Universidad de Zaragoza

Bioinformática

Edición genética: cuestiones éticas y retos bioinformáticos

Trabajo de la asignatura

Autor:

José Manuel Sánchez Aquilué, 759267

Zaragoza, Aragón

abril, 2021

Escuela de
Ingeniería y Arquitectura
Universidad Zaragoza

1. Exordio

En este documento no se pretende profundizar en exceso, no se trata de una memoria como podría ser la de un trabajo de fin de grado. Lo que se busca es recapitular todo aquello que será expuesto en las sesiones del 5 y 6 de mayo. Asimismo, se referenciará la bibliografía que sirvió de base para elaborar este trabajo, de manera que quien esté interesado en ahondar en la materia tenga acceso a fuentes de calidad.

Como ya se anunció en la presentación, se tratará el tema de la edición genética desde una perspectiva bioética y bioinformática. Comenzaremos con un repaso de la evolución de las técnicas de edición a lo largo de la historia y se comentará su base científica, aunque sin entrar en demasiados detalles técnicos. Después se abordarán los distintos aspectos bioéticos que conciernen a esas técnicas y la legislación actual al respecto. En último lugar, trataremos el gran reto bioinformático que supone, más concretamente nos centraremos en el problema de off-target y veremos una serie de algoritmos de machine learning y de deeplearning para abordarlo.

2. Técnicas de manipulación y selección genéticas

Antes de nada, huelga decir que la manipulación genética es una disciplina de aparición reciente, aunque solo sea porque conocemos la estructura del DNA desde mediados del siglo pasado. En el cuadro 1 dejo una serie de fechas que fueron cruciales en la historia de esta materia.

Hito	**Año**
George Mendel sienta las bases de al genética	1863
Watson y Crick examinan la estructura del DNA	1953
primer DNA recombinado	1970s
desarrollo de la PCR	1980s
Proyecto Genoma Humano	2003

Cuadro 1: Hitos de la genética

En esencia existen 3 métodos de edición de DNA por nucleasas. El primero de ellos, basado en dedos de zinc, fue de los primeros en descubrirse, allá por los ochenta. Supuso una revolución ya que esta técnica permitía reconocer una secuencia concreta de DNA y alterarla. No obstante, dado el excesivo coste que supone, fue reemplazada por otras técnicas más modernas. Por otro lado tenemos los procedimientos basados en la nucleasa TALE, otra enzima de restricción como los dedos de zinc. Por innumerables razones las quimeras de TALE se pueden construir con mayor especificidad y cuentan con una mayor capacidad de focalización que ZFP. Por último tenemos el archiconocido CRISPR-CAS. No entraremos demasiado en detalle de la tecnología que hay detrás porque se ha presentado un trabajo dedicado exclusivamente a ese tema, solo se recordará lo que ya sabemos: que depende de un RNA guía, que permite, al igual que las tecnologías anteriores, detectar una parte de un gen y manipularla o anularla y que es el método preferido por excelencia ya que es significativamente más barato y eficiente de producir (Cuadro 2).

Nombre	Tiempo	Coste
ZFPs/ZFNs	-	30.000€
TALEs/TALENs	3-4 meses	10.000€
CRISPR-CAS	2-3 semanaas	20-30€

Cuadro 2: Métodos de edición por nucleasas

La selección genética se da de manera natural, es el mecanismo que impulsa la evolución darwiniana. Históricamente se ha practicado la selección de manera artificial con animales y plantas. Actualmente la selección de embriones se da en tratamientos de fecundación in vitro. Podemos

distinguir entre PGD (Preimplantation Genetic Diagnosis), diagnosticar enfermedades genéticas, y PGS (Preimplantation Genetic Screening), detectar alteraciones cromosómicas. Para analizar el DNA en PGD se emplean PCRs y en PGS, FISH. A continuación veremos porqué no siempre se puede realizar un PGD y qué debates éticos trae consigo.

3. Cuestiones bioéticas

Como comprenderán no siempre se puede realizar una PGD para satisfacer el capricho de unos padres de tener niños a la carta, eso es eugenesia. Por esa razón en técnicas de reproducción asistida el donante es anónimo o se trata de la pareja de la madre. Los embriones sólo pueden ser rechazados si se detecta una enfermedad genética rara de aparición temprana, incurable y potencialmente mortal. Aquellos embriones rechazados se donan a otras parejas o a la investigación. La investigación con embriones de seres humanos también está muy condicionada. Por otro lado, en España no se permite la gestación subrogada.

Naturalmente, en el momento en que alguien considera que los embriones, al tener DNA propio, son individuos en sí mismos, estos temas no están exentos de polémicas. En su opinión se estaría utilizando a "seres humanos" como medios, en vez de como fines en sí mismos, en detrimento de un principio fundamental de la ética y la dignidad humana. Además de que no estamos ante una técnica perfecta, con lo cual siempre habrá un porcentaje de embriones que mueran en el proceso. Asimismo, si conseguimos desarrollar métodos que predigan la posibilidad de enfermedad a futuro, estaríamos poniendo en riesgo el derecho a la intimidad de las personas y generando cierto estigma. A modo de ejemplo, habría aseguradoras, que si no se lo impiden, podrían negarse a dar cobertura a las personas con mayor riesgo.

Dadas esas circunstancias, aparenta ser más sensato editar, con las técnicas mencionadas anteriormente, el genoma de los embriones enfermos en vez de descartarlos.

Antes de nada comentar que la razón principal por la que no se realizan este tipo de prácticas comúnmente en seres humanos es porque causan daños colaterales. Se ha conseguido inactivar genes y cromosomas enteros, si bien, todavía estamos muy lejos de elegir las características a nuestro antojo. Como ya sabrán, CRISPR/Cas9 identifica una secuencia de DNA y la modifica o anula. Lo que ocurre es que una misma secuencia puede aparecer en otra parte del gen u otro gen, dando lugar a mutaciones inesperadas. Es lo que se conoce como el problema off-target. Una máxima fundamental de la medicina es no hacer daño al paciente (Primum non nocere), pilar elemental del juramento hipocrático. Por esa razón no se debería considerar estas prácticas bajo ningún concepto en el momento actual. Ahora bien, es posible que el perfeccionamiento de esta tecnología esté más cerca de lo que nos parece, y se debería elaborar una legislación que sea tajante en los aspectos más controvertidos.

La ley española es muy clara en ese aspecto, considera una infracción muy grave la manipulación genética con fines no terapéuticos o terapéuticos no autorizados. Es importante que estos grandes avances sean utilizados como herramientas de progreso, igualdad y bienestar y no como instrumentos denigrantes para el hombre. A principios del siglo pasado, con la pérdida de poder de las religiones y el auge de corrientes nihilistas, se popularizó el darwinismo social (no tiene nada que ver con Darwin) y malas interpretaciones de la filosofía de Nietzsche. Además, a causa de los movimientos migratorios, el racismo estaba muy presente en la población de occidente. Se pensaba que la sociedad había alterado el curso de la evolución y el Estado tenía que intervenir para encarrilarlo. En consecuencia, se promulgaron leyes eugenésicas en toda Europa y se llevaron a cabo los peores crímenes de la humanidad. Con la caída de la Alemania Nazi, se comprendió que estas prácticas eran de lo más aberrantes y se fueron prohibiendo progresivamente en el resto de Europa. La historia nos ha aleccionado para que la manipulación genética con fines terapéuticos no se convierta en el caballo de Troya del diseño de seres humanos.

Es aterrador pensar en la brecha social que se formaría si se diera vía libre a la edición génica, ya que esta tecnología solo sería asequible para un pequeño porcentaje de la población del primer mundo. Por mucho que se quisiera hacer accesible a todo el mundo, no todos los países tienen tan asentado el estado del bienestar como nosotros y tampoco se puede sufragar con dinero público una intervención sin fines terapéuticos. No solo hablaríamos de desigualdad social, sino que las mujeres se verían bastante afectadas al abrirse nuevos modelos de negocio de explotación reproductiva. Otro aspecto por el que debe preocuparse la ética es el estigma que se puede crear hacia las personas a las que se les haya realizado un proceso de manipulación. Finalmente, se debe aclarar que no todo depende de la genética, los defensores de esta práctica pueden llegar a pensar que se podrían mejorar aquellos rasgos de su personalidad e inteligencia con los que no están conformes. No obstante, esas características son el resultado de la interacción de un conjunto de genes y el ambiente.

4. Historia de He Jiankui

Por último, antes de pasar a la cuestión bioinformática, me gustaría mencionar un caso práctico en el que se llevó a cabo una manipulación mediante CRISPR/Cas9 sin tener en cuenta los principios éticos ni valorar los riesgos que podría acarrear. Se trata del científico He Jiankui, quien intentó ayudar a un padre con VIH dándole inmunidad a sus hijas. Así que editó el gen CCR5 y las consecuencias fueron nefastas por diversas razones. En primer lugar, una de las dos chicas no adquirió inmunidad al SIDA y se piensa que podrían ser quiméricas, es decir, unas células serían inmunes y otras no. Además es probable que CRISPS afectará a otros genes, responsables de curar enfermedades autoinmunes. Por no hablar de que CCR5 tiene otras funciones como proteger al linfocito del virus de la gripe. Por tanto, su supresión conlleva un aumento de la probabilidad de contraer gripes más graves y otro tipo de infecciones como el virus del Nilo. Como ya imaginarán, He Jiankui fue condenado a cárcel por realizar estos experimentos.

5. Técnicas de ML y DL para resolver el problema off-target

Durante el resto de la sesión se expondrán distintas soluciones para el problema de off-target a través de métodos de aprendizaje automático. Recordemos que los off-target son mutaciones genéticas no deseadas que se producen en la edición, cuando la secuencia que queremos editar coincide con otra parte del gen (o de otro gen). Nuestro objetivo será encontrar las secuencias de ARN guía que minimicen los daños colaterales.

5.1. Experimento 1: técnicas clásicas de aprendizaje automático

El primer experimento que analizaremos viene de la mano de Shixiong Zhan y sus compañeros, quienes trataron de resolver este problema con técnicas clásicas de aprendizaje automático.

En minería de datos es de vital importancia conocer el tipo de datos con los que estamos trabajando. En esta ocasión contamos con un conjunto de datos de 25332 secuencias, de las cuales 152 provocan off-target. En lo que respecta a predictores tenemos de tres clases: los métodos de puntuación de off-target, la conservación evolutiva y las anotaciones de cromatina. Los primeros se calculan en base a la posición, identidades y desemparejamientos entre la secuencia de DNA y el RNA guía. El resto son características cromáticas, que se ha demostrado que intervienen en la eficiencia de CRISPR/Cas9. No vamos a detenernos en averiguar como se calcula cada puntuación, pero en el paper del autor, que se encuentra referenciado al final del documento, están desarrollados los cálculos.

Puntuación de off-target
Puntuación CFD
Puntuación CCTop
Puntuación Cropit
MIT de la web
Puntuación MIT
Conservación evolutiva
Puntuación PhyloP
Puntuación PhastCons
Anotaciones de cromatina
ChromHMM
Segway

Cuadro 3: Predictores de los modelos de Shixiong Zhan

Los métodos de clasificación (Cuadro 5) utilizados están implementados en python y pertenecen a la biblioteca scikit learn. Con el fin de prevenir sobre-ajuste, se realizará una validación cruzada de los datos con 5 iteraciones. Como criterios de evaluación tendremos en cuenta el área bajo la curva ROC (AUC_{ROC}) y la curva de precisión-exhaustividad (AUC_{PRC}).

De este experimento podemos sacar en claro las siguientes conclusiones: en primer lugar, como parecía apuntar, funciona mejor la combinación de puntuaciones que atender a un único criterio (Cuadro 4). Por otro lado, añadir el predictor PhyloP supone una mejora en la eficiencia, mientras que el PhastCons y las anotaciones cromáticas no afectan al incremento del rendimiento. Por último, en base a a la comparación entre clasificadores (Cuadro 5), podemos inferir que AdanBoost es la mejor manera de resolver este problema, en tanto que SVM es la peor.

Puntuación	AUC_{ROC}	AUC_{PRC}
Todos juntos	0.938	0.300
Puntuación CFD	0.915	0.139
Puntuación CCTop	0.765	0.108
Puntuación Cropit	0.806	0.113
MIT de la web	0.732	0.168
Puntuación MIT	0.872	0.192

Cuadro 4: Comparación entre los cinco métodos por separado y todos juntos.

Clasificador	AUC_{ROC}	AUC_{PRC}
AdaBoost	0.9383	0.2998
Random Forest	0.8504	0.2236
Multilayer Perceptron (MLP)	0.9297	0.2431
Support Vector Machine (SVM)	0.6785	0.2038
Arboles de decisión	0.7981	0.2313

Cuadro 5: Comparación entre clasificadores.

5.2. Experimento 2: redes neuronales prealimentadas y convolucionales

A continuación sintetizaremos los experimentos de Jiecong Lin y Ka-Chun Wong con redes neuronales.

Comentar que, a diferencia del proyecto anterior, ellos no trabajan con una combinación de puntuaciones, sino que le pasan a la red neuronal la secuencia de DNA y RNA guía. Representan esas cadenas en matrices de 23x4, siendo 23 la longitud y 4 la codificación de los nucleótidos en formato one-hot. Al tratarse de dos secuencias, se aplica la operación de or (Figura 1). En cuanto al

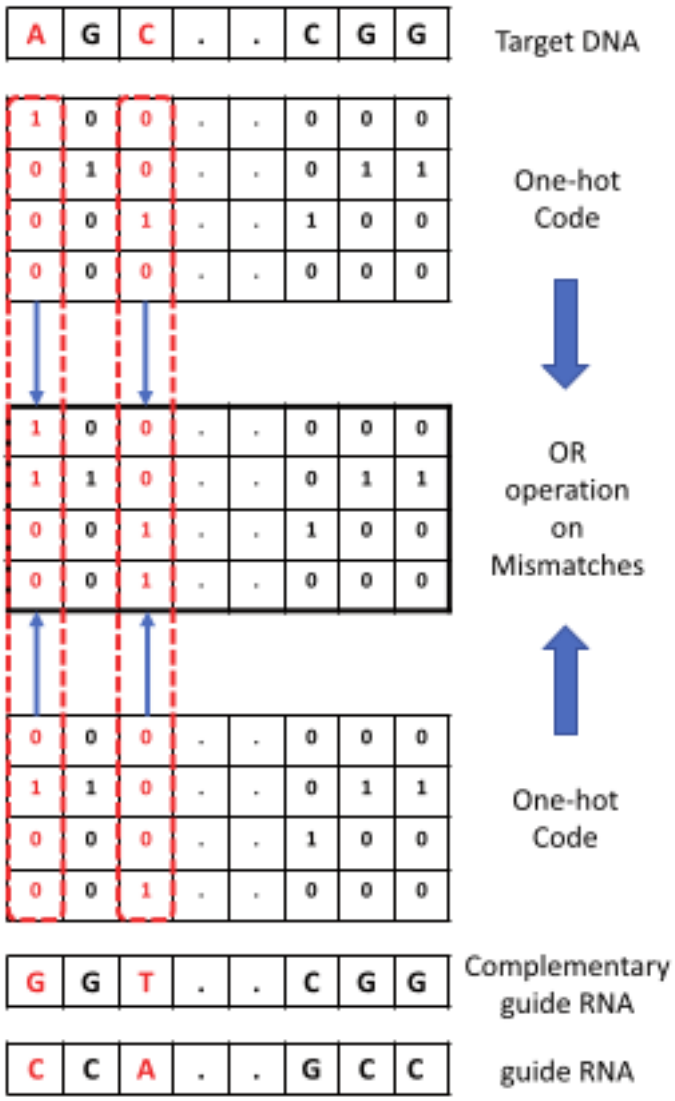

Figura 1: Codificación del par de secuencias sgRNA-DNA

conjunto de datos, hay 26034 secuencias, de las cuales 143 son offtarget, y se aplican, nuevamente, cinco iteraciones de validación cruzada.

En el repositorio podemos encontrar modelos de redes neuronales prealimentadas, es decir sin ciclos, y convolucionales. Además se comparó la eficacia de estas con modelos clásicos de aprendizaje automático y con las puntuaciones, que son las mismas que en el subapartado anterior. Evidentemente se pusieron a prueba distintas variantes de cada modelo. Entre las redes neuronales prealimentadas podemos encontrar variaciones en el número de capas, y por ende neuronas por capa. A su vez, contamos con una red neuronal convolucional estándar (Figura 2) y versiones: sin normalizar los lotes, sin dropout, sin max-pooling o cambiando el tamaño de la ventana de max-pooling. Para quien no lo sepa, la normalización es un proceso en el que se establece una escala común entre datos. Drop-out es una estrategia para desactivar neuronas de la red con el fin de evitar sobre-ajuste. Por último, max-pooling es utilizar el valor máximo de una ventana como resumen de sus elementos.

El criterio a seguir es el área de la curva ROC. Como se puede apreciar en la comparativa de modelos (Cuadro 7), el número idóneo de capas para red neuronal es 3. El AUC de ambos tipos de

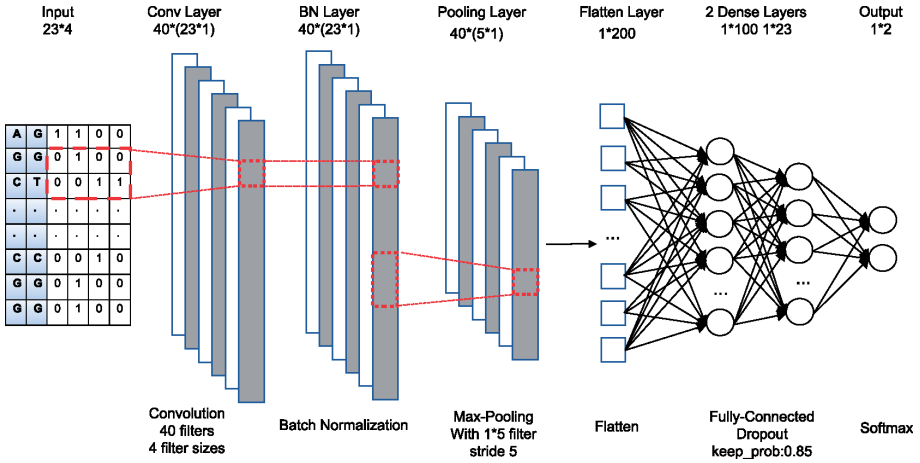

Figura 2: Estructura de la red neuronal convolucional

modelos es prácticamente el mismo. Tal y como se podría imaginar, la red neuronal convolucional funciona mejor con batch normalization, drop-out y max-pooling. Sin el dropout apenas se aprecia la diferencia. Ahora bien, los otros dos elementos son esenciales y el tamaño adecuado de la ventana es de cinco.

Modelo	AUC
FNN_2layer	0.842
FNN_3layer	**0.970**
FNN_4layer	0.954
CNN_std	**0.972**
CNN_nbn	0.954
CNN_nd	0.969
CNN_np	0.899
CNN_pool_win3	0.903
CNN_pool_win7	0.967

Cuadro 6: Comparación entre distintas variantes de redes neuronales.

Estas puntuaciones tan altas parecen indicarnos que las redes neuronales son la mejor técnica para resolver el problema de los off-targert. De hecho, si entramos a compararlas con modelos tradicionales de la estadística, percibimos una mejora sustancial. Ni que decir tiene que no hay parangón entre el calculo de un tipo de puntuación y un modelo de aprendizaje profundo. Tal vez el único tipo de puntuación que podría considerarse es la determinación de frecuencias de corte (CFD).

Modelo	AUC
FNN_3layer	**0.970**
CNN_std	**0.972**
Regresión Logistica	0.931
Gradient boosting	0.914

Modelo	AUC
FNN_3layer	**0.970**
CNN_std	**0.972**
puntuación CROP-IT	0.807
puntuación CFD	0.912
puntuación MIT	0.865
puntuación CCTop	0.776
puntuación MIT-web	0.728

Cuadro 7: Comparación entre redes neuronales, modelos clásicos y puntuaciones.

5.3. Experimento 3: más redes neuronales convolucionales

En la subsección anterior hemos comprobado que los modelos de deep-learning superan con creces a los clásicos, así que vamos a seguir profundizando en estos métodos con la esperanza de mejorar el rendimiento. La estructura que veremos acto seguido no es para nada sencilla y quizá no sea fácil de comprender a priori. Fue elaborada por Qiaoyue Liu, Xiang Cheng, Gan Liu, Bohao Li y Xiunqin Liu. En esta ocasión cada par de secuencias sgRNA-DNA se codifica como un vector de 23 índices de palabras. Ya que hay 4 nucleótidos, existen 16 combinaciones de parejas de nucleótidos, 16 palabras.

Una vez que tenemos calculados los indices de palabras, se las pasamos a un modelo GloVe que nos devolverá vectores de representación de palabras. Siendo que hay 16 palabras y el Glove está configurado para devolver vectores de dimensión 100, tras pasar el proceso de embedding tenemos una matriz de 16x100 que será el input de una Long short-term memory (LSTM) bidireccional, un tipo de red neuronal recurrente. El output de esta red alimenta 5 capas convolucionales, con distinto número de kernel y tamaño de kernel. Entre capas hay batch normalization y la función de activación es una relu, salvo al final que es softmax, como en cualquier clasificador. El resultado de las capas convolucionales se «aplana» y atraviesa un drop-out de 0.3 y dos capas densas de 20 y 2 neuronas.

Al igual que en el apartado anterior, se han probado distintas versiones del modelo estándar (Cuadro 8). Contamos con un modelo sin la LSTM bidireccional, otro que revierte el orden (primero aplica la convolución y luego la LSTM). También se ha probado a desprendernos del batch normalization y el drop-out. En primer lugar, destacar que sin batch normalization tiene la misma precisión que un modelo aleatorio. La eliminación del drop-out supone una mejora de AUC y una baja de la exhaustividad. Por otra parte, deshacernos de la LSTM, o revertir el orden, sí que supone una gran perdida de exhaustividad y aumenta ligeramente el AUC.

Modelo	Exhaustividad	AUC_{ROC}	AUC_{PRC}
CnnCrispr	**0.857**	0.975	**0.679**
CnnCrispr_NoLSTM	0.611	**0.987**	0.651
CnnCrispr_Conv_LSTM	0.643	0.986	0.67
CnnCrispr_NoBatchNor	-	0.5	0.504
CnnCrispr_NoDropOut	0.810	0.985	0.625

Cuadro 8: Comparación entre modelos de redes neuronales convolucionales.

Finalmente observamos que esta red neuronal convolucional cuenta con el mayor AUC_{PRC} con diferencia frente a cualquier otro método. Curiosamente el área bajo la curva que se le otorga a la puntuación CFD es mayor que en el paper del experimento anterior, de modo que deja a la red neuronal convolucional anterior por debajo de la puntuación. Supongo que se deberá a los data sets que han utilizado en ambos experimentos, aunque llama la atención semejante diferencia. El modelo DeepCrispr[5] , implementado por otros autores, cuyo trabajo no se ha podido analizar con detenimiento en esta memoria debido a las limitaciones de extensión, cuenta con el mejor AUC_{ROC} y el segundo mejor AUC_{PRC}. Se trata esencialmente de una red neuronal convolucional profunda.

Modelo	AUC_{ROC}	AUC_{PRC}
CnnCrispr	0.975	**0.679**
CFD	0.942	0.316
MIT	0.77	0.044
CNN_std	0.947	0.208
DeepCrispr	**0.981**	0.497

Cuadro 9: Comparación entre el modelo de Qiaoyue Liu y otros métodos.

6. Peroratio

Antes de pasar al turno de preguntas, me gustaría agradecerles su atención y espero haber sido lo suficientemente claro, ya que es un tema complejo y no tengo la costumbre de exponer trabajos tan extensos. Confío además en que hayan disfrutado de esta charla y haber sido capaz de despertar cierto interés sobre las cuestiones bioéticas y bioinformaticas que aguarda la edición genética. Si les ha quedado cualquier duda sobre aquello que he expuesto o les interesa que profundice sobre algún fenómeno que haya nombrado de pasada, a continuación trataré de resolver esas cuestiones. Muchas gracias.

Referencias

[1] Ley 14/2006, de 26 de mayo 2003, sobre técnicas de reproducción humana asistida. («BOE» núm. 126).

[2] Ley 35/1988, de 22 de noviembre 1988, sobre técnicas de reproducción asistida. («BOE» núm. 282).

[3] Ley 45/2003, de 21 de noviembre 2003, por la que se modifica la ley 35/1988 sobre técnicas de reproducción asistida. («BOE» núm. 280).

[4] Ley orgánica 10/1995, de 23 de noviembre 1995, del código penal. («BOE» núm. 281).

[5] Yan J Chen M Hong N Xue D Zhou C Zhu C Chen K Duan B Chuai G, Ma H. Deepcrispr : optimized crispr guide rna design by deep learning. *Genome Biol.*

[6] Adam P Cribbs and Sumeth M W Perera. Science and bioethics of crispr-cas9 gene editing: An analysis towards separating facts and fiction. *The Yale journal of biology and medicine*, 90(4):625–634, 2017.

[7] Rothschild J. Ethical considerations of gene editing and genetic selection. *Journal of general and family medicine*, 21:37–47, 2020.

[8] Juan-Ramón Lacadena. Edición genómica: ciencia y ética. *Revista Iberoamericana de Bioética*, (3):1–16, jun. 2017.

[9] Jiecong Lin and Ka-Chun Wong. Off-target predictions in CRISPR-Cas9 gene editing using deep learning. *Bioinformatics*, 34(17):i656–i663, 09 2018.

[10] Cheng X. Liu G. et al Liu, Q. Deep learning improves the ability of sgrna off-target propensity prediction. *BMC Bioinformatics*, 21(51), 2020.

[11] Alexis Molina and Laura Serra. Genetic scissors for a dystopian future. *ESCI news*, 2021.

[12] Shixiong Zhang, Xiangtao Li, Qiuzhen Lin, and Ka-Chun Wong. Synergizing CRISPR/Cas9 off-target predictions for ensemble insights and practical applications. *Bioinformatics*, 35(7):1108–1115, 08 2018.