Preguntes

Bon dia Marc,

Amb aquest missatge iniciem la defensa del teu treball final de master.

El tribunal està format per la Yolanda Guillén, en Ferran Prados (en condició de coordinador de l'àrea de l'assignatura) i per mi mateix (en condició de tutor).

Durant la defensa, els i les integrants del tribunal faran diferents rondes de preguntes que hauràs de contestar en un plaç de 48 hores.

Tenint en compte que els altres membres del tribunal han tingut menys temps per repassar la memòria, jo mateix obriré el torn de preguntes.

Abans de començar amb les preguntes, m'agradaria donar-te l'enhorabona per el treball realitzat. Cal remarcar la dificultat de fer un treball que cobreix una pipeline d'inici a fi: començant per la seqüenciació de les regions editades per la llibreria, realitzant passos intermitjos com l'anàlisi del coverage o la creació d'un data set, i acabant en la implementació i validació d'un model de machine learning.

– En el treball has hagut de generar un dataset sintètic per a poder entrenar els classificadors, ja que el coverage de la seqüenciació ha esta baix. Creus que els resultats del treball haguessin estat molt diferent si haguessis pogut utilitzar un dataset real? Què s'hauria de canviar en el disseny de l'experiment per a millorar el coverage?

– Per a representar les zones d'interès de les seqüències, has optat per fer servir one-hot encoding. ¿Podies haver representat les regions d'alguna altra manera? ¿Com creus que hauria afectat els resultats del model?

– En la teva presentació i en la memòria parles d'altres models computacionals que intenten predir els resultats d'edicions CRISPR. Quines son les semblances i diferències amb el teu model? Comparativament, quin rendiment presenten?

Atentament,

Albert Pla

Respostes

Bones Albert,

Moltes gràcies per l'enhorabona en el treball, i gràcies a tu per supervisar-lo i ajudar-me en l'enfocament dels models! A continuació, responc les teves preguntes.

– En el treball has hagut de generar un dataset sintètic per a poder entrenar els classificadors, ja que el coverage de la seqüenciació ha esta baix. Creus que els resultats del treball haguessin estat molt diferent si haguessis pogut utilitzar un dataset real? Què s'hauria de canviar en el disseny de l'experiment per a millorar el coverage?

Sí, crec que els resultats haguessin canviat molt amb unes dades experimentals reals i no simulades. En primer lloc, els models utilitzats per la simulació no es basen en edicions al genoma sinó en seqüències sintètiques, mentre que les dades experimentals que haguéssim obtingut provindrien directament del genoma. Així, segurament, hagués estat necessari incorporar descriptors d'accessibilitat a la seqüència genòmica, com l'estat de la cromatina en la regió d'interès, per aconseguir un elevat poder de predicció. Aquests no s'han incorporat en les dades simulades, ja que els models utilitzats per fer-les ometen aquest factor.

En segon lloc, les dades haguessin tingut una qualitat diferent en cada regió d'edició. Així doncs, segurament, caldria haver realitzat un procés de filtrat per descartar aquelles regions per les quals s'ha obtingut informació insuficient. Com que el procés de simulació de dades generava 5000 *reads* per cada regió, no tenia sentit realitzar-ho amb dades simulades.

Finalment, les dades experimentals contindrien errors de seqüenciació i soroll atribuïble a l'error experimental. Així, caldria realitzar un anàlisi molt més detallat per quantificar els productes d'edició (i distingir-los d'errors de seqüenciació). A més, els models no assolirien capacitats de predicció tan altes, tal i com es menciona a la memòria. Cal tenir en compte que els descriptors utilitzats per a la simulació de dades d'eficiència i l'entrenament de models a partir d'aquestes són gairebé idèntics, fet que explica l'elevada capacitat de predicció dels algoritmes (en *validation set*, l'*accuracy* arribava a 1).

La millora del *coverage* es podria obtenir en dos punts del procés. El primer és el procés d'enriquiment, que és en el que es seleccionen específicament les regions editades del genoma i es descarta la resta. Quan més eficient és aquest procés, major part de les lectures de seqüenciació es destinen a les regions d'interès i menor part s'obté de regions genòmiques que no són d'interès. Per valorar aquest procés, es pot realitzar un anàlisi que valora l'*on-target rate*. En aquest anàlisi, es compara el nombre de *reads* que pertanyen a una regió d'interès (*on-target*) i el nombre de *reads* que no (off-target). Tot i això, la inspecció visual dels resultats de seqüenciació amb el visualitzador IGV permet veure que la major part dels *reads* es concentren en les regions d'interès. Així, el procés d'enriquiment es considera correcte, i caldria optimitzar l'altre punt del procés que es comenta a continuació.

El segon punt del procés en el que es pot millorar el *coverage* és el pas de seqüenciació. La mida del *run* de seqüenciació determina el nombre de seqüències que es poden observar. Si s'utilitzés un *run* més gran, es podria haver obtingut un major nombre de seqüències, així que es podrien haver observat *reads* que fossin de cèl·lules editades. Idealment, es podria observar totes les regions genòmiques d'interès de totes les cèl·lules editades. Tot i això, no cal assolir una mida tant elevada perquè tan sols cal obtenir suficients resultats d'edició com per poder entrenar un model capaç de predir la freqüència de cada un. Degut a la diversitat d'aquests resultats d'edició, caldria observar un mínim d'entre 100 i 1000 *reads* **editats** per regió. Com que els *reads* editats

són una minoria (s'estima que siguin un 0.16% dels *reads* totals per la regió), caldria incrementar el *coverage* a través del procés de seqüenciació fins a uns 100,000-1,000,000 *reads* per regió d'interès.

– Per a representar les zones d'interès de les seqüències, has optat per fer servir one-hot encoding. ¿Podies haver representat les regions d'alguna altra manera? ¿Com creus que hauria afectat els resultats del model?

Crec que en aquest cas és important utilitzar *one-hot encoding*, ja que sinó els resultats del model haguessin estat diferents. Les seqüències de DNA es solen representar en els estudis que utilitzen Machine learning a través de *one-hot encoding*. El motiu és que es considera una seqüència de nucleòtids en que cada un dels llocs pot prendre 4 categories diferents. Com que aquestes categories no estan ordenades té sentit utilitzar categories binàries (present/absent) en comptes de valors numèrics enters per representar-les.

Si s'utilitzessin valors enters per representar categories en comptes de *one-hot*, aleshores el nombre de descriptors seria menor. Per exemple, si s'utilitzés el codi (A-> 1, C-> 2, T-> 3, G-> 4) per un gRNA de 20 nucleòtids, tan sols caldrien 20 descriptors (cada un amb 4 categories) per representar-ho. En canvi, amb *one-hot* calen 4x20=80 descriptors binaris. Tot això, crec que utilitzar valors numèrics enters implica que l'algoritme consideraria l'ordre d'aquestes categories, quan en realitat aquest ordre és arbitrari. Si és així, és possible que intrínsecament el model no pogués aconseguir bona capacitat de predicció, i un canvi en l'ordre arbitrari definit canviaria les prediccions del model totalment. Així doncs, com que en aquest cas no hi ha un ordre entre els quatre nucleòtids, la millor opció per representar la seqüència és *one-hot*.

Pel que tinc entès, es podria comparar el tipus de codificació *one-hot* utilitzat amb la dels algoritmes que utilitzen text com a descriptors. Aquests, acaben codificant el text a través de vectors *one-hot* indicant la paraula que hi ha assignant-li un 1 en un vector de 0s per la resta de paraules del diccionari. En aquest cas, el diccionari consta tan sols dels quatre nucleòtids.

També m'agradaria comentar que la codificació de nucleòtids amb *one-hot* encoding no sempre es duu a terme en forma d'un vector unidimensional. En els models que utilitzen una xarxa neuronal convolucional, les seqüències es poden representar en dues dimensions, on la primera dimensió correspon a cada posició de la seqüència, i la segona al nucleòtid present. Dos exemples en són els treballs de Chuai et al 2018[1] i Xue et al 2019[2].

– En la teva presentació i en la memòria parles d'altres models computacionals que intenten predir els resultats d'edicions CRISPR. Quines son les semblances i diferències amb el teu model? Comparativament, quin rendiment presenten?

Pel que fa al model de predicció de l'eficiència, hi ha molts estudis i s'han utilitzat aproximacions molt diverses. Els més similars al meu model es comenten a la memòria. Breument, els treballs de *Doench et al 2014*[3] i *Chari et al 2015*[4] utilitzen un classificador binari similar al meu model. Tots incorporem com a descriptors els nucleòtids, dinucleòtids i contingut GC. El meu, però, hi afegeix el nombre total de cada tipus de nucleòtid (A, C, G, o T) independentment de la posició. La diferència entre aquests models són les dades experimentals. Per *Doench et al 2014*, les dades provenen d'un anàlisi fenotípic. Per *Chari et al 2015*, venen d'un anàlisi genotípic com en el nostre cas. Tot i això, *Chari et al 2015* utilitza regions d'interès sintètiques (com la resta d'estudis citats aquí), mentre que el meu hauria d'utilitzar (si hi hagués prou coverage) regions genòmiques.

Un treball molt interessant és el de *Doench et al 2016*[5], en el que millora el model previ de *Doench et al 2014*. La millora ve en part d'afegir descriptors com el recompte dels tipus de nucleòtid (que ja es té en compte en el meu model) i la posició del gRNA en el gen editat (només té sentit en el seu estudi, ja que és fenotípic i els varis gRNAs es troben en diferents posicions del mateix gen, mentre que en el nostre s'editen gens diferents). També millora el model utilitzant linear regression en comptes de logistic regression. Cal mencionar que en l'estudi previ que utilitzava logsitic regression dividia les dades en porcions desiguals (classes binàries amb el 20% més eficient i 80% no eficient), així que linear regression evitava la pèrdua d'informació al dividir-ho en classes binàries. Seria interessant explorar aquesta aproximació en el nostre treball.

Els models previs assoleixen valors de AUC propers a 0.8, que oscil·len aproximadament entre 0.75 i 0.95. El meu model assoleix una AUC major, de 0.98. Tot i això, no s'hauria de comparar els seus models amb els meus, ja que els meus utilitzen dades simulades a partir de models computacionals. Així, la bondat de l'ajust dels meus models és a les dades simulades dels seus models, mentre que pels altres models, la bondat d'ajust és amb les dades experimentals. Tal i com s'ha comentat anteriorment, les dades experimentals tenen més soroll. A més, al realitzar les edicions al genoma en el nostre cas, es podria esperar que la bondat de l'ajust fos lleugerament més baixa per factors genòmics que no es poden tenir en compte. Tot i això, seria un càlcul més realista perquè les condicions són més semblants a les que utilitzarien els usuaris del model per dissenyar experiments d'edició genètica. Per tant, segurament el meu model tindria un ajust a les dades lleugerament inferior que els valors d'AUC=0.8 dels altres.

Els models que no s'ha comentat en la memòria són aquells que utilitzen xarxes neuronals. Degut a la versatilitat d'aquestes xarxes, hi ha molts estudis que utilitzen modificacions relativament petites en l'estructura de la xarxa neuronal per incrementar la capacitat de predicció. Cal remarcar que alguns d'ells assoleixen rendiments semblants, amb AUC=0.8, mentre que d'altres assoleixen AUC=0.95. Alguns d'aquests estudis són els de *Zhang et al 2020*[6], *Chuai et al 2018*[1] i *Xue et al 2019*[2]. El fet d'afegir diverses capes de convolució es pot adaptar bé a identificar relacions complexes entre les nucleòtids de la seqüència del gRNA, fet que els hi hauria de concedir un major valor de predicció.

Pel que fa al model de predicció de la freqüència dels resultats, hi ha menys de 10 estudis però també utilitzen aproximacions diverses. En la memòria, tan sols s'ha plantejat dues aproximacions interessants per realitzar aquest tipus de prediccions. Aquestes aproximacions són noves fins on arriba el meu coneixement (no hi ha literatura publicada que conegui que en parli), així que no es poden comparar amb altres models. Tot i això, la capacitat de predicció obtinguda ha estat molt baixa, i es pot concloure que es deu a la falta de descriptors de micro-homologia. La resta d'estudis que es centren en la predicció de la freqüència dels productes utilitzen aquests descriptors, que juguen un paper essencial en la determinació dels productes d'edició. Els tres estudis més rellevants en aquesta àrea (Allen et al 2019[7], Shen et al 2018[8] i Chen et al 2019[9]) es comparen en el treball de Chen et al 2019, concloent que el model de Chen et al 2019 és el més precís amb un *mean squared error* de 0.01.

Referències utilitzades:

1. Chuai, G. *et al.* DeepCRISPR: optimized CRISPR guide RNA design by deep learning. *Genome Biol* **19**, 80 (2018).
2. Xue, L., Tang, B., Chen, W. & Luo, J. Prediction of CRISPR sgRNA Activity Using a Deep Convolutional Neural Network. *J. Chem. Inf. Model.* **59**, 615–624 (2019).
3. Doench, J. G. *et al.* Rational design of highly active sgRNAs for CRISPR-Cas9–mediated gene inactivation. *Nat Biotechnol* **32**, 1262–1267 (2014).
4. 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).
5. 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).
6. Zhang, G., Dai, Z. & Dai, X. C-RNNCrispr: Prediction of CRISPR/Cas9 sgRNA activity using convolutional and recurrent neural networks. *Computational and Structural Biotechnology Journal* **18**, 344–354 (2020).
7. Allen, F. *et al.* Predicting the mutations generated by repair of Cas9-induced double-strand breaks. *Nat Biotechnol* **37**, 64–72 (2019).
8. Shen, M. W. *et al.* Predictable and precise template-free CRISPR editing of pathogenic variants. *Nature* **563**, 646–651 (2018).
9. 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).