συχνότητες από το NHEJ, και δεδομένου ότι δεν συμβαίνει τυπικά σε μετα-μιτωτικά κύτταρα, το HDR δεν χρησιμοποιείται τόσο συχνά σε DMD ανάπτυξη θεραπείας.



Εικ.10 Απεικόνιση της ανάπλασης του DNA μετά από το κόψημο της αλυσίδας με τη χρήση NHEJ**.** Μηχανισμοί non-homologous end joining (NHEJ) για τη θεραπεία μέσω CRISPR/Cas9. Στην πρώτη γραμμή βλέπουμε ένα μεταλλαγμένο γονίδιο ασθενή που οδηγεί σε μεταλαγμένο πλαίσιο ανάγνωσης του γονιδίου. Ο μηχανισμός CRISPR/Cas9, με NHEJ repair, μπορεί να διορθώσει το πλαίσιο ανάγνωσης με τρεις τρόπους: reframing, classical exon skipping, or direct exon skipping.

Ο ακόλουθος πίνακας συνοψίζει συνοπτικά τις μελέτες που έχουν μέχρι στιγμής αναπτυχθεί στρατηγικές CRISPR για τη θεραπεία της DMD. Είναι αξιοσημείωτο ότι η πλειονότητα αυτών των μελετών χρησιμοποίησαν το σύστημα CRISPR / Cas9. Όπως φαίνεται στον πίνακα, κοινές στρατηγικές για την επεξεργασία της δυστροφίνης γονιδίου in vitro περιλαμβάνουν τους τρεις μηχανισμούς του NHEJ που περιγράφηκαν προηγουμένως. Οι δυνατότητες αυτών των προσεγγίσεων, καθώς και μερικές μοναδικές άλλες, έχουν καταδειχθεί χρησιμοποιώντας μια ποικιλία in vitro μοντέλα, συμπεριλαμβανομένων των διαχρονικών μυϊκών κυττάρων ασθενών DMD, πρωτογενών κυττάρων ασθενών DMD, mdx δορυφορικών κυττάρων ποντικιών και, ιδιαίτερα, DMD ανθρώπινα επαγόμενα πολυδύναμα αρχέγονα κύτταρα (hiPSCs) και τα παράγωγά τους. Είναι ενδιαφέρον, η χρήση των hiPSCs που προέρχονται από ασθενείς με DMD για την προκλινική μελέτη των θεραπειών CRISPR / Cas9 κερδίζει αυτή τη στιγμή έλξη.
Τα hiPSCs μπορούν να εξαχθούν από εύκολα προσβάσιμους ιστούς ασθενών και, λόγω της πολυπλοκότητας τους, μπορούν να κατευθυνθούν για να διαφοροποιηθούν σε σχετικούς τύπους κυττάρων για τη δοκιμασία θεραπείας με CRISPR / Cas 9 DMD. Καρδιομυοκύτταρα που προέρχονται από hiPSCs μπορούν να μοντελοποιήσουν με μεγαλύτερη ακρίβεια την αναπτυξιακή εξέλιξη και τη φυσιολογία της ανθρώπινης καρδιάς, που δεν εκπροσωπείται επαρκώς στα διάφορα ζωικά μοντέλα.

Για παράδειγμα, η καρδιομυοπάθεια που παρουσιάζεται από τα mdx ποντίκια είναι σημαντικά λιγότερο σοβαρή από αυτή που παρατηρείται στους ασθενείς, επίσης, η ρύθμιση του καρδιακού αγγειακού συστήματος, η δραστηριότητα των διαύλων ιόντων και η λειτουργία μυοσίνης ρυθμίζονται διαφορετικά σε ποντίκια. Υπάρχει επίσης το γεγονός ότι τα ποντίκια και τα σκυλιά έχουν συχνά αυξημένη ανοχή και αντοχή στις καρδιοτοξικές επιδράσεις των φαρμάκων και συνεπώς μπορεί να οδηγήσει σε ανακριβή αντιπροσώπευση των επιδράσεων αυτών των φαρμάκων στους ανθρώπους.

Όπως αναφέρθηκε και σε προηγούμενα, τα hiPSCs προσφέρουν επίσης το πλεονέκτημα της αξιόπιστης αντιγραφής μεταλλάξεων ασθενών σε καρδιομυοκύτταρα in vitro χωρίς την ανάγκη για διηθητική λήψη καρδιακών βιοψιών. Επί του παρόντος, καρδιακά μυϊκά κύτταρα λαμβάνονται από μεταμοσχεύσεις, ανεπανόρθωτα κατεστραμμένες καρδιές ή μέσω της διαδιαφοροποίησης των ινοβλαστών. Ωστόσο, τα hiPSCs μπορούν να ληφθούν από τα κύτταρα στο δέρμα ή τα ούρα (μετά τον επαναπρογραμματισμό) μοντελοποιώντας περισσότερα από 7000 τύπους μετάλλαξης DMD. [10]

Cas Enzyme	Strategy	Target Gene Region(s)	Model(s)	Delivery	Study Highlights
SpCas9	NHEJ reframing, HDR exon correction	*Dmd* exon 23	*mdx* mice	1-cell embryo injection	Dystrophin restoration observed by IHC (up to 100%) and WB; 17% *Dmd* HDR correction resulted in 47–60% dystrophin-positive fibers in skeletal muscles and the heart
SpCas9	NHEJ reframing, exon skipping, HDR exon knock-in	*DMD* intron 44/exon 45	DMD hiPSCs, hiPSC-derived skeletal muscle cells (ex44 del.)	Electroporation	Dystrophin restoration in derived skeletal muscle cells observed by WB and IHC for all strategies; CRISPR was as effective as using TALEN
SpCas9	NHEJ reframing, single/multiple exon deletion	*DMD* exons 45–55 (for reframing each exon), introns 50 and 51 (ex51 del.), introns 44 and 55 (ex45–55 del.)	immortalized DMD patient muscle cells (ex48–50 del.), immunodeficient NSG mice	Electroporation	Generated targeted deletions of exon/s in vitro, particularly of the large exon 45–55 region which led to dystrophin rescue by WB; mice transplanted with treated myoblasts (exon 51-deleted) showed dystrophin-positive fibers by IHC
dSpCas9-VP16	Utrophin upregulation	*UTRN* A/B promoter	immortalized DMD patient muscle cells (ex45–52 del.)	Electroporation	1.7–6.9-fold upregulation of utrophin achieved; restored β-dystroglycan expression observed by WB with as little as 1.7-fold upregulation
SpCas9	Duplicated exons removal	*DMD* intron 27	primary DMD patient fibroblasts (ex18–30 dup.)	LV transduction, with Adeno-MyoD	4.42% full-length dystrophin production achieved post-treatment, accompanied with α-dystroglycan restoration
SpCas9	Single exon deletion	*Dmd* exon 23, introns 22 and 23 (ex23 del.)	*mdx* mice	AAV9 delivery (i.m., i.p., i.v.)	All modes of injection led to appearance of dystrophin-positive fibers as evaluated by IHC: ~25.5% 6 wks post-i.m., ~4.6% and ~9.6% in skeletal and cardiac muscles respectively 12 wks post-i.v., ~1.8% and ~3.2% in skeletal and cardiac muscles respectively 8 wks post-i.p.
SaCas9	Single exon deletion	*Dmd* introns 22 and 23 (ex23 del.)	*mdx* mice	AAV8 delivery (i.m., i.p., i.v.)	Intramuscular injections led to ~59% of transcripts with exon 23 deleted, which restored about 8% dystrophin of healthy levels by WB, proper relocalization of DGC proteins, and muscle function improvement; systemic injections restored dystrophin production in the heart and skeletal muscles
SpCas9, SaCas9	Single exon deletion	*Dmd* introns 22 and 23 (ex23 del.)	*mdx* mice, *mdx* satellite cells	AAV9 delivery (i.m., i.p., i.v.)	Dual-vector (Cas9 and gRNAs on separate constructs) had higher cutting efficiency than a single-vector system (Cas9 and gRNAs on the same construct) in vitro; dystrophin restoration >10% observed in the heart and skeletal muscles upon systemic treatment; correction also possible in satellite cells
SpCas9	Hybrid exon formation via internal exon deletion	*DMD* exons 50 and 54	immortalized DMD patient muscle cells (ex51–53 del.), hDMD/*mdx* mice	Lipotransfection (in vitro)/ electroporation (in vivo)	Dystrophin restoration successful in vitro by WB, not shown in vivo; hybrid exon formation thought to preserve dystrophin rod domain structure better

Cas Enzyme	Strategy	Target Gene Region(s)	Model(s)	Delivery	Study Highlights
SpCas9	NHEJ reframing, single/multiple exon deletion	*DMD* exons 51, 53, introns 52 and 53 (ex53 del.), 43 and 54 (ex44–54 del.)	immortalized DMD patient muscle cells (ex48–50, or 45–52 del.)	Sequential LV then AdV transduction/AdV transduction	Study showed the possibility of combining both TALEN and CRISPR approaches in one gene editing strategy; also, comparable editing was obtained with Cas9 and gRNA delivered either together or separately in AdV
SpCas9	Multiple exon deletion	*Dmd* introns 20 and 23 (ex21–23 del.)	*mdx* mice	Electroporation/AdV transduction	Treatment restored proper calcium dynamics in muscle (electroporation), and restored dystrophin to 50% of wild-type levels, as well as dystrophin-associated complex sarcolemmal localization and muscle membrane integrity (transduction)
SpCas9	Multiple exon deletion	*DMD* introns 44 and 55 (ex45–55 del.)	DMD hiPSCs, hiPSC-derived skeletal and cardiac muscle cells (ex46–51 or 46–47 del., ex50 dup.), immunodeficient NSG-*mdx* mice	Nucleofection	CRISPR-mediated deletion of the large exon 45–55 region achieved, restored membrane function and dystrophin, β-dystroglycan expression by WB and IHC; mice transplanted with hiPSC-derived skeletal muscle cells showed dystrophin-positive fibers by IHC
SpCas9	NHEJ reframing, single/multiple exon deletion	*DMD* exons 51, 53 (for reframing) introns 52 and 53 (ex53 del.), introns 43 and 54 (ex44–54 del.)	immortalized DMD patient muscle cells (ex48–50, or 45–52 del.)	AdV transduction	AdV with 2gRNA-SpCas9 constructs work as good as those with 1gRNA-SpCas9 constructs in terms of corrective ability and dystrophin restoration
SpCas9, SaCas9	Multiple exon deletion, HDR exon correction	*Dmd* exon 53, introns 51 and 53 (ex52–53 del.)	*mdx4cv* mice (nonsense ex53 mutation)	AAV6 delivery (i.m., i.v.)	Dual vector approach (SpCas9 and gRNA separate) yielded higher correction efficiency than single vector approach (SaCas9 and gRNA together); systemic treatment restored dystrophin expression in the heart (~34% dystrophin-positive fibers) and skeletal muscles (~10–50% dystrophin-positive fibers)
LbCpf1, AsCpf1	NHEJ reframing, single exon skipping, HDR exon correction	*DMD* exon 51, intron 50	DMD hiPSCs, hiPSC-derived cardiac muscle cells (ex48–50 del.), *mdx* mice	Nucleofection (in vitro)/ 1-cell embryo injection (in vivo)	Cpf1 editing successfully restored dystrophin expression and improved mitochondrial function in cardiomyocytes; 5/24 pups (injected at the embryo stage) showed HDR correction and had ameliorated dystrophic phenotypes
SpCas9	Duplicated exon removal	*DMD* exon 2, intron 2	immortalized DMD patient muscle cells (ex2 dup.)	PEI transfection/LV transduction	Use of a single gRNA can delete a duplicated exon, resulting in slight dystrophin rescue by WB and IHC
SpCas9	HDR exon correction	*Dmd* exon 23	*mdx* mice, *mdx* satellite cells	Lipotransfection (template, gRNA), AdV transduction (Cas9)/AdV transduction	Higher transduction efficiency obtained when AdVs were used for both Cas9 and gRNA-HDR template delivery; mice transplanted with corrected satellite cells showed dystrophin-positive fibers by IHC

Cas Enzyme	Strategy	Target Gene Region(s)	Model(s)	Delivery	Study Highlights
SpCas9	Multiple exon deletion	*DMD* introns 44 and 55 (ex45–55 del.)	humanized *mdx* mice with *DMD* exon 45 del.	Electroporation	Exon 45–55 deletion by CRISPR possible in vivo; first use of the humanized DMD mouse model with exon 45 del. for CRISPR studies
SpCas9	Multiple exon deletion	*DMD* introns 2 and 7 (ex3–9 del.), introns 5 and 7 (ex6–7 del.), introns 6 and 11 (ex7–11 del.)	DMD hiPSCs, hiPSC-derived cardiac muscle cells (ex8–9 or ex3–7 del.)	Nucleofection	Dystrophin with ex7–11 del. showed the least functionality, while those with ex3–9 del. had the highest functionality in terms of assessing iPSC-derived cardiomyocyte calcium cycling
SpCas9	HDR correction	*Dmd* exon 23	*mdx* primary muscle cells, *mdx* mice	CRISPR-Gold nanoparticles (i.m.)	5.4% HDR correction of the *Dmd* mutation in *mdx* was observed after CRISPR treatment and cardiotoxin injection, dystrophin-positive fibers found by IHC; 0.8% HDR correction observed without cardiotoxin co-injection, which led to significantly improved hanging test performance
SpCas9	NHEJ reframing, single exon skipping	*Dmd* exon 51	mice with *Dmd* exon 50 del.	AAV9 delivery (i.m., i.p.)	Successful dystrophin restoration in the heart and skeletal muscles; systemic injections led to improved muscle function; first application of CRISPR in the ex50 del. mouse model
SpCas9	Single exon deletion	*Dmd* introns 50 and 51 (ex51 del.)	primary human skeletal muscle cells	HCAdV delivery	Up to 93.3% exon 51 deletion observed in vitro upon delivery of CRISPR agents by HCAdV
SpCas9	NHEJ reframing, exon skipping	*DMD* exon 51, introns 47, 50, 54	DMD hiPSCs, hiPSC-derived cardiac muscle cells (ex48–50 del., pseudo-ex47, ex55–59 dup.)	Nucleofection	All strategies corrected the respective patient mutations and restored dystrophin production in iPSC-derived cardiomyocytes; 3D-engineered heart muscle produced from treated iPSC-derived cardiomyocytes showed improved contractile force
CjCas9	NHEJ reframing	*Dmd* exon 23	mice with deletions in *Dmd* exon 23	AAV9 delivery (i.m.)	CjCas9 displayed higher targeting specificity than SpCas9; use of CjCas9-based CRISPR can lead to successful dystrophin restoration and improvement in muscle function as well
SaCas9	Hybrid exon formation via multiple exon deletion	*DMD* exons 47 and 58	DMD skeletal muscle cells (ex51–53 del., ex49–50 del., ex51–56 del., ex50–52 del.), humanized *mdx* mice with *DMD* ex52 del.	LV transduction (in vitro)/AAV9 delivery (in vivo; i.v.)	gRNAs designed to produce exon deletions that best preserved dystrophin protein structure were able to show dystrophin restoration in vitro and in vivo (slight rescue in the heart)
SpCas9	NHEJ reframing, exon skipping	*Dystrophin* exon 51	deltaE50-MD canine model (ex50 del.)	AAV9 delivery (i.m., i.v.)	First published study on dystrophin gene correction in a dog model; ~3–70% dystrophin restoration of healthy levels in skeletal muscles and ~92% in the heart found by WB
nSpCas9-ABE7.10	Base editing to correct a nonsense mutation	*Dmd* exon 20	mice with a nonsense mutation in *Dmd* exon 20	trans-splicing AAV2/9 delivery (i.m.)	~3.3% base editing frequency achieved 8 weeks post-treatment with no detectable off-target effects; ~17% dystrophin-positive fibers and restored localization of nNOS observed by IHC
dSa/SpCas9-TAM	Base editing to induce exon skipping	*DMD* intron 50 5′ splice site	DMD hiPSCs, hiPSC-derived cardiac muscle cells (ex51 del.)	Lipotransfection	~100% base editing efficiency achieved; corrected iPSC-derived cardiomyocytes had restored dystrophin protein, low CK and miR-31 levels, and restoration of β-dystroglycan expression

3 ΣΥΛΛΟΓΗ ΔΕΔΟΜΕΝΩΝ ΚΑΙ ΕΡΓΑΛΕΙΑ

3.1 Εργαλεία Για Την Αξιολόγηση sgRNas

CCTOP

Το CCTop είναι ένα εργαλείο για τον προσδιορισμό των κατάλληλων θέσεων στόχου CRISPR / Cas9 σε μια δεδομένη αλληλουχία ερωτήσεων και την πρόβλεψη των πιθανών τοποθεσιών εκτός στόχου. Υπάρχει διαθέσημο Online αλλά και ξεχωριστά γραμμένος σε Python.
Χρησιμοποιεί τον ευθυγραμμιστή ανάγνωσης Bowtie 1 για τον εντοπισμό των τοποθεσιών εκτός στόχου(Off-targets).

Το εργαλείο αυτό χρησιμοποιήθηκε για τον εντοπισμό Off-targets αλλά και για την αξιολόγηση της αλυσίδας οδηγού(sgRNA)

Για τον υπολογισμό του score χρησιμοποιείται ο παρακάτω αλγόριθμος:

- Δημιουργούμε 10 χαρακτηριστικά της αλυσίδας. Κατα σειρά είναι:
  1. Τη περιεκτικότητα σε γουανίνη και κυτοσίνη
  2. Εάν ο στόχος μας στο DNA έχει στη θέση 19 γουανίνη παίρνουμε βάρος 1 αλλιώς 0
  3. Εάν ο στόχος μας στο DNA έχει στη θέση 2 αδενίνη ή θυμίνη παίρνουμε βάρος 1 αλλιώς 0
  4. Εάν ο στόχος μας στο DNA έχει στη θέση 11 αδενίνη ή γουανίνη παίρνουμε βάρος 1 αλλιώς 0
  5. Εάν ο στόχος μας στο DNA έχει στη θέση 5 γουανίνη παίρνουμε βάρος 1 αλλιώς 0
  6. Εάν ο στόχος μας στο DNA έχει στη θέση 3 αδενίνη ή θυμίνη παίρνουμε βάρος 1 αλλιώς 0
  7. Εάν ο στόχος μας στο DNA έχει στη θέση 17 αδενίνη ή γουανίνη παίρνουμε βάρος 1 αλλιώς 0
  8. Εάν ο στόχος μας στο DNA έχει στη θέση 4 αδενίνη ή κυτοσίνη παίρνουμε βάρος 1 αλλιώς 0
  9. Εάν ο στόχος μας στο DNA έχει στη θέση 13 γουανίνη παίρνουμε βάρος 1 αλλιώς 0
  10. Εάν ο στόχος μας στο DNA έχει στη θέση 14 αδενίνη παίρνουμε βάρος 1 αλλιώς 0

Τέλος παίρνουμε το εσωτερικό γινόμενο του διανύσματος χαρακτηριστικών μας με ένα διάνυσμα βαρών (συγκεκριμένων) και σε αυτό προσθέτουμε μια σταθερά (model offset). [16]

3.3 Συλλογή και Ανάλυση Δεδομένων

Αν και υπάρχουν πολλά δεδομένα για CRISPR στο γονίδιο της δυστροφίνης, εμάς μας ενδιαφέρουν μόνο αυτά που έχουν χρησιμοποιηθεί πειραματικά για exon-skipping και έχουν την ίδια Cas προτεΐνη. Επιλέξαμε δεδομένα μόνο από μελέτες σε ανθρώπινα κύτταρα(HPSC). Αυτό που μας ενδιαφέρει δεν είναι το αν η αλληλουχία sgRNA θα οδηγήσει σε κόψιμο της αλυσίδας DNA αλλά εάν θα πραγματοποιηθεί exon-skipping. Το ποσοστό επιτυχίας του exon-skipping δίνεται από τους ερευνητές.[18-36]

Τα δεδομένα μας αποτελούνταν από 4 βασικές πληροφορίες:

- Την αλληλουχία DNA που στοχεύουμε
- Την κατεύθυνση της αλυσίδας
- Το εάν η αλληλουχία που στοχεύουμε βρίσκεται σε εσόνιο ή εξόνιο
- Το ποσοστό των κυττάρων στο οποίο έγινε επιτυχώς το exon-skipping

Από αυτά μετατρέψαμε τη μεταβλητή του ποσοστού των κυττάρων στο οποίο έγινε το exon-skipping σε κατηγορική. Πιο συγκεκριμένα δημιουργήσαμε 2 κλάσεις, τα «καλά» sgRNA ως αυτά που είχαν ποσοστό exon-skipping μεγαλύτερο από 15% και τα «κακά» sgRNA ως αυτά που είχαν ποσοστό exon-skipping κάτω από 15%.

Με βάση την αλληλουχία πήραμε όσο το δυνατόν περισσότερα δεδομένα μπορούσαμε ώστε να τα αξιοποιήσουμε στη δημιουργία ενός αλγορίθμου ταξινόμησης. Αυτά είναι:

- Η θερμοκρασία στην οποία σπάνε οι δεσμοί μεταξύ της αλυσίδας σύμφωνα με το κανόνα του Wallace[13]. Δίνεται από τον τύπο:

$$Tm = 64.9 + 41(yG + zC - 16.4)(wA + xT + yG + zC)$$

- Η θερμοκρασία στην οποία σπάνε οι δεσμοί μεταξύ της αλυσίδας με βάση το ποσοστό της γουανίνης και κυτοσίνης[14]. Δίνεται από τον τύπο:

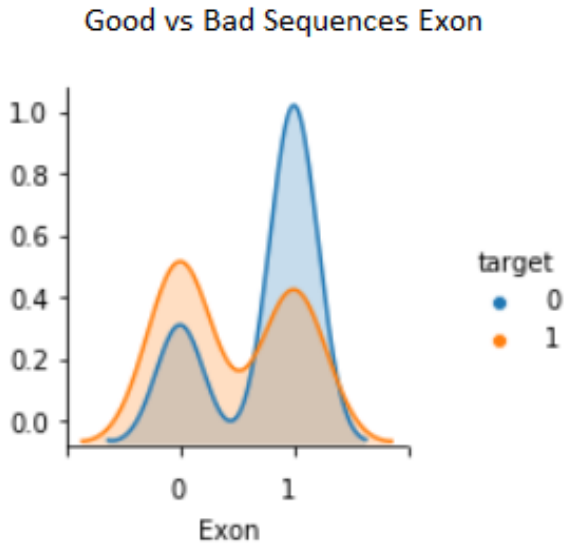
$$Tm = 4GC + 2TA$$

- Η θερμοκρασία στην οποία σπάνε οι δεσμοί μεταξύ της αλυσίδας με βάση το κανόνα του κοντινότερου γείτονα. Η κύρια διαφορά είναι ότι το μοντέλο πλησιέστερου γείτονα αντιμετωπίζει μια έλικα DNA ως μια σειρά αλληλεπιδράσεων μεταξύ «γειτονικών» ζευγών βάσης. [15]

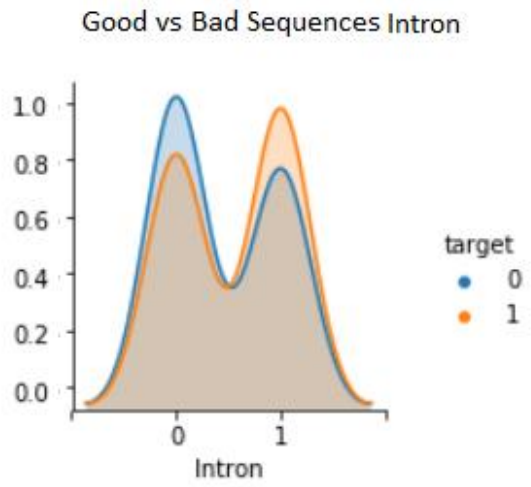
Sequence	ΔH° kcal/mol	ΔS° cal/k·mol
AA/TT	−7.9	−22.2
AT/TA	−7.2	−20.4
TA/AT	−7.2	−21.3
CA/GT	−8.5	−22.7
GT/CA	−8.4	−22.4
CT/GA	−7.8	−21.0
GA/CT	−8.2	−22.2
CG/GC	−10.6	−27.2
GC/CG	−9.8	−24.4
GG/CC	−8.0	−19.9
Init. w/term. G·C	0.1	−2.8
Init. w/term. A·T	2.3	4.1
Symmetry correction	0	−1.4

- CCTOP score
- sgRNA designer score
- Off targets
- Το περιεχόμενο σε γουανίνη και κυτοσσίνη

Εξόνια

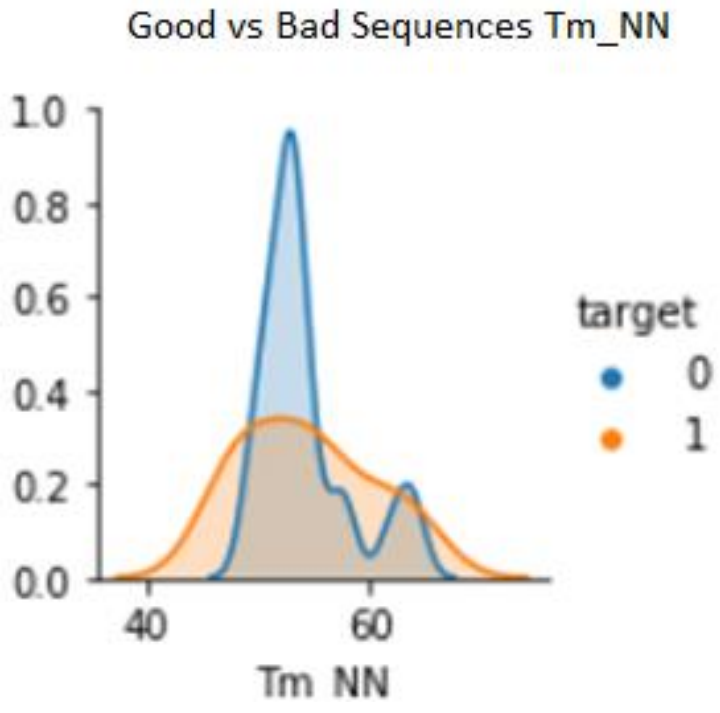


Ιντρόνια

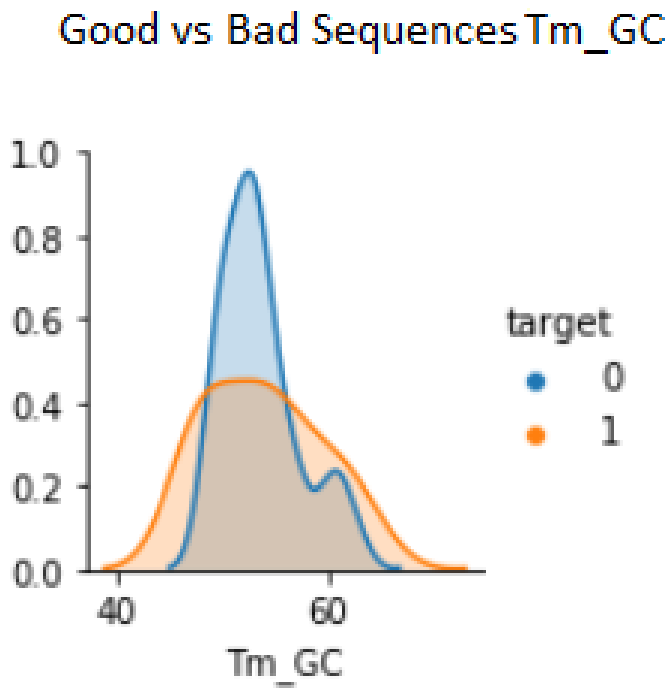


Παρατηρούμε ότι οι «κακές» αλληλουχίες (0) είναι πιο πιθανό να βρίσκονται σε εξόνια ενώ οι «καλές» (1) όχι. Επίσης οι καλές αλληλουχίες είναι πιο πιθανό να βρίσκονται σε εσόνιο από τις κακές. Αυτό μπορεί να αιτιολογηθεί από το γεγονός ότι στα intronia υπάρχουν τα exonic splicing enhancers που βοηθάνε στην ένωση μεταξύ τον εξονίων που κωδικοποιούνται. Έτσι με την τροποποίηση της αλληλουχίας σε αυτό το σημείο μπορεί να επιτευχθεί το exon skipping.

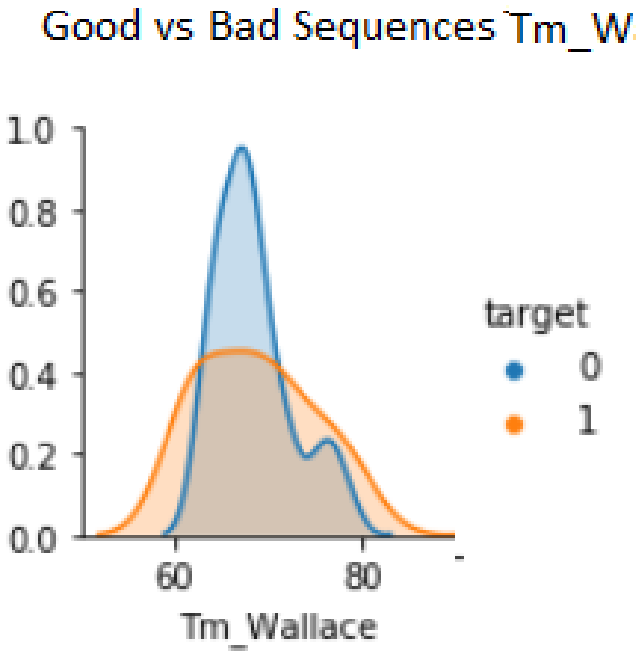
Melting Temperature Nearest Neighbour



Melting Temperature GC Content



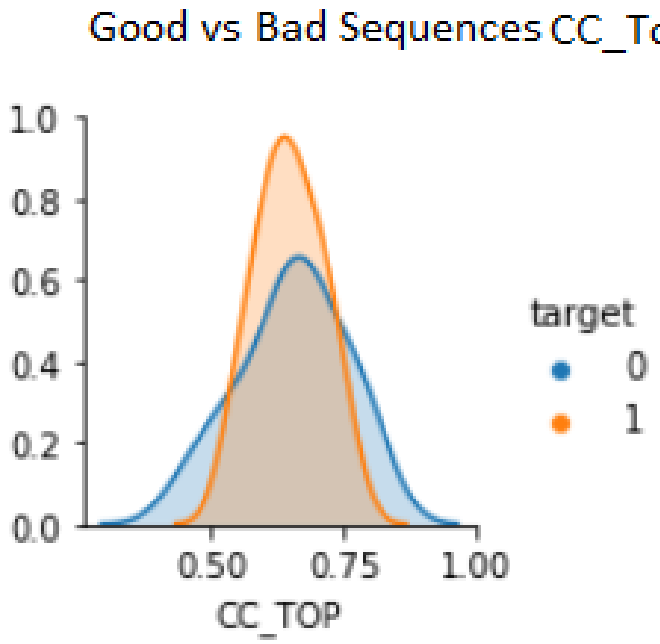
Melting Temperature Wallace Rule



Για τις παραπάνω μεταβλητές παρατηρούμε πως ακολουθούν και οι τρείς μια παρόμοια συνάρτηση κατανομής για τις δυο κλάσεις. Οι καλές αλληλουχίες φαίνεται να απλώνονται και σε μεγαλύτερες και σε μικρότερες θερμοκρασίες, ενώ οι κακές

περιορίζονται κοντά στο διάστημα 50-60. Θα μπορούσαμε να συμπεράνουμε λοιπόν ότι οι πιο ακραίες τιμές (είτε χαμηλές είτε υ υψηλές) ευνοούν το exon-skipping.

CC Top

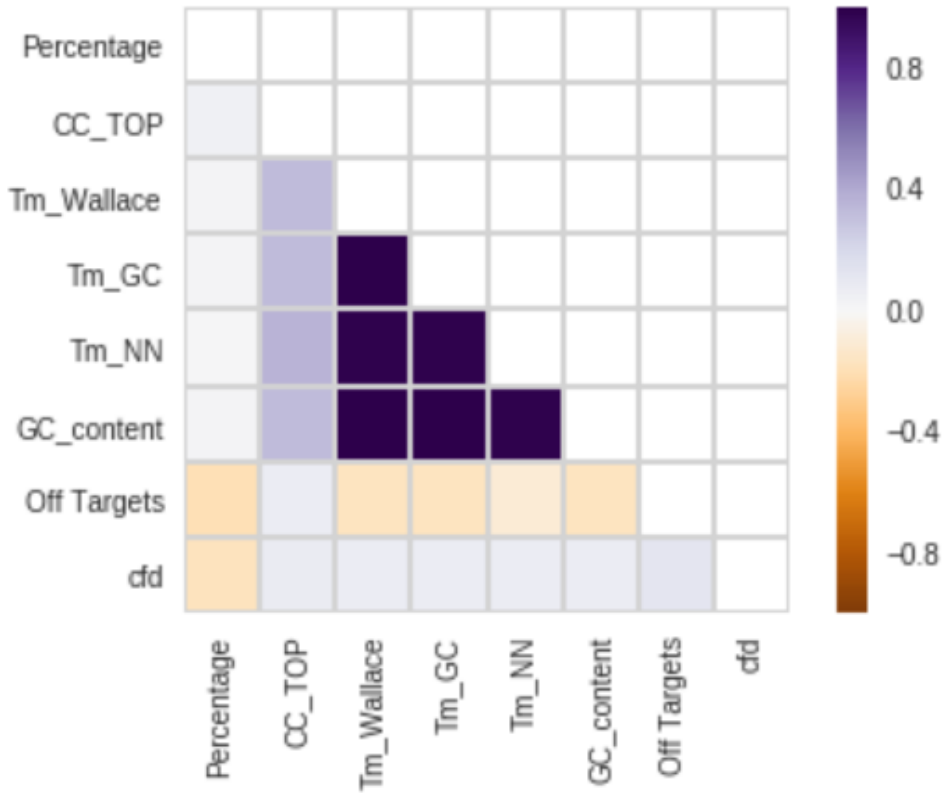


Παρατηρούμε πως οι καλές αλληλουχίες έχουν σκορ από το εργαλείο CC Top κυρίως συγκεντρωμένο στο 65% με μικρή διασπορά ενώ οι κακές έχουν μεγαλύτερη. Συμπεραίνουμε λοιπόν ότι το εργαλείο CC Top μπορεί να μας διευκολύνει στο να βρούμε τις κακές αλληλουχίες πιο εύκολα απ' ότι τις καλές.

Off Targets

Παρατηρούμε ότι τα off Targets στις καλές αλληλουχίες τείνουν να είναι λιγότερα σε σύγκριση με τις κακές. Αυτό θα μπορούσε να υποστηρίξει ότι οι καλές αλληλουχίες τείνουν να είναι μοναδικές στο DNA. Λόγω αυτού πολλές φορές η πρωτεΐνη cas9 μπορεί να δεθεί στην λάθος αλληλουχία (off target) αλλά εν τέλει να μην κόψει και σε εκείνο το σημείο το DNA.

Πινακας Συσχετήσεων



Παρατηρούμε ότι υπάρχει μεγάλη συσχέτιση μεταξύ των τριών μεταβλητών για τη θερμοκρασία στην οποία σπάνε οι δεσμοί μεταξύ των δεοξυριβονουκλεοτιδίων, κάτι λογικό, όμως κάθε μεταβλητή περιέχει διαφορετικές πληροφορίες και έτσι τις κρατάμε όλες για την εκπαίδευση του μοντέλου.

3.4 Αλγόριθμοι Μηχανικής Εκμάθησης

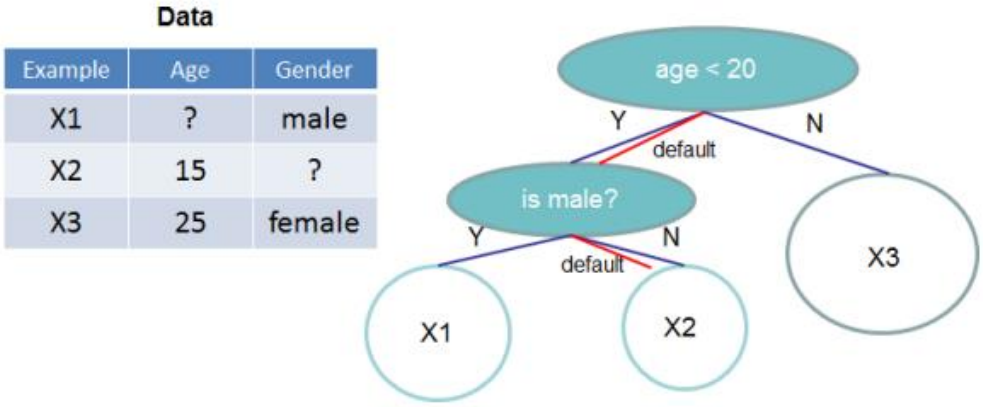
Η Μηχανική Μάθηση (machine learning) είναι κλάδος της Τεχνητής Νοημοσύνης και αφορά στη δημιουργία μοντέλων ή προτύπων από ένα σύνολο δεδομένων(training set) με τη χρήση υπολογιστικών συστημάτων. Υπάρχουν δύο κατηγορίες μηχανικής μάθησης:

- μάθηση με επίβλεψη (supervised learning)
- μάθηση χωρίς επίβλεψη (unsupervised learning)

Στην πρώτη περίπτωση ο αλγόριθμος κατασκευάζει μια συνάρτηση που απεικονίζει το σύνολο δεδομένων σε κλάσεις οι οποίες είναι γνωστές a priori, με απώτερο στόχο τη γενίκευση της συνάρτησης αυτής και για δεδομένα που δε γνωρίζουμε τη κλάση τους. Υπάρχουν δύο κατηγορίες προβλημάτων, τα προβλήματα ταξινόμησης (classification), δημιουργίας δηλαδή μοντέλων πρόβλεψης διακριτών τάξεων (κλάσεων) και τα προβλήματα παλινδρόμησης ή παρεμβολής(regression), δημιουργίας μοντέλων πρόβλεψης αριθμητικών τιμών. Στη δεύτερη κατηγορία, της μη επιβλεπόμενης μάθησης, ο αλγόριθμος κατασκευάζει ένα μοντέλο με βάση τα δεδομένα εισόδου χωρίς να γνωρίζει ποιες και πόσες είναι οι τιμές εξόδου, χρησιμοποιείται σε προβλήματα ομαδοποίησης και ανάλυσης συσχετισμών. Το δικό μας πρόβλημα είναι πρόβλημα ταξινόμησης, ανήκει στη πρώτη κατηγορία της επιβλεπόμενης μάθησης, αφού οι κλάσεις στις οποίες ανήκει το κάθε sgRNA είναι γνωστές και διακριτές. Υπάρχουν αρκετοί αλγόριθμοι ταξινόμησης, επιλέχθηκαν οι XGBoost, Decision Tree, Logistic Regression, Boosted Decision Tree και Random Forest για τα δεδομένα μας.

Δεντρο Απόφασης (Decision Tree)

Τα δέντρα απόφασης είναι ένας απλός αλγόριθμός όπου γίνεται η ταξινόμηση με βάση μια αλληλουχία δυαδικών αποφάσεων. Κάθε διαχωρισμός γίνεται με βάση το αν αυξάνεται η ομοιογένεια (συνηθέστερος ορισμός της ομοιογένειας είναι το κρητήριο Gini) με εφαρμογή του διαχωρισμού. Εάν γίνεται έχουμε ένα Node (διαχωρισμός) ενώ εάν δε γίνεται έχουμε ένα leaf. Ο αλγοριθμός μας ταξινομεί με βάση τα leaf. Λόγω της απλότητάς του είναι ένας απτούς πιο κατανοητούς αλγόριθμους και παρά την απλότητά του είναι πολλές φορές πολύ αποτελεσματικός. Στα άλλα πλεονεκτηματά του είναι ότι δεν επιρεάζεται από ασθενείς μεταβλητές, ενώ στα μειονεκτήματά του η απόδοση του βασίζεται πάρα πολύ στο σύνολο εκπαίδευσης και στη διακύμανση των δεδομένων εκπαίδευσης.



Εικ.11 Απεικόνιση Αλγορίθμου δέντρου απόφασης

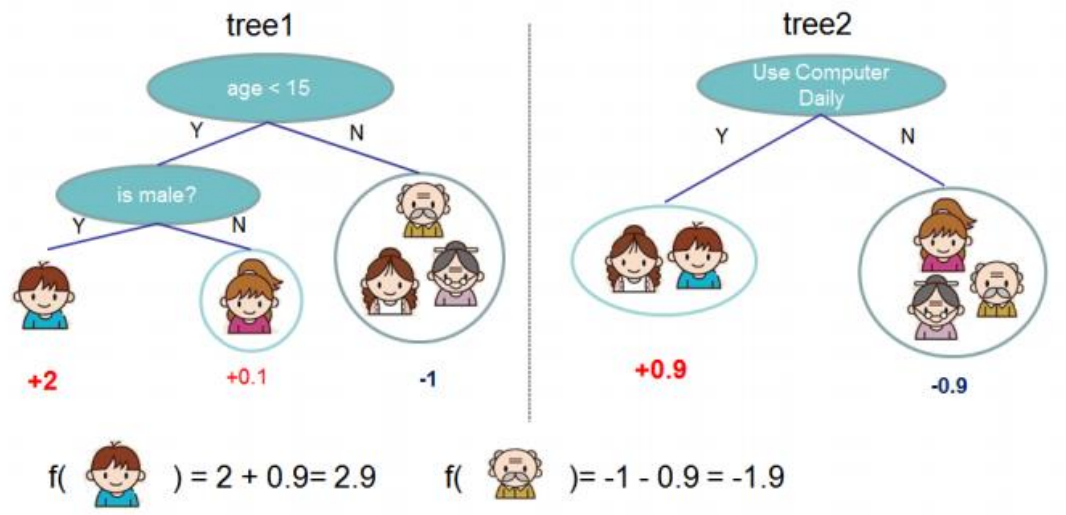
Λογιστική Παλινδρόμιση (Logistic Regression)

Η λογιστική παλινδρόμηση είναι ένα μοντέλο για την πρόβλεψη μιας κατηγορικής εξαρτημένης τιμής (κλάση) από 1 ή περισσότερες ανεξάρτητες μεταβλητές (features). Βασίζεται στη λογιστική ή σιγμοειδή συνάρτηση :

$$y = \frac{e^{\beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k}}{1 + e^{\beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k}}$$

όπου έστω ότι έχουμε 2 κλάσεις υπολογίζεται η πιθανότητα για την μια κλάση (0 ως 1) και με βάση το κατώφλι που έχει τεθεί (συνήθως 0.5) κατατάσσεται σε αυτήν αν >0.5 ή στη δεύτερη κλάση αν <= 0.5. Η εκτίμηση των παραμέτρων γίνεται με τη μέθοδο της μέγιστης πιθανοφάνειας (maximum likelihood). Εφαρμόζεται μόνο σε προβλήματα ταξινόμησης και ιδιαίτερα σε προβλήματα δυαδικής ταξινόμησης.

Δάσος δέντρων απόφασης (Random Forest)



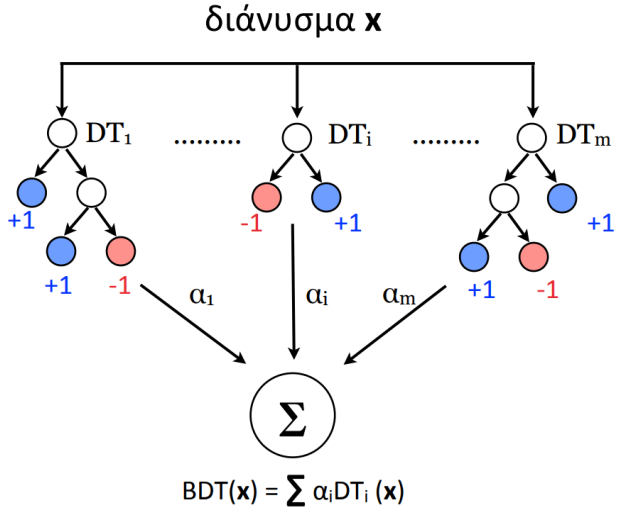
Εικ.12 Απεικόνιση αλγορίθμου Random Forest

Ο αλγόριθμος Random Forest ανήκει στην κατηγορία της συλλογικής μάθησης (ensemble). Κατά την εκπαίδευση του αλγορίθμου, τα δεδομένα εκπαίδευσης χωρίζονται σε υποσύνολα και προκύπτει ένα δέντρο απόφασης (decision tree) για κάθε υποσύνολο. Η τελική απόφαση παίρνεται από όλα τα δέντρα που δημιουργήθηκαν μέσω ψήφου πλειοψηφίας (majority vote). Ο συγκεκριμένος αλγόριθμος έχει πολλά πλεονεκτήματα όπως η διαχείριση μεγάλου όγκου δεδομένων και η απουσία overfitting, υπερπροσαρμογής δηλαδή πάνω στα δεδομένα εκπαίδευσης και η δυσκολία πρόβλεψης των νέων δεδομένων. Επίσης δεν επηρεάζεται από μεταβλητές που είναι ισχυρά συσχετισμένες μεταξύ τους. Χρησιμοποιείται σε προβλήματα ταξινόμησης και παλινδρόμησης ωστόσο είναι πιο αποτελεσματικός στην ταξινόμηση. Για την κατασκευή ενός δένδρου απόφασης (decision tree) αρχικά όλα τα δεδομένα είναι σε ένα κόμβο και ακολουθεί διάσπαση με βάση μια συνθήκη (if – then) για κάποιο από τα γνωρίσματα. Αναδρομική κλήση σε κάθε κόμβο ώστε τα δεδομένα που θα υπάρχουν σε ένα τελικό κόμβο (φύλλο – leaf) να ανήκουν σε μια κλάση.

Ενδυναμωμένο δέντρο απόφασης (Boosted Decision Tree)

Ο αλγόριθμος Boosted Decision Tree δημιουργεί ένα δάσος από δέντρα με την παρακάτω λογική:

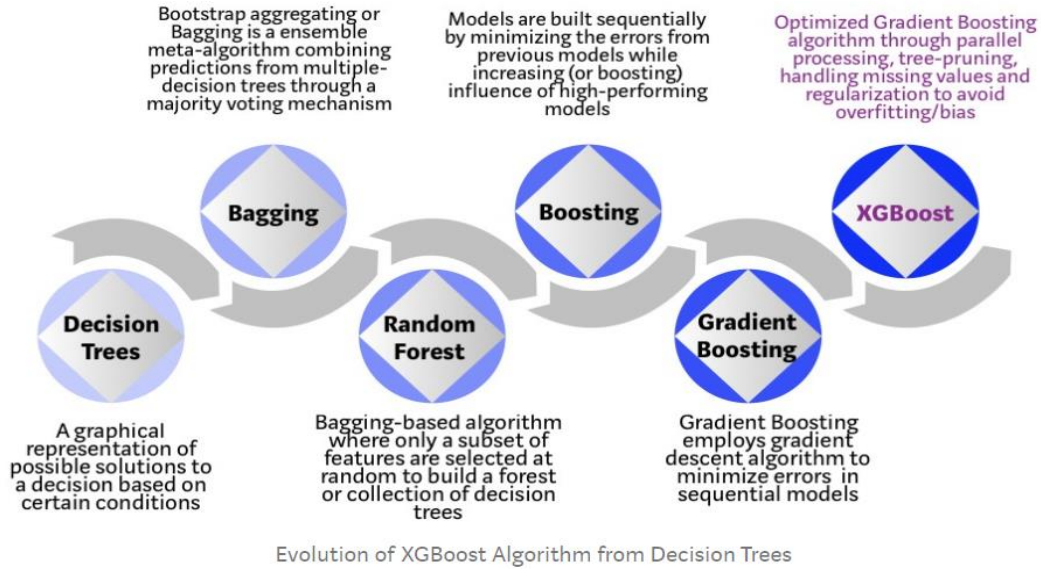
1. Αρχικοποιούμε τα βάρη ($w_i=1/N$)
2. Εκπαιδεύουμε το πρώτο δέντρο. Αφού το εκπαιδεύσουμε δίνουμε μεγαλύτερο βάρος στα διανύσματα που ταξινόμησε λάθος.
3. Κατασκευάσουμε ένα νέο δέντρο προσπαθεί να ταξινομήσει σωστά τα διανύσματα με το μεγαλύτερο βάρος. Αφού γίνει η ταξινόμηση ξαναμεταβάλουμε τα βάρη έτσι ώστε να έχουν μεγαλύτερο βάρος τα διανύσματα που ταξινόμησε λάθος.
4. Επαναλαμβάνουμε αυτή τη διαδικασία για όσα δέντρα ορίσουμε και ο τελικός ταξινομητής προκύπτει από το βεβαρυμμένο άθροισμα των ασθενών ταξινομητών



Εικ.13 Απεικόνιση αλγορίθμου Boosted Decision Tree

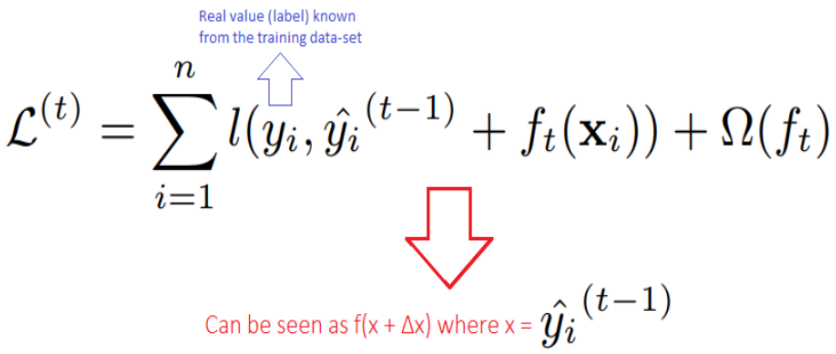
XGBoost

Ο αλγόριθμος XGBoost αποτελεί τη "εξέλιξη" των δέντρων απόφασης. Χρησιμοποιεί μεθόδους boosting και βελτιστοποίησης ώστε να πετύχει το καλύτερο δυνατό αποτέλεσμα σε λιγότερο χρόνο εκπαίδευσης.



Εικ.14 Εξέλιξη των αλγορίθμων δέντρων απόφασης

Η συνάρτηση κόστους του δίνεται από τον τύπο:



όπου ο όρος $f_t(x_i)$ είναι αυτός που βελτιώνει περισσότερο το προηγούμενο μοντέλο μας και ο $\Omega(f_t)$ είναι ο όρος τακτοποίησης (Regularization). Η συνάρτηση αυτή δεν βελτιστοποιείται στον ευκλείδειο χώρο. Έτσι κάνοντας χρήση του θεωρήματος του Taylor αλλάζουμε την αντικειμενική συνάρτηση έτσι ώστε να μπορούμε να χρησιμοποιήσουμε αλγορίθμους βελτιστοποίησης. Πιο συγκεκριμένα έχουμε:

$$f(x) \approx f(a) + f'(a)(x-a)$$

$$\Delta x = f_t(\mathbf{x}_i)$$

όπου f(x) είναι η αντικειμενική συνάρτηση f(α) είναι η πρόβλεψη του προηγούμενου βήματος και το υπόλοιπο μέρος είναι ο νέος μαθητευόμενος (learner). Έτσι προκήπτει:

$$\sum_{i=1}^{n} \left[g_i f_t(x_i) + \frac{1}{2} h_i f_t^2(x_i) \right] + \Omega(f_t)$$

$$g_i = \partial_{\hat{y}^{(t-1)}} l(y_i, \hat{y}^{(t-1)}), \quad h_i = \partial^2_{\hat{y}^{(t-1)}} l(y_i, \hat{y}^{(t-1)})$$

Ορίζεται το σύνολο των στιγμιότυπων στο φύλλο j ως:

- ανασυντάσσουμε την αντικειμενική συνάρτηση σε κάθε φύλλο

$$\begin{aligned} Obj^{(t)} &\simeq \sum_{i=1}^{n} \left[g_i f_t(x_i) + \frac{1}{2} h_i f_t^2(x_i) \right] + \Omega(f_t) \\ &= \sum_{i=1}^{n} \left[g_i w_{q(x_i)} + \frac{1}{2} h_i w^2_{q(x_i)} \right] + \gamma T + \lambda \frac{1}{2} \sum_{j=1}^{T} w_j^2 \\ &= \sum_{j=1}^{T} \left[(\sum_{i \in I_j} g_i) w_j + \frac{1}{2} (\sum_{i \in I_j} h_i + \lambda) w_j^2 \right] + \gamma T \end{aligned}$$

Το οποίο είναι το άθροισμα Τ ανεξάρτητων τετραγωνικών συναρτήσεων.

Στη συνέχεια ορίζεται το σκορ δομής (structure score).

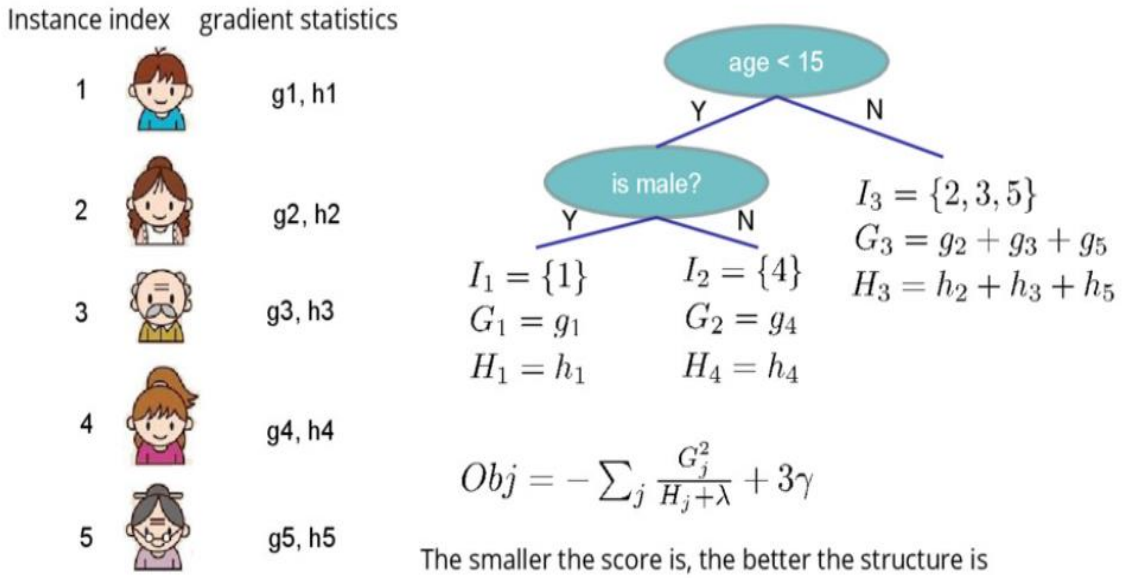
$$argmin_x\ Gx + \tfrac{1}{2}Hx^2 = -\tfrac{G}{H},\ H > 0 \qquad \min_x\ Gx + \tfrac{1}{2}Hx^2 = -\tfrac{1}{2}\tfrac{G^2}{H}$$

$$G_j = \sum_{i \in I_j} g_i \quad H_j = \sum_{i \in I_j} h_i$$

$$\begin{aligned} Obj^{(t)} &= \sum_{j=1}^{T} \left[(\sum_{i \in I_j} g_i) w_j + \tfrac{1}{2}(\sum_{i \in I_j} h_i + \lambda) w_j^2 \right] + \gamma T \\ &= \sum_{j=1}^{T} \left[G_j w_j + \tfrac{1}{2}(H_j + \lambda) w_j^2 \right] + \gamma T \end{aligned}$$

Υποθέτοντας πως η δομή του δέντρου είναι σταθερή, το βέλτιστο βάρος σε κάθε φύλο και η τελική αντικειμενική τιμή είναι:

$$w_j^* = -\frac{G_j}{H_j + \lambda} \qquad Obj = -\tfrac{1}{2} \sum_{j=1}^{T} \frac{G_j^2}{H_j + \lambda} + \gamma T$$



Εικ.15 Απεικόνιση σκορ δομής

Ο αλγόριθμος δομής του δέντρου φαίνεται παρακάτω:

- Ξεκινάμε από ένα δέντρο με βάθος 0 Για κάθε κόμβο φύλλων του δέντρου προσπαθούμε να προσθέσουμε ένα διαχωρισμό. Η αλλαγή του αποτελέσματος μετά την προσθήκη του διαχωρισμού είναι:

$$Gain = \frac{G_L^2}{H_L + \lambda} + \frac{G_R^2}{H_R + \lambda} - \frac{(G_L + G_R)^2}{H_L + H_R + \lambda} - \gamma$$

όπου ο πρώτος όρος είναι το σκορ του αριστερού φύλου, ο δεύτερος το σκορ του δεξιού φύλου και ο τρίτος είναι το κόστος πολυπλοκότητας λόγω του επιπλέον φύλου.

- Εάν το gain είναι αρνητικό τερματίζουμε

- Αυξάνουμε το βάθος των δέντρων στο μέγιστο και διώχνουμε τους διαχωρισμούς που δεν βελτιώνουν την ομοιογένεια.

Μεταβλητές αλγορίθμου

Οι μεταβλητές του αλγορίθμου αναλυτικά είναι:

1. Learning rate

Κάνει το μοντέλο πιο ισχυρό με τη συρρίκνωση των βαρών σε κάθε βήμα

Τυπικές τιμές που χρησιμοποιούνται: 0,01-0,2

2. min_child_weight

Ορίζει το ελάχιστο άθροισμα των βαρών όλων των παρατηρήσεων που απαιτούνται σε ένα παιδί.

Χρησιμοποιείται για τον έλεγχο over-fitting. Οι υψηλότερες τιμές εμποδίζουν ένα μοντέλο να μάθει σχέσεις που μπορεί να είναι ιδιαίτερα εξειδικευμένες στο συγκεκριμένο δείγμα που επιλέγεται για την εκπαίδευση του δέντρου.

Πολύ υψηλές τιμές μπορεί να οδηγήσουν σε under-fitting, επομένως, θα πρέπει να ρυθμηστεί με τη χρήση Cross Validation.

3. max_depth

Το μέγιστο βάθος του δέντρου

Χρησιμοποιείται για τον έλεγχο over-fitting αφού πολύ μεγάλο βάθος θα εκπαιδεύσει το μοντέλο σε σχέσεις του δείγματος εκπαίδευσης.

Τυπικές τιμές που χρησιμοποιούνται: 3-10

4. max_leaf_nodes

Ο μέγιστος αριθμός φύλλων του δέντρου. (2^{ν} όπου ν το βάθος του δέντρου)

5. gamma

Η ελάχιστη μείωση στη συνάρτηση κόστους ώστε να γίνει ένας διαχωρισμός.

Αναλόγως τη συνάρτηση κόστους έχουμε και διαφορετικές τιμές της gamma μεταβλητής

6. max_delta_step

Στο μέγιστο βήμα δέλτα επιτρέπουμε την εκτίμηση βάρους κάθε δέντρου.

Εάν έχει οριστεί σε θετική τιμή, μπορεί να συμβάλει στη μεγαλύτερη συντηρητικότητα του βήματος ενημέρωσης.

Συνήθως αυτή η παράμετρος δεν είναι απαραίτητη, αλλά μπορεί να βοηθήσει

στην λογιστική παλινδρόμηση όταν η κλάση είναι εξαιρετικά ανισορροπημένη.

This is generally not used but you can explore further if you wish.

7. subsample

Το μέγεθος του δείγματος που θα είναι τυχαίο δείγμα για κάθε δέντρο.

Οι χαμηλότερες τιμές καθιστούν τον αλγόριθμο πιο συντηρητικό και αποτρέπουν το overfitting, αλλά οι πολύ μικρές τιμές ενδέχεται να οδηγήσουν σε under-fitting.

Συνήθεις τιμές: 0.5-1

8. colsample_bytree [default=1]

Το πλήθος των στηλών να είναι τυχαία δείγματα για κάθε δέντρο.

Συνήθεις τιμές: 0.5-1

9. colsample_bylevel

Δείχνει την αναλογία υπο-δειγμάτων των στηλών για κάθε διαχωρισμό, σε κάθε επίπεδο.

10. lambda

Ο L2 όρος τακτοποίησης για τα βάρη

11. alpha

Ο L1 όρος τακτοποίησης για τα βάρη

Μπορεί να χρησιμοποιηθεί σε περίπτωση πολύ μεγάλης διαστάσεων, έτσι ώστε ο αλγόριθμος να τρέχει γρηγορότερα όταν υλοποιείται

12. scale_pos_weight

Μια τιμή μεγαλύτερη από 0 θα πρέπει να χρησιμοποιείται σε περίπτωση υψηλής ανισορροπίας δειγμάτων κλάσεων, καθώς βοηθάει στην ταχύτερη σύγκλιση.[21]

Overfitting

Ένα βασικό πρόβλημα στα προβλήματα μηχανικής εκμάθησης είναι το overfitting, δηλαδή το να έχουμε εκπαιδεύσει το μοντέλο μας αυστηρά πάνω στις διαφορές μεταξύ των δεδομένων εκπαίδευσης με αποτέλεσμα να χάσει την γενικότερη εικόνα του προβλήματος . Για αυτό το λόγο είναι απαραίτητο τα αξιολογήσουμε τα μοντέλα μας ώστε να ήμαστε σίγουροι πως δεν έχουμε φαινόμενο overfitting.Έτσι λοιπόν χρειάζεται να φτιάξουμε μεθόδους αξιολόγησης. Στα προβλήματα ταξινόμησης μπορούμε να χρησιμοποιήσουμε διάφορες μεθόδους. Οι βασικότερες είναι

- Η ακρίβεια (accuracy) όπου στο σύνολο αξιολόγησης μετράει το ποσοστό σωστών αξιολογήσεων.

- Το Recall όπου για το σύνολο αξιολόγησης υπολογίζει σε κάθε κλάση το λόγο των σωστά ταξινομήσεων παρατηρήσεων προς τον αριθμό των παρατηρήσεων που κανονικά έπρεπε να ταξινομηθούν σωστά.

- Το Precision όπου για το σύνολο αξιολόγησης σε κάθε κλάση υπολογίζει γτο λόγο των σωστά ταξινομήσεων παρατηρήσεων προς τον αριθμών που το μοντέλο ταξινόμησε ως τη συγκεκριμένη κλάση.

Μία λύση είναι το Cross Validation όπου χωρίζουμε τα δεδομένα και χρησιμοποιήσουμε τα μισά για εκπαίδευση και τα υπόλοιπα για αξιολόγηση.

Ένας άλλος τρόπος αξιολόγησης είναι να χωρήσουμε τα δεδομένα μας και να το εκπαιδεύσουμε και να το αξιολογήσουμε πολλές φορές(k-Fold Cross Validation).

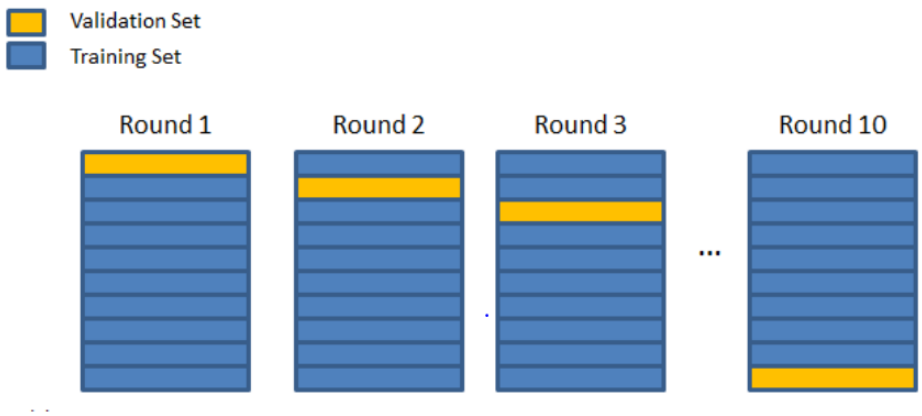
- Χωρίζουμε τα δεδομένα μας σε k ίσα κομμάτια

- Κρατάμε έξω από τα δεδομένα της ένα κομμάτι κάθε φορά και χρησιμοποιούμε τα υπόλοιπα για εκπαίδευση.

- Αξιολογούμε το μοντέλο στο κομμάτι που άφησαμε από έξω.

Επαναμβάνουμε αυτή τη διαδικασία μέχρι να έχουμε χρησιμοποιήσει όλα τα κομμάτια για εκπαίδευση από μία φορά μόνο. Έπειτα υπολογίζουμε το μέσο όρο

των αξιολογήσεων αυτών. Αυτή η μέθοδος λοιπόν είναι πιο γενικευμένη με αποτέλεσμα να είναι πιο αξιόπιστη.



the process of cross validation in general

Leave One Out

Με αυτόν τον τρόπο αξιολόγησης αφήνουμε μία παρατήρηση έξω από το σύνολο εκπαίδευσης. Εάν υπήρχαν n παρατηρήσεις θα χρησιμοποιούσαμε τις n-1 για εκπαίδευση και 1 για αξιολόγηση. Αυτό επαναλαμβάνεται n φορές ώστε να χρησιμοποιηθούν όλες οι παρατηρήσεις και για αξιολόγηση και για εκπαίδευση. Έπειτα παίρνουμε το μέσο όρο της αξιολόγησης των n εκπαιδευμένων μοντέλων.

Η καμπύλη λειτουργικών χαρακτηριστικών (ROC curve)

Η Η καμπύλη ROC εκφράζει τη σχέση μεταξή των αληθώς θετικών (True Positive-TP) και ψευδώς θετικών(False Positive-FP) αποτελεσμάτων της ταξινόμισης, καθώς μεταβάλλεται προοδευτικά προς μια κατεύθυνση το διαχωριστικό όριο.

Η θέση κάθε σημείου της καμπύλης ROC προσδιορίζεται από ένα ορισμένο % TP και % FP αποτελεσμάτων της δοκιμασίας που αντιστοιχεί στο συγκεκριμένο διαχωριστικό όριο. Η καμπύλη αυτή εγγράφεται μέσα σε ένα τετράγωνο, στις τέσσερις γωνίες του οποίου αντιστοιχούν οι ακραίες τιμές (0 και 1) του % TP και του %FN αποτελεσμάτων, καθώς και των συμπληρωματικών αυτών ποσοστων.

Επιλογή Μεταβλητών

Σε πολλά προβλήματα μηχανικής εκμάθησης έχουμε πολλές μεταβλητές στα δεδομένα μας. Τις περισσότερες φορές όμως δεν έχουν όλες οι μεταβλητές καλή διαχωριστική ικανότητα. Υπάρχουν διάφοροι μέθοδοι στατιστικής για να επιλέξει κανείς μεταβλητές όπως το επίπεδο σημαντικότητας της μεταβλητής (feature importance) των δέντρων απόφασης ή το recursive feature elimination. Στη συγκεκριμένη εργασία χρησιμοποιήσαμε το recursive feature elimination.

Recursive Feature Elimination

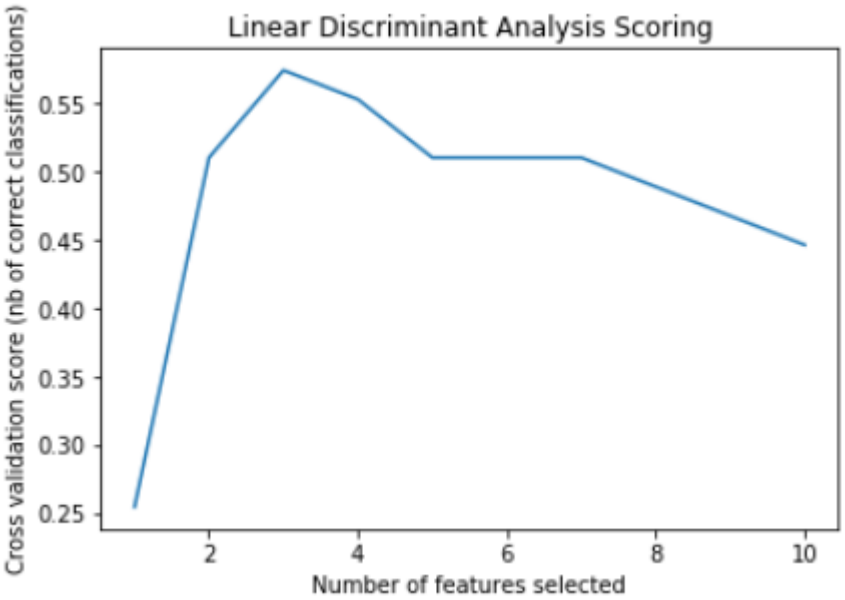
Το recursive feature elimination λειτουργεί ως εξής:

1. Ξεκινάμε και εκπαιδεύουμε το μοντέλο μας με μια μόνο μεταβλητή και αξιολογούμε την απόδοσή του.
2. Κρατάμε τη μεταβλητή με τη καλύτερη απόδοση.
3. Κρατώντας τη μια μεταβλητή από το προηγούμενο βήμα, εκπαιδεύουμε το μοντέλο μας με όλα τα πιθανά ζευγάρια μεταβλητών και αξιολογούμε την απόδοσή του.
4. Κρατάμε το ζευγάρι με την καλύτερη απόδοση.
5. Επαναλαμβάνουμε τα παραπάνω μέχρι να φτάσουμε στο σημείο που χρησιμοποιούνται όλες οι μεταβλητές για εκπαίδευση και αξιολογούμε το μοντέλο μας
6. Βλέπουμε ποιος συνδυασμός μεταβλητών είχε τη καλύτερη αξιολόγηση και χρησιμοποιούμε αυτόν για την εκπαίδευση του αλγορίθμου.

4 ΣΥΓΚΡΙΣΗ ΜΟΝΤΕΛΩΝ ΚΑΙ ΑΞΙΟΛΟΓΗΣΗ

4.1 Σύγκριση Αλγορίθμων

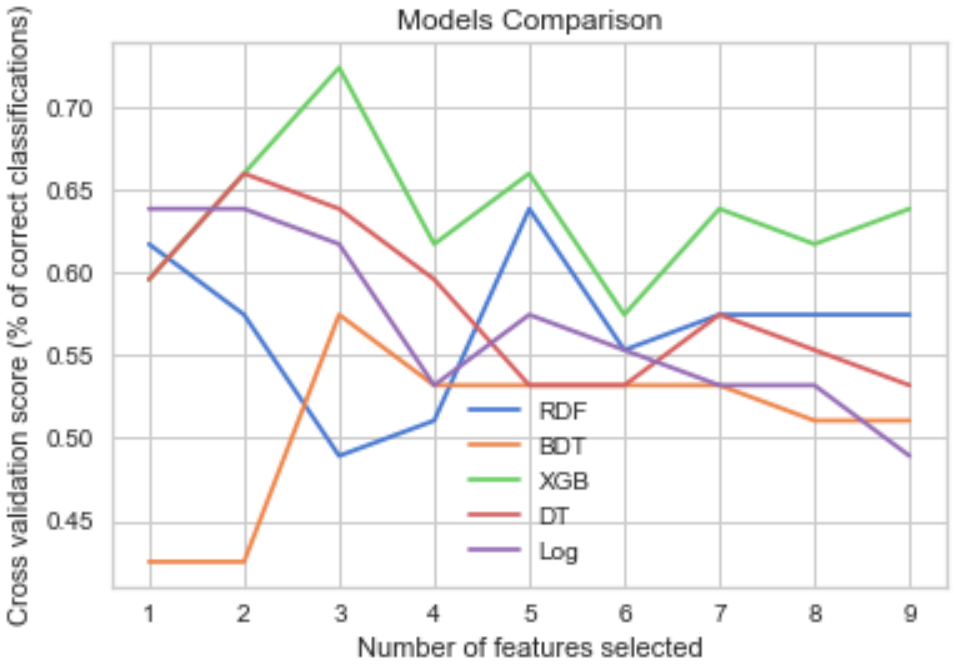
Για να υπάρχει μια βάση για τη συνέχεια εξετάστηκε ένας απλός γραμμικός ταξινομητής Linear Discriminant Analysis[29].



Χρησιμοποιώντας recursive feature elimination με αξιολόγηση σε Leave One Out η υψηλότερη απόδοση που πήραμε με τον αλγόριθμο Linear Discriminant Analysis ήταν 57%.

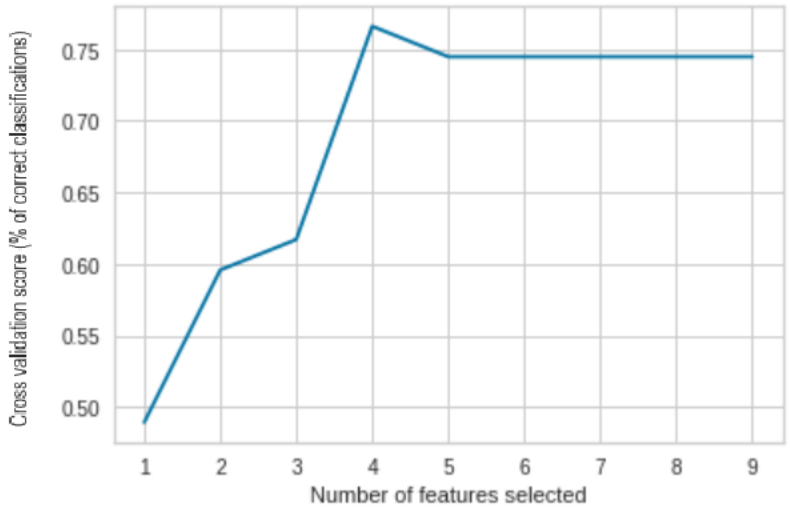
Στη συνέχεια εξετάσαμε 5 αλγορίθμους,

- XGBoost με αρχικοποίηση βαρών 0.5, με 100 δέντρα με βάθος 3 και learning rate 0.1
- Boosted decision tree με 50 δέντρα και βάθος 3
- Random Forest με 50 δέντρα και βάθος 3
- Decision Tree με βάθος 3
- Logistic Regression



Κάνοντας Recursive Feature Elimination για όλους του αλγορίθμους και αξιολογόντας τα μοντέλα με leave one out για κάθε σετ δεδομένων, πήραμε το παραπάνω διάγραμμα. Ο αλγόρθυμος που αποδίδει καλύτερα είναι ο XGBoost, το decision tree και ο Random Forest.

Επιλέγοντας κατάλληλες υπερπαραμέτρους για τον XGBoost (base_score=0.4, gamma=0.4, learning_rate=0.001, max_depth=7, n_estimators=1000) και επαναλανμβάνοντας το Recursive Feature Elimination, πήραμε το παρακάτω διάγραμμα.



Σύμφωνα με το παραπάνω ο αλγόριθμος αποδίδει καλύτερα αν κρατήσουμε 4 από τα εννέα χαρακτηριστικά μας. Πιο συγκεκριμένα αρκεί να κρατήσουμε τις μεταβλητές:

- Exon
- Tm_NN

- Tm_GC
- CC_Top

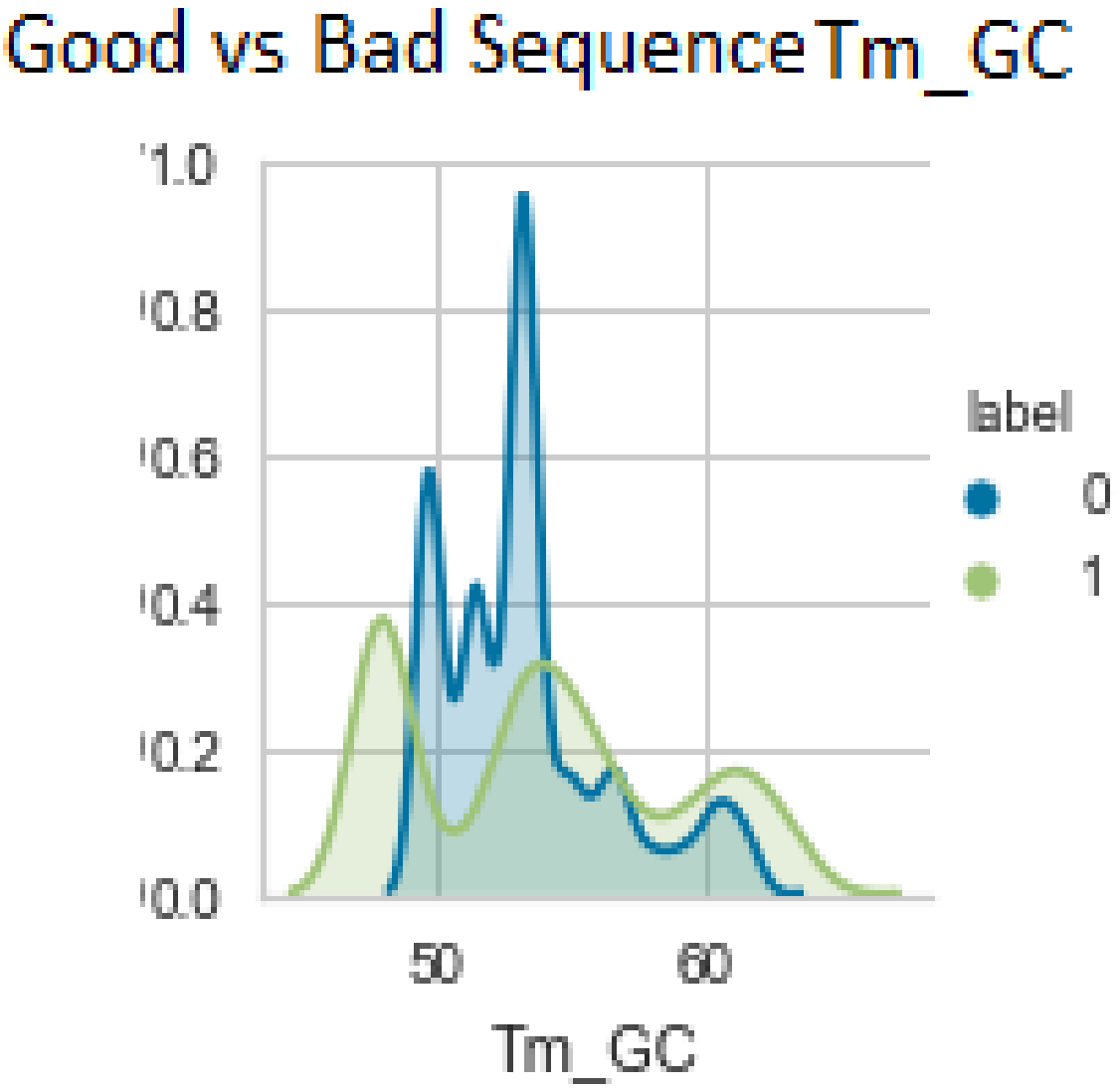
Οι μεταβλητές Tm_NN και Tm_GC αν και είναι πολύ συσχετισμένες μεταξύ τους, περιέχουν διαφορετικές πληροφορίες και έτσι κρατάμε και τις δύο.

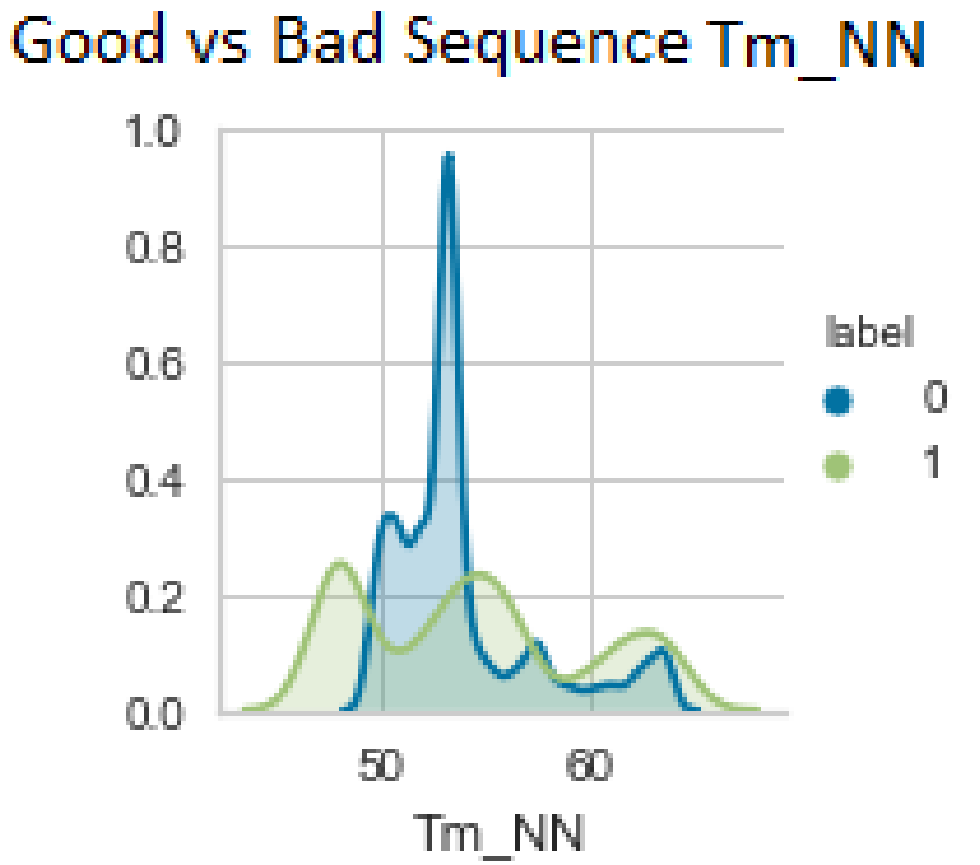
4.2 Προσομοίωση Σε Τεχνητά Δεδομένα

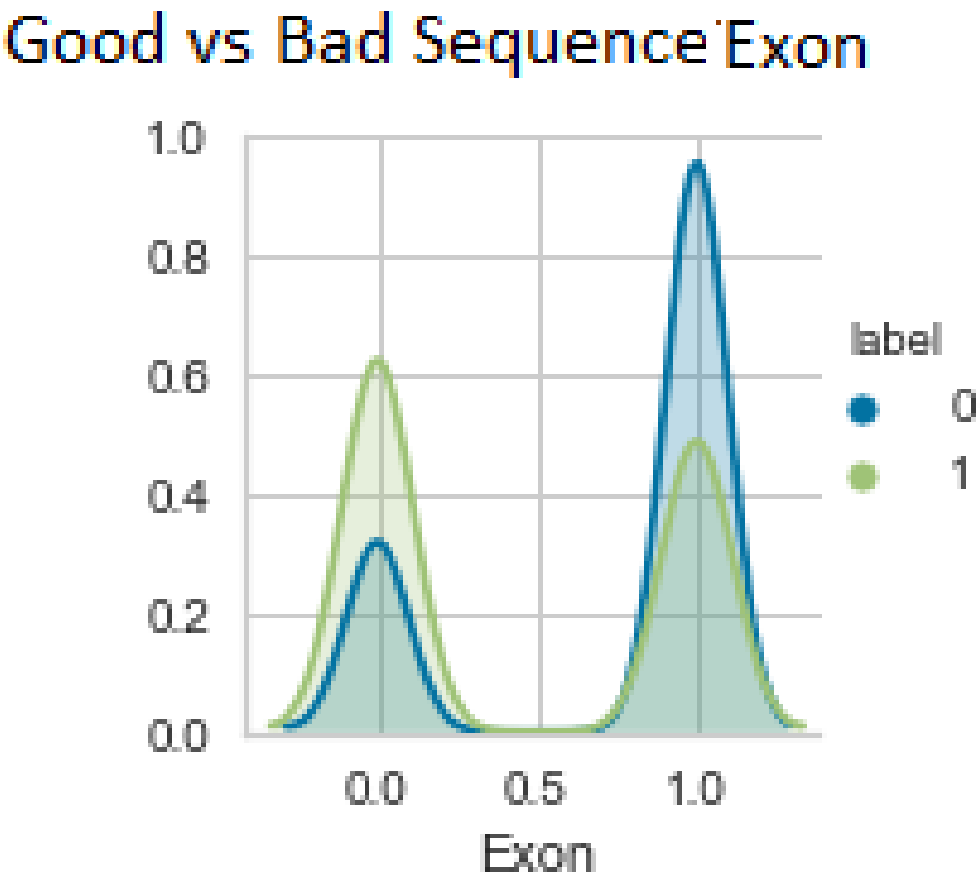
Ο αλγόριθμος SMOTE

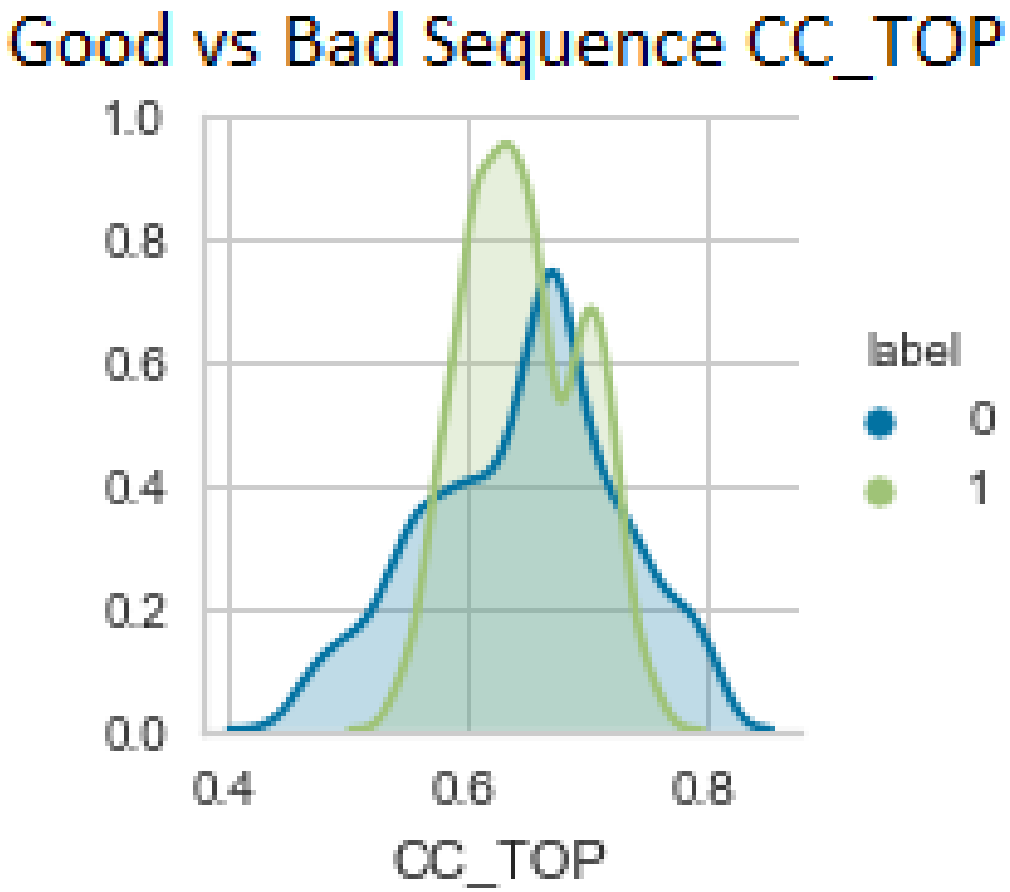
Ο αλγόριθμος SMOTE συνήθως χρησιμοποιείται ως μέθοδος upsampling. Ο SMOTE δημιουργεί γραμμές μεταξύ της κλάσης που έχει λιγότερα δεδομένα. Για να δημιουργήσουμε διανύσματα με τον SMOTE δημιουργήσαμε πολλές φορές ίδια διανύσματα από κάθε κλάση και χρησιμοποιήσαμε τον αλγόριθμο δύο φορές.

Συγκεκριμένα δημιουργήσαμε 2368 χαρακτηριστικά αλληλουχιών για τις καλές αλληλουχίες και 2377 για τις κακές. Οι συναρτήσεις πυκνότητας πιθανότητας των τεχνητών δεδομένων φαίνονται παρακάτω:

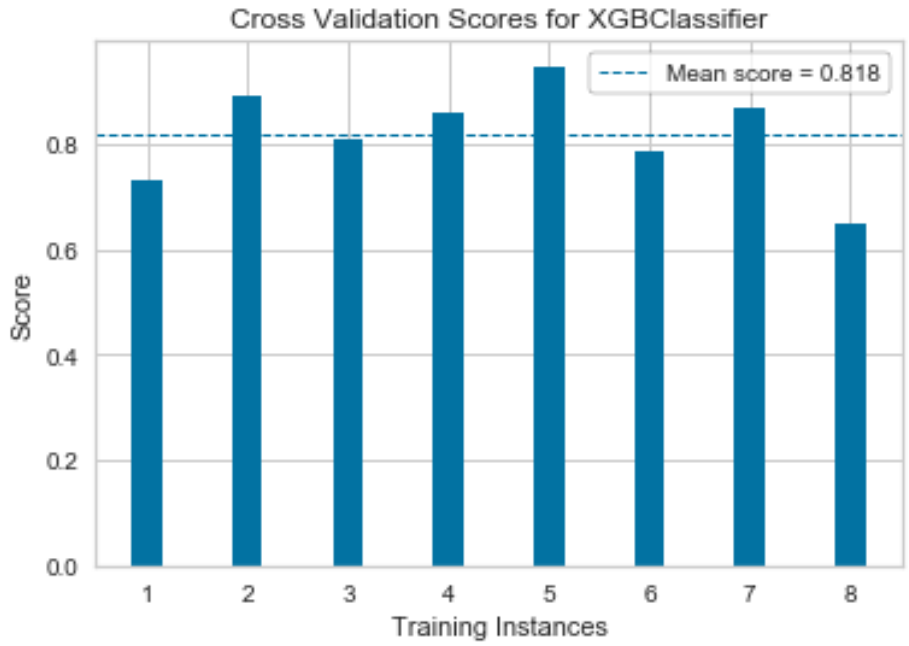


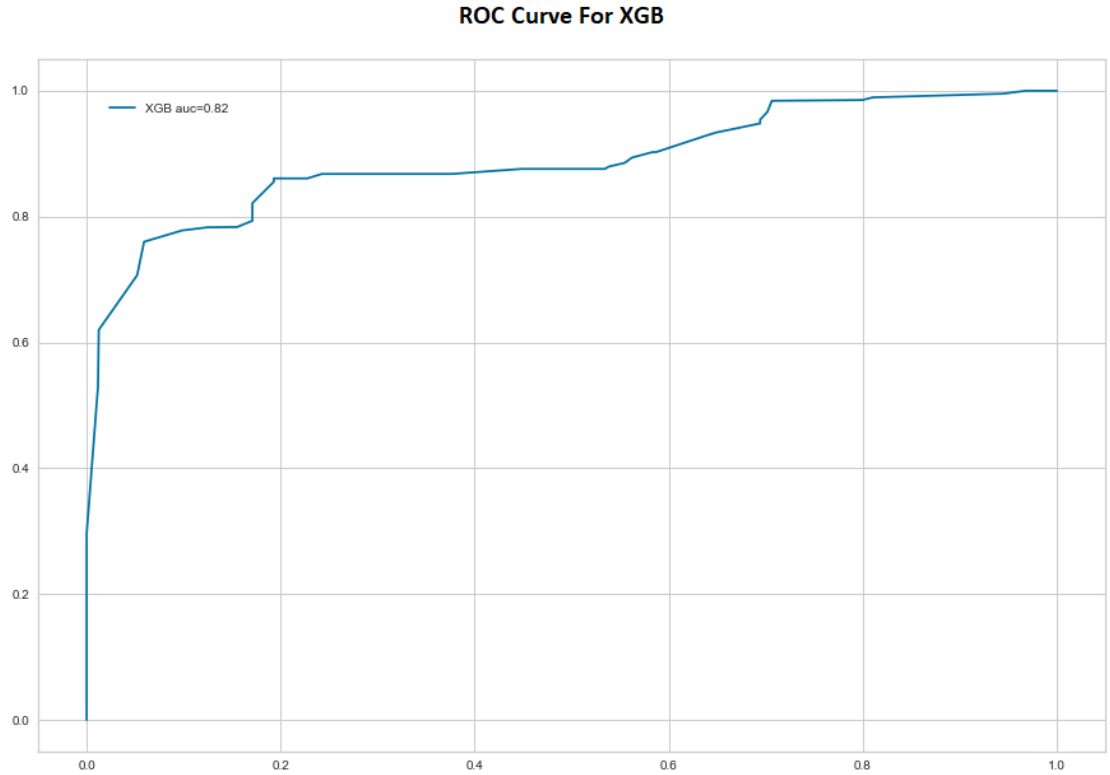

Good vs Bad Sequence Tm_NN
1.0
0.8
0.6
0.4
0.2
0.0
50
60
Tm_NN
label
0
1


Good vs Bad Sequence Exon
1.0
0.8
0.6
0.4
0.2
0.0
0.0
0.5
1.0
Exon
label
0
1


Good vs Bad Sequence CC_TOP
1.0
0.8
0.6
0.4
0.2
0.0
0.4
0.6
0.8
CC_TOP
label
0
1

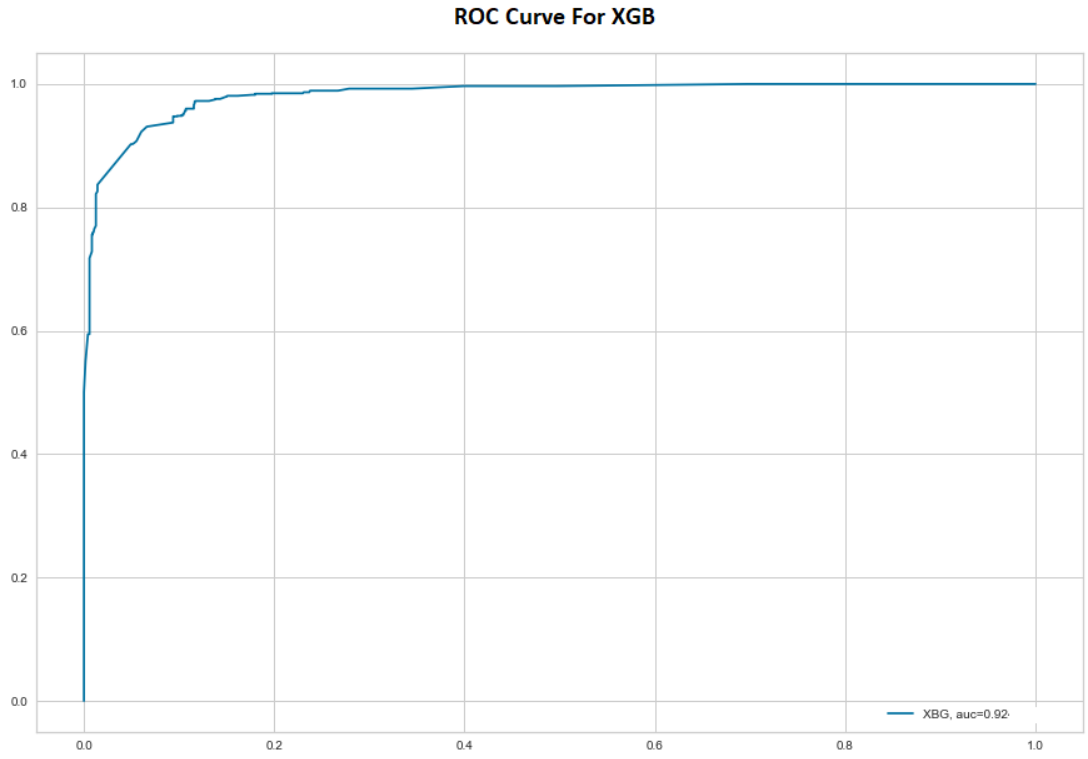
Εκπαιδεύσαμε τον αλγόριθμο στα αρχικά δεδομένα και χρησιμοποιήσαμε 8-fold cross-validation ώστε να αξιολογήσουμε το μοντέλο σε όλα τα τεχνητά δεδομένα. Τα αποτελέσματα φαίνονται στο παρακάτω διάγραμμα. Το μοντέλο μας έχει μέσο accuracy 81.8%. Υπάρχει αρκετά μεγάλη διασπορά μεταξύ των αποτελεσμάτων κάτι που οφείλεται στο μικρό όγκο δεδομένων εκπαίδευσης.





Η καμπύλη Roc του μοντέλου μας είναι αρκετά καλή αλλά δεν είναι ομαλή λόγω μικρού όγκου των δεδομένων εκπαίδευσης. Χωρίζοντας τα τεχνητά δεδομένα σε δύο ισοπληθηκές ομάδες μία για εκπαίδευση και μια για αξιολόγηση μπορούμε

να σχεδιάσουμε την παρακάτω καμπύλη ROC. Η παρακάτω καμπύλη μας δείχνει ότι το μοντέλο μας έχει πολύ καλή διαχωριστική ικανότητα.



5 Περίληψη Αποτελεσμάτων

Με βάση τον αλγόριθμο τα καλύτερα χαρακτηριστικά που μπορούν να χρησιμοποιηθούν για τη πρόβλεψη της καταλληλόλητας της αλυσίδας είναι η θέση του στόχου μας, εάν δηλαδή βρίσκεται σε εξόνιο ή όχι, η θερμοκρασία λιωσίματος της αλυσίδας με βάση τη περιεκτικότητα σε γουανίνη και κυτοσίνη και με το κανόνα του πλησιέστερου γείτονα καθώς και το σκορ που μας δίνει το εργαλείο CC top που βασίζεται στις θέσεις των νουκλεοτιδίων στην αλυσίδα που στοχεύουμε.

Καταφέραμε να φτιάξουμε ένα μοντέλο που μπορεί να αποδώσει και στο μικρό όγκο δεδομένων που αρχικά είχαμε, αλλά και σε μια προσομοίωση. Έτσι λοιπόν μπορούμε να χρησιμοποιήσουμε το μοντέλο όπως είναι παρά τον μικρό όγκο δεδομένων που χρησιμοποιήσαμε για αυτό. Έπειτα αφού εκπαιδεύοντας το και αξιολογώντας το στα τεχνητά δεδομένα καταφέραμε να δείξουμε ότι αν ο ίδιος αλγόριθμος είχε περισσότερα δεδομένα για εκπαίδευση θα απέδιδε καλύτερα. Έτσι το εργαλείο μπορεί να χρησιμοποιηθεί άμεσα, αλλά με την προσθήκη νέων δεδομένων στο σύνολο εκπαίδευσης θα μπορέσει να βελτιωθεί και η απόδοσή του.

6 ΒΙΒΛΙΟΓΡΑΦΙΑ

[1] National Institute of Neurological Disorders and Stroke (NINDS)

[2] Duchenne de Boulogne GBA. De l'electrisation Localisee et de Son Application a la Pathologie et a la Therapeutique . 2nd ed. Paris: Bailiere & Fils; 1861. 4

[3] Little W. On the nature and treatment of the deformities of the human frame: being a course of lectures delivered at the Royal Orthopaedic Hospital in 1843: With numerous notes and additions, to the present time. London: Longman, Brown, Green and Longmans; 1853. pp. 14 16. 3

[4] Darras, B. T., Menache-starobinski, C. C., Hinton, V., & Kunkel, L. M. (2015). Dystrophinopathies

[5] https://genes-genetics.weebly.com/inheritance.html

[6] CRISPR-Mediated AdaptiveImmune Systems in Bacteriaand Archaea(2013) 2. Doudna lab CRISR systems

[7] Diverse evolutionary roots and mechanistic variations of the CRISPR-Cas systems (2016)

[8] RNA and DNA Targeting by a Reconstituted *Thermus thermophilus* Type III-A CRISPR-Cas System(2017)

[9] https://www.addgene.org/guides/crispr/

[10] Applications of CRISPR/Cas9 for the Treatment of Duchenne Muscular Dystrophy. Lim, Yoon, Yokota

[11] CCTop: An Intuitive, Flexible and Reliable CRISPR/Cas9 Target Prediction Tool. Stemmer M1, Thumberger T1, Del Sol Keyer M1, Wittbrodt J1, Mateo JL1.

[12a] Doench, J. G., Fusi, N., Sullender, M., Hegde, M., Vaimberg, E. W., Donovan, K. F., Root, D. E. (2016). Optimized sgRNA design to maximize activity and minimize off-target effects of CRISPR-Cas9. Nature biotechnology, 34(2), 184-191. doi:10.1038/nbt.3437 [Nat Biotechnol]

[12b] Sanson, K. R., Hanna, R. E., Hegde, M., Donovan, K. F., Strand, C., Sullender, M. E., Doench, J. G. (2018). Optimized libraries for CRISPR-Cas9 genetic screens with multiple modalities. Nature communications, 9(1), 5416. doi:10.1038/s41467-018-07901-8 [Nat Commun]

[13] Wallace RB et al. (1979) Nucleic Acids Res 6:3543-3557, PMID 158748

[14] Marmur J and Doty P (1962) J Mol Biol 5:109-118

[15] John SantaLucia Jr. (1998). "A unified view of polymer, dumbbell, and oligonucleotide DNA nearest-neighbor thermodynamics".

[16] Stemmer, M., Thumberger, T., del Sol Keyer, M., Wittbrodt, J. and Mateo, J.L. " CCTop: an intuitive, flexible and reliable CRISPR/Cas9 target prediction tool. PLOS ONE (2015). doi:10.1371/journal.pone.0124633

[17] Li HL, Fujimoto N, Sasakawa N, Shirai S, Ohkame T, Sakuma T, Tanaka M, Amano N, Watanabe A, Sakurai H, Yamamoto T, Yamanaka S, Hotta A. Precise correction of the dystrophin gene in duchenne muscular dystrophy patient induced pluripotent stem cells by TALEN and CRISPR-Cas9. (2015)

[18] Young CS, Hicks MR, Ermolova NV, Nakano H, Jan M, Younesi S, Karumbayaram S, Kumagai-Cresse C, Wang D, Zack JA, Kohn DB, Nakano A, Nelson SF, Miceli MC, Spencer MJ, Pyle AD. A Single CRISPR-Cas9 Deletion Strategy that Targets the Majority of DMD Patients Restores Dystrophin Function in hiPSC-Derived Muscle Cells. (2016)

[19] Ousterout DG, Kabadi AM, Thakore PI, Majoros WH, Reddy TE, Gersbach CA. Multiplex CRISPR/Cas9-based genome editing for correction of dystrophin mutations that cause Duchenne muscular dystrophy.

[20] http://doudnalab.org/research_areas/crispr-systems/

[21] Tianqi Chen, Carlos Guestrin, XGBoost: A Scalable Tree Boosting System (2016)

[22] https://towardsdatascience.com/https-medium-com-vishalmorde-xgboost-algorithm-long-she-may-rein-edd9f99be63d

[23] https://horizondiscovery.com/en/products/gene-editing/gene-editing-reagents/hdr-knock-in-templates

[24] U.S National Library of Medicine
https://ghr.nlm.nih.gov/gene/DMD#location

[25] https://genes-genetics.weebly.com/inheritance.html

[26] Long C, Li H, Tiburcy M, Rodriguez-Caycedo C, Kyrychenko V, Zhou H, Zhang Y, Min YL, Shelton JM, Mammen PPA, Liaw NY, Zimmermann WH, Bassel-Duby R, Schneider JW, Olson EN, Correction of diverse muscular dystrophy mutations in human engineered heart muscle by single-site genome editing.

[27] Tina Liu- poster on DNA and RNA-targeting Type III CRISPR system.

[28] Διαλέξεις Κ. Κουσουρή (2017) Αναγνώριση Προτύπων και Νευρωνικά Δίκτυα. Εθνικό Μετσόβιο Πολυτεχνείο

[29] https://machinelearningmastery.com/linear-discriminant-analysis-for-machine-learning/