Predicció dels resultats d'edició genòmica amb CRISPR-Cas9 i *base editors* a partir de la seqüència de la regió modificada

Marc Expòsit Goy
Màster Universitari en Bioinformàtica i Bioestadística UOC-UB
Àrea 3.3. Estudi de dades òmiques amb tècniques d'intel·ligència artificial

Consultor: Albert Plà Planas
Consultor extern: Marc Güell Cargol
Professor responsable de l'assignatura: Ferran Prados Carrasco

24 de Juny de 2020

FITXA DEL TREBALL FINAL

Títol del treball:	*Predicció dels resultats d'edició genòmica amb CRISPR-Cas9 i base editors a partir de la seqüència de la regió modificada*
Nom de l'autor:	*Marc Expòsit Goy*
Nom del consultor/a:	*Albert Plà Planas*
Nom del PRA:	*Ferran Prados Carrasco*
Data de lliurament (mm/aaaa):	*06/2020*
Titulació o programa:	*Máster Universitari en Bioinformàtica i Bioestadística UOC-UB*
Àrea del Treball Final:	*Àrea 3.3. Estudi de dades òmiques amb tècniques d'intel·ligència artificial*
Idioma del treball:	*Català*
Paraules clau	*CRISPR gene editing outcomes, classification models, machine learning*

Resum del Treball (màxim 250 paraules): *Amb la finalitat, context d'aplicació, metodologia, resultats i conclusions del treball*

L'ús de les tècniques d'edició genòmica com a teràpia està limitat per un control gairebé nul dels productes d'edició genètica. La seqüència de la regió editada determina en gran part les mutacions introduïdes. En aquest treball, s'utilitzen models d'aprenentatge automàtic per predir els productes d'edició genètica de CRISPR-Cas9 a partir de la seqüència del gRNA. Així, es podria fer un disseny intel·ligent de la regió a editar per controlar els productes d'edició genètica, acostant aquestes tècniques a la pràctica clínica.

A diferència dels estudis previs, que introdueixen modificacions en seqüències sintètiques, en aquest estudi es realitzen edicions en 1785 regions úniques del genoma. Per tant, les dades experimentals reflecteixen de forma més realista les condicions clíniques.

A través de l'anàlisi de les regions genòmiques d'interès per seqüenciació de nova generació es conclou que falta profunditat de seqüenciació per observar edicions genètiques en les dades experimentals. Per això, es simulen les dades a partir de models computacionals ja existents.

El model de predicció de l'eficiència es planteja com un classificador binari, i l'algoritme que aconsegueix major exactitud és el logistic regression. Aquest model recrea les eficiències del model utilitzat per simular les dades de forma eficaç. El problema de predicció dels resultats d'edició es planteja en dues aproximacions diferents que cal seguir desenvolupant.

En resum, aquest treball planteja l'aproximació que cal seguir i desenvolupa tots els processos necessaris per passar de les dades genòmiques experimentals a l'entrenament d'un model computacional per predir els resultats d'edició genètica a partir de la seqüència.

Abstract (in English, 250 words or less):

The potential use of gene editing technologies as therapeutics is limited by the lack of control in the outcomes of gene editing. These outcomes are determined, in part, by the sequence of the edited region. In this work, a machine learning model is used to predict the outcomes of CRISPR-Cas9 gene editing from the sequence of the gRNA. This model could be used to improve gRNA design so that gene editing outcomes are controlled.

While previous studies introduce mutations in synthetic target sequences, in this work insertions are done in 1785 unique regions of the genome. Hence, experimental data reflect more closely the conditions in which the techniques would be applied in the clinic.

Analyzing the target genomic regions reveals that sequencing coverage is not enough to quantify gene editing outcomes. Hence, these are simulated using previously developed models. Simulated data is treated in the same was as it would be done with experimental data.

The gRNA efficiency prediction model is developed as a binary classifier, and logistic regression is the algorithm with the higher accuracy. The predictions are similar between this model and the original model used to simulate the data. The model to predict gene editing outcomes is planned using two different approaches that require further development.

In brief, this work defines the steps and develops all the processes needed to go from experimental genomic data to the training of a computational model that predicts gene editing outcomes from the gRNA sequence.

Índex

1. Introducció 1
1.1 Context i justificació del Treball 1
1.2 Objectius del Treball 2
1.3 Enfocament i mètode seguit 4
1.4 Planificació del Treball 5
1.5 Breu sumari de productes obtinguts 8
1.6 Breu descripció dels altres capítols de la memòria 10
2. Context 11
2.1. L'edició genètica 11
2.2. Estudis previs 14
3. Mètodes 16
3.1. Experiment d'edició genètica 16
3.2. Preparació i anàlisi descriptiu de les dades de seqüenciació 17
3.3. Model computacional 20
4. Resultats i discussió 30
4.1. Anàlisi descriptiu de les dades de seqüenciació 30
4.2. Model computacional 37
5. Conclusions 53
6. Glossari 54
7. Bibliografia 55

Llista de figures

Figura 1. Cronograma del treball plantejat inicialment. 8
Figura 2. Cronograma del treball una vegada finalitzat. 8
Figura 3. Sistema d'edició genètica CRISPR-Cas9. 12
Figura 4. Sistemes de base editing. 12
Figura 5. Productes d'edició segons la tècnica utilitzada. 13
Figura 6. Etiquetes assignades per classificar els productes d'edició genètica. 23
Figura 7. Esquema de la construcció dels descriptors a partir de la seqüència del gRNA. 25
Figura 8. Eficiència d'electroporació valorada per flow cytometry. 31
Figura 9. Distribució del nombre de reads de cada gRNA en les quatre mostres. 32
Figura 10. Distribució del nombre de reads de les llibreries perfect i mismatch en les mostres estudiades. 33
Figura 11. Distribució del coverage mitjà de les regions d'interès segons la mostra estudiada. 34
Figura 12. Visualització de dos regions d'interès amb IGV. 36
Figura 13. Distribució de l'eficiència d'edició de les regions d'interès segons les dades simulades. 37
Figura 14. Comparació del nombre de delecions i insercions en les dades simulades. 38
Figura 15. Freqüència de les delecions segons la seva mida. 38
Figura 16. Freqüència de les insercions d'una sola base segons el nucleòtid inserit. 39
Figura 17. Evolució de l'accuracy del classificador kNN segons el nombre de veïns considerats (k) i si es té en compte la distància d'aquests o no. 40
Figura 18. Evolució de l'accuracy de la logstic regression en el conjunt de validació segons el cost de regularització (C) i el tipus de regularització. 41
Figura 19. Heatmaps de l'accuracy del model SVM segons el tipus de kernel utilitzat i el cost de regularització (C). 41
Figura 20. Accuracy segons els paràmetres del model Random Forest. 42
Figura 21. Primers quatre nivells de decisió d'un dels 220 arbres del Random Forest. 42
Figura 22. Importància dels 15 descriptors més importants en l'algoritme de Random Forest. 43

Figura 23. Corba ROC comparant les prediccions del conjunt de test segons els models entrenats. 44

Figura 24. Comparació entre les prediccions i els valors reals d'eficiència dels gRNAs segons els quartils. 45

Figura 25. Comparació entre els valors predits i reals d'eficiència dels gRNAs amb un model de regressió isomètrica. 46

Figura 26. Freqüències en el conjunt de dades simulades dels grups de productes d'edició definits. 47

Figura 27. Preparació de les dades pel model de predicció del producte majoritari d'edició. 49

Figura 28. Matriu de confusió dels models per predir el producte d'edició genètica majoritari. 50

Figura 29. Preparació de les dades pel model de predicció dels productes més abundants. 50

Figura 30. Distribució del nombre de coincidències entre les dades reals i dels models. Es mostra les coincidències segons si les prediccions es fan amb el model Random Forest o el classificador dummy. 51

1. Introducció

1.1 Context i justificació del Treball

Les tècniques d'edició genòmica permetrien corregir la majoria de les més de 75,000 variants genètiques associades a una malaltia humana[1]. Recentment, s'ha popularitzat l'ús de la nucleasa programable CRISPR-Cas9 i els *base editors* ABE i CBE. Totes elles permeten introduir mutacions en un punt concret del genoma. Mentre que CRISPR-Cas9 actua causant un tall de doble cadena (*double stranded break, DSB*) per introduir insercions i delecions, els *base editors* canvien l'estructura química dels nucleòtids per provocar mutacions puntuals de canvi de base. A través de qualssevol d'aquestes modificacions es canvia una regió concreta del genoma, que pot ser interessant tant en recerca com per fins terapèutics.

Tot i això, el potencial terapèutic d'aquestes eines d'edició genòmica està limitat perquè les mutacions produïdes depenen del mecanisme cel·lular de reparació de l'ADN i aquest no es pot controlar. Per tant, tot i que aquestes eines d'edició genòmica permeten modificar regions concretes de genoma, el tipus de modificació introduït, que és del que depèn el resultat final d'edició, no es pot controlar. Per exemple, si s'utilitzés CRISPR-Cas9 per tractar una malaltia causada per la inserció d'un sol nucleòtid, es podria controlar amb alta precisió la regió del genoma editada per introduir modificacions en el gen concret. Ara bé, no es podria assegurar que la modificació causada correspon a la deleció de la base responsable de la malaltia, fet que limita l'aplicació d'aquestes tècniques d'edició genòmica.

Estudis previs indiquen que el producte obtingut de l'edició genètica depèn en gran mesura de la seqüència editada[2,3]. Mentre algunes seqüències d'ADN afavoreixen les insercions, d'altres afavoreixen les delecions o canvis de base. La freqüència amb la que afavoreixen un resultat o un altre també depèn de la seqüència. És a dir, hi ha seqüencies que afavoreixen un resultat en concret (*e.g.* les insercions d'una sola base són molt més probables que les delecions o insercions grans) mentre que d'altres no n'afavoreixen cap en particular. En el primer cas, els resultats de l'edició són relativament homogenis, mentre que en el segon cas els resultats són diversos.

Per tant, es podria millorar la predictibilitat del procés d'edició genòmica a través de l'elecció de la regió a editar, ja que els resultats obtinguts depenen de la seqüència a editar. Així doncs, sempre que es desitgi editar un gen sense que la modificació hagi de ser en una regió concreta d'aquest, es podria dirigir l'edició genètica a una regió del gen que afavoreixi un sol resultat d'edició i que aquest sigui el resultat desitjat. Per exemple, si es desitja canviar el marc de lectura d'un gen per fer-lo no funcional, es poden introduir canvis en el marc de lectura que causin la terminació de la transcripció en qualssevol lloc de la seqüència del gen. Per això, es pot millorar l'eficiència del procés d'edició escollint una regió a editar en la que el resultat majoritari sigui una inserció o deleció que causi el canvi en el marc de lectura desitjat.

Com que els resultats d'edició depenen de la seqüència editada, en aquest treball es crea un model computacional d'aprenentatge automàtic (*machine learning*) que relaciona els resultats d'edició amb característiques de la seqüència. Així, el model es podria utilitzar per predir la freqüència de cada resultat possible de l'edició genètica a partir d'una seqüència d'ADN determinada. L'ús d'aquest model permetria escollir la regió del genoma a editar que tingui una seqüència tal que el resultat d'edició genòmica sigui el desitjat. Així, el seu ús en recerca permetria millorar l'eficiència dels experiments de modificació genètica i accelerar el descobriment de diverses funcions genètiques. El seu ús en teràpia genètica permetria assegurar que les modificacions introduïdes tenen l'efecte terapèutic desitjat i no resultats inesperats.

Els models computacionals desenvolupats fins ara utilitzen dades experimentals obtingudes en condicions que difereixen molt de les que s'utilitzarien per a teràpia. En aquestes estudis, les edicions es realitzen en plasmidis amb seqüències a modificar sintètiques, en comptes d'editar el genoma en si, fet que exclou dels models la complexitat i estructura del genoma. A més, els experiments s'han realitzat en cèl·lules HEK293F, una línia cel·lular tumoral immortalitzada. Aquestes cèl·lules tenen una estructura genètica marcadament diferent a les cèl·lules humanes somàtiques, així que els mecanismes de reparació cel·lular són diferents que a les cèl·lules *in vivo*. Així, els models computacionals desenvolupats fins ara no es recreen correctament les condicions terapèutiques. Per tant, les seves prediccions podrien allunyar-se de la realitat al aplicar-se per teràpies *in vivo*, que és precisament on les eines d'edició genètica tindrien un major impacte.

En aquest treball s'introduiran unes 1,700 modificacions al genoma de cèl·lules neuromusculars de ratolí C2C12 per predir el producte d'edició genètica a partir de la seqüència. Per tant, els resultats d'edició observats en les dades experimentals reflecteixen el context cel·lular i factors com la distribució genòmica i els mecanismes de reparació específics de les cèl·lules del teixit muscular, que té rellevància clínica. Així, el model permetria millorar la precisió de les prediccions respecte els models desenvolupats fins al moment, ja que utilitza condicions experimentals més similars a les de la pràctica clínica, acostant així les eines d'edició genètica al seu ús terapèutic.

1.2 Objectius del Treball

Els objectius s'han formulat seguint el criteri SMART (específic, mesurable, assignable, rellevant, i programat en el temps). El treball es pot dividir en dos objectius generals:

1. Preparar les dades experimentals de seqüenciació massiva i comprovar que són adequades per entrenar un model computacional a través d'un anàlisi descriptiu i exploratori. En cas que hi hagi biaix, el model s'adequa per tenir-ho en compte.
2. Entrenament i validació d'un model computacional capaç de predir la freqüència de cada resultat d'edició genòmica a partir de la seqüència de nucleòtids de la regió editada. El model és adequat si pot determinar quin resultat (inserció o deleció) és més probable donada una seqüència qualssevol del genoma.

Cada un d'aquests objectius generals es pot dividir al seu torn en diferents objectius específics:

1. Anàlisi descriptiu i exploratori de les dades experimentals de seqüenciació massiva:
    1.1. Assegurar que el conjunt de dades utilitzades són de qualitat eliminant aquelles lectures (*reads*) de baixa qualitat, comprovant que el nombre de seqüències eliminat és baix (menys d'un 10% del conjunt de dades).
    1.2. Alinear les seqüències obtingudes al genoma de referència per a poder identificar les regions del genoma editades i comprovar que la profunditat de seqüenciació és suficient per poder observar edició genòmica en aquestes regions (almenys calen 2,000 *reads* de cada regió editada).
    1.3. Preparar les dades en un format adequat per entrenar i validar el model. Comptabilitzar la freqüència dels resultats d'edició genètica (insercions, delecions, canvis de base, etc.) en cada regió genètica editada i relacionar-la amb la seqüència d'aquella regió.
    1.4. Garantir que les regions estudiades tenen una distribució d'edició semblant entre elles perquè el model no tingui biaix per a certes seqüències.
2. Desenvolupar un model computacional per predir el resultat de l'edició genètica a partir de la seqüència.
    2.1. Escollir el mètode d'anàlisi i predicció més adequat per a les dades obtingudes i la tasca a realitzar.
    2.2. Emprar un algoritme per entrenar el model escollit de forma eficient, cal entrenar un model diferent per cada mètode d'edició utilitzat (CRISPR-Cas9, ABE, i CBE).
    2.3. Validar el funcionament del model creat: ajustar els paràmetres del model perquè generalitzi correctament i es pugui aplicar per a predir el resultat de l'edició genètica a partir de la seqüència.
    2.4. Mostrar exemples de com aplicar aquests models: escollir una o més mutacions amb rellevància mèdica i il·lustrar com el model creat permetria saber si una edició genètica plantejada seria eficaç per resoldre el problema clínic.

Els objectius inicials mostrats aquí s'han vist afectats per alguns contratemps i desviacions. La causa d'aquestes i l'efecte que han tingut en els objectius inicials es comenta amb detall a la secció *1.4.2. Desviacions en la temporalització*.

1.3 Enfocament i mètode seguit

Prèviament a l'inici d'aquest treball, s'ha utilitzat una llibreria de gRNA únics per editar unes 1,700 regions del genoma de cèl·lules C2C12. El treball comença amb la definició dels mètodes d'enriquiment i seqüenciació massiva que permeten estudiar la diversitat de productes finals de les regions editades.

Abans d'iniciar el model computacional, es processen les dades de seqüenciació. Primer, es descarten les lectures de baixa qualitat a partir de la qualitat assignada a cada seqüència en els fitxers FASTQ. A continuació, s'indexa el genoma de referència de les cèl·lules de ratolí C2C12 (que difereix del genoma de referència de ratolí mm10). Finalment, s'alineen les dades de seqüenciació massiva al genoma de referència utilitzant Burrows-Wheeler Aligner (BWA) perquè aquest és adequat per alinear seqüències curtes i poc diferents (tal i com múltiples reads d'una regió) amb una seqüencia llarga com un genoma de referència.

Per l'anàlisi exploratori i descriptiu de les dades obtingudes, es visualitzen les dades utilitzant l'Integrative Genomics Viewer (IGV) i s'utilitza Python 3 per calcular paràmetres d'interès com la profunditat de seqüenciació de les regions editades. També s'empra un algoritme per extreure la freqüència de cada resultat possible d'edició en cada regió i relacionar-lo amb la seva seqüència. Es tindrà especial cura d'adequar el format de sortida d'aquest anàlisi perquè les dades es puguin utilitzar per entrenar el model de forma eficient. Paral·lelament, s'analitza la distribució del nombre de còpies de cada gRNA present en la llibreria utilitzada per editar les cèl·lules, amb l'objectiu de comprovar si hi ha biaix en les dades d'edició genètic entre les diferents regions.

L'elecció del model d'intel·ligència artificial adequat per predir el resultat de l'edició segons la seqüència genètica es basa en la literatura existent sobre el tema. Per això, es condueix una revisió bibliogràfica sobre estudis similars, comparant les semblances i diferències entre aquests. A més dels models utilitzats prèviament, es proven diversos algoritmes per identificar el més adequat segons les dades obtingudes.

La implementació dels models computacionals es fa utilitzant la llibreria *scikit-learn* de Python. L'entrenament del model es fa utilitzant *Jupyter Notebooks* a la plataforma *Google Colaboratory* per facilitar la presentació i revisió dels *scripts* utilitzats per terceres persones. Per el processament de les dades genòmiques només s'utilitza *Google Colaboratory* si la tasca a realitzar requereix poca computació. Per a tasques intensives, s'utilitza un servidor remot amb el que es comunica a través d'una interfície UNIX. Per això, per aquestes tasques intensives només es mostren els *scripts* utilitzats a *Google Colaboratory* però no es poden executar. Tot el codi utilitzat es troba en diferents *Jupyter Notebooks* en una carpeta compartida de *Google Drive* perquè es puguin visualitzar i executar amb *Google Colaboratory*. Els fitxers de dades utilitzats per aquestes llibretes es mantenen en un repositori de *GitHub*, que també conté una copia de les *Jupyter Notebooks*.

1.4 Planificació del Treball

1.4.1 Tasques i fites

A continuació es mostren les tasques planificades a partir dels objectius del treball, agrupades en tres fites que corresponen a les proves d'avaluació continuada (PAC). Aquestes fites canvien a mesura que es desenvolupa el treball. Així, tal i com es comenta en el següent apartat *1.4.2. Desviacions en la temporalització*, la fita 2 es formula de nou per tenir en compte les desviacions del treball.

0. Fita 0. Desenvolupar el pla de treball, incloent objectius, tasques, planificació temporal i possibles desviacions (**PAC 1**).
1. Fita 1. Anàlisi descriptiu i exploratori de les dades experimentals de seqüenciació massiva (**PAC 2**).
    1.1. Control de qualitat de les dades: assegurar que el conjunt de dades utilitzades són de qualitat eliminant aquelles lectures (reads) de baixa qualitat, comprovant que el nombre de seqüències eliminat és baix (menys d'un 10% del conjunt de dades).
    1.2. Alineament de les dades amb el genoma de referència i anàlisi de les regions editades:
        1.2.1. Obtenir el genoma de referència adequat per les cèl·lules utilitzades (C2C12).
        1.2.2. Indexar el genoma de referència i alinear-hi les dades obtingudes utilitzant BWA.
        1.2.3. Identificar les regions del genoma editades, alineant-hi els gRNA de la llibreria.
        1.2.4. Comprovar que la profunditat (*coverage*) de seqüenciació en aquestes regions sigui suficient per observar diversos resultats d'edició en cada una (almenys calen 2,000 reads de cada regió editada).
    1.3. Preparar les dades en un format adequat per entrenar i validar el model. Comptabilitzar la freqüència dels resultats d'edició genètica (insercions, delecions, canvis de base, etc.) en cada regió genètica editada i relacionar-la amb la seqüència d'aquella regió.
    1.4. Garantir que les regions estudiades tenen una distribució d'edició semblant entre elles perquè el model no tingui biaix per a certes seqüències.
2. Fita 2 (*Pla inicial*). Model computacional per predir el resultat de l'edició genètica a partir de la seqüència (**PAC 3**).
    2.1. Escollir el mètode d'anàlisi i predicció més adequat per a les dades: revisar la literatura per identificar models utilitzats en altres estudis amb una finalitat similar. Requereix entendre les peculiaritats de cada estudi, informar-se sobre la teoria del model i veure quina seria la seva possible implementació en Python.
    2.2. Emprar un algoritme per entrenar el model escollit de forma eficient: utilitzar la llibreria scikit-learn de Python per entrenar el model

computacional escollit, cal entrenar un model diferent per cada mètode d'edició utilitzat (CRISPR-Cas9, ABE, i CBE).

2.3. Validar el funcionament del model creat: ajustar els paràmetres del model perquè es generalitzi correctament i es pugui aplicar per a predir el resultat de l'edició genètica a partir de la seqüència.

2.4. Mostrar exemples de com aplicar aquests models: escollir una o més mutacions amb rellevància mèdica i il·lustrar com el model creat permetria saber si una edició genètica plantejada seria eficaç per resoldre el problema clínic.

1.4.2. Desviacions en la temporalització

En l'anàlisi de riscs del pla de treball (PAC 1) es preveien algunes desviacions que podrien afectar el desenvolupament dels objectius del treball. A més dels riscs plantejats inicialment, alguns imprevistos han allargat els períodes de temps assignats a cada tasca. En general, les desviacions en la temporalització es deuen a:

- Cal convertir les coordenades d'edició dissenyades a partir del genoma de referència *mm10* al genoma de la línia cel·lular C3H, que és el que s'ha utilitzat (desviació no prevista).
- La baixa profunditat de seqüenciació en les regions editades requereix la simulació de dades (desviació prevista).
- Es descarta entrenar un model diferent per a cada enzim estudiat, i el treball es centra en dos models diferents per CRISPR-Cas9 (desviació no prevista).

El primer factor inesperat té a veure amb la diferència entre el genoma utilitzat per dissenyar la llibreria de gRNAs i el genoma editat. Per una banda, com que l'experiment inicialment s'havia de realitzar *in vivo*, la llibreria de gRNAs es va dissenyar per editar regions del genoma de ratolí (*mus musculus*) de la soca C57BL/6, utilitzant el genoma de referència *mm10* (versió GRCm38). Per l'altra, finalment es va decidir editar cèl·lules C2C12 (musculars) de ratolí *in vitro*, que és una línia cel·lular que deriva de la soca de ratolins C3H. Aquesta, té certes variacions genòmiques respecte la soca C57BL/6.

Per tant, el genoma en el que s'han realitzat les edicions és diferent al genoma pel que s'han dissenyat les seqüències a editar. Això implica que pot haver-hi mutacions en les seqüències seleccionades com a diana per edició, que impedeixin que part de la llibreria de gRNAs sigui funcional. A més, com que les coordenades genòmiques per una mateixa posició canvien en cada versió genòmica, és necessari reconvertir les coordenades diana definides prèviament amb el genoma *mm10* a les coordenades del genoma de C3H. La conversió de les coordenades diana d'edició entre el genoma *mm10* i el genoma C3H no s'havia previst inicialment, ja que no s'esperava l'elevat nombre de mutacions entre les seqüències d'aquests dos genomes. La importància de conèixer del cert la seqüència no editada per poder estudiar les edicions d'una sola base generades per CRISPR va fer palesa la necessitat de treballar amb el genoma C3H com a referència en comptes de *mm10*.

El procés de conversió de les coordenades es va assolir correctament, però les complicacions que va haver-hi en aquest procés van allargar la durada que s'havia planificat per aquesta tasca. Com que aquesta tasca és necessària per continuar amb l'avanç del projecte, aquest contratemps causa canvis en els períodes previstos per cada tasca i redueix el temps disponible per a treballar en la tasca 2.

A diferència de la primera desviació, la segona es podia preveure en l'anàlisi de risc. Al obtenir les dades preparades per l'anàlisi s'observa que la freqüència de modificacions genètiques és massa baixa per entrenar un model computacional. La causa segurament són factors experimentals com una baixa eficiència de transfecció (el procés d'introducció del material genètic necessari per a l'edició a les cèl·lules a editar), o una profunditat de seqüenciació insuficient.

Per mitigar l'impacte de la segona desviació s'havia previst incloure el procés de simulació de dades per a l'entrenament del model computacional. Així doncs, la fita 2 es modifica per incloure el procés de simulació de dades. Com que cada enzim d'edició genòmica (CRISPR-Cas9, ABE, i CBE) genera un tipus de mutacions diferent, caldria realitzar tres simulacions diferents per cada un (utilitzant mecanismes i principis totalment diferents per a cada un) i entrenar tres models diferents per cada un. A causa d'aquest increment de tasques en la fita 2, es limita l'abast del projecte a la simulació i predicció dels resultats d'edició genètica per CRISPR-Cas9, descartant l'entrenament de models per ABE i CBE previst inicialment. Degut a l'ordre dels fets, en la fita 1 s'inclou informació dels resultant tant de CRISPR-Cas9 com d'ABE i CBE, mentre que la fita 2 es centra tan sols en CRISPR-Cas9.

L'elecció de CRISPR-Cas9 en comptes de ABE o CBE es deu al fet que els resultats de CRISPR-Cas9 son més diversos que els d'ABE i CBE. Al centrar-se en més detall en CRISPR-Cas9, es planteja entrenar dos models diferents per predir els seus resultats d'edició en comptes d'un. Per una banda, un dels dos models realitza prediccions sobre l'eficiència o activitat segons la seqüència. Per l'altra banda, l'altre model realitza prediccions sobre quins són els resultats/tipus d'edició majoritaris. Per tant, la part d'entrenament dels models es divideix en l'entrenament de dos models diferents, un d'eficiència i l'altre de resultat d'edició.

Finalment, l'endarreriment i l'increment de tasques en la fita 2 són suficients per simular les dades, quantificar les mutacions, i entrenar el model d'eficiència de CRISPR-Cas9. Tot i això, falta temps per completar el model del resultat d'edició de CRISPR-Cas9. D'aquest darrer, es plantegen dues aproximacions que permetrien abordar la tasca de predicció dels diferents resultats d'edició des de perspectives diferents però complementàries. També es descarta l'objectiu 2.4 centrat en mostrar exemples del funcionament del model amb aplicació terapèutica.

El conjunt de desviacions afecten principalment la fita 2 del projecte, així que es torna a formular la fita 2 per adaptar-la al desenvolupament del treball:

2. Fita 2 (*Desenvolupament final*). Model computacional per predir l'eficiència i resultats d'edició genètica de CRISPR-Cas9 a partir de la seqüència (**PAC 3**).
    2.1. Simulació de dades d'edició de CRISPR-Cas9 i caracteritzar-les.
    2.2. Caracteritzar (*featurization*) la seqüència de les regions a editar.

2.3. Entrenament d'un model per predir l'eficiència de CRISPR-Cas9 segons la seqüència.

2.4. Plantejar dues aproximacions per predir els resultats d'edició de CRISPR-Cas9 segons la seqüència.

1.4.3. Cronograma

Tal i com s'ha comentat en la secció anterior, algunes desviacions en la temporalització han fet necessari ajustar els períodes destinats a cada tasca. Així, es contrasten els objectius inicials segons la planificació temporal plantejada inicialment (Figura 1) amb les tasques realitzades al final i els períodes dedicats a cada una (Figura 2). Mentre que en el primer cronograma el color de les tasques correspon a la seva prioritat, en el segon cronograma el color correspon al seu grau d'assoliment.

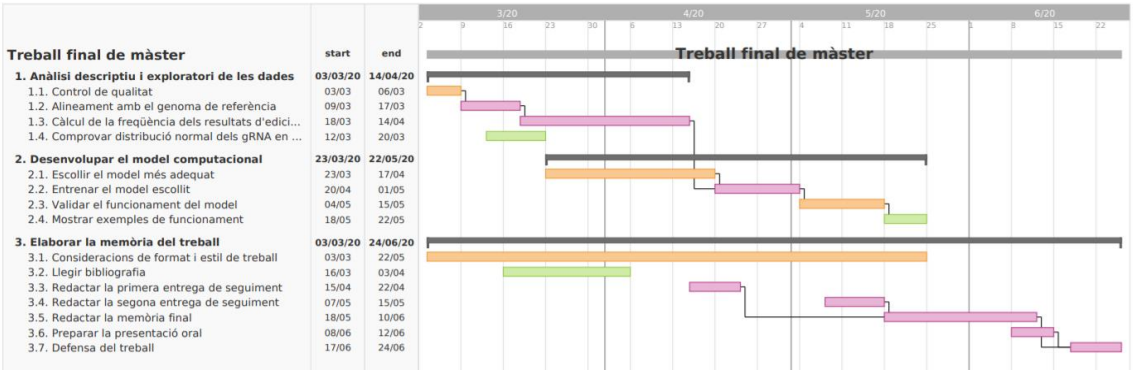

Figura 1. Cronograma del treball plantejat inicialment.

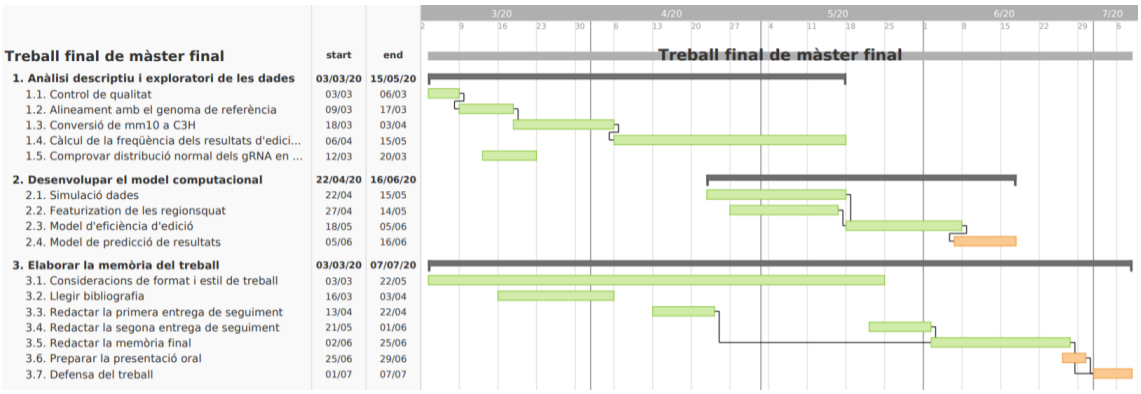

Figura 2. Cronograma del treball una vegada finalitzat.

1.5 Breu sumari de productes obtinguts

Els productes obtinguts en aquest treball són els *scripts* utilitzats en cada una de les tasques a realitzar. Aquests s'han desenvolupat tenint en compte que s'haurien de poder reutilitzar per realitzar l'anàlisi de dades d'una tasca semblant. Així, quan es repeteixi l'experiment per incrementar l'eficiència d'edició genètica, es podran utilitzar els *scripts* desenvolupats en aquest treball per processar les dades i entrenar els models computacionals.

Cada una de les tasques dutes a terme i descrites prèviament (*1.4.1. Tasques i fites*) es descriu en una *Jupyter Notebook* creada a *Google Colaboratory*. Per tant, els resultats són una col·lecció de llibretes amb el codi per reproduir cada una de les tasques. Algunes de les tasques s'han realitzat en el servidor remot, així que en aquest cas es mostren els *scripts* utilitzats però no es poden executar.

Tot el codi utilitzat en aquest treball, que inclou el conjunt de llibretes i els fitxers necessaris per executar-les, està disponible en dues plataformes: *Google Drive* i *GitHub*. La carpeta de *GitHub* conté els fitxers de dades utilitzats per l'anàlisi excepte els fitxers massa grans i que s'han processat amb el servidor remot. També conté les llibretes ja executades perquè es puguin visualitzar de forma estàtica. Així, s'hi inclouen les figures generades per la tesi.

La carpeta compartida de *Google Drive* conté les *Jupyter Notebooks* de cada una de les tasques adaptades per reproduir els anàlisis utilitzant *Google Colaboratory*. D'aquesta forma, es poden retocar els *scripts* i executar en directe per comprovar el bon funcionament del codi.

L'enllaç a *GitHub* és: https://github.com/marcexpositg/CRISPRed

L'enllaç a la carpeta compartida de *Google Drive* és:

https://drive.google.com/drive/folders/1LnE60PEwfueoES_KNFeh3ox7QxCFBEp_?usp=sharing

Nota: per visualitzar les *Jupyter Notebooks* de *Google Drive* a *Google Colaboratory* cal fer clic al fitxer, a "obre amb" i escollir "*Google Colaboratory*".

A més d'incloure els enllaços a les carpetes conjuntes, a continuació s'inclou l'enllaç de cada llibreta directament a *Google Colaboratory* per si obrir-les des de *Google Drive* és complicat:

1. DescriptiveAnalysis/
    1.1. *SequencingDataProcessing.ipynb*. Conté els *scripts* utilitzats per el processament de qualitat de les dades genòmiques i per l'alineament amb el genoma de referència. Disponible aquí.
    1.2. gRNALibDistribution.ipynb. Conté els *scripts* utilitzats per el processament de les dades de *target sequencing* en el control de qualitat de la llibreria. Conté codi que es pot executar per visualitzar les figures generades. Disponible aquí.
    1.3. CoordiantesToC3H.ipynb. Conté la informació del procés utilitzat per convertir les coordenades d'interès del genoma *mm10* al genoma C3H. Disponible aquí.
    1.4. CoverageAnalysis.ipynb. Conté els *scripts* utilitzats en l'anàlisi del *coverage*, es pot executar per generar les figures mostrades a la memòria (cal esperar uns 10 minuts). Disponible aquí.
2. Model/
    2.1. *DataSimulation.ipynb*. Conté els *scripts* utilitzats per a simular les dades, combinant el model d'eficiència de *Doench et.al.*[4] i el model Indelphi[2] per predir les freqüències dels productes d'edició. Es pot executar un exemple per simular 5000 *reads* de 3 regions d'interès. Disponible aquí.

2.2. *LabelGen.ipynb*. Conté l'*script* utilitzat per analitzar les dades simulades i quantificar l'eficiència de cada regió i la freqüència de les categories de productes d'edició genètica segons la regió. Es mostra la quantificació de 3 regions d'interès com a exemple. Disponible aquí.

2.3. *OutcomesProfiling.ipynb*. Mostra l'ús de *Seaborn* i *Matplotlib* per representar visualment els resultats de l'anàlisi descriptiu de les dades simulades. Disponible aquí.

2.4. *Featurization.ipynb*. Conté l'*script* utilitzat per convertir cada regió d'interès en un seguit de descriptors de la seva seqüència i utilitzat per entrenar els models d'aprenentatge automàtic. Disponible aquí.

2.5. *EffModel.ipynb.* Conté el procés d'entrenament dels classificadors per predir l'eficiència d'edició, incloent la preparació de les dades, l'entrenament dels quatre models de classificació i la comparació entre ells i amb el model de *Doench et.al.*[4]. Es pot executar íntegrament per reproduir els resultats mostrats a la memòria. Disponible aquí.

2.6. *OutcomesModel.ipynb.* Conté la planificació i entrenament dels classificadors per predir la freqüència dels resultats d'edició. Es pot executar íntegrament per reproduir les figures de la memòria. Disponible aquí.

1.6 Breu descripció dels altres capítols de la memòria

El següent capítol, 2. Context, adreça detalls i explica el mecanisme de funcionament de les tècniques d'edició genètica utilitzades en aquest treball. En la segona part d'aquest capítol, es fa un repàs bibliogràfic a estudis similars al d'aquesta memòria, tot mencionant les diferències experimentals i del model computacional. Així, l'objectiu del capítol 2. Context és proporcionar al lector els coneixements biològics necessaris per entendre el treball i informació sobre els models computacionals existents fins el moment. A més, pretén destacar el factor innovador d'aquest treball.

El capítol 3. Mètodes descriu els processos utilitzats per a generar els resultats d'aquest treball. L'enfoc del capítol es centra en indicar les dades que s'utilitzen per cada procés, com es transformen, i quins resultats se'n obtenen. Es complementa cada explicació indicant la llibreta de *Jupyter Notebook* que conté els *scripts* per executar el procés, que és on s'hauria de mirar per veure els detalls gràcies als comentaris del codi. A més, en aquest capítol es fa incisió en el perquè s'han triat les aproximacions utilitzades i perquè s'han descartat les alternatives. També es comenten algunes limitacions que podria tenir els processos per generar els resultats.

El capítol 4. Resultats i discussió mostra els resultats dels anàlisis i els models computacionals entrenats. Es fa especial incisió en les conseqüències dels resultats obtinguts en re-ajustar la planificació inicial del treball. Tant aquest com el capítol 3 es divideixen en dues parts clares que coincideixen amb les dues fites del treball: el processat de dades genòmiques per una banda, i l'entrenament del model computacional per l'altra.

Finalment, el treball acaba amb el capítol 5, que presenta les conclusions del treball de forma resumida, i la bibliografia referenciada en aquest treball.

2. Context

2.1. L'edició genètica

Les tècniques d'edició genètica permeten introduir mutacions al genoma de forma controlada. En els darrers anys, la caracterització i aplicació del sistema CRISPR-Cas9 en edició genòmica ha facilitat la possibilitat de modificar el genoma de forma precisa, programable, i econòmica[5]. Així, el sistema CRISPR-Cas9 ha permès avançar la recerca en l'àmbit de les bio-ciències en àrees d'especialització tant diverses com la microbiologia, l'estudi del desenvolupament, o l'enginyeria de teixits. Les seves aplicacions no es limiten només a recerca, ja que s'ha utilitzat per a crear noves varietats vegetals i soques microbianes d'interès industrial[6]. Tot i això, l'aplicació que podria tenir un major impacte és l'estudi i tractament de malalties. Fins ara, la utilització de CRISPR-Cas9 en l'àmbit de la salut s'ha vist limitada per la falta de vectors que facilitin l'entrada del sistema a la cèl·lula i per la falta de control en el resultat de l'edició genètica[7,8].

El sistema CRISPR-Cas9 està composat per dos components, un RNA guia (*guide RNA,* gRNA) i una endonucleasa Cas9. El sistema utilitzat més freqüentment utilitza la proteïna Cas9 de *Streptococcus pyogenes* (SpCas9). La proteïna Cas9 s'uneix al DNA genòmic gràcies al complex format entre el gRNA i el DNA genòmic, i introdueix un tall de doble cadena en una regió específica. El gRNA és una seqüència de 20 nucleòtids de llargada dissenyada per complementar la regió genòmica a editar. La seqüència de 20 nucleòtids complementària al gRNA en el genoma va seguida per una seqüència curta de DNA que s'anomena *protospacer adjacent motif* (PAM), i és essencial per permetre la unió de la proteïna Cas9 utilitza a la regió de tall. En el cas de SpCas9, la seqüència PAM és sempre "NGG", és a dir, qualssevol nucleòtid ("N") seguit per dues guanines ("GG").

Per a l'actuació del sistema CRISPR-Cas9, el gRNA s'uneix a la regió d'interès a editar a través de l'aparellament de bases complementàries entre el gRNA i la regió genòmica d'interès. El complex format entre el gRNA i l'endonucleasa Cas9 situa l'enzim Cas9 en la regió específica del genoma a editar, de tal forma que Cas9 talla les dues cadenes de DNA genòmic pel mateix punt, introduint un tall de doble cadena (Figura 3).

El tall de doble cadena introduït per CRISPR-Cas9 activa els mecanismes de reparació de la cèl·lula, que comencen a arreglar el genoma per mantenir la seva integritat. Els mecanismes de reparació més comuns inclouen el *non-homologous end joining* (NHEJ) i el *micro-homology mediated end joining* (MMEJ), responsables de la inserció o deleció d'alguns nucleòtids en el punt de tall (*indels*). La seqüència reparada es coneix com al producte de l'edició genòmica. Com que les edicions genòmiques es realitzen sempre en un conjunt de cèl·lules, i els talls de doble cadena es poden reparar de múltiples maneres, els resultats obtinguts en l'experiment d'edició genètica són diversos. És a dir, algunes cèl·lules contindran un producte com una inserció d'una base, mentre que d'altres cèl·lules contindran insercions diferents o delecions de varies mides.

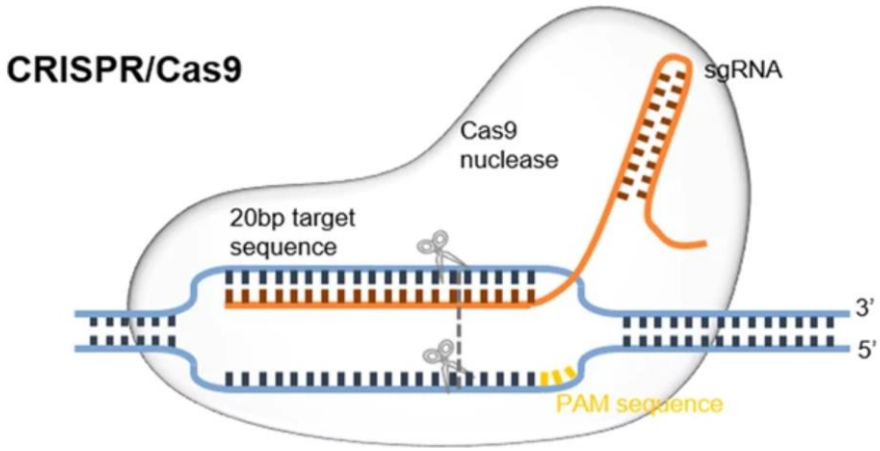

Figura 3. Sistema d'edició genètica CRISPR-Cas9. Es mostra la seqüència del gRNA (taronja) de 20 nucleòtids (bp) complementària a la regió genòmica d'interès (blau). L'aparellament entre gRNA i genoma obre la doble cadena de DNA per tal que l'endonucleasa Cas9 provoqui un tall en les dues cadenes de DNA. Extreta de Li et. al.[8].

Recentment, s'ha modificat els components del sistema CRISPR-Cas9 per crear noves tècniques d'edició genètica, com els *base editors* i els *prime editors*[9]. Els *base editors* utilitzen un enzim Cas9 inactivat per tal que no tingui activitat endonucleasa. És a dir, que s'uneix a la regió d'interès però no provoca un tall de doble cadena. Aquesta Cas9 inactivada es coneix com a Cas9n. Així, Cas9n es pot utilitzar per dirigir altres enzims cap a una regió d'interès del genoma, aprofitant l'alta precisió i modularitat pròpia del sistema CRISPR-Cas9. Aquesta capacitat s'aprofita en els *base editors* per dirigir enzims que modifiquen l'estructura química dels nucleòtids cap a la regió de tall.

Actualment es disposen de dos *base editors* diferents. Per una banda, els *cytosine base editors* (CBE) fusionen Cas9n amb una desaminasa de citosines que converteix citosines (C) en la regió de l'*R-loop* en uracils (U, equivalent a T en RNA). Així, pot canviar parells de bases C-G a T-A[10]. Per l'altra banda, els *adenine base editors* (ABE) fusionen Cas9n a una desaminasa d'adenosina, que catalitza la conversió d'adenosina (A) a guanosines (G), canviant parells de bases A-T a G-C (Figura 4)[11]. Els *base editors* s'utilitzen extensivament en recerca, i diverses modificacions en la seva estructura els acosten progressivament a la pràctica clínica com a possible teràpia gènica[12,13].

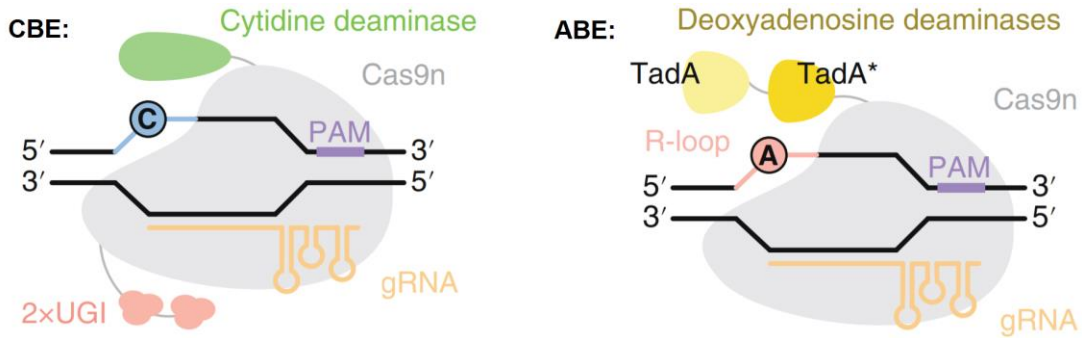

Figura 4. Sistemes de base editing. Es mostren els components dels CBE (esquerra) i ABE (dreta), incloent Cas9n, el gRNA, i diversos enzims per catalitzar la conversió dels nucleòtids. Extreta de Liu et. al.[9] .

Els productes de l'edició genètica obtinguts amb cadascuna de les tècniques presentades són diferents. Mentre que el sistema CRISPR-Cas9 genera insercions o delecions de diferents mides al voltant del punt de tall, els *base editors* no canvien la longitud de la seqüència editada sinó que modifiquen la composició dels nucleòtids d'aquesta, de C->T en CBE i de A->G en ABE. Tot i això, en tots els casos, els productes de l'edició genètica són diversos perquè s'edita un conjunt de cèl·lules. Algunes d'aquestes no presentaran modificacions en la regió d'interès, mentre que d'altres poden presentar diferents *indels* si s'ha utilitzat CRISPR-Cas9 o diferents canvis de base si s'ha utilitzat algun dels *base editors* (Figura 5). Per tant, sol ser necessari utilitzar algun mètode d'*screening* per identificar aquelles cèl·lules que contenen la modificació d'interès.

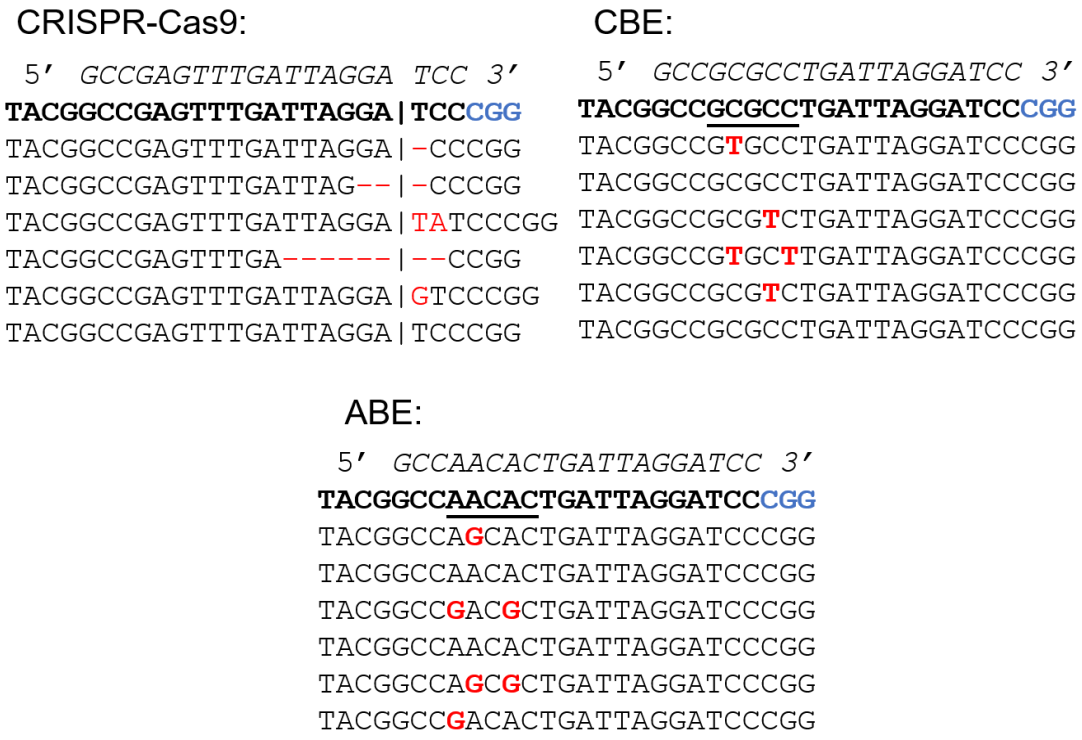

Figura 5. Productes d'edició segons la tècnica utilitzada. En tots els casos es mostra els 20 nucleòtids de gRNA orientats de 5' a 3', la seqüència genòmica en negreta incloent la PAM en blau. Per CRISPR-Cas9, s'indica el punt de tall de doble cadena amb (|) i s'indiquen les insercions o delecions en vermell. Per CBE i ABE es subratlla la regió editable i s'indiquen els canvis de base en vermell. En tots els casos s'inclou un exemple de seqüència no editada, ja que l'eficiència de les tècniques no és del 100%.

La diversitat dels productes obtinguts limita l'aplicació de les tecnologies d'edició genètica a la pràctica clínica[14]. Estudis previs demostren que els productes obtinguts depenen de la seqüència de la regió editada, tant per CRISPR-Cas9 com per *base editors*[2,3]. Per tant, s'ha desenvolupat models predictius per predir els resultats d'edició segons la seqüència que es pretén editar. Així, aquests models es poden utilitzar per identificar les seqüències que permeten aconseguir els resultats d'edició genètica desitjats. Per tant, s'utilitzen per millorar el disseny del gRNA per aconseguir que els resultats d'edició siguin els esperats. El control sobre el producte d'edició obtingut permetria avançar en el desenvolupament de tractaments de teràpia gènica.

2.2. Estudis previs

Fins al moment, s'ha desenvolupat diversos models computacionals que permeten predir els resultats d'edició genètica. La majoria dels models computacionals es centren en predir l'eficiència de l'edició genètica, amb l'objectiu de millorar el disseny del gRNA per aconseguir una major proporció de cèl·lules modificades genèticament. Tan sols un nombre petit d'estudis (~3) es centra en predir la freqüència de cada producte d'edició genètica. En el moment d'inici del treball, tots els estudis publicats es centraven en el sistema CRISPR-Cas9, però durant el transcurs del treball s'ha publicat els 2 primers models de predicció dels resultats d'edició genètica de *base editors*[15,16]. Aquests, no es discuteixen en aquesta secció ja que els models predictius entrenats en aquest treball es basen en CRISPR-Cas9.

El primer estudi en el que es desenvolupa un model computacional per modelar l'activitat del gRNA en funció de la seva seqüència es va publicar per *Doench et al.* el 2014[4]. En aquest estudi s'utilitza una llibreria de gRNAs (molts gRNA diferents) per modificar diverses posicions de 3 gens del genoma de línies cel·lulars tumorals humanes. L'activitat dels gRNAs es valora a través d'una prova fenotípica. En concret, s'utilitza *Fluorescent Activtated Cell Sorting (FACS)* per separar les cèl·lules en les que hi ha hagut edició de les que no han estat editades. A continuació, s'utilitza seqüenciació de nova generació per veure quins gRNAs estan enriquits en la població en la que hi ha hagut edició. Per tant, aquells gRNAs més actius són aquells que estan més enriquits en la població de cèl·lules editades respecte la població no editada.

Tot i que el disseny experimental proposat per *Doench et. al*. és simple i efectiu, el fet d'editar tan sols 3 gens del genoma amb múltiples gRNAs limita l'espai de seqüències que es pot explorar. Així, el model computacional entrenat podria estar lleugerament esbiaixat per algunes seqüències. L'estudi publicat per *Chari et. al.* el 2015[17] soluciona aquest problema utilitzant una llibreria de gRNAs per introduir mutacions en un conjunt de regions a editar sintètiques. El més interessant de l'estudi és que l'activitat dels gRNAs es valora a través del genotip editat. En concret, una vegada realitzada l'edició, s'utilitza seqüenciació de nova generació per seqüenciar les regions a editar sintètiques. Això permet obtenir el conjunt de les seqüències editades, i valorar l'activitat dels gRNAs en funció del percentatge de seqüències modificades. Així, es millora la sensibilitat de l'assaig en no dependre de proves fenotípiques.

La darrera aproximació descrita s'ha utilitzat en múltiples estudis centrats en millorar els models de predicció a través de la implementació d'altres algoritmes d'aprenentatge automàtic (*machine learning*) com xarxes neuronals. Aquests, no es discuteixen en detall perquè en aquest treball no utilitza xarxes neuronals com a models de predicció.

La utilització de regions a editar sintètiques i posterior seqüenciació d'aquestes per veure els resultats d'edició genòmica també ha permès el desenvolupament de models per predir la freqüència de cada resultat d'edició genètica. En aquests models, no es prediu si el gRNA és actiu o no, sinó que es prediu la freqüència en la que els resultats contindran delecions o insercions i de quina mida. El primer estudi que es centra en la predicció dels resultats d'edició genètica es va publicar per *Shen et. al.* el 2017[2]. En aquest, a més d'utilitzar la seqüència com a predictor, es caracteritza el

procés de micro-homologia, un dels principals mecanismes de reparació del DNA. S'utilitza una xarxa neuronal per modelitzar el procés de NHEJ, una altra per MMEJ i un classificador de *k-Nearest Neighbors* (kNN) per les insercions.

L'altre estudi que es centra en la predicció dels resultats d'edició genètica es va publicar per *Chen et. al.* el 2019[3]. Aquest, utilitza *logistic regression* com a classificador, un model computacional més semblant als utilitzats en aquest treball, i es podria considerar com l'estudi més proper a aquesta secció del treball.

Cada un dels estudis mencionats utilitza un algoritme diferent per a realitzar les prediccions. Al comparar-los, es pot veure que els models que prediuen l'eficiència majoritàriament utilitzen només característiques de la seqüència, mentre que models de resultats d'edició incorporen característiques de micro-homologia de la seqüència (Taula 1).

Taula 1. Comparació dels models existents de predicció en edició genètica.

Estudi	Predicció	Descriptors	Algoritme	Mida llibreria	Línia cel·lular
Doench et. al. (2014)[4]	Eficiència d'edició	Nucleòtids, dinucleòtids i contingut GC	SVM per *feature selection* i *Logistic Regression*	1278 gRNAs	Human Embrionic Kidney (HEK293F)
Chari et. al. (2015)[17]	Eficiència d'edició	Nucleòtids i dinucleòtids	SVM	~1400 gRNAs	Human Monocytic Leukemia (MOLM13)
Shen et. al. (2017)[2]	Freqüència dels productes	Nucleòtids, dinucleòtids i potencial de micro-homologia	*Feedforward Neural Networks* i kNN	1872 gRNAs	*Mouse Embryonic Stem Cells* (mESC)
Chen et. al. (2019)[3]	Freqüència dels productes	Nucleòtids, dinucleòtids i potencial de micro-homologia	*Logistic Regression*	6872 gRNAs	Human Embrionic Kidney (HEK293F)

L'aproximació experimental seguida en aquest estudi és diferent a la de la resta d'estudis publicats fins el moment. L'estudi de *Doench et. al.*[4] introdueix les edicions genètiques al genoma, però calcula els resultats fenotípicament. La resta, extreu els resultats seqüenciant la regió editada però són regions sintètiques i no genòmiques. En canvi, en aquest estudi s'introdueixen les edicions genètiques en el genoma i es valoren els resultats genotípicament. Al editar el genoma, el model té en compte factors com l'organització del genoma i l'estat d'activació de la cromatina, que els altres models passen per alt. Al seqüenciar els resultats en comptes de valorar-ho fenotípicament, es pot caracteritzar cada un dels resultats i això permet entrenar un model per valorar l'eficiència.

A més, es pretén entrenar tant un model de predicció d'eficiència com un model de predicció dels resultats d'edició. Per tant, si es combinen els dos models, seria la primera plataforma a proporcionar tant l'eficiència com la freqüència de cada resultat, totes dos informacions rellevants per millorar el disseny de gRNAs.

3. Mètodes

3.1. Experiment d'edició genètica

Per relacionar la seqüència de la regió d'interès amb els productes d'edició genètica cal estudiar una diversitat de seqüències. Per això, en comptes d'utilitzar un sol gRNA per estudiar una sola seqüència, s'utilitza una llibreria de 1785 gRNAs diferents. Aquesta llibreria es dissenya abans de l'inici del treball per contenir una elevada diversitat de composició de la seqüència i es dissenya per tal que les regions d'interès corresponguin a més de 600 gens de ratolí involucrats en el desenvolupament muscular.

La llibreria es sintetitza químicament (*Twist Bioscience*, USA) i s'amplifica per PCR per clonar-la en un vector pUC19 adaptat per a l'expressió de gRNAs. Aquest pUC19 s'adapta prèviament per tal d'expressar els gRNAs de forma constitutiva amb el promotor U6 endogen de ratolí, i per tal que cada vector contingui un sol gRNA.

Per a modificar genèticament les cèl·lules, cal utilitzar un vector plasmídic amb un enzim d'edició del DNA i un altre vector amb els gRNA que dirigeixin l'enzim d'edició a la regió del DNA que es vol modificar. Per poder caracteritzar cada tècnica d'edició, s'utilitzen diferents condicions amb un enzim d'edició diferent però la mateixa llibreria. A més, es prova d'utilitzar una llibreria *mismatch*, que és semblant a la llibreria de gRNAs (anomenada *perfect* per diferenciar-la), en l'experiment de CRISPR-Cas9. També s'inclou un control de transfecció utilitzant el plasmidi pSICO que expressa GFP. Per tant, les condicions realitzades de forma independent són 4:

- Cas9P: CRISPR-Cas9 + llibreria de gRNAs *perfect*
- Cas9M: CRISPR-Cas9 + llibreria de gRNAs *mismatch*
- ABE: ABE + llibreria de gRNAs *perfect*
- CBE: CBE + llibreria de gRNAs *perfect*
- GFP: CRISPR-Cas9 + GFP plasmid (control de transfecció, no hi ha edició)

Per cada una d'aquestes condicions, s'electroporen 2,000,000 de cèl·lules C2C12 al 70-90% de confluència amb 20µg d'una barreja equimolar dels dos components (enzim edició + llibreria gRNA) corresponents. Per a l'electroporació, es divideixen les cèl·lules en dos vials de 1,000,000 als que s'afegeixen 10µg del DNA a electroporar, i es sotmeten a 1,650V en 3 pols de 10ms. Just després de l'electroporació, es transfereixen en plaques de cultius de 6 pous amb medi DMEM i s'incuben a 37ºC en un incubador de CO2 humidificat durant 48 hores.

A continuació, la mostra GFP s'utilitza per comprovar l'eficiència de transfecció utilitzant *flow cytometry*. Per la resta de mostres, s'extreu el DNA amb l'objectiu d'enviar-lo a seqüenciar. Cada una de les quatre mostres restants es seqüencia utilitzant dos processos diferents:

- *Shotgun sequencing*: el DNA genòmic es fragmenta, i s'utilitzen unes sondes biotinilades per seleccionar i seqüenciar només aquelles regions genòmiques d'interès (les regions que tenien un gRNA que hi dirigia una edició). Per tant, s'enriqueix la mostra genòmica per obtenir més profunditat de seqüenciació en aquelles regions d'interès on s'espera observar l'edició.

- *Targeted sequencing*: les seqüencies de gRNA en els plasmidis de la llibreria de gRNA s'amplifiquen per PCR i es seqüencien, amb l'objectiu d'observar la qualitat de la llibreria utilitzada i l'efecte de la quantitat relativa de gRNAs en cada mostra.

Així doncs, el *shotgun sequencing* s'utilitza per estudiar l'eficiència i resultats d'edició en el genoma, mentre que en *targeted sequencing* no es seqüencia el genoma, sinó els plasmidis que contenien la llibreria de gRNA com a control de qualitat de la llibreria.

Cada mostra i per cada tècnica s'amplifiquen les seqüències utilitzant un identificador o *barcode* únic que permet ajuntar totes les mostres per seqüenciar-les juntes. La seqüenciació es duu a terme per *illumina* utilitzant el kit de reactius v3 amb una *read length* de 2x300bp i 600 cicles per arribar a un total de 25 milions de *reads*.

3.2. Preparació i anàlisi descriptiu de les dades de seqüenciació

En aquesta secció es descriuen els mètodes utilitzats per a l'anàlisi i el tractament de les dades obtingudes en el procés de seqüenciació. En el cas que sigui possible, es menciona la llibreta de *Jupyter Notebook* que conté els *scripts* utilitzats en aquella secció, per tal que es pugui identificar en la secció *1.5. Resultats obtinguts* i comprovar el funcionament del codi. Cal comentar que algunes d'aquestes tasques de processament de les dades de seqüenciació requereixen fitxers massa grans o temps d'execució massa llargs per executar-les en directe des de *Google Colaboratory*. Totes les llibretes mencionades en la secció 3.2. estan dins la carpeta *1.DescriptiveAnalysis*.

3.2.1. Control de qualitat de les dades

Els *reads* de seqüenciació es divideixen en vuit mostres a partir de l'identificador o *barcode* utilitzat per a cada una de les vuit mostres (Cas9P, Cas9M, ABE i CBE, seqüenciades per *targeted* i per *shotgun*). Per cada una de les mostres s'obtenen dos fitxers FASTQ, cada un amb els *reads* d'una sola direcció (R1 o R2). Aquests fitxers contenen un identificador de cada *read*, la seqüència obtinguda, i la qualitat de seqüenciació (*phred quality score*) de cada nucleòtid. S'eliminen els *reads* amb baixa qualitat de seqüenciació en el procés conegut com a *trimming*. Per a fer-ho, es crea un petit *script* de Bash que utilitza Trimmomatic per ordenar les lectures i eliminar totes aquelles amb poca qualitat de seqüenciació. Aquest *script* s'executa en el servidor remot però es pot visualitzar a la llibreta *1.2.SequencingDataProcessing.ipynb*. Els fitxers de seqüenciació abans i després de processar són massa grans per compartir-los a *GitHub*.

3.2.2. Distribució de la llibreria de gRNAs

Quan es treballa amb llibreries, és important assegurar que cada component de la llibreria està representat en l'experiment. A més, cal valorar si hi ha diferències en l'abundància relativa dels components de la llibreria per saber si això podria causar diferències en l'activitat mesurada dels gRNAs. Per això, es visualitza la distribució de l'abundància de cada gRNA de la llibreria en les mostres de *targeted sequencing*. Aquestes mostres tan sols s'utilitzen per aquest anàlisis.

L'anàlisi comença amb les dades de *targeted sequencing* un cop han passat el control de qualitat descrit a la secció *3.2.1. Control de qualitat de les dades*. També s'utilitza un fitxer amb la seqüència dels gRNAs de la llibreria (1785 seqüències úniques). Per començar, es crea un *script* que cerca la seqüència de cada gRNA de la llibreria en el conjunt de *reads* de *targeted sequencing* de cada una de les dues mostres. Aquest, conta el nombre de *reads* que contenen la seqüència (tant directa com *reverse-complement*) dels gRNAs. Aquest nombre permet comparar l'abundància relativa dels gRNAs en la llibreria, partint de la premissa que si un gRNA està més present que un altre en la llibreria inicial, s'haurà seqüenciat un nombre major de vegades.

Cal mencionar que per cada una de les quatre mostres es cerquen les seqüències dels gRNA de la llibreria *perfect* i de la llibreria *mismatch*. Això és així perquè originalment les dues llibreries venen d'un mateix tub en el que s'amplifiquen diferencialment per PCR. Tot i això, es possible que alguns gRNAs de la llibreria *mismatch* s'hagin clonat en els vectors de la llibreria considerada *perfect* i viceversa. Per això, és interessant mesurar si el nombre de gRNAs de la llibreria incorrecte és gaire alt en cada una de les mostres.

Per a visualitzar els resultats, es creen gràfics de violí per comparar la presència de les dues llibreries en cada mostra i un gràfic que mostra la distribució del nombre de *reads* identificats dels gRNA entre les mostres.

3.2.3. Alineament amb el genoma de referència

Les cèl·lules utilitzades per l'experiment són cèl·lules de la línia C2C12, que provenen de la soca de ratolí C3H. Per tant, el seu genoma és diferent del de la soca de ratolí C57BL/6, que és la que s'utilitza com a genoma de referència per ratolí (GRCm38/mm10). Per això, cal obtenir un altre genoma de referència propi de la soca C3H. Aquest genoma es descarrega en format FASTA des de la base de dades de UCSC, amb l'identificador GCA_001632575.1_C3H_HeJ5.

A continuació, s'escriu un petit *script* per indexar i alinear cada mostra genòmica de *shotgun sequencing* amb el genoma de C3H. Cal remarcar que només s'alinea amb el genoma les mostres enriquides de *shotgun sequencing* de cada una de les 4 condicions, ja que les mostres de *targeted sequecing* contenen la llibreria dels gRNA i no informació genòmica. L'*script* utilitzat per alinear les mostres utilitza BWA per indexar el genoma de referència C3H i alinear-hi cada mostra. Breument, el procés d'indexació del genoma crea una estructura d'arbre de decisió que facilita l'algoritme de BWA identificar la regió genòmica amb la que aparella cada *read*. Els *reads* alineats al genoma es poden consultar en els fitxers (.sam) creats per BWA

Finalment, s'utilitza SAMtools per convertir el fitxer alineat (.sam) en forma binària (.bam), ordenar les lectures per coordenades i indexar-lo (.bai) per facilitar-ne la visualització amb IGV. El procés d'alineament amb el genoma de referència requereix més recursos computacionals que la resta del treball, però és un procés ben descrit i que requereix eines que el fan relativament senzill d'implementar.

L'*script* per a l'alineament amb el genoma s'executa en el servidor remot, i per tant només es pot visualitzar a la llibreta *1.2.SequencingDataProcessing.ipynb*. Els fitxers de seqüenciació alineats són massa grans per compartir-los a *GitHub*.

3.2.4. Conversió de les coordenades del genoma *mm10* a C3H

Per a poder estudiar si hi ha hagut edició, és essencial definir les zones del genoma que s'estudien. Les posicions genòmiques s'indiquen amb coordenades a l'estil "chr1:1,403,234-1,404,212", que indicaria el conjunt de nucleòtids del cromosoma 1 entre la posició 1,403,234 i la posició 1,404,212. Les regions d'interès s'havien dissenyat en el genoma de referència mm10, que deriva de la soca de ratolí C57BL/6, mentre que l'experiment finalment es va realitzar en la línia cel·lular C2C12, que deriva de la soca de ratolí C3H. Per això, cal convertir les coordenades de les regions d'interès entre el genoma mm10 i el genoma C3H.

Per fer-ho, s'utilitza l'eina Remap de NCBI, seleccionant com a genoma d'origen el GRCm38/mm10 i com a destí GCA_001632575.1_C3H_HeJ5. El resultat obtingut no és perfecte per a totes les regions d'interès. Per això, es refinen els resultats utilitzant comandes de Bash per obtenir informació dels resultats. Per una banda, hi ha 3 regions d'interès que s'han identificat múltiples vegades en el genoma C3H i no es pot decidir en concret quina és la més interessant. També hi ha 30 regions d'interès que no s'han localitzat en el genoma C3H automàticament. A més, hi ha 20 zones d'interès amb mutacions entre mm10 i C3H. Tot i que s'intenten corregir aquestes desviacions manualment utilitzant BLAT, s'opta per descartar les regions d'interès que no s'han convertit correctament utilitzant Remap. També es descarten 13 gRNAs per tenir nucleòtids no identificats al voltant de la seva seqüència genòmica en C3H. Per tant, del conjunt de 1785 regions d'interès úniques inicial, el procés de conversió redueix el nombre de regions d'interès a estudiar al grup de 1722 que s'han pogut identificar correctament amb Remap.

El procés de conversió de les coordenades genòmiques s'explica en detall a la llibreta *1.3.CoordiantesToC3H.ipynb*. Tot i que al final s'opti per descartar aquelles regions que no s'han convertit correctament, es realitzen molts intents diferents (tal com demostra la llibreta amb els detalls del procés) i aquest imprevist allarga el període planificat inicialment per la fita 1. A més, cal mencionar que el procés es va repetir dues vegades, ja que la primera vegada es va fer amb un conjunt de coordenades que no pertanyien a les regions d'edició exactament, sinó àrees properes a aquestes.

3.2.5. Profunditat de seqüenciació en les regions d'interès

Per tal de caracteritzar els productes d'edició genòmica és necessari observar un gran nombre d'edicions genètiques en les regions d'interès. Per tant, cal que la profunditat o *coverage* en aquestes regions sigui prou com per visualitzar varis productes d'edició genètica en cada una de les regions estudiades. Per saber si la profunditat de seqüenciació és suficient, s'utilitzen les coordenades de les regions d'interès en el genoma C3H identificades prèviament per buscar el *coverage* en aquestes.

Com que aquest procés requereix els *reads* alineats al genoma de cada mostra, es desenvolupa al servidor remot. Per fer-ho, es crea un *script* que utilitza l'eina "mpileup" de la suite SAMtools. Aquesta eina utilitza el fitxer de les coordenades d'interès en format BED per cercar en els fitxers de les mostres alineades (.bam) i

retornar el nombre de vegades que cada nucleòtid comprès en les àrees d'interès ha estat seqüenciat. El resultat s'emmagatzema en format Pileup.

A continuació, es podrien escollir dues opcions per interpretar el *coverage*. La primera constaria en determinar el *coverage* com la mitjana global de *coverage* de tot el conjunt de nucleòtids de les regions d'interès. Aquesta opció s'ha descartat perquè la profunditat de *coverage* sol ser constant en una regió d'interès però canviar entre regions d'interès. Per això, si es seguís aquesta aproximació s'obtindria un valor mitjà de *coverage* però no hi hauria informació sobre si algunes regions d'interès s'han seqüenciat correctament o no.

En canvi, s'ha escollit calcular el *coverage* mitjà de cada regió d'interès i visualitzar el resultat amb un gràfic de distribució. Així, es pot veure si hi ha un elevat nombre de regions d'interès amb *coverage* alt o baix, i veure si aquest és homogeni o canvia entre regions d'interès. A més, proporciona informació útil. Tenir en compte si una regió d'interès té poc *coverage* evitaria l'error al considerar que l'edició en aquella regió és poc eficient, ja que es sabria que no és que no hi hagi edicions sinó que no s'han pogut observar per *coverage* insuficient.

Per calcular el *coverage* mitjà de cada regió d'interès s'ha creat un *script* de Python, que es pot executar a la llibreta *1.4. CoverageAnalysis.ipynb*. Breument, aquest *script* itera sobre cada regió d'interès per cercar els nucleòtids en les coordenades d'aquesta regió, emmagatzema els valors de *coverage* individuals d'aquests nucleòtids, i calcula la mitjana de la regió d'interès com la mitjana de *coverage* dels nucleòtids en la regió d'interès. Finalment, guarda els resultats en un DataFrame de Pandas o exportats en format .csv perquè es puguin utilitzar posteriorment.

El darrer pas és visualitzar la distribució del *coverage* amb un gràfic de distribucions. Per això, es crea un *script* senzill que representa el *coverage* mitjà de cada regió d'interès amb una corba de densitat per cada mostra analitzada. Aquest *script* es pot trobar i executar a la llibreta *1.4. CoverageAnalysis.ipynb*.

Per complementar les conclusions de l'anàlisi de *coverage*, es visualitzen manualment algunes regions d'interès amb el navegador genòmic integrat IGV del Broad Institute.

3.3. Model computacional

3.3.1. Simulació de dades

Com que no s'observen resultats d'edició genètica en les dades experimentals, cal simular-los per recrear el procés d'entrenament del model que es duria a terme amb les dades experimentals. Tal i com s'ha comentat anteriorment, com que s'han de simular les dades es limita l'abast del projecte a la simulació de CRISPR-Cas9, així que ja no es torna a considerar l'activitat dels ABE, CBE, ni es té en compte la llibreria *mismatch*. Les dades simulades es basen en altres models computacionals similar al que es vol entrenar aquí, i encara que no permeten extreure conclusions noves, permeten preparar l'enfoc de l'entrenament del model computacional.

Com que el model a entrenar hauria de predir tant l'eficiència com la freqüència dels productes d'edició genètica, es simulen les dades tenint en compte tots dos paràmetres. Cap model computacional disponible fins ara realitza prediccions tant de l'eficiència com dels productes d'edició genètica, però hi ha varis models que realitzen aquestes prediccions de forma individual. Per tant, es decideix utilitzar un model d'eficiència per simular l'eficiència dels gRNAs i un model de freqüència dels resultats per simular els diferents productes d'edició genètica.

El model escollit per simular l'eficiència és el model creat per *Doench et. al.*[4]. Aquest model té una etapa de selecció de descriptors o *features*, així que el resultat d'eficiència es basa tan sols en un conjunt reduït de pesos que s'assignen a descriptors de la seqüència del gRNA. Degut al baix nombre de descriptors, és senzill adaptar les prediccions realitzades per aquest model en un *script* de Python i combinar-lo amb qualssevol altre procés en la *pipeline* per simular les dades.

El model escollit per simular la freqüència dels resultats d'edició genètica és el model creat per *Shen et. al.*[2], anomenat Indelphi. L'elecció d'aquest model en comptes del model de *Chen et. al.*[3] es basa en que el model Indelphi es pot importar com un mòdul de Python per utilitzar varies de les seves funcions. Aquestes funcions faciliten molt la simulació de les dades, així que es considera el model més adequat. Un inconvenient d'escollir aquest model és que només realitza prediccions d'insercions d'una base. Tot i que les insercions d'una base són el tipus d'inserció més abundant, seria interessant modelitzar insercions de més d'una base, ja que també poden ser rellevants per la mutació que es desitja aconseguir.

La simulació de dades es podria haver simplificat considerablement si s'haguessin predit directament els valors numèrics de cada producte d'edició i l'eficiència de cada gRNA. En canvi, es va decidir realitzar la predicció de les seqüències, simulant les dades en brut que s'obtindrien dels resultats de seqüenciació. Així, s'ha pogut desenvolupar un *script* per quantificar els resultats d'edició i l'eficiència a partir de les seqüències obtingudes, simulant tant com és possible el procés que es duria a terme al analitzar els resultats experimentals de seqüenciació.

Prèviament a la simulació de dades s'obté la seqüència genòmica de les regions d'interès, que serveix com a base per generar els resultats d'edició i és necessària pels models de predicció. Aquesta seqüència s'obté a partir de les coordenades convertides al genoma C3H, tal i com es descriu al final de la llibreta *1.3.CoordiantesToC3H.ipynb*. Així, el procés de simulació de dades comença amb 1785 seqüències genòmiques que corresponen a les regió d'interès. Aquestes seqüències tenen una llargada de 120 nucleòtids, i el punt de tall determinat pel gRNA es troba exactament a la posició 60 de cada seqüència. Aquest format d'entrada és el que utilitza Indelphi per realitzar les prediccions.

El primer pas del procés de simulació és calcular l'eficiència d'edició de cada regió d'interès. Per això s'adapten els paràmetres del model de *Doench et. al.*[4] i el valor d'eficiència obtingut s'ajusta a un valor entre 0 i 1. Quan més proper a 1, més eficient és el gRNA. Per recrear les dades experimentals, el valor en l'escala de 0-1 indica el percentatge de seqüències que es modificaran en la simulació de dades. Com que s'escull simular 5000 *reads* per cada regió, si la regió té un valor d'eficiència de 0.6, es simularan 2000 *reads* no modificats idèntics a l'original i 3000 amb mutacions.

A continuació, Indelphi utilitza la seqüència de les regions d'interès per predir la freqüència de cada resultat d'edició. El format en el que genera les prediccions Indelphi inclou la creació de la seqüència resultant de cada producte d'edició del qual prediu l'eficiència. Aquesta funció s'adapta manualment perquè les seqüències de les delecions incloguin *gaps* simbolitzats amb "-" i facilitar la identificació de les delecions. Així, s'obté una llista de les seqüències de cada un dels resultats d'edició i la probabilitat d'obtenir cada un. Després de normalitzar les probabilitats perquè sumin 1, s'utilitza la funció randomchoice de Numpy perquè escolleixi els resultats d'edició d'aquesta llista segons les probabilitats de cada un.

Així, per exemple, d'una regió amb eficiència 0.6, hi hauria 3000 *reads* amb mutacions. A continuació, Indelphi prediria, per exemple que la freqüència del resultat "deleció de 2 bases en la posició 3" és de 0.2, i generaria 3000*0.2=600 *reads* que contenen exactament aquest resultat. Per tant, les dades simulades per aquesta regió contindrien 2000 *reads* sense editar (perquè té una eficiència de 0.6), 600 *reads* amb una "deleció de 2 bases en la posició 3", i la resta de *reads* correspondrien a la resta de productes d'edició que prediu Indelhpi tenint en compte les freqüències esperades de cada un.

Per tant, el resultat del procés de simulació consta de 5000 *reads* simulats per cada una de les 1785 regions d'interès. La única simplificació significativa que es genera respecte a unes dades experimentals autèntiques és que hi ha el mateix nombre de *reads* de cada regió. A més, cal tenir en compte que aquestes dades no contenen soroll experimental, així que es pot preveure que els models entrenats tinguin una exactitud més baixa en dades reals. L'*script* utilitzat per simular les dades es troba a la llibreta *2.1.DataSimulation.ipynb*. L'execució d'aquest *script* requereix cert temps, així que en la plataforma *Google Colaboratory* es prepara per executar un exemple en el que es simulen 5000 *reads* per 3 regions d'interès, en comptes de 1785. Es pot veure el format del fitxer final amb les dades simulades, que és similar al que s'obtindria amb dades experimentals.

3.3.2. Quantificació dels resultats d'edició genètica

Si s'utilitzessin dades experimentals, caldria analitzar l'alineament dels *reads* al genoma de referència per caracteritzar l'eficiència i el perfil dels resultats d'edició de cada una de les regions d'interès. Al fer-ho, es compararia la seqüència de cada *read* amb la seqüència original per veure si és idèntica o no, i en el cas que sigui mutada, assignar-la a una categoria de resultat d'edició.

En el procés de simulació de dades es podrien haver obtingut directament els valors numèrics d'eficiència i freqüència de resultats, però s'opta per simular les dades experimentals abans de quantificar els resultats. El procés de quantificació dels resultats analitza cada regió d'interès individualment, i valora els resultats d'edició de diferents formes. Per una banda, identifica el percentatge de *reads* que són idèntics a la regió original, i per l'altra, classifica els *reads* editats en els diferents productes possibles d'edició genètica. De totes les formes possibles en les que es podria realitzar aquest procés, s'opta per escollir un mètode robust, modular i versàtil. La modularitat de l'*script* és essencial per adaptar-lo a diferents categories en les que es podria desitjar predir els resultats d'edició genètica.

Així doncs, l'aproximació escollida parteix de la seqüència original de cada regió d'interès i genera un diccionari que conté les seqüències dels resultats possibles d'edició genètica categoritzats segons el tipus d'edició genètica. Amb aquest diccionari, es busca el nombre de *reads* que corresponen a cada tipus de producte genètic. Finalment, es calcula la freqüència relativa de cada producte d'edició. En tot aquest procés, és essencial definir una classificació dels resultats d'edició genètica. S'opta per classificar els resultats al màxim detall, ja que si es volen agrupar varis productes en una sola categoria es podria implementar després fàcilment.

Per tant, les insercions d'una sola base es caracteritzen en quatre categories, segons el nucleòtid afegit en el punt de tall. Si hi hagués insercions de més d'una base, aleshores caldria afegir tantes categories com combinacions de nucleòtids possibles hi hagués. Per exemple, si es consideressin les insercions de dues bases, caldria afegir-hi 16 categories (4^2), corresponents a: AA, AT, AC, AG, TT, TA, TC, TG, etc…Tot i això, com que Indelphi s'ha utilitzat per la simulació i només prediu insercions d'una base, tan sols es tenen en compte 4 categories d'insercions.

Pel que fa les delecions, es té en compte la mida i la posició d'inici d'aquestes. Seguint l'exemple de l'estudi de *Chen et. al.*[3], es consideren totes aquelles delecions menors de 30 nucleòtids i que cauen en la regió (-3,+2) respecte el punt de tall. El nombre total de delecions possibles ascendeix a 536 categories diferents. A més, s'inclouen categories per classificar totes aquelles insercions majors de certa mida i totes les delecions majors a certa mida. Per defecte, s'agrupen en la categoria d'insercions grans totes aquelles insercions majors d'un nucleòtid i en la categoria de delecions grans totes aquelles iguals o majors a 30 nucleòtids.

Per identificar cada classe de resultat d'edició es defineix un codi amb tota la informació necessària. La primera lletra del codi indica si es tracta d'una inserció o una deleció. Els dos dígits següents indiquen la mida de la inserció o deleció. Finalment, els tres dígits següents indiquen la posició d'inici (5') de la deleció o la posició en la que s'ha realitzat la inserció. Per indicar les categories d'insercions o delecions grans s'utilitzen les sigles "EMT" en comptes dels tres dígits finals, que venen de "Equal or More Than". Així, és senzill identificar cada producte d'edició diferent a partir d'aquest codi de 5 dígits (Figura 6). Per exemple, la categoria D01-02 indica les delecions de mida 1 en la posició -2 del punt de tall. La categoria I0100A indica la inserció d'una A en la posició 1 respecte el punt de tall. La categoria I02EMT totes les insercions iguals o majors a 2 nucleòtids.

```
          Signe: ----- ++
        Posició: 54321 01
 5' GCCGAGTTTGATTAGGA TCC 3'
TACGGCCGAGTTTGATTAGGA|TCCCGG  genom.
TACGGCCGAGTTTGATTA-GA|TCCCGG  D01-03
TACGGCCGAGTTTGATTAG-A|TCCCGG  D01-02
TACGGCCGAGTTTGATTAGG-|TCCCGG  D01-01
TACGGCCGAGTTTGATTAGGA|-CCCGG  D01+00
TACGGCCGAGTTTGATTAGGA|T-CCGG  D01+01
TACGGCCGAGTTTGATT--GA|TCCCGG  D02-04
TACGGCCGAGTTTGATTA--A|TCCCGG  D02-03
TACGGCCGAGTTTGATTAG--|TCCCGG  D02-02
TACGGCCGAGTTTGATTAGGATTCCCGG  I0100A
TACGGCCGAGTTTGATTAGGATGTCCCGG I02EMT
```

Figura 6. Etiquetes assignades per classificar els productes d'edició genètica.

La versatilitat de l'*script* permetria adaptar-lo fàcilment a altres categoritzacions. Per exemple, s'ha inclòs variables per definir la mida màxima de delecions o insercions a estudiar, i les etiquetes es generen automàticament en resposta a aquestes. També permetria agrupar varis tipus d'*outcomes* en una sola categoria.

Una vegada s'ha definit l'etiqueta i seqüència de cadascun dels productes d'edició genètica de les regions a editar, s'utilitza aquest diccionari per contar el nombre de vegades que s'identifica cadascun en les seqüències originals. També s'inclou una etiqueta per contar el nombre de *reads* sense modificar.

Finalment, el nombre de *reads* modificats es representa en percentatge i correspon a l'eficiència que s'intentarà predir en el model de predicció de l'eficiència dels gRNAs. En canvi, el nombre de *reads* de cada producte d'edició es converteix a freqüència relativa i s'exporta per utilitzar en el model de predicció dels resultats d'edició genètica.

L'*script* utilitzat per quantificar cada resultat d'edició es pot trobar i executar a la llibreta *2.2.LabelGen.ipynb*. Tal com en el cas de les dades simulades, com que és un *script* que requereix cert temps, es presenta un exemple reduït per quantificar els productes d'edició genètica de les 3 regions d'interès simulades en l'apartat anterior.

Ara que es poden assignar categories a cada un dels productes d'edició genètica, s'analitzen les dades simulades. Així, s'identifica el percentatge de delecions i insercions introduïdes, la distribució de la mida de les delecions, i el percentatge de cada tipus de base introduïda. Els *scripts* utilitzats per analitzar les dades simulades es poden trobar a la llibreta *2.3.OutcomesProfiling.ipynb*, i es poden executar directament, ja que consten d'un seguit d'*scripts* per visualitzar dades de forma senzilla.

3.3.3. Caracterització dels descriptors

Per relacionar les seqüències de les regions d'interès amb els productes d'edició genètica, s'extreuen un conjunt de descriptors o *features* que descriuen la seqüència de la regió d'interès. Els descriptors utilitzats en aquest cas inclouen la posició i identitat dels nucleòtids que formen el gRNA estudiats individualment i en grups de dos (dinucleòtids). També s'inclou com a descriptor el contingut GC del gRNA, que indica el percentatge de Citosines (C) i Guanidines (G) en la seqüència d'aquest.

L'elecció d'aquests predictors es basa en que són els mateixos que utilitzen per predir l'eficiència dels gRNAs els models de *Doench et. al*[4]*.* i *Chari et. al.*[17]. Altres estudis indiquen que descriptors com l'estat de la cromatina, que regula l'accessibilitat a les regions del genoma, i paràmetres termodinàmics del gRNA poden millorar les prediccions d'eficiència[18]. Tot i això, la millora en l'exactitud es poca i la complexitat en la que incrementa la generació dels descriptors incrementa. Per això, s'opta per limitar els descriptors a la composició de nucleòtids i dinucleòtids de la seqüència i el seu contingut GC. També s'hi inclou quatre descriptors amb el percentatge de cada nucleòtid en la seqüència del gRNA, que recullen informació de la composició de nucleòtids del gRNA sense tenir en compte la posició d'aquests.

La composició de nucleòtids es representa en vectors *one-hot* tal i com es faria si es tractés d'un diccionari de paraules. Així, per a cada posició del gRNA hi ha 4 caselles (una per cada nucleòtid), amb un valor 1 indicant la base que hi ha en aquella posició i valor 0 per la resta de nucleòtids absents en aquella posició (Figura 7). Els descriptors pels dinucleòtids també es realitzen de forma similar. En canvi, el contingut GC i el contingut de cada tipus de nucleòtid sense tenir en compte la posició són valors numèrics entre 0 i 1 que indiquen percentatges.

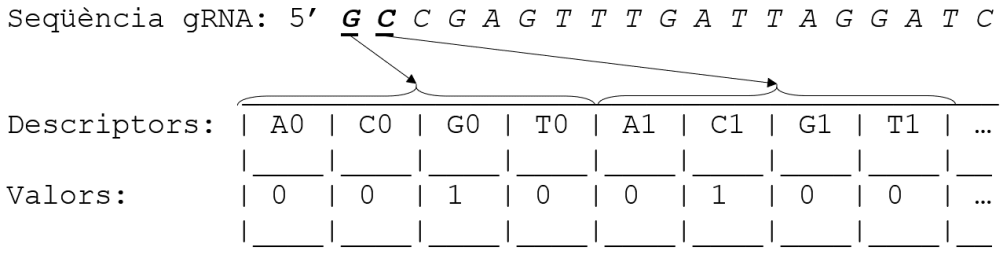

Figura 7. Esquema de la construcció dels descriptors a partir de la seqüència del gRNA. La composició dels nucleòtids del gRNA es converteix a vectors one-hot amb la informació del tipus i posició de nucleòtid.

L'*script* utilitzat per generar els descriptors de cada una de les 1785 regions d'interès es pot executar a la llibreta *2.4.Featurization.ipynb*. Com en els altres casos, el processat de les 1785 regions és relativament lent i per això es mostra un exemple reduït per mostrar-ne el funcionament.

3.3.4. Model de predicció de l'eficiència d'edició

El model de predicció de l'eficiència dels gRNAs es planteja com un classificador binari amb dues classes, -1 (no eficient) i +1 (eficient). Les dades d'eficiència són valors continus entre 0 i 1 que indiquen el percentatge de *reads* editat. Per dividir-les en aquests dos grups, s'ordenen de menor a major, i s'assigna al quartil més baix la classe de -1, i al quartil més alt la classe de +1. Per tant, la meitat de les dades, corresponent als gRNAs als quartils centrals, no s'utilitza per el modelatge de l'activitat.

Cal mencionar que la divisió entre els dos grups es fa seguint el treball de referència de *Chari et. al.*[17]. Prèviament a aquesta divisió, es va provar la divisió proposada per *Doench et. al.*[4], que després d'ordenar les dades assigna la classe -1 al 80% de gRNAs amb menor activitat i la classe +1 al 20% de gRNAs amb més activitats. El resultat d'aquesta divisió són dues classes de mides diferents. Encara que amb aquesta aproximació s'utilitza el total de les dades, l'entrenament dels algoritmes amb aquests grups de mida diferent no va assolir uns bons nivells de predicció. Per això, es va canviar a la divisió proposada per *Chari et. al.*[17] per tenir dos grups de la mateixa mida encara que es perdi la meitat de les dades.

El conjunt de dades de les classes +1 i -1 es divideix estratificadament (mantenint la proporció de cada classe) en un conjunt de *training* per fer l'entrenament i validació i un conjunt de *test* per valorar l'ajust dels models. El conjunt de *training* té el 72% de les dades i el de *test* el 28% restant. Aquesta divisió es planteja a l'inici, així que és comuna per tots els models entrenats. No es valora l'ajust diverses vegades canviant aquests grups com es faria en el procés de *nested cross-validation*.

Per l'entrenament dels models, es proven diversos paràmetres. Per identificar el valor òptim d'aquests paràmetres, es realitza 5-*fold cross-validation* en les dades de *training*. És a dir, per cada un dels paràmetres estudiats, s'entrenen 5 models utiltizant 4/5 de les dades de *training*, l'ajust de cadascun dels quals es valora fent prediccions del 1/5 restant de dades de *training* en el conjunt anomenat com a *validation*. Aquests conjunts canvien en cada un dels models estudiats. La mitjana dels valors d'*accuracy* que assoleix el model en el conjunt de *validation* de cada un dels 5 models entrenats per cada paràmetre s'utilitza per identificar el model amb millors paràmetres.

Una vegada el model amb els millors paràmetres s'ha identificat, es valora l'ajust d'aquest model a les dades del conjunt *test*. La divisió total entre les dades de *training*+*validation* i el conjunt *test* assegura que les dades de *test* no intervenen en cap moment en el procés d'entrenament i permeten simular l'ajust que tindria el model al realitzar prediccions sobre seqüències de gRNA noves. El procés de *5-fold cross-validation* durant l'entrenament permet dividir el conjunt d'entrenament entre validació i entrenament, aconseguint així models que es generalitzin bé i siguin menys sensibles a les peculiaritats de soroll i error aleatori de les dades d'entrenament.

Per implementar aquesta estratègia d'entrenament en diferents models, s'utilitza la llibreria *scikit-learn* de Python3. Així, s'utilitza *GridSearchCV* per cercar diferents paràmetres de cada model fent *5-fold cross-validation* amb les dades de *training* per cada un d'ells. Aquesta llibreria també permet entrenar 4 classificadors basats en algoritmes diferents, incloent *k-Nearest Neighbors (kNN), logistic regression, Support Vector Machine (SVM)* i *Random Forests*.

El classificador kNN es basa en la distància euclidiana per identificar els k-veïns més propers al punt que es vol classificar. Aleshores, assigna a la mostra a classificar el valor més prevalent entre els k-veïns més propers a aquesta. Per exemple, si es decideix estudiar els 5-veïns més propers, s'identifiquen les 5 observacions més properes segons la distància euclidiana al punt a predir, i si 3 o més d'aquests pertanyen a la classe +1, aleshores la predicció es classifica com a +1. Aquest algoritme es considera un *lazy learner*, ja que no té fase d'entrenament *per se* en la que s'ajustin certs pesos, sinó que la identificació i comparació amb els veïns es realitza íntegrament en el procés de generar prediccions.

Pel classificador kNN s'han optimitzat dos paràmetres: el nombre de veïns a considerar per assignar la classe (k), i si s'aplica un mètode de pesos en aquests veïns. El primer paràmetre regula si es tenen en compte més o menys observacions properes a la dada a predir per assignar una classe. El segon paràmetre regula si les observacions dels k-veïns tenen la mateixa importància al assignar la classe (*uniform*) o bé es pesen i es consideren proporcionals segons el valor de la distància amb la mostra a predir (*distance*). Si es consideren uniformes, la classe dels veïns més llunyans té el mateix pes que els veïns més propers, i si es consideren proporcionals a la distància, la importància canvia segons la distància.

La *logistic regression* és un classificador binari que es basa en un model lineal amb pesos per a cada *feature* per predir els *log odds* d'obtenir una classe i no l'altre. És a dir, calcula la probabilitat de tenir una classe donats els valors de les *features* de la seqüència a predir. L'entrenament del model consisteix en ajustar els paràmetres de pesos de cada *feature* perquè les probabilitats siguin més properes al valor de classe.

En el model de *logistic regression* s'optimitzen dos paràmetres diferents, tots dos relacionats amb la regularització per permetre una millor generalització del model. El primer és el tipus de *penalty* considerat, que pot ser L1 o L2. La *penalty* L1 correspon al valor absolut de la magnitud dels coeficients, així que quan es minimitzen es poden obtenir *features* amb un coeficient zero. En canvi, la *penalty* L2 és el quadrat de la magnitud dels coeficients, així que al minimitzar-los es redueix la magnitud dels coeficients dels *features* en la mateixa proporció. Quan s'utilitza la *penalty* L1 es podria considerar que es duu a terme una selecció dels descriptors d'interès pel model, ja que el valor del coeficient d'alguns models serà zero i, per tant, no intervindran en la predicció. En canvi, quan s'utilitza la *penalty* L2 es té en compte el valor de tots els predictors.

L'altre paràmetre important en la *logistic regression* és el valor del cost (C) de regularització. Els models d'aprenentatge automàtic es sol utilitzar la regularització per millorar la generalització del model entrenat. Sense regularització, el model intentaria classificar a la perfecció totes les mostres d'entrenament. Ara bé, quan s'apliqués el model a mostres noves per fer prediccions, el model estaria sobre-ajustat (*overfitted*) a les dades d'entrenament. En aquest cas, les prediccions del model serien pobres perquè s'ha tornat excessivament sensible a l'error experimental que podria ser característic de les dades d'entrenament. Per això, s'aplica un paràmetre de regularització que redueix l'ajust a les dades d'entrenament amb l'objectiu d'obtenir millors resultats al realitzar prediccions en dades noves (generalitzar millor). El valor de C és la inversa de la força de regularització (λ). Per tant, a valors baixos de C la regularització és fluixa, el model s'ajusta millor a les dades de *training* però es corre el risc d'*overfitting*. En canvi, a valors alts de C, la regularització és forta i el model s'ajusta pitjor a les dades de *training* perquè les prediccions siguin més realistes, tot i que si el valor és massa alt el model podria estar *underfitted*.

El classificador SVM separa dues classes binàries definint un hiperplà que separa els dos grups en l'espai de tantes dimensions com descriptors hi ha. Aquest hiperplà es coneix com a *decision boundary* perquè els valors en una banda són d'una classe i els valors en l'altre són de l'altra. El procés d'entrenament es centra en identificar l'hiperplà que separi les classes més adequat i el procés de predicció en valorar en quin dels dos costats de l'hiperplà es troba la dada a predir.

Els paràmetres estudiats per SVM inclouen el paràmetre C, que realitza la mateixa funció de controlar la regularització que en *logistci regression*, i paràmetres que tenen a veure amb el *kernel* utilitzat per transformar les dades. En els classificadors SVM és habitual transformar les dades per convertir-les en separables linealment. Per això, es proven tres *kernels* de transformació diferents. El *kernel* linear equival a no transformar les dades, i s'utilitzaria en cas que l'espai dimensional de les dades (equivalent al nombre de *descriptors*) sigui suficientment gran com perquè siguin separables linealment, ja que és el més senzill d'entrenar. El *kernel* polinòmic afegeix termes que tenen a veure amb la interacció entre els descriptors, que en aquest cas podrien ser importants per captar les relacions entre els elements de la seqüència de gRNA. Quan s'utilitza el *kernel* polinòmic, s'ajusta el grau del polinomi. Finalment, el *kernel Radial Basis Function* (RBF) o gaussià defineix regions d'influència de cada punt, la mida de les quals s'ajusta amb el paràmetre *gamma*.

El classificador *Random Forest* combina múltiples arbres de decisions (*Decision Trees*) diferents per realitzar les prediccions. Es basa en el principi que al utilitzar diversos arbres de decisions diferents, els errors dels uns es compensaran amb els altres i la predicció realitzada seguirà el patró subjacent de les dades. Per tant, en la seva optimització s'utilitzen tres paràmetres diferents. El paràmetre *n_estimators* defineix el nombre d'arbres de decisió diferents que s'entrenaran per fer les prediccions. El paràmetre *max_features* defineix el nombre de descriptors que considerarà cada un dels arbres de decisió. Limitar els descriptors aleatòriament abans de construir cada arbre assegura que els arbres entrenats són diferents entre ells. El paràmetre *max_depth* indica el nombre de decisions que pot tenir cada arbre per classificar les dades. Si aquest valor fos massa gran, incrementaria l'*overfitting*.

Per tots els classificadors, els paràmetres s'han inicialitzat cobrint un rang ampli de possibilitats i s'han anat centrant en un conjunt reduït en base els paràmetres dels models amb més *accuracy*. Per exemple, si un paràmetre es prova entre 1 i 10 i el model amb major *accuracy* és el model amb valor 10, s'expandiria el rang per cobrir de 1 a 50 i poder identificar un màxim centrat. Si al fer-ho es troba un màxim al voltant de un valor de 20, es reduiria el rang provat entre 15 i 25 per poder ajustar el punt concret del màxim.

Amb els millors paràmetres per cada model s'han realitzat les prediccions de les dades del conjunt de test. A partir dels valors probabilístics entre 0 i 1 de pertànyer a la classe activa (+1), s'han recreat les corbes ROC de cada un dels models i calculat l'àrea sota la corba ROC (AUC) per valorar-los. Així, es poden comparar els diferents classificadors i escollir aquell que és més interessant per predir l'eficiència dels gRNAs.

La comparació entre el model entrenat i els valors predits per el model de *Doench et. al.* es fa ordenant els gRNAs per activitat segons la probabilitat de pertànyer a la classe +1. Es defineixen grups de quintils d'aquests gRNAs. Els grups de quintils també es creen per els valors reals d'eficiència. Finalment, es representa la freqüència en la que els gRNAs dels quintils reals es troben en els quintils predits. També s'ajusta un model de regressió isomètrica per comparar si la posició dels gRNAs segons activitat és semblant entre els dos models.

Els resultats obtinguts en l'entrenament dels models es poden visualitzar en la llibreta *2.5.EffModel.ipynb*. Aquesta llibreta conté tots els *scripts* utilitzats en la preparació de les dades i l'entrenament del model. Els models SVM i *Random Forest* requereixen un temps d'entrenament que s'apropa als 5 minuts a causa dels diferents paràmetres provats. Tot i això, es pot executar tot el codi d'aquesta llibreta per validar la reproductibilitat dels resultats de la memòria.

3.3.5. Model de predicció dels resultats d'edició

Per predir la freqüència els diferents productes d'edició genètica, es simplifica el problema a la predicció de 8 grups que tenen en compte la mida de les delecions o les insercions. Aquests 8 grups es determinen a partir de la composició de les dades simulades per tal que la seva freqüència en les dades sigui similar i no hi hagi cap grup amb moltes més dades que un altre.

Tal i com es detalla a la secció de resultats, el problema de predicció es simplifica en dues aproximacions diferents, la predicció del resultat majoritari d'edició i la predicció de diversos productes relativament abundants.

Per a la predicció del resultat majoritari d'edició, es crea una sola etiqueta (*label*) per cada regió d'interès. Aquesta correspon al producte d'edició que té una major freqüència en les dades simulades per aquella regió. Com que hi ha 8 grups possibles i es vol predir la majoritària, les classes són excloents entre elles i es tracta d'un problema de predicció multi-classe.

Es proven els algoritmes de *logistic regression* i SVM. En tots dos casos, s'utilitza la tècnica de *5-fold cross-validation* per escollir els millors paràmetres entre un conjunt com els que s'han comentat en la secció de models de predicció de l'eficiència. A diferència del procés de predicció d'eficiència, aquest problema no té només 2 classes sinó que en té 8. Per això, s'entrenen 8 classificadors binaris *one-vs.-rest* que calculen la probabilitat que el producte d'edició majoritari sigui aquell en front de qualssevol altre. Per tant, les prediccions són en realitat la probabilitat d'obtenir cada un dels resultats d'edició possibles, i s'escull el majoritari a partir del que té més probabilitat. Així, per a cada regió d'interès es fa una sola predicció del producte majoritari.

En l'altra aproximació, es prediuen els resultats d'edició més abundants per cada regió, en comptes de tan sols el resultat majoritari. Per això, es decideix arbitràriament aplicar un *threshold* a la freqüència dels productes de 0.13, per sobre del qual es considera que aquell producte d'edició és abundant. Així, per cada regió s'obté una llista dels grups de productes que són més abundants en els *reads* simulats per aquesta regió. Per tant, el problema no s'enfoca com una predicció multi-classe, sinó com una predicció de la presència o absència de múltiples etiquetes.

Per aquesta aproximació s'utilitza *Random Forest* amb *5-fold cross-validation* per escollir els paràmetres adequats. El model de *Random Forest* permet predir la probabilitat de diverses classes, així que s'adapta perfectament a aquesta aproximació. El resultat per cada regió estudiada és un indicador binari de la presència o absència de cada un dels possibles resultats d'edició, que indica els productes que s'observarien majoritàriament. Per comparar aquest model amb prediccions aleatòries, s'entrena un classificador *dummy* que realitza prediccions aleatòries d'acord amb la freqüència de cada un dels productes d'edició. Les prediccions es comparen amb els valors utilitzats per entrenar el model per veure el nombre de coincidències entre els resultats predits per cada regió.

Els resultats obtinguts en aquests models es pot visualitzar i reproduir en la llibreta *2.6.OutcomesModel.ipynb*. Aquesta llibreta conté tots els *scripts* utilitzats en la preparació de les dades i l'entrenament del model.

4. Resultats i discussió

4.1. Anàlisi descriptiu de les dades de seqüenciació

4.1.1. Resum experimental

Breument, per poder predir els resultats d'edició genètica a partir de la seqüència de les regions editades, s'ha utilitzat una llibreria de 1785 gRNAs per editar el mateix nombre de regions diferents. Així, s'assegura que el model s'entrena amb una diversitat de seqüències prou gran com per identificar els patrons responsables dels resultats d'edició genètica.

S'han realitzat quatre experiments independents per valorar les tècniques utilitzades, CRISPR-Cas9, ABE i CBE. Hi ha dues mostres de CRISPR-Cas9, ja que es prova d'utilitzar una llibreria *mismatch*. Tot i que l'objectiu de la llibreria *mismatch* és fora l'abast d'aquest projecte, es comenta l'anàlisi dels seus resultats perquè té implicacions en les proves realitzades aquí. Per distingir la llibreria *mismatch* de la llibreria de 1785 gRNAs utilitzada per introduir modificacions genètiques, s'anomena a la darrera com a llibreria *perfect*.

El procés d'electroporació determina en gran part la qualitat de les dades obtingudes per entrenar el model. Si l'electroporació no és eficient, aleshores els vectors amb els enzims d'edició i la llibreria no entren a les cèl·lules. En aquest cas, poques cèl·lules estarien editades i per tant, s'observaria un menor nombre de productes d'edició genètic. Així, podria ser que el nombre de productes d'edició fos insuficient per entrenar un model de predicció.

Per valorar el procés d'electroporació, s'ha electroporat una cinquena mostra amb un plasmidi pSICO que expressa GFP de forma constitutiva. Per tant, les cèl·lules que hagin estat electroporades mostraran expressió de GFP i tindran valors de fluorescència elevada al examinar-les en un *Flow Cytometer*. Per l'anàlisi de *Flow Cytometry* també s'utilitza un control negatiu de cèl·lules no electroporades per identificar el nivell basal de fluorescència.

L'anàlisi de 10,000 cèl·lules per *Flow Cytometry* indica que aproximadament un 60% de les cèl·lules mostren fluorescència de GFP. Per tant, un 60% de la població estudiada haurà rebut els vectors necessaris per a l'edició genètica. Aquest valor indica que un 40% de les cèl·lules no hauran estat editades pel simple fet de no haver rebut els vectors d'edició genètica. Com que no es seleccionen selectivament les cèl·lules que han rebut els vectors d'edició, les dades de seqüenciació mostraran aproximadament un 40% de seqüències de les regions d'interès que no han estat editades per pertànyer a cèl·lules sense el material d'edició. A aquest valor, caldrà sumar-hi el percentatge de seqüències no editades per la relativament mitjana eficiència d'edició intrínseca d'aquestes tècniques.

En general, es considera que un 60% d'eficiència d'electroporació és un valor acceptable. Òptimament, aquest valor seria del 80%, ja que una eficiència del 100% és impracticable. Tot i això, si el *coverage* obtingut és suficient, un 60% d'eficiència hauria de ser suficient per observar múltiples productes d'edició genètica.

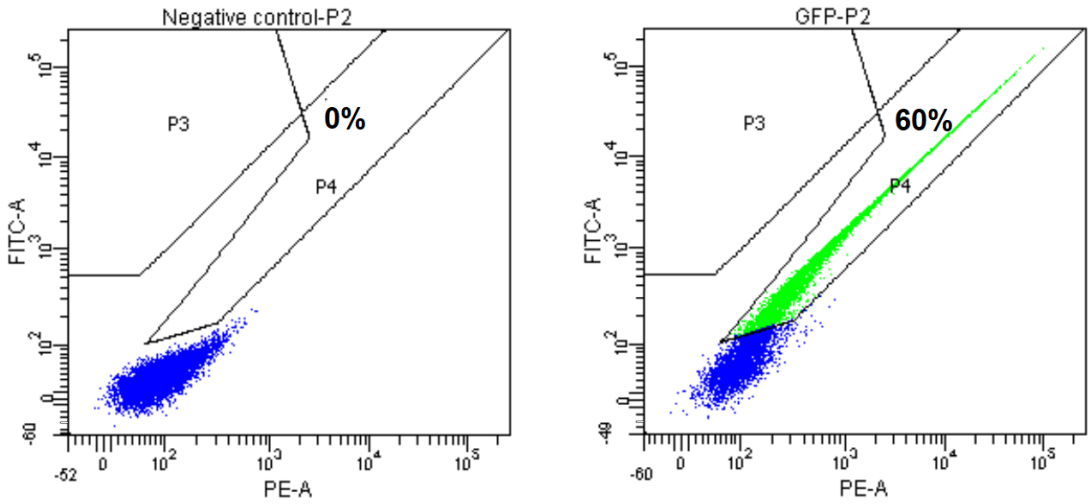

Figura 8. Eficiència d'electroporació valorada per flow cytometry. Es mostra la disposició de les cèl·lules del control no electroporat (esquerra) i electroporat amb pSICO-GFP (dreta) segons la seva fluorescència (eixos vertical i horitzontal). Les cèl·lules verdes en l'àrea P4 són les cèl·lules amb alts valors de fluorescència, que corresponen a les transfectades amb GFP. Les cèl·lules de l'àrea P4 corresponen a exactament el 60% de les 10,000 cèl·lules analitzades.

4.1.2. Distribució de la llibreria de gRNAs

Una presència desigual dels gRNAs de la llibreria podria produir diferències en les eficiències d'edició observades. Si un gRNA és present a la llibreria en més còpies que un altre és més probable que aquest entri a un major nombre de cèl·lules i s'observi un major nombre d'edicions a causa de la major abundància. Això, introduiria error en les relacions entre seqüències i eficiència observada. A més, si alguns gRNAs hi són en molta més abundància que altres, és possible que els gRNAs menys abundants entrin a un nombre més baix de cèl·lules i el nombre de productes observats sigui insuficient per entrenar un model.

Per valorar si cal tenir en compte diferències en l'abundància relativa dels gRNAs de la llibreria, s'ha seqüenciat els gRNAs de les llibreries electroporades a cada una de les condicions. A continuació, s'ha contat el nombre de *reads* de cada gRNA de la llibreria en les dades de seqüenciació. Finalment, s'ha representat la distribució del nombre de *reads* de cada gRNA per valorar si hi ha diferències entre les abundàncies d'alguns gRNAs.

Els resultats indiquen que la variació del nombre de *reads* dels gRNAs és relativament petita i centrada al voltant de 150 *reads* per gRNA en les mostres de la llibreria *perfect* (Cas9P, ABE i CBE) i al voltant de 85 per la mostra Cas9M que utilitza la llibreria *mismatch* (Figura 9). Per les mostres de la llibreria *perfect*, la majoria de gRNAs hi són presents entre 200 i 50 vegades, amb l'excepció d'alguns *outliers* amb 400 *reads*. Per la mostra de la llibreria *mismatch*, la majoria de GRNAs hi són entre 120 i 30 vegades. Tot i que la variació és major en les mostres de la llibreria *perfect*, la variació no arriba ni tan sols a un ordre de magnitud. Aquesta variació és relativament petita a les variacions que hi ha en altres llibreries, que podrien arribar als dos o tres ordres de magnitud. Per això, es considera com una variació acceptable en el nombre de *reads* per gRNA.

Cal tenir en compte que aquesta variació podria afectar l'activitat assignada a cada gRNA. Ara bé, és relativament petita, aquesta variació no depèn de la seqüència del gRNA, i es disposa d'un nombre elevat de gRNAs diferents. Per això, es pot considerar que algunes observacions compensaran les altres i per tant, la variació en el nombre de *reads* és un factor de soroll que no arribarà a afectar greument les prediccions del model.

També és interessant observar que la distribució entre les mostres que utilitzen la mateixa llibreria (totes excepte Cas9M) és la mateixa, mentre que canvia depenent de la llibreria utilitzada inicialment. Per tant, la distribució de la llibreria que s'observa depèn de la composició inicial (abans de la transfecció) de la llibreria. Això indica que el procés d'electroporació no altera la composició de la llibreria, fet que és positiu per aconseguir resultats no esbiaixats.

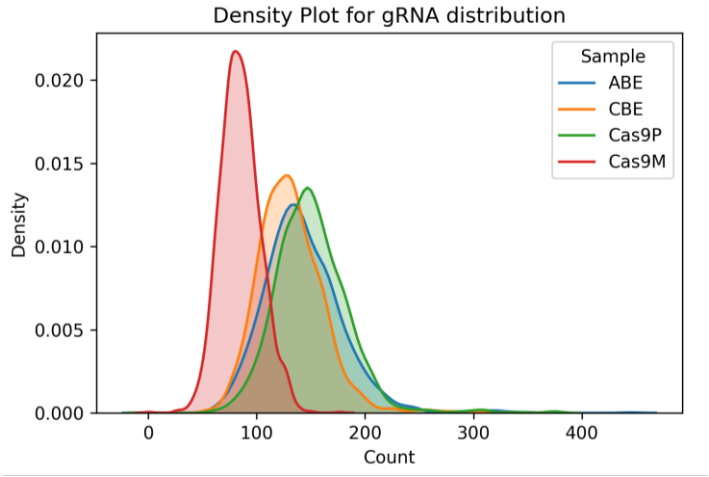

Figura 9. Distribució del nombre de reads de cada gRNA en les quatre mostres. Cal tenir en compte que es representen només els reads de la llibreria que es va transfectar a cada mostra (perfect per ABE, CBE i Cas9P i mismatch per Cas9M).

Tal i com s'ha comentat prèviament, les llibreries *perfect* i *mismatch* es sintetitzen juntes i es separen amb una PCR selectiva. Tot i això, hi ha la possibilitat que alguns components de la llibreria *mismatch* s'incloguin en la llibreria *pefect*, i viceversa. Aquest fet empitjoraria l'eficiència d'edició genòmica ja que es veuria reduïda la proporció de gRNAs de la llibreria correcte que permeten l'edició. Per això, s'ha contat el nombre de *reads* dels gRNAs de la llibreria *mismatch* en les mostres que utilitzen la *perfect* i viceversa.

El resultat indica que la proporció de gRNAs de la llibreria incorrecte tan sols és del 6% del total de *reads*. Gràficament, la diferència entre les abundàncies de les llibreries és clara (Figura 10). Les mostres Cas9P, ABE i CBE es van electroporat amb la llibreria *perfect* i per això la majoria de *reads* corresponen a la llibreria *perfect* i només un 6% dels reads són de la llibreria *mismatch*. En canvi, la mostra Cas9M es va electroporat amb la llibreria *mismatch* i per això la majoria dels *reads* corresponen a *mismatch* i tan sols uns pocs a *perfect*. Així doncs, es conclou que la separació entre les dues llibreries és clara i no suposa un problema a tenir en compte.

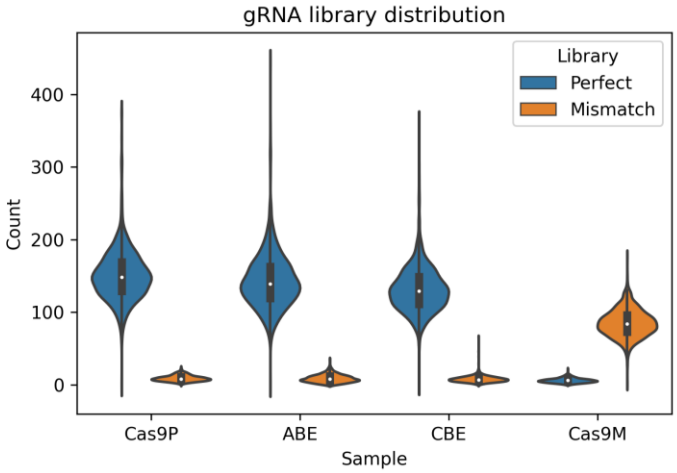

Figura 10. Distribució del nombre de reads de les llibreries perfect i mismatch en les mostres estudiades. Cal tenir en compte que les mostres Cas9P, ABE i CBE s'electroporen amb la llibreria perfect i la mostra Cas9M amb la mismatch.

Finalment, cal comentar la presència d'un gRNA que no s'ha observat en cap de les mostres. Com que el nombre mínim de *reads* de la resta de gRNAs és de 50, no és factible pensar que és un gRNA que no ha estat representat en la llibreria. En canvi, es podria tractar d'un gRNA que no es va poder sintetitzar correctament de forma química. Així doncs, caldria tenir en compte que no es pot esperar observar edicions genòmiques en la regió complementària al gRNA absent en la llibreria.

En resum, l'anàlisi de la distribució de *reads* de la llibreria indica que la variació entre les abundàncies relatives dels gRNAs és relativament petita i, per tant, no cal tenir aquest factor en compte al valorar l'eficiència dels gRNAs. Amb l'excepció d'un gRNA que no està en cap mostra ni en cap de les llibreries, la resta de gRNAs estan representats a les llibreries. Per tant, caldria esperar observar edició genètica en totes les regions genòmiques complementàries als gRNAs de la llibreria. El percentatge de gRNAs de la llibreria *mismatch* a la llibreria *perfect* i viceversa és del 6%, així que l'eficiència d'edició no es veu reduïda per aquest fet.

4.1.3. Profunditat de seqüenciació en les regions d'interès

Per entrenar un model capaç de predir els diferents resultats d'edició és necessari observar una alta quantitat de productes d'edició per cada regió editada, que no tan sols permeti valorar la diversitat dels productes sinó també quins són més freqüents. Per això, abans de procedir amb l'anàlisi de les freqüències dels productes d'edició, cal determinar si s'ha obtingut suficient profunditat de seqüenciació o *coverage* de les regions d'interès per observar-hi edicions.

Abans de l'anàlisi del *coverage* cal fer un alineament amb el genoma i identificar les coordenades d'interès per extreure selectivament el *coverage* d'aquestes regions. Mentre que l'alineament al genoma de ratolí C3H és senzill, en el procés de conversió de les coordenades d'interès surt un imprevist. Les coordenades d'interès són les regions complementàries als gRNAs de la llibreria, i com que es van dissenyar amb el genoma mm10, aquestes no coincideixen amb les del genoma de C3H.

Tal i com s'ha descrit a la secció de mètodes, l'ús de Remap de NCBI permet convertir les coordenades d'un genoma a un altre. Ara bé, els resultats del procés de conversió no són complets. Les seqüències de 30 gRNAs de la llibreria no s'han pogut identificar al genoma C3H, segurament a causa de mutacions o falta de *coverage* en l'*assembly* del genoma C3H. Fins i tot realitzant cerques manuals d'aquestes regions amb BLAT en el servidor remot no ha permès identificar les coordenades equivalents a aquestes regions en el genoma C3H. També s'ha identificat 20 gRNAs amb petites mutacions en el genoma de C3H, que es descarten perquè podrien confondre una mutació introduïda amb una mutació present al genoma C3H però no en *mm10*. Finalment, es descarten també 13 gRNAs per tenir nucleòtids no identificats al voltant de la seva seqüència genòmica en C3H. Aquestes impossibilitarien la quantificació de mutacions perquè la seqüència original és desconeguda.

Per tant, el procés de conversió redueix el conjunt de 1785 regions d'interès a un grup de 1722 que es poden identificar clarament en el genoma de C3H. Caldria tenir en compte que aquests 65 gRNAs que no s'ha identificat en C3H no s'utilitzen per entrenar el model. Tot i que no s'esperava la pèrdua de gRNAs en el procés de conversió, el nombre de seqüències no identificades es relativament petit, i per tant, es pot continuar amb l'anàlisi descartant aquestes regions.

Amb les dades alineades al genoma i les coordenades de cada regió identificades, es procedeix a l'anàlisi del *coverage*. Tal i com s'ha discutit en la secció de mètodes detalladament, cal valorar el *coverage* de cada regió d'interès i valorar la distribució d'aquest. Seria incorrecte obtenir un valor global de *coverage* per cada mostra perquè no tindria en compte que cada regió pot tenir un *coverage* diferent. Per tant, a partir dels fitxers *pileup* i les coordenades, es calcula el *coverage* mitjà de cada regió d'interès. Per visualitzar-ho, es representa la distribució d'aquest valor per les 1722 regions d'interès segons la mostra.

El resultat obtingut indica que la distribució de *coverage* canvia considerablement entre regions d'interès (Figura 11). També s'observa que la distribució està esbiaixada a la dreta, fet que indica que algunes regions tenen poc *coverage* però la majoria tenen valors alts i similars de *coverage*. La distribució del *coverage* és similar entre les mostres, confirmant que en cap d'elles hi ha un error experimental en la fase de seqüenciació que pugui afectar les dades obtingudes.

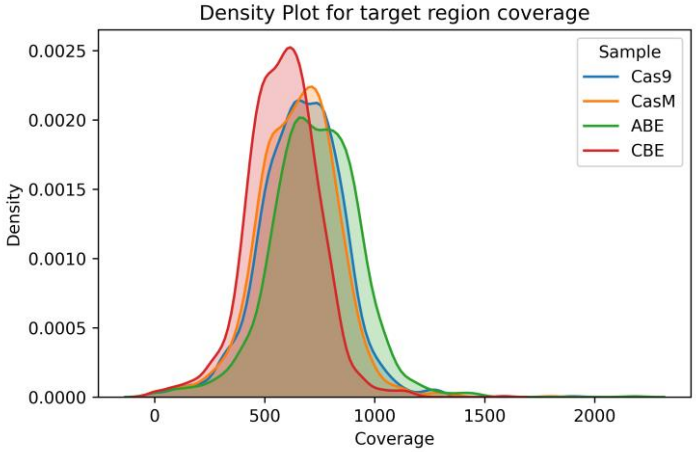

Figura 11. Distribució del coverage mitjà de les regions d'interès segons la mostra estudiada.

Poques regions d'interès tenen *coverage* proper al zero, confirmant que el procés d'enriquiment per seqüenciar selectivament les regions d'interès ha funcionat. Ara bé, fins i tot les regions amb *coverage* alt no arriben als 1000 *reads* de *coverage*, i s'eperava obtenir resultats més alts. És possible que tot i que el procés d'enriquiment hagi funcionat, la seva eficiència no hagi estat suficient per reduir al mínim la seqüenciació de regions genòmiques que no són d'interès (perquè no hi ha cap gRNA complementari a elles). Si aquest fos el cas, la obtenció de *reads* fora de les regions d'interès disminueix la proporció de *reads* de les regions d'interès, que explicaria els valors baixos de *coverage*.

El càlcul de la mitjana de *coverage* per regió d'interès de cada mostra (Cas9P=669, Cas9M=650, ABE=729, CBE=583) confirma que el *coverage* de les regions és insuficient per obtenir la diversitat de productes d'edicions esperada. Això compromet l'entrenament del model, ja que falten dades dels diferents productes d'edició que permetin identificar relacions entre seqüència de la regió i resultats d'edició.

Cal tenir en compte que, per exemple, si s'obtenen 1000 *reads* d'una regió d'interès, aproximadament 400 corresponen a *reads* no modificats perquè són de cèl·lules que no han rebut els vectors d'edició (ja que l'eficiència de transfecció és del 60%). Dels 600 *reads* de cèl·lules amb el material d'edició, l'eficiència d'edició per si és propera al 10%, així que s'esperaria observar 60 *reads* modificats. Finalment, però, cal tenir en compte que no totes les cèl·lules contenen la llibreria sencera, ja que aproximadament s'espera que tan sols entre 5 i 10 gRNAs diferents entren a cada cèl·lula. Per això, hi ha una probabilitat de 10/1785=0.5% que algun d'aquests 60 *reads* que podrien estar editats corresponguin a cèl·lules que tenen precisament un gRNA per la regió genòmica observada. Així doncs, 1000 *reads* s'espera que siguin insuficients per observar ni tan sols algun producte d'edició genètica.

Les dades genòmiques es poden visualitzar utilitzant IGV i cercant amb les coordenades convertides a C3H les regions d'interès. La inspecció visual dels resultats permet observar que les regions d'interès s'enriqueixen selectivament, ja que només s'observen *reads* al voltant de les regions d'interès (Figura 12). Això confirma que el procés d'enriquiment és correcte, perquè es seqüencien tan sols les regions d'interès i no la resta del genoma. Tot i això, s'hauria d'observar un nombre elevat de mutacions en el punt exacte de tall, i tal sols s'observen mutacions puntuals fora de la zona que es podria editar amb el gRNA. Per tant, es considera que aquestes mutacions són errors de seqüenciació i no productes d'edició genètica.

Tot i que cada regió d'interès té una probabilitat baixa d'obtenir *reads* editats pel *coverage* insuficient, com que s'analitzen fins a 1722 regions diferents caldria esperar almenys alguna edició en alguna de les regions. La identificació d'aquestes edicions permetria confirmar que l'experiment ha funcionat correctament i tots els enzims d'edició són actius. Tot i això, el procés d'identificació d'edicions és complex, i les dades de *coverage* ja permeten intuir que les dades obtingudes són insuficients per entrenar un model computacional. Per això, i a causa dels imprevistos en la fase 1 del treball, no es procedeix a crear un *script* per analitzar totes les regions d'interès i identificar possibles mutacions. Per fer-ho, tan sols caldria tenir en compte la seqüència original del genoma i comparar-hi la seqüència dels *reads* alineats.

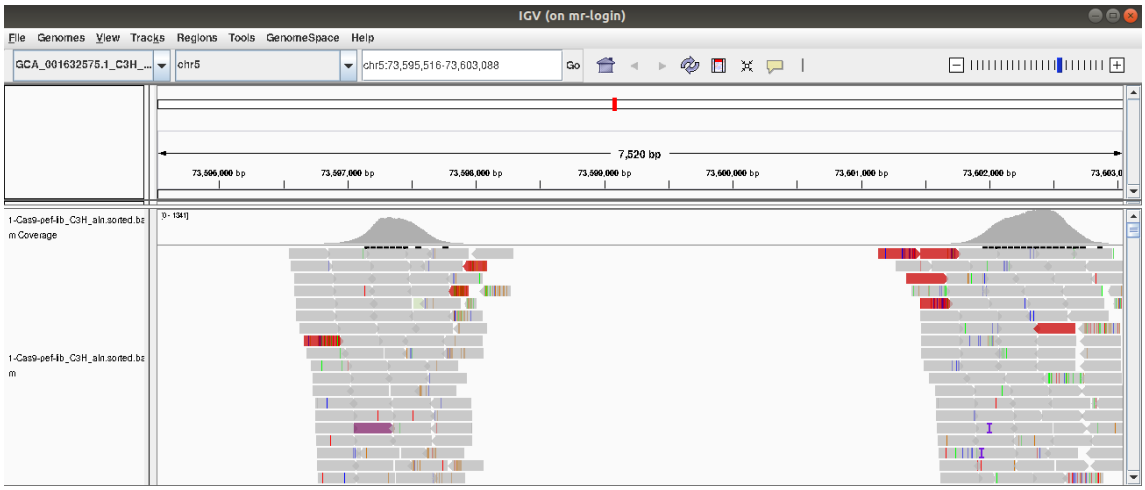

Figura 12. Visualització de dos regions d'interès amb IGV. Es mostren els reads de la mostra Cas9P que alineen en dues regions d'interès al cromosoma 5. La barra superior horitzontal representa el genoma de referència de C3H i els gràfics de distribució indiquen el coverage dels nucleòtids en aquesta regió. A la part inferior es poden veure els reads que alineen en aquesta regió d'interés, i les mutacions es representen mutacions amb un codi de colors.

En resum, el *coverage* de les regions d'interès no és suficient com per observar productes d'edició. El procés d'enriquiment sembla haver funcionat perquè s'han seqüenciat selectivament les regions d'interès. Per obtenir suficient *coverage* en les regions d'interès es podria optimitzar el procés d'enriquiment o bé incrementar la profunditat de seqüenciació utilitzada. Com que l'enriquiment sembla correcte, l'experiment es repetirà incrementant la profunditat de seqüenciació per aconseguir més *coverage* i observar varis resultats d'edició.

Per descartar que els components utilitzats per l'edició genètica siguin l'error, s'ha realitzat un experiment en que en comptes d'una llibreria s'utilitzen 5 gRNAs. Així, es poden seqüenciar aquestes regions per *targeted sequencing* en comptes d'enriquir-les per fer-ho amb *shotgun*. D'aquesta manera, el *coverage* incrementarà molt i es podrà comprovar si hi ha edició genètica o no. Tot i això, els resultats d'aquest experiment no s'han pogut obtenir a temps per la memòria pels imprevistos derivats de la pandèmia de COVID19.

En la planificació inicial del treball ja s'havia contemplat el risc d'obtenir un *coverage* insuficient per entrenar un model. La solució indicada en aquest cas és la de simular les dades d'edició a partir dels models ja existents. Tot i que la simulació de dades no permetria obtenir conclusions ni entrenar un model més proper a les condicions *in vivo* que els creats fins ara, permet recrear el procés que es seguirà un cop les dades experimentals siguin correctes. Així, la simulació de dades permet veure diferents formes d'enfocar el problema de predicció, identificar els potencials problemes que pot haver-hi en l'entrenament, i fins i tot crear *scripts* que es podrien reutilitzar directament quan les noves dades estiguin disponibles. Per tant, la simulació de dades permet seguir amb el treball i facilitarà l'avenç de la fase d'entrenament del model un cop hi hagi dades experimentals correctes.

4.2. Model computacional

4.2.1. Anàlisi descriptiu dels resultats d'edició

En el procés de simulació de dades es recreen les condicions experimentals simulant 5000 *reads* de cada una de les 1785 regions d'interès. Els *reads* simulats tenen en compte l'eficiència d'edició d'aquella regió predita utilitzant el model de *Doench et. al.*[4]. Els *reads* editats contenen la distribució dels productes d'edició predita utilitzant el model Indelphi descrit per *Chen et. al.*[3]. Les dades simulades s'analitzen per quantificar l'eficiència d'edició i la freqüència de cada producte. Així, es pot realitzar un anàlisi descriptiu de les dades simulades, tal i com es faria si les dades fossin experimentals. Aquest anàlisi aporta informació que és útil en el moment de plantejar l'enfoc adequat dels models d'aprenentatge automàtic.

Primer, s'analitza la distribució de l'eficiència d'edició de les dades. El valor d'eficiència indica el percentatge de *reads* de la regió d'interès que es troben modificades respecte l'original. Els resultats indiquen que l'eficiència dels gRNAs és majoritàriament inferior a 0.5 (*Figura 13*). Tan sols hi ha alguns gRNAs amb eficiències que superen el 0.6, i encara menys que superin eficiència 0.8. La distribució obtinguda a partir de les dades simulades coincideix amb el que caldria esperar. Al generar seqüències aleatòries de gRNAs, és més probable que els gRNAs no tinguin activitat que no pas que sí que en tinguin. Per això, la llibreria conté més gRNAs poc actius que amb alta activitat, tal i com indica la figura. Això també indica que el model entrenat seria capaç de diferenciar diversos graus d'activitat entre els gRNAs actius.

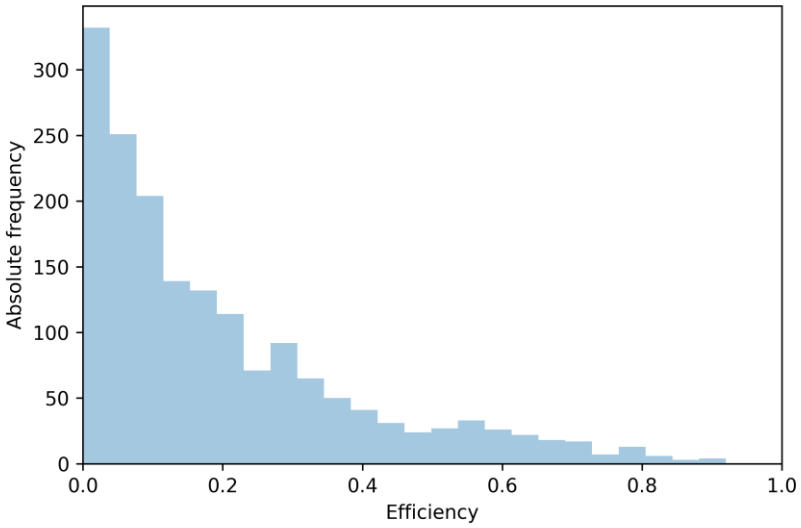

Figura 13. Distribució de l'eficiència d'edició de les regions d'interès segons les dades simulades. L'eficiència es mesura com el percentatge de reads simulats editats, i es mostra la freqüència absoluta de les eficiències de les 1785 regions d'interès.

Els següents anàlisis es centren en la caracterització dels diferents productes d'edició genètica. Al comparar el nombre de *reads* que contenen insercions i els que contenen delecions, es veu clarament que les delecions són molt més freqüents que les insercions (Figura 14). Aquest resultat coincideix amb les dades experimentals d'Indelphi[3], fet que no és d'estranyar perquè s'utilitza aquest model per simular les dades. En qualssevol cas, indica que s'han simulat correctament.

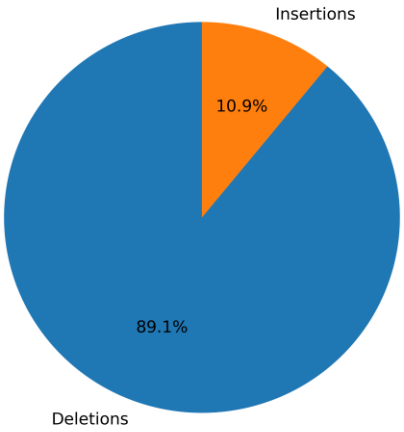

Figura 14. Comparació del nombre de delecions i insercions en les dades simulades.

La freqüència de les delecions decreix al incrementar la mida de la deleció (Figura 15). La distribució regular que segueix la baixada de freqüència es deu a la utilització d'una llibreria amb un elevat nombre de seqüències que es compensen uns amb els altres. Si en comptes d'observar el conjunt s'observés una sola regió d'interès, s'observaria una distribució particular de la freqüència de cada mida de deleció que dependria de la seqüència d'aquella regió. El fet d'obtenir una distribució regular al ajuntar totes les dades, es neutralitza l'efecte de la seqüència i s'obté un gràfic que indica que, independentment de la seqüència, les delecions més petites són més freqüents.

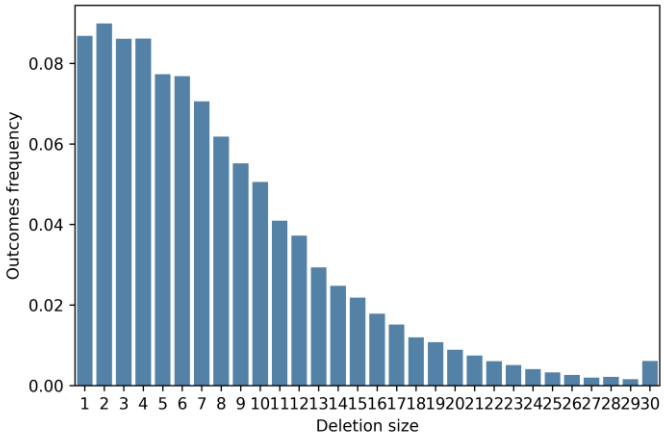

Figura 15. Freqüència de les delecions segons la seva mida.

L'anàlisi de les insercions indica que la inserció d'una adenina o timina és més freqüent que la inserció d'una citosina o guanosina (Figura 16). Segurament aquest fet té a veure amb l'estructura química dels enllaços dels parells de bases, ja que l'enllaç adenina-timina està format per 2 ponts d'hidrogen i és més dèbil que l'enllaç citosina-guanosina format per tres ponts d'hidrogen. El mecanisme d'acció de les insercions d'una base es coneix, i concorda amb el fet que la presència d'una A o T adjacent al punt de tall facilita l'aparellament incorrecte, generant així insercions d'una sola base[3].

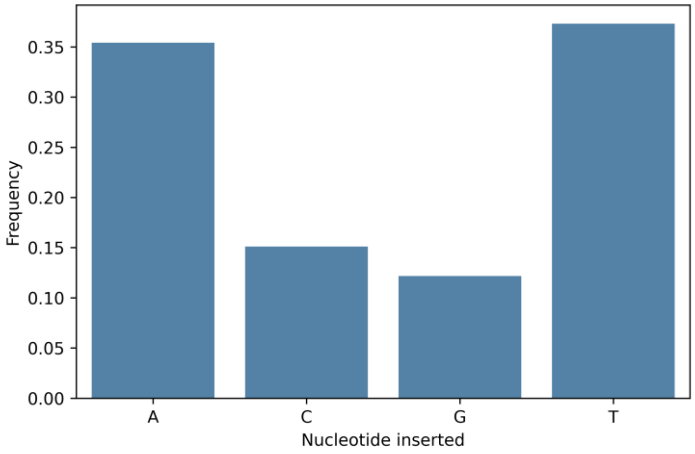

Figura 16. Freqüència de les insercions d'una sola base segons el nucleòtid inserit.

En resum, els resultats obtinguts analitzant el conjunt de dades simulades reflecteixen els determinants absoluts que depenen dels mecanismes de reparació més que de la seqüència del gRNA. Sobre l'eficiència, s'extreu la conclusió que seqüències aleatòries de gRNA és més probable que no tinguin activitat. A més, la distribució irregular indica que el model entrenat permetrà diferenciar amb precisió el grau d'eficiència dels gRNAs que tinguin certa eficiència. Sobre els resultats d'edició genètica, es pot veure que, en general, les delecions són més comunes que les insercions, les delecions més petites són més freqüents que les grans, i les insercions d'una sola base de A o T són més comunes que les de C o G.

Totes aquestes conclusions han estat descrites prèviament en altres estudis, i es relacionen amb característiques intrínseques del procés d'edició. Tot i això, l'objectiu del model és predir l'eficiència i els resultats d'edició d'una seqüència particular de gRNA. Aquesta, no té perquè seguir els patrons generals d'edició, ja que particularitats de la seqüència canvien la freqüència dels resultats d'edició obtinguts. Per exemple, per un gRNA concret seria més probable obtenir delecions de 5 nucleòtids que delecions de 2 nucleòtids, encara que la tendència general indiqués el contrari. Aquest fet indica la importància de l'entrenament d'un model que realitzi les prediccions per a seqüències concretes i no generals.

El fet que les dades simulades recreïn els patrons generals que s'esperarien és positiu, ja que indica que la llibreria de gRNAs conté una diversitat prou elevada de seqüències com per obtenir una representació completa de les tendències d'edició. Així, la caracterització d'aquesta llibreria de gRNAs seria representativa de qualssevol seqüència que es desitgi utilitzar aleshores per realitzar prediccions.

Els resultats mostrats en aquesta secció es poden reproduir si es simulen les dades tal i com es descriu a la llibreta *2.1.DataSimulation.ipynb*, es quantifiquen els seus resultats d'edició tal i com es descriu a *2.2.LabelGen.ipynb*, i es generen les figures d'aquesta secció amb la llibreta *2.3.OutcomesProfiling.ipynb*.

4.2.2. Model de predicció de l'eficiència d'edició

Pel model d'eficiència s'entrena un classificador binari per distingir entre la classe no activa (-1) i la classe activa (+1). Com que els valors d'eficiència indiquen el percentatge de *reads* editats, cal convertir els valors continus en dos grups binaris. Tenint en compte que la distribució de l'eficiència de les dades és desigual, la millor estratègia és ordenar els gRNAs de major a menor activitat i utilitzar els quartils per definir grups d'activitat de mides iguals, tal i com proposa *Chari et. al.*[17].

Prèviament, s'ha provat d'utilitzar dos grups d'activitat de mida desigual però els models tenien poca capacitat de predicció per l'elevat nombre de dades de la classe -1 respecte la classe +1. També s'ha desestimat separar les dades en dos grups definint un valor arbitrari d'eficiència per sobre el qual serien d'una classe o d'una altra, ja que es crearien grups desiguals per la distribució esbiaixada.

Els models s'han entrenat utilitzant *5-fold cross-validation* en les dades de *training* per escollir els paràmetres que resulten en una major exactitud (*accuracy*) del model. La mètrica utilitzada per optimitzar s'ha escollit com l'*accuracy*, ja que correspon al percentatge de dades classificades correctament. És important mencionar que la implementació de *5-fold cross-validation* permet calcular l'*accuracy* en el conjunt de *validation*, que és diferent de les dades utilitzades per entrenar el model, i que el valor d'*accuracy* obtingut és la mitjana de les 5 fraccions diferents de *validation*.

Els paràmetres determinats com a òptims pel model de kNN és utilitzar k=16 veïns propers i tenir en compte la distància d'aquests veïns amb la mostra a predir per identificar la classe a la que pertany (Figura 17). L'evolució d'aquests paràmetres deixa veure que quan el nombre de veïns considerats és parell, assignar pesos als veïns segons les distàncies millora l'*accuracy*. En canvi, si el nombre de veïns és imparell l'*accuracy* del model és la mateixa si no es tenen en compte les distàncies. Al considerar un nombre de veïns menor o major que 16 baixa l'*accuracy*. Normalment es tria un nombre de veïns imparell per evitar empats, però com que aquí s'utilitza la distància per assignar pesos, la probabilitat que hi hagi empats és pràcticament nul·la.

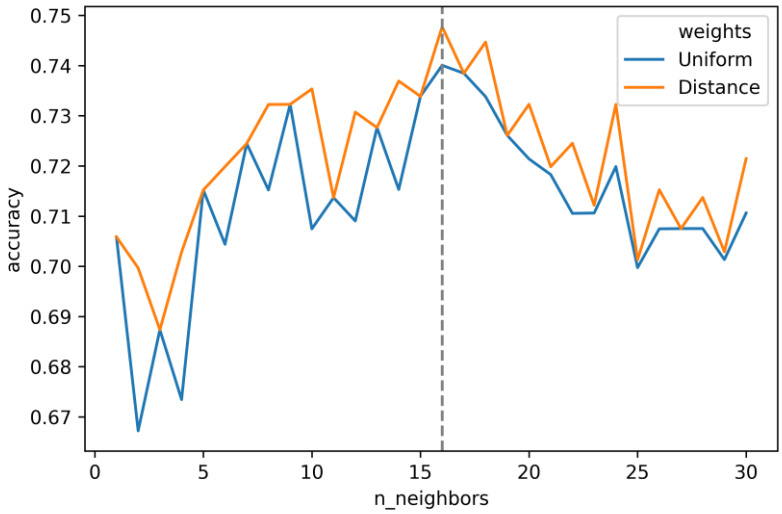

Figura 17. Evolució de l'accuracy del classificador kNN segons el nombre de veïns considerats (k) i si es té en compte la distància d'aquests o no.

Els paràmetres que maximitzen l'*accuracy* en el model de *logistic regression* és utilitzar la regularització L1 amb un cost de regularització C=1 (Figura 18). Es pot veure que baixos valors de cost de regularització disminueixen molt l'*accuracy* del model. Els valors baixos de C redueixen el pes de la regularització, així que el model s'ajusta molt a les dades d'entrenament i la seva *accuracy* en la predicció de dades noves baixa molt perquè està sobre-ajustat (*overfitted*). No es podria observar aquesta tendència si no s'implementés la *cross-validation* per valorar l'ajust en dades de validació i no d'entrenament.

També és interessant observar que la regularització L2 s'ajusta millor a valors baixos de C que en la regularització L1. Tot i això, quan s'ajusta la regularització perquè el model generalitzi bé la regularització L1 s'ajusta una mica millor. Com que es disposen de molts descriptors, segurament no tots són necessaris, així que la regularització L1 que actua com a *feature selection* s'ajusta més al problema plantejat.

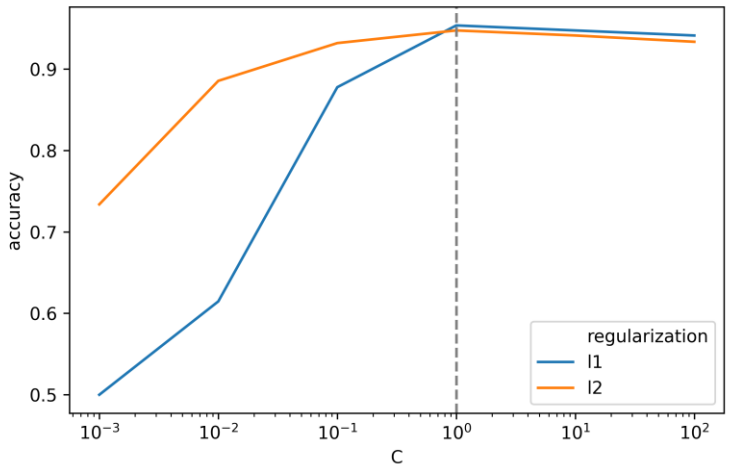

Figura 18. Evolució de l'accuracy de la logistic regression en el conjunt de validació segons el cost de regularització (C) i el tipus de regularització.

El classificador SVM que ofereix una major *accuracy* és si s'utilitza un *kernel* polinòmic de grau 2 amb un paràmetre de regularització de 100 (Figura 19). El *kernel* de segon ordre té en compte les interaccions entre els descriptors, i el fet que s'identifiqui com el model amb més *accuracy* indica que les interaccions de segon ordre entre els descriptors podrien ser importants. També cal mencionar que les diferències entre l'*accuracy* són bastant notables segons el paràmetre de regularització. El *kernel* lineal no canvia segons el paràmetre *gamma* perquè és una transformació que no té paràmetres a ajustar.

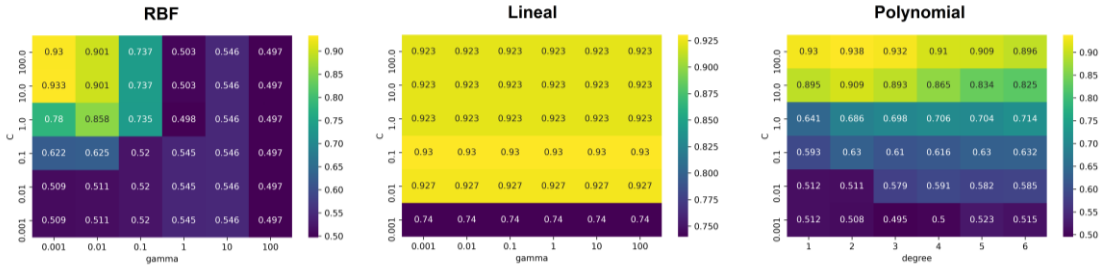

Figura 19. Heatmaps de l'accuracy del model SVM segons el tipus de kernel utilitzat i el cost de regularització (C).

El classificador *Random Forest* amb major *accuracy* és el que té en compte 280 arbres de decisió diferents (*n_estimators*), cadascun dels quals considera 8 *features* (*max_features*), i té una profunditat de 15 (*max_depth*) (Figura 20). Els valors d'*accuracy* mostrats varien molt poc, ja que es realitzen 4 iteracions en diferents rangs de paràmetres. Crida l'atenció que la direcció d'optimització en el cas de *Random Forest* no ha estat clara al llarg de les iteracions, així que s'ha identificat un màxim local que es podria millorar si es segueixen fent proves.

Un avantatge de l'algoritme *Random Forest* és que és relativament senzill interpretar el model entrenat. Per exemplificar-ho, es mostra part d'un dels 220 arbres de decisió que conté el model entrenat (Figura 21). Aquest arbre en concret primer es fixa en la presència d'una guanina a la posició 19, i depenent de si hi és o no, analitza la presència de citosina en la posició 14 o del dinucleòtid TG en la posició 17. L'arbre complet conté 15 nodes de decisió perquè *max_depth=15*, dels que tan sols es mostren 4. Cal tenir en compte que la predicció final depèn de 220 arbres com aquest.

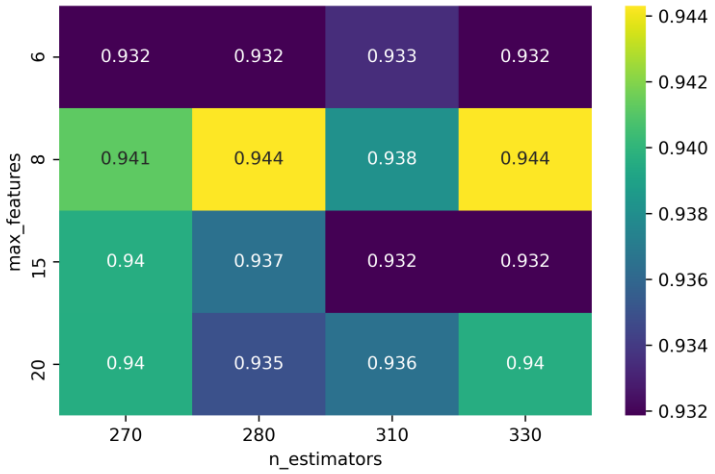

Figura 20. Accuracy segons els paràmetres del model Random Forest.

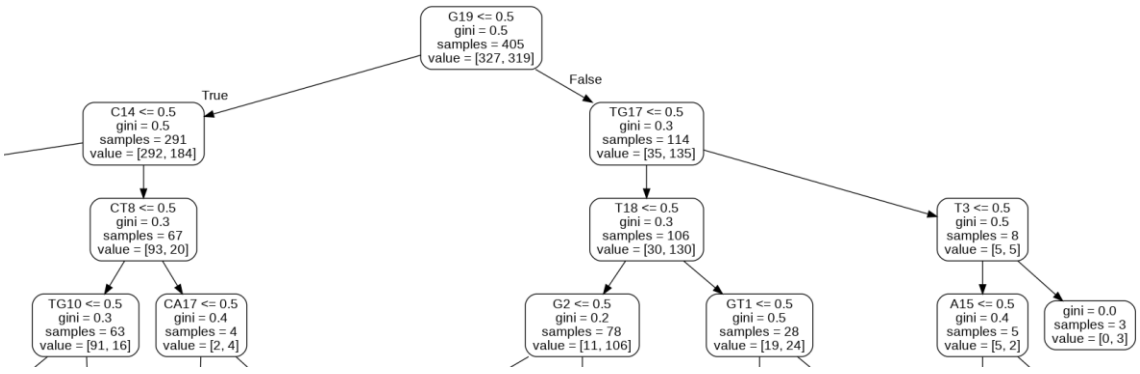

Figura 21. Primers quatre nivells de decisió d'un dels 220 arbres del Random Forest.

A partir de la presència dels descriptors en els arbres de decisió de *Random Forest* es pot calcular la importància relativa de cada descriptor. Visualitzant els 15 descriptors amb major importància, es veu que s'hi inclou la composició dels 4 nucleòtids sense tenir en compte la posició (A, G, C, T) i el contingut GC (GC_content) (Figura 22). La resta de descriptors importants pertanyen tant a nucleòtids com dinucleòtids en posicions particulars de la seqüència. El fet que s'observin tots els tipus diferents de predictors utilitzats entre els 15 predictors amb major importància en *Random Forest* indica que l'elecció d'aquests tipus de prediccions és adequada per la tasca realitzada. Cal mencionar que la importància dels descriptors és similar i decreix progressivament, sense que s'observin alguns predictors que siguin molt més importants que els altres. Aquest fet era d'esperar, ja que l'eficiència és el resultat de la composició del conjunt de la seqüència, així que no es pot atribuir a un conjunt molt concret de característiques d'aquesta.

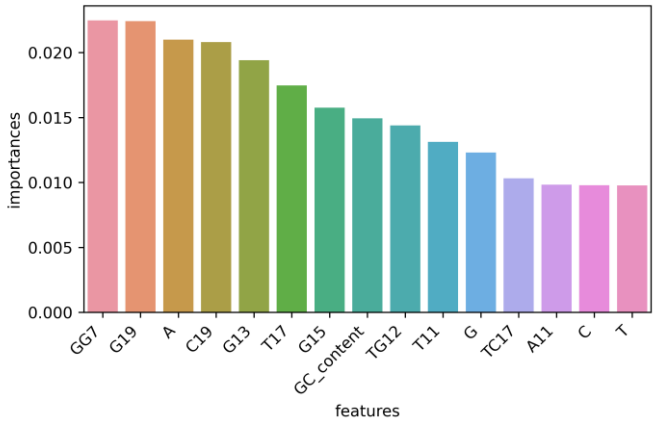

Figura 22. Importància dels 15 descriptors més importants en l'algoritme de Random Forest.

La comparació dels diferents classificadors es fa amb les dades de test, que no s'han utilitzat en cap procés d'entrenament i són les mateixes per a tots els classificadors. Tot i que els models s'hagin optimitzat per l'*accuracy*, es comparen els models utilitzant la corba ROC i l'àrea sota aquesta corba, AUC (Figura 23). En aquesta corba es mostra la proporció de falsos positius a l'eix horitzontal i la proporció de positius verdaders a l'eix vertical. Valors més alts de verdaders positius indiquen una major capacitat del model de detectar positius, és a dir, major sensitivitat o *recall*. Valors més baixos de falsos negatius indiquen una major especificitat o *precision*.

Un classificador òptim tindria el màxim de sensibilitat i el màxim d'especificitat, ja que així tots els positius es detectarien com a positius i no hi hauria falsos positius. Per això, la corba del classificador òptim estaria enganxada a la part superior esquerra de la corba ROC, mentre que un classificador que realitza prediccions aleatòries seria una diagonal. En aquest cas, és important mencionar que la classe activa (+1) dels gRNA es considera positiva i la classe de gRNAs no actius (-1) la negativa. Pel problema en qüestió, és més interessant incrementar la sensitivitat del model que l'especificitat. Així, es redueix el nombre de gRNAs que tindrien activitat i es classifiquen sense activitat, encara que això suposo incrementar els gRNAs sense activitat classificats com a actius. Aquesta decisió es basa en que la majoria de seqüències genòmiques tindran poca activitat i, per tant, és important classificar correctament els pocs gRNAs que podrien tenir activitat.

El model kNN és el que té pitjor capacitat de predicció, i és l'únic que es descartaria per realitzar les prediccions. La raó segurament és que el model kNN no és gaire adequat quan hi ha un nombre elevat de descriptors i aquests són dispersos. En aquest cas, hi ha molts descriptors que tenen valors 0 perquè corresponen als nucleòtids o dinucleòtids no presents en la seqüència. Per tant, era d'esperar que fos el model amb menys exactitud en les seves prediccions.

La resta de models tenen valors d'*accuracy*, AUC i perfils de la corba ROC similar. Valors d'AUC majors indiquen una major capacitat de predicció i major balanç entre especificitat i sensibilitat. Tot i que el valor d'AUC és lleugerament superior amb SVM, s'escull el model de *logistic regression* com el més adequat. El cercle negre indica la posició del *threshold* de 0.5 per decidir a partir de quina probabilitat es considera una o altra classe. Si es manté aquest *threshold*, la sensitivitat de *logistic regression* és més gran que la de SVM. Tot i que es podria ajustar el *threshold* d'SVM per reduir la diferència, el valor de sensitivitat segueix sent major amb *logistic regression*. Així doncs, s'escull *logistic regression* com a òptim, tot i que tant SVM com Random Forests serien alternatives vàlides.

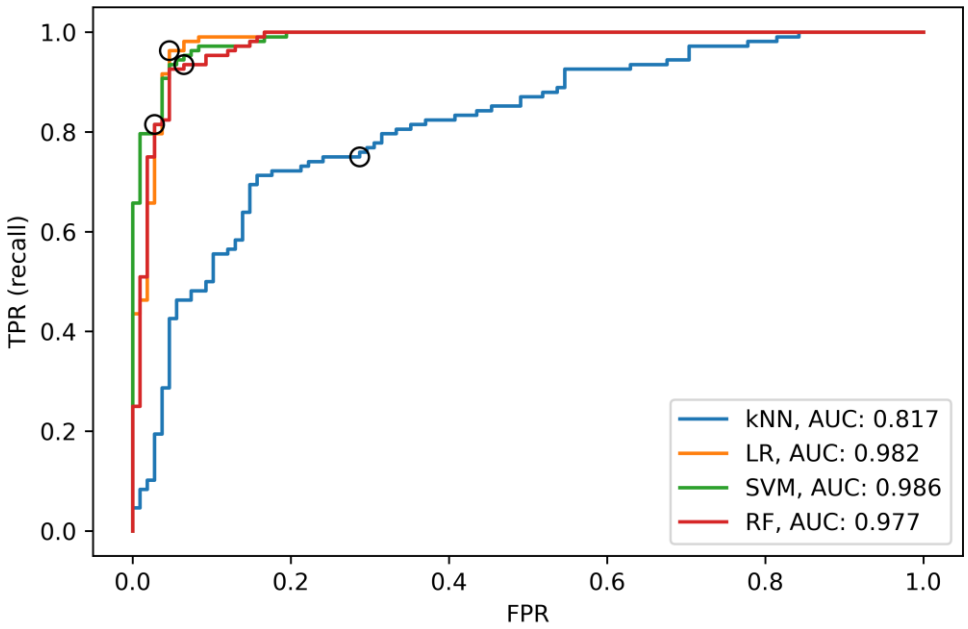

Figura 23. Corba ROC comparant les prediccions del conjunt de test segons els models entrenats. Els models utilitzats de cada classificador són els que tenen els paràmetres amb major accuracy. Es mostra l'àrea sota la corba (AUC) per a cada model. El cercle negre indica la posició de la corba si s'utilitza el threshold de 0.5 per separar les dues classes.

Havent escollit *logistic regression* com al classificador més adequat, es compara el valor d'eficiència predit per el model entrenat amb el valor d'eficiència dels gRNAs real, que correspon al que s'ha calculat amb el model de *Doench et. al.*[4] per simular les dades. Per fer la comparació, s'ordenen els gRNAs segons l'activitat i es creen grups de quintils. El valor continu d'eficiència de les prediccions correspon a la probabilitat que un gRNA pertanyi a la classe activa (+1). Es representa el percentatge de gRNAs que pertanyen a un quintil en els valors d'eficiència reals (simulats) segons el quintil predit.

La representació indica que els gRNAs del quintil amb més activitat segons la predicció conté més d'un 60% de gRNAs que també pertanyien al quintil amb més activitat segons els valors reals (Figura 24). Fins a un 90% dels gRNAs predits amb major activitat tenen una eficiència real major del 0.6. Això indica que els gRNAs que es prediguin amb alta activitat realment tindran alta activitat, perquè les prediccions concorden amb el valor real.

També es pot veure que el quintil amb els gRNAs de menor activitat té fins a un 80% de gRNAs que pertanyen a gRNAs amb una eficiència real (simulada) menor al 0.2. Per tant, els gRNAs que el model predigui com a no actius tenen un 80% de probabilitats de tenir una eficiència realment menor a 0.2. La composició dels quintils del centre canvia seguint la tendència que caldria esperar, ja que els grups predits amb més eficiència contenen un percentatge major de gRNAs que realment tenen més eficiència. Aquesta visualització indica que les prediccions del model concorden amb les dades reals d'eficiència. Cal mencionar que com que les dades reals són simulades, no hi ha errors experimentals que podrien empitjorar l'ajustament dels models.

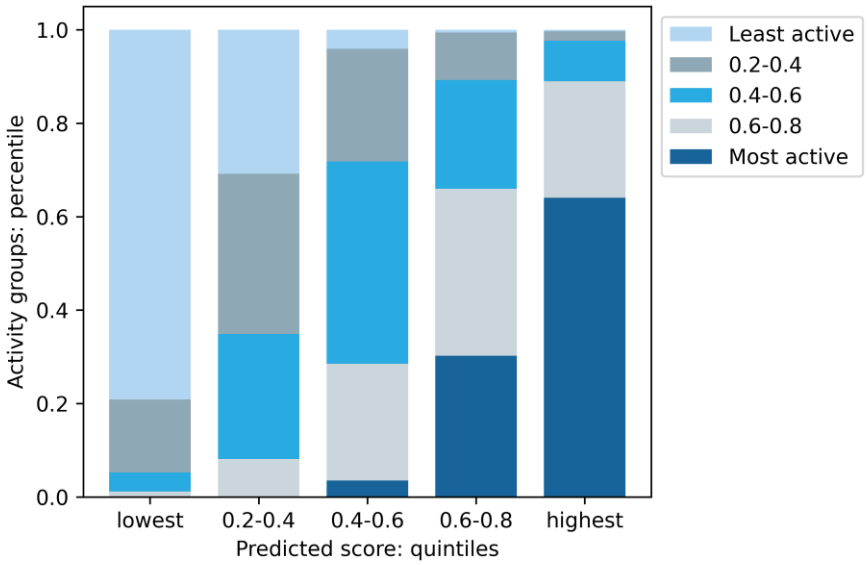

Figura 24. Comparació entre les prediccions i els valors reals d'eficiència dels gRNAs segons els quartils. L'eix vertical mostra agrupacions dels gRNAs en quintils segons la seva eficiència. Els colors de les barres indiquen el quintil d'eficiència real al que pertanyen els gRNAs dels grups verticals. L'eix horitzontal mostra el percentatge de composició de cada quintil.

Un altre mètode de comparació de les prediccions amb els valors reals és utilitzar una regressió isomètrica. Aquesta regressió ajusta un model que valora si l'ordre de les dades, ordenades de menor a major valor d'eficiència, és el mateix en les prediccions que en les dades reals. Per tant, el model valora si els gRNAs predits com a més eficients també són més eficients en els valors reals i viceversa. Aquest model difereix d'una regressió lineal perquè no té en compte la magnitud del canvi d'eficiència entre predicció i real. Aquesta es pot obviar, ja que al final, l'*score* d'eficiència és només un indicador que permet comparar relativament els gRNAs, així que la magnitud absoluta d'aquesta no és important. Per això, precisament, cal utilitzar regressió isomètrica en comptes de regressió lineal.

El resultat indica que la tendència general és similar entre les prediccions i els valors reals (Figura 25). El coeficient de determinació de la regressió és de R^2=0.671. L'ajust podria ser més alt, però la relació directe entre prediccions i valor real és clara. Gràficament es veu que el model assigna valors alts de predicció a alguns gRNAs amb un valor d'eficiència real baix i, en canvi, no assigna valors de predicció alts a gRNAs que siguin baixos. Això indica que la sensitivitat del model és major que l'especificitat d'aquest, ja que part dels gRNAs predits com a actius en realitat no ho seran. Tal i com s'ha discutit anteriorment, aquest és el balanç sensitivitat-especificitat més adequat per aquest tipus de problema.

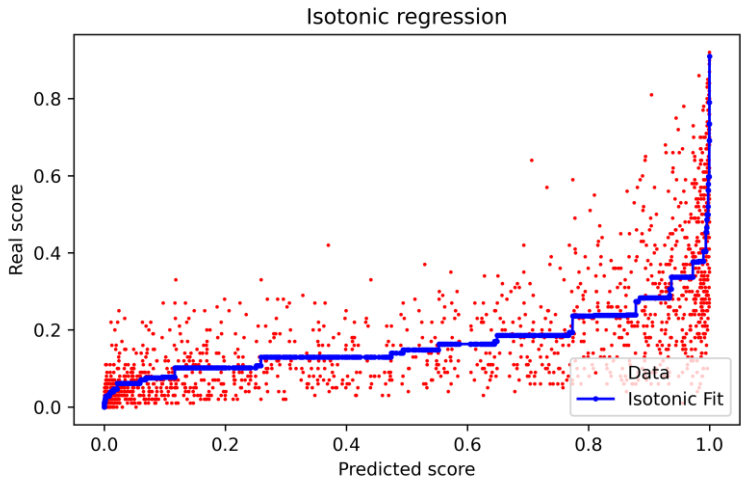

Figura 25. Comparació entre els valors predits i reals d'eficiència dels gRNAs amb un model de regressió isomètrica.

En resum, després d'optimitzar els paràmetres de classificadors binaris kNN, *logistic regression*, SVM i *Random Forests* s'han comparat per veure que tots realitzen prediccions acurades excepte kNN. S'ha escollit *logistic regression* com a model més adequat perquè incrementa la sensitivitat al màxim sense perjudicar gaire l'eficiència. Les prediccions d'aquest model segueixen la tendència dels valors d'eficiència determinats en la simulació de dades. Cal tenir en compte que com que s'ha utilitzat dades simulades, no hi ha errors experimentals que afectin la capacitat de predicció dels models. Si hi hagués errors experimentals, l'*accuracy* dels models baixaria i potser el model ideal seria un altre. Tot i això, els *scripts* creats es podrien adaptar ràpidament a uns nous valors d'eficiència experimental per identificar el model més adequat per les dades experimentals.

4.3.3. Model de predicció dels resultats d'edició

Les dades simulades contenen 542 tipus de productes d'edició, incloent 4 insercions d'una sola base segons el nucleòtid inserit, 536 delecions entre 1 i 29 nucleòtids en les posicions -3 a +2 respecte el punt de tall, i 2 categories per agrupar les insercions o delecions més grans. Per simplificar l'entrenament d'un model computacional per predir la freqüència d'aquests resultats, s'han agrupat aquests tipus d'edició en categories més àmplies.

A partir de l'anàlisi descriptiu de les dades simulades, s'ha definit 8 grups de tipus d'edició genètica. Aquests 8 grups es troben en una freqüència similar en el conjunt de les dades simulades, fet que permet entrenar un model amb una quantitat de dades similar per cada grup. Ara bé, la freqüència dels grups és diferent en cada regió d'interès, ja que depèn de la seqüència d'aquesta regió. Un dels grups creats agrupa totes les insercions, ja que aquestes són relativament poc freqüents. Els 7 grups restants agrupen delecions segons la seva mida, sense tenir en compte la posició respecte el punt de tall (Figura 26). Es pot veure que els grups amb delecions de mida petita agrupen menys mides de delecions que els grups amb delecions grans, ja que les delecions petites són més freqüents i amb l'agrupament es pretén aconseguir grups d'igual freqüència.

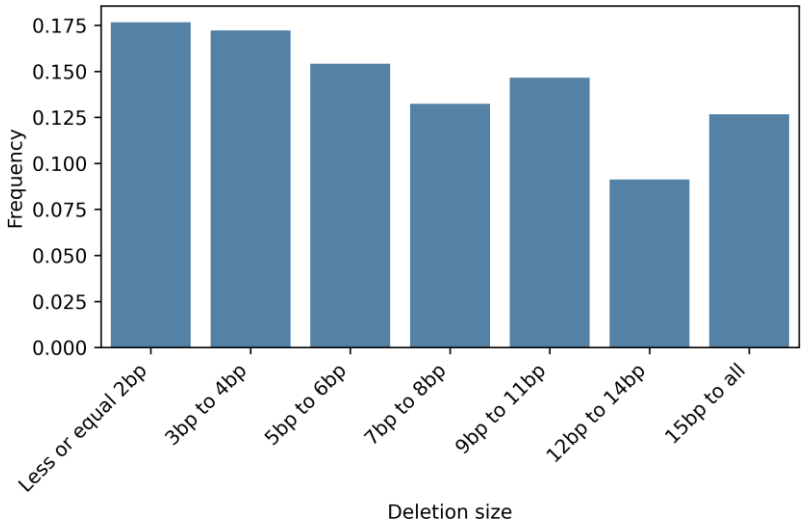

Figura 26. Freqüències en el conjunt de dades simulades dels grups de productes d'edició definits. És important recalcar que les dades analitzen totes les dades simulades, i que la freqüència de cada producte canvia segons la regió editada.

Per tant, les dades inicials en aquesta fase consisteixen en la freqüència observada de cada un dels 8 grups de productes per a cada regió. L'objectiu del model seria recrear exactament les freqüències observades, per poder realitzar prediccions de la freqüència de cada producte segons la seqüència de la regió. Per fer-ho, seria necessari utilitzar un classificador de múltiples etiquetes amb *soft labels*, ja que els resultats no són mútuament excloents ni la seva presència és binària (present/no present), sinó que cada un hi és present amb certa freqüència. Tot i això, *scikit learn* no disposa de classificadors que es puguin adaptar fàcilment a aquesta situació, així que s'ha optat per enfocar el problema en dues aproximacions diferents.

La primera aproximació es basa en predir el resultat majoritari de l'edició genètica segons cada seqüència. En aquesta, el problema es simplifica a una tasca de classificació multi-classe, ja que tan sols s'ha de predir quina de les classes és més probable d'obtenir. En aquest cas, les classes són mútuament excloents perquè tan sols es prediu la classe majoritària. Per tant, el problema es torna un problema de predicció d'una sola etiqueta o *label* d'entre un conjunt de 8 classes diferents.

La segona aproximació es basa en predir quins resultats d'edició són relativament abundants segons cada seqüència. En aquesta, el problema es simplifica utilitzant múltiples *labels* que poden prendre un valor binari de categoria, present o absent. Així doncs, com que els diferents productes són etiquetes diferents, cadascuna de les quals pot ser-hi o no, no són mútuament excloents entre elles. Per tant, es torna un problema de classificació binària de la presència o absència de 8 etiquetes diferents.

Entendre les diferències de cada una de les aproximacions és important. Mentre que en la primera aproximació hi ha 1 etiqueta amb 8 classes diferents, en la segona aproximació hi ha 8 etiquetes que poden prendre 2 classes diferents. Per tant, el primer és un problema multi-classe i el segon un problema multi-*label* (Taula 2). La informació de cada predicció també és diferent. La primera aproximació permet predir quin resultat és majoritari, però no es sap si hi ha altres resultats que també són molt abundants o el majoritari ho és amb diferència. La segona aproximació permet predir quins resultats es podria esperar observar amb certa abundància, però no permet identificar si algun d'ells és prevalent. Per tant, cada aproximació aborda el problema d'una forma diferent i complementària. Combinant les dues aproximacions es podria obtenir informació sobre els resultats més abundants i quin d'ells és majoritari. Aquesta informació s'apropa molt a la quantitat d'informació que es voldria aconseguir amb el model plantejat inicialment, que quantificar la freqüència de cada resultat d'edició.

Taula 2. Comparació de les aproximacions per predir els resultats d'edició genètica.

	Etiquetes a predir	Possibles valors de l'etiqueta
Multi-classe	1	>1 (8)
Multi-*label*	>1 (8)	2 (0 o 1)

Per implementar la primera aproximació, s'ha obtingut el resultat majoritari de cada edició a partir del producte més freqüent. Així, les 8 etiquetes inicials s'han reduït a una sola etiqueta amb 8 classes diferents i mútuament exclusives (*Figura 27*). Per identificar les classes majoritàries, s'entrenen 8 classificadors binaris *one-vs.-rest* que indiquen la probabilitat d'obtenir cada una de les classes en front la resta. Per això, és important que cada classe estigui representada en una freqüència similar, tal i com s'ha assegurat prèviament. Aquestes prediccions es simplifiquen a una sola *label* considerant aquella que té més probabilitat de ser-hi en front la resta. Cal mencionar que la probabilitat calculada que cada classe hi sigui en front la resta no correspon a la freqüència que s'esperaria obtenir d'aquella classe. El motiu és que el model s'entrena proporcionant informació només sobre la classe majoritària, i per tant, només es pot aplicar per aquesta finalitat. Seria erroni considerar que les probabilitats de cada classe podrien correspondre a les seves freqüències respectives, ja que aquestes dades no es proporcionen per l'entrenament.

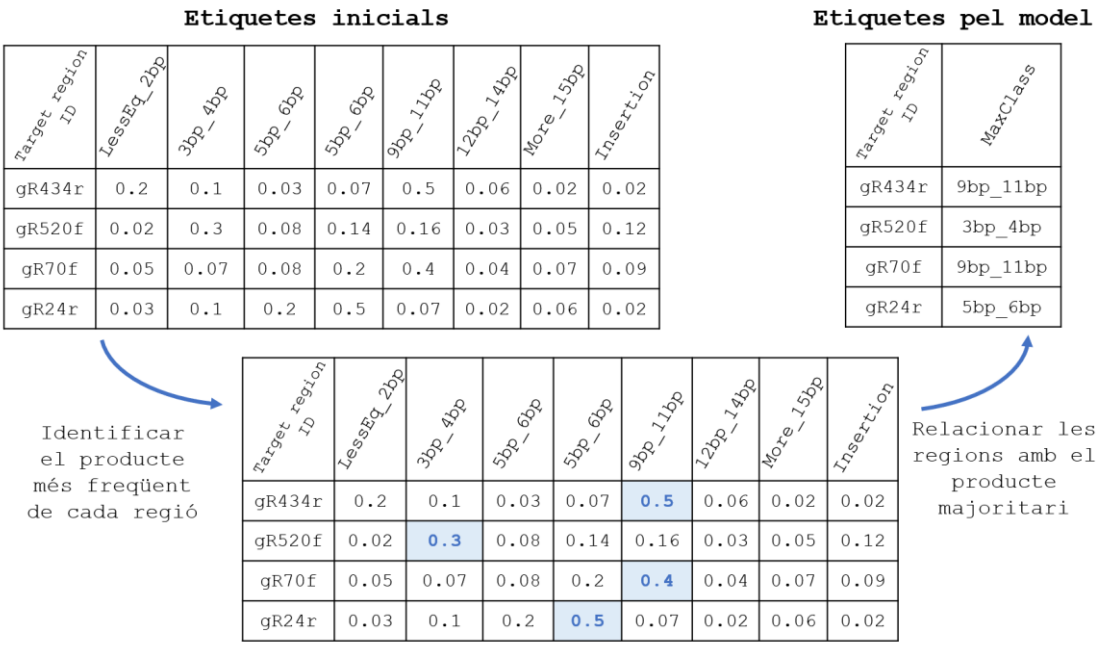

Figura 27. Preparació de les dades pel model de predicció del producte majoritari d'edició. A partir de les freqüències dels resultats observades en la simulació, s'identifica el producte amb una freqüència més alta. El model s'entrena per predir una sola etiqueta amb la classe més freqüent d'entre el conjunt de 8 classes possibles.

Utilitzant aquesta aproximació, s'entrenen un model de *logistic regression* i SVM. Per valorar l'ajust del model, es realitzen prediccions en el conjunt de dades de *test*, i s'utilitza una matriu de confusió per veure si la classe majoritària predita concorda amb la real. També es podria construir una corba ROC per cada classe, però no cal entrar en tant detall. La matriu de confusió dels dos models mostra un alt nombre de prediccions incorrectes, és a dir, que estan fora la diagonal (Figura 28). El patró descrit en les matrius és similar en els dos models, indicant que més que l'entrenament del model, les prediccions incorrectes es deuen a l'enfoc del problema o als descriptors utilitzats. Es pot veure que el major nombre de prediccions correctes són les delecions de 2 nucleòtids o menys. Tot i això, el model tendeix a predir aquest tipus de classe com a majoritària quan en les dades reals hi ha altres tipus d'insercions. Així, en *logistic regression* s'arriben a classificar incorrectament fins a 45 regions en les que el producte majoritari són delecions de 3-4 nucleòtids que es prediuen com regions amb un producte majoritari de 1-2 nucleòtids de delecions.

En el cas de *logistic regression* el model està clarament esbiaixat, ja que no realitza cap predicció en que la classe majoritària fos una inserció o deleció major de 12 nucleòtids. Segurament, les petites diferències entre les freqüències relatives de cada producte d'edició són responsables que el model tingui poca capacitat per predir-les. L'*accuracy* del model *logistic regression* és del 0.34, i per l'SVM és de 0.27, fet que indica que la majoria de prediccions són incorrectes. Per tant, caldria revaluar aquesta aproximació tal i com es descriu més endavant.

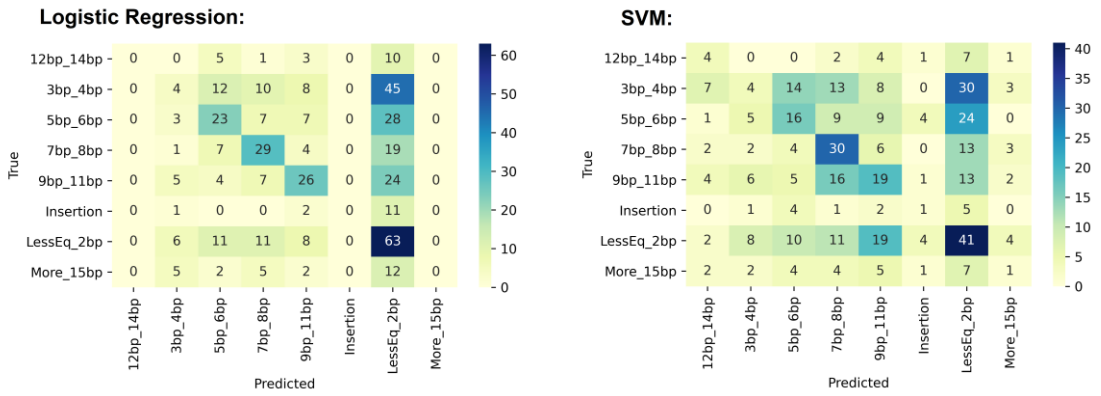

Figura 28. Matriu de confusió dels models per predir el producte d'edició genètica majoritari.

Per implementar la segona aproximació, s'estableix un llindar de 0.13 en la freqüència observada de cada producte d'edició en les dades simulades. Si la freqüència és superior a 0.13 es considera que el producte és abundant, mentre que si és inferior es considera el producte no abundant. D'aquesta forma, es converteixen els *soft labels* que indiquen la freqüència en *hard labels* binaris que indiquen si el producte és abundant entre els resultats o no (Figura 29). El valor de 0.13 s'ha triat arbitràriament, però tenint en compte que el llindar permeti considerar una diversitat de classes abundants representativa de la població.

El classificador *Random Forest* s'entrena per predir l'abundància o no de cada una de les 8 classes. Així, les prediccions finals corresponen a la probabilitat que cada tipus de producte sigui abundant, o no, segons la regió d'interès. Altre cop, és important recalcar que aquestes probabilitats no corresponen a la freqüència que s'observaria de cada resultat, com en el model que es plantejava entrenar inicialment. Per obtenir-ho, caldria utilitzar les freqüències de cada classe per l'entrenament del model, mentre que en aquesta aproximació s'utilitza un valor binari.

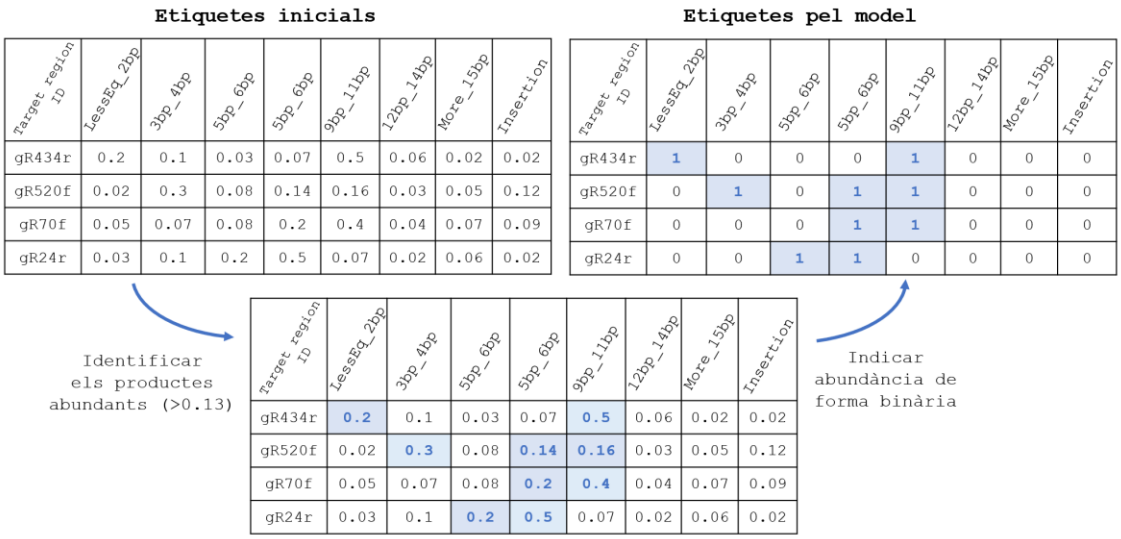

Figura 29. Preparació de les dades pel model de predicció dels productes més abundants. S'identifiquen les classes amb una freqüència major al 0.13 en les dades de simulació. Aquestes es consideren abundants (1) entre els productes d'edició de cada regió, i la resta es consideren no abundants (0). Així, el classificador prediu 8 etiquetes binàries.

Per valorar si aquest model té alguna capacitat de predicció, s'entrena un classificador *dummy* que realitza prediccions aleatòries de la presència o absència de cada classe tenint en compte la freqüència observada en el conjunt de dades simulades. Les prediccions d'aquest model no tenen en compte els descriptors, així que al comparar el *Random Forest* amb el classificador *dummy* s'hauria de veure si les prediccions de *Random Forest* són més acurades. Si fos així, es confirmaria que el classificador *Random Forest* realitza prediccions a partir de la seqüència i, per tant, estaria aprenent dels descriptors.

La comparació dels dos models es fa realitzant prediccions del conjunt de dades de *test*, que no s'han utilitzat en l'entrenament del model. Es valora si aquestes prediccions són correctes mesurant el grau en el que prediuen els valors de cada tipus de resultat per cada seqüència. Per exemple, les dades de *test* podrien indicar que per certa regió d'interès els productes d'edició abundants són "Insercions" i "Delecions de menys de 2 nucleòtids". Si el model prediu que els productes d'edició abundants per aquesta regió són "Insercions" i "Delecions de 5 a 6 nucleòtids", es considera que el model ha encertat el valor de 6 de les 8 classes a predir. El motiu és que ha predit incorrectament que les "Delecions de menys de 2 nucleòtids" no són abundants, quan en realitat sí que ho són, i que les "Delecions de 5 a 6 nucleòtids" són abundants quan en realitat no ho són.

Així, es valora el nombre d'encerts de cada model per cada una de les regions en el conjunt de *test*. Al representar la distribució del nombre d'encerts al llarg de les regions del conjunt *test*, s'observa que el model *Random Forest* tendeix a encertar un nombre major de resultats que el model *dummy* (Figura 30). Això indica que les prediccions de *Random Forest* són més semblants a les reals que les del classificador *dummy*. Per tant, es conclou que el classificador *Random Forest* extreu algunes característiques dels descriptors de la seqüència de les regions i els utilitza per fer prediccions acurades sobre l'abundància dels productes d'edició. Tot i això, tan sols una petita part de les regions tenen prediccions acurades de totes les classes, i el més comú és que el classificador encerti correctament només l'abundància o absència de 6 dels 8 productes possibles. Per tant, es pot considerar que el model *Random Forest* relaciona la seqüència i el producte i això li permet realitzar prediccions relativament acurades. Tot i això, entre la simplificació dels grups de productes i la exactitud insuficient de les prediccions, es considera que l'aproximació no permet assolir els objectius plantejats inicialment.

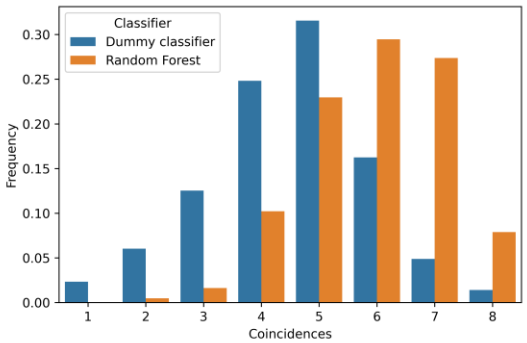

Figura 30. Distribució del nombre de coincidències entre les dades reals i dels models. Es mostra les coincidències segons si les prediccions es fan amb el model Random Forest o el classificador dummy.

Per permetre la precisió de la predicció dels resultats que seria necessària per acostar les tècniques d'edició com CRISPR-Cas9 a la clínica, es poden plantejar dues millores al model de predicció. La primera és utilitzar un model que permeti tenir en compte la freqüència de cada resultat d'edició. Aquest milloraria les aproximacions utilitzades aquí perquè s'ha vist que cada una d'elles, al simplificar el procés, impedeix establir una relació clara entre producte d'edició i seqüència editada. Un model que es podria implementar per treballar amb aquestes *soft labels* seria un model de regressió, les prediccions del qual s'ajustin per obtenir freqüències relatives de cada producte.

La segona millora a introduir és afegir descriptors de micro-homologia. Tal i com s'ha comentat a la secció *2.2. Estudis previs*, els models de predicció de l'eficiència només tenen compte la composició de la seqüència, però els models per predir els resultats d'edició hi afegeixen característiques de micro-homologia de les seqüències. S'ha vist que el resultat de l'edició està en relacionat en gran part pel procés de micro-homologia. Per tant, no haver inclòs descriptors de micro-homologia en la predicció dels productes d'edició podria ser el principal factor que explica el baix poder predictiu dels models entrenats. Les característiques de micro-homologia són relativament senzilles de calcular, així que s'està treballant per implementar un model que les tingui en compte amb l'objectiu de valorar si són essencials per predir els resultats d'edició.

5. Conclusions

En aquest treball s'ha desenvolupat completament el procés per entrenar un model computacional per predir els resultats d'edició genètica a partir de dades experimentals de seqüenciació.

En la part d'anàlisi de dades de seqüenciació, es confirma que la llibreria de 1785 gRNA diferents utilitzada no conté biaix i, per tant, es pot utilitzar per editar totes les regions d'interès. A continuació, s'alineen les dades genòmiques de les regions editades s'alineen el genoma de ratolí i s'utilitzen les seves coordenades per valorar el *coverage* de seqüenciació. La falta de *coverage* no permet identificar els resultats d'edició que s'esperava observar, així que es recorre a la simulació de dades.

En la part d'entrenament de models computacionals, es simulen 5000 *reads* per cada regió d'interès combinant un model per predir l'eficiència i un altre per predir els resultats. Seguidament, es quantifica cada resultat d'edició a partir de la seqüència, tal i com es faria experimentalment. S'entrenen quatre classificadors binaris per predir l'eficiència a partir de la seqüència. El classificador kNN és el que té una exactitud menor, i dels altres classificadors s'opta per escollir *logistic regression* tot i que tan sols té un punt més de sensitivitat que SVN o *Random Forest*. Les prediccions d'aquests models segueixen la mateixa tendència que les dades simulades, així que l'exactitud dels models és molt elevada. Cal tenir en compte que es tracten de dades simulades que no contenen error experimental intrínsecament, així que quan s'apliquin a dades experimentals l'exactitud podria disminuir.

Les dues aproximacions plantejades en el model de productes d'edició genètica han permès explorar la complexitat del problema i fer palesa la necessitat d'afegir descriptors de micro-homologia al model. Així, s'està treballant en el desenvolupament d'un model complet que pugui realitzar aquestes prediccions acuradament.

Al llarg del treball hi ha hagut inconvenients previstos, com la simulació de dades, i no previstos, com la conversió de coordenades del genoma *mm10* a C3H. Aquestes desviacions han reduït l'abast del treball a un model computacional per CRISPR-Cas9, deixant de banda els *base editors*. Tot i això, la planificació inicial era encertada perquè s'ha anat seguint, amb petites desviacions, al llarg del treball. A més, ha resultat essencial per limitar les tasques en el temps tot i els imprevistos i poder completar l'entrenament de models computacionals.

En general, tot i que els resultats experimentals no permeten establir noves relacions entre seqüències i productes d'edició, es considera que l'objectiu del treball s'ha assolit. Els *scripts* desenvolupats es poden executar des de *Google Colaboratory* i es podrien adaptar fàcilment a noves dades. Gràcies a les conclusions de l'anàlisi de dades genòmiques, l'experiment s'està repetint amb una capacitat de seqüenciació major que hauria de permetre obtenir dades d'edició. Un cop aquestes estiguin disponibles, els *scripts* d'aquest treball permetran accelerar el procés d'anàlisi i entrenament dels models. A més, les aproximacions plantejades en els models gràcies a les dades simulades han permès descartar models com kNN i plantejar la millor aproximació pel model de simulació de predicció de resultats.

6. Glossari

ABE, de l'anglès Adenine Base Editor

Accuracy, exactitud dels models de predicció, és a dir, el percentatge de prediccions que concorden amb els resultats reals

AUC, de l'anglès *Area Under the Curve*, que indica l'àrea sota la curva ROC

BWA, Burrows-Wheeler Aligner

Base editors, editors de nucleòtids

C2C12, línia cel·lular de mioblastoma de ratolí immortalitzada

CBE, de l'anglès Cytosine Base Editor

CRISPR-Cas9, tècnica d'edició genètica basada en l'enzim Cas9 del sistema CRISPR, de l'anglès Clustered Regularly Interspaced Short Palindromic Repeats

Coverage, profunditat de seqüenciació, nombre de vegades que una posició del genoma apareix en les lectures de seqüenciació

DNA, de l'anglès desoxiribonucleic acid

DSB, de l'anglès double stranded break

Dummy classifier, model de classificació que realitza prediccions aleatòries

Electroporació, procés d'introducció de DNA dins una cèl·lula a través de pols elèctrics

Endonucleasa, enzim que talla DNA

Features, descriptors utilitzats per l'entrenament de models de predicció

Flow cytometry, citometria de flux, per comptabilitzar la fluorescència de cèl·lules individuals

Hard labels, etiquetes de dades assignades a classes en que formar-ne part és binari, és a dir, o bé l'element pertany a la classe o bé no hi pertany.

HEK293F, de l'anglès Human Embryonic Kidney 293 cells

IGV, Integrative Genomic Viewer

Labels, etiquetes, característiques que es volen predir amb el model de predicció

Reads, lectures de seqüenciació, són les seqüències obtingudes en l'anàlisi de seqüenciació de nova generació

RNA, de l'anglès ribonucleic acid

ROC, de l'anglès receiver operating characteristic curve

Soft labels, etiquetes de dades associades a una probabilitat, que permeten que un element pertanyi a més d'una classe

7. Bibliografia

1. Landrum, M. J. *et al.* ClinVar: public archive of interpretations of clinically relevant variants. *Nucleic Acids Res* **44**, D862–D868 (2016).
2. Shen, M. W. *et al.* Predictable and precise template-free CRISPR editing of pathogenic variants. *Nature* **563**, 646–651 (2018).
3. Chen, W. *et al.* Massively parallel profiling and predictive modeling of the outcomes of CRISPR/Cas9-mediated double-strand break repair. *Nucleic Acids Research* **47**, 7989–8003 (2019).
4. Doench, J. G. *et al.* Rational design of highly active sgRNAs for CRISPR-Cas9–mediated gene inactivation. *Nat Biotechnol* **32**, 1262–1267 (2014).
5. Ran, F. A. *et al.* Genome engineering using the CRISPR-Cas9 system. *Nature Protocols* **8**, 2281–2308 (2013).
6. National Academies of Sciences. *Human Genome Editing: Science, Ethics, and Governance*. (National Academies Press (US), 2017).
7. Rodriguez-Rodriguez, D., Ramirez-Solis, R., Garza-Elizondo, M., Garza-Rodríguez, M. & Barrera-Saldaia, H. Genome editing: A perspective on the application of CRISPR/Cas9 to study human diseases (Review). *Int J Mol Med* (2019) doi:10.3892/ijmm.2019.4112.
8. Li, H. *et al.* Applications of genome editing technology in the targeted therapy of human diseases: mechanisms, advances and prospects. *Sig Transduct Target Ther* **5**, 1 (2020).
9. Anzalone, A. V., Koblan, L. W. & Liu, D. R. Genome editing with CRISPR–Cas nucleases, base editors, transposases and prime editors. *Nat Biotechnol* (2020) doi:10.1038/s41587-020-0561-9.
10. Anzalone, A. V. *et al.* Search-and-replace genome editing without double-strand breaks or donor DNA. *Nature* **576**, 149–157 (2019).
11. Gaudelli, N. M. *et al.* Programmable base editing of A•T to G•C in genomic DNA without DNA cleavage. *Nature* **551**, 464–471 (2017).
12. Gaudelli, N. M. *et al.* Directed evolution of adenine base editors with increased activity and therapeutic application. *Nat Biotechnol* (2020) doi:10.1038/s41587-020-0491-6.
13. Levy, J. M. *et al.* Cytosine and adenine base editing of the brain, liver, retina, heart and skeletal muscle of mice via adeno-associated viruses. *Nat Biomed Eng* **4**, 97–110 (2020).
14. Teboul, L., Herault, Y., Wells, S., Qasim, W. & Pavlovic, G. Variability in Genome Editing Outcomes: Challenges for Research Reproducibility and Clinical Safety. *Molecular Therapy* **28**, 1422–1431 (2020).
15. Arbab, M. *et al.* Determinants of Base Editing Outcomes from Target Library Analysis and Machine Learning. *Cell* S0092867420306322 (2020) doi:10.1016/j.cell.2020.05.037.

16. Xiang, X. *et al. Massively parallel quantification of CRISPR editing in cells by TRAP-seq enables better design of Cas9, ABE, CBE gRNAs of high efficiency and accuracy*. http://biorxiv.org/lookup/doi/10.1101/2020.05.20.103614 (2020) doi:10.1101/2020.05.20.103614.

17. Chari, R., Mali, P., Moosburner, M. & Church, G. M. Unraveling CRISPR-Cas9 genome engineering parameters via a library-on-library approach. *Nat Methods* **12**, 823–826 (2015).

18. Doench, J. G. *et al.* Optimized sgRNA design to maximize activity and minimize off-target effects of CRISPR-Cas9. *Nat Biotechnol* **34**, 184–191 (2016).