Preguntes

Hola Marc,

Felicitats! Aquí van les meves preguntes.

– Coneixies prèviament totes les eines que has utilitzat en el TFM? Quina ha estat la curva d'aprenentatge? Com ha estat la experiència amb Google Colab?

– Per què no has provat l'algorisme del decision tree? Hauria funcionat?

– Quin serien els passos a seguir per noves dades?

– Què et sembla el TFM del teu company Alfonso Aguado que aborda un tema similar? Com valores els seus resultats?

Una segona lectura amb calma de certes parts del text hauria millorat l'acabat global del mateix (3r paràgraf pel final de la intro) o corregit certs petits errors (per exemple al començament pàgina 12 o a la pàgina 47 «a sigut difiícil» en comptes de: ha sigut difícil) que desvirtuen el gran treball fet. I ara les 4 preguntes que faig a tots els alumnes:

– Quines assignatures i coneixements apressos durant el máster t'han servit per realitzar aquest TFM?

– Quin seria el cost d'aquest projecte si tenim en compte les hores dedicades?

– Quines són les parts que t'han costat més i menys del TFM?

– Quin serà el següent pas en la teva carrera professional?

Salutacions.

Ferran

Respostes

– Coneixies prèviament totes les eines que has utilitzat en el TFM? Quina ha estat la curva d'aprenentatge? Com ha estat la experiència amb Google Colab?

Abans del màster de la UOC no coneixia la majoria d'eines utilitzades en el TFM. Cal mencionar que he realitzat el màster en un any, així que les assignatures de Machine Learning o Anàlisis de Dades Òmiques les vaig fer durant el segon semestre, paral·lelament al TFM. Per això, el treball ha requerit molta feina d'autoaprenentatge. Durant el primer semestre, vaig realitzar unes pràctiques extra-curriculars al laboratori que em va permetre obtenir les dades experimentals, alhora que familiaritzar-me amb Linux i a comunicar-me amb el clúster Marvin (on es realitza la part d'anàlisi genòmica) a través de ssh.

Tenia coneixements bàsics de programació en R, bash, i Fortran, però estava decidit a realitzar el TFM utilitzant Python i Scikit Learn per aprendre'n i tenir l'oportunitat de fer un projecte d'inici a fi. Per això, durant el primer semestre em vaig preparar utilitzant MOOCs per adquirir coneixements bàsics de [Python](https://www.edx.org/course/introduction-to-computer-science-and-programming-7) i [computació](https://www.edx.org/course/introduction-to-computational-thinking-and-data-4#!). Com que l'assignatura de machine learning de la UOC l'havia de realitzar en paral·lel al TFM, durant el primer semestre i principis del segon vaig aprendre els fonaments dels algoritmes de machine learning utilitzant també un [MOOC](https://www.edx.org/course/machine-learning-with-python-from-linear-models-to). Aquest darrer es centra en utilitzar Python per codificar cada un dels algoritmes des de zero. En canvi, en el treball, he utilitzat la llibreria Scikit Learn. Per aprendre el funcionament de la llibreria, vaig utilitzar la documentació i un [llibre](https://www.oreilly.com/library/view/introduction-to-machine/9781449369880/).

Així, diria que gran part de la feina ha estat formar-me en aquests temes i per això la guia del tutor, Albert Plà, ha estat tant important. En general, realitzar la resta d'assignatures del màster m'ha acostumat a cercar errors en el codi i solucionar-los, que ha sigut molt útil pel TFM. Crec que el fet de preparar-me durant el primer semestre a través de llibres i MOOCs ha facilitat que pogués realitzar un treball extens en múltiples àmbits. Com que el treball és la primera vegada que posava aquests coneixements en pràctica, ha requerit una bona dosi de perseverança i prova-error-correció-aprenentatge.

L'Albert Plà, tutor del treball, va suggerir-me l'ús de Google Colab. Coneixia el concepte de les Jupyter Notebooks i el fet de fer fàcil la col·laboració em va convèncer que era una bona eina per presentar el treball. Per la part de dades de seqüenciació, crec que no són la plataforma ideal, ja que requereix treballar amb fitxers grans i realitzar anàlisis de varies hores. A més, hi ha alternatives com Galaxy que són més adequades per aquestes tasques. Tot i això, l'ús de Google Colab m'ha permès mostrar clarament el codi utilitzar per reproduir les figures que es deriven d'aquests anàlisis, i també il·lustrar els processos realitzats al servidor extern a partir de la sintaxi Markdown. En canvi, per la part d'entrenament del model crec que són molt convenients. Es podrien canviar les dades per unes altres i ràpidament adaptar els anàlisis que he realitzat a noves dades. A més, permeten intercalar text i codi, facilitant la comprensió al lector.

En general, valoro molt positivament l'experiència amb Google Colab, i espero poder-les utilitzar en un futur per mostrar de forma reproduïble les investigacions que realitzi i facilitar-ne la comprensió.

– Per què no has provat l'algorisme del decision tree? Hauria funcionat?

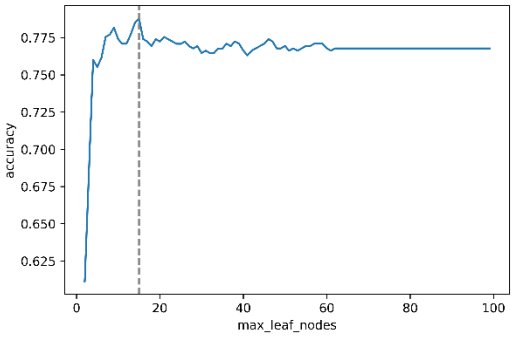
La decisió de provar un *random forest* directament sense provar un *decision tree* abans es deu a varies raons. Els *random forest* són menys sensibles a canvis en les dades, ja que la combinació de múltiples arbres compensa els canvis que poden haver-hi en un de sol. En aquest cas s'utilitzen dades simulades, però els models s'entrenen pensant en dades experimentals. Les dades experimentals tenen més soroll que les simulades, així que s'opta per escollir directament *random forest* que segurament milloraria *decision tree* en una situació amb més soroll.

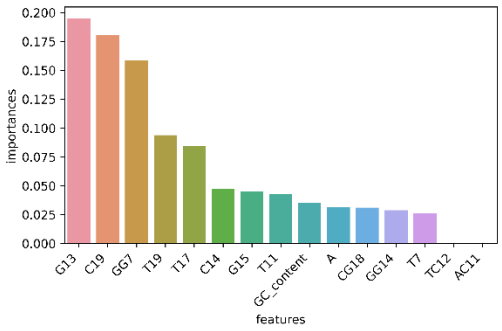
Un dels avantatges de *decision tree* és una millor interpretabilitat de les dades. Amb un sol *decision tree* es pot visualitzar tant la importància dels predictors com l'arbre que s'utilitza per les condicions. Per exemple, es podria veure quin nucleòtid en concret és responsable de determinar els resultats en major part. En canvi, en *random forest* es poden visualitzar arbres individuals però com que el resultat depèn de més de 200 arbres (en aquest treball), la descripció visual no representa el model en sí. Tot i això, si que permeten calcular la importància dels descriptors.

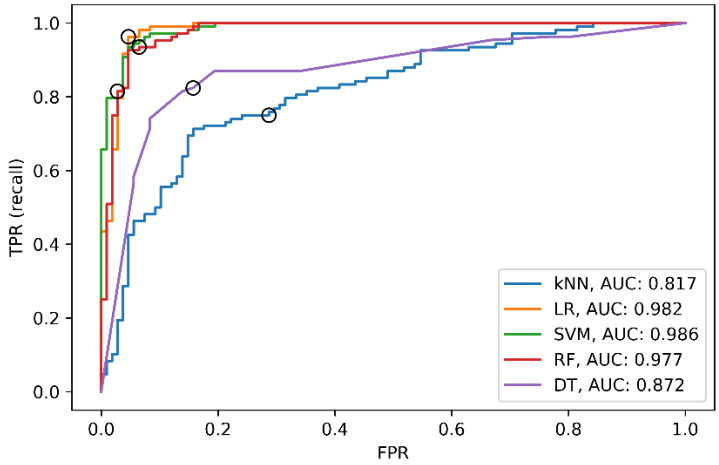
En aquest cas, s'espera que no tots els descriptors siguin importants però que les interaccions entre aquests sí que ho siguin, ja que el resultat és un efecte combinat de varis nucleòtids. Per tant, la visualització que es podria obtenir de visualitzar un sol arbre de decisions no seria gaire valuosa biològicament, ja que no són descriptors individuals el que importa. En canvi, veure el conjunt de descriptors importants sí que és d'interès, cosa que també es pot fer amb *random forest*. Per això, l'avantatge de la interpretabilitat de *decision tree* no representa un avantatge en aquest treball.

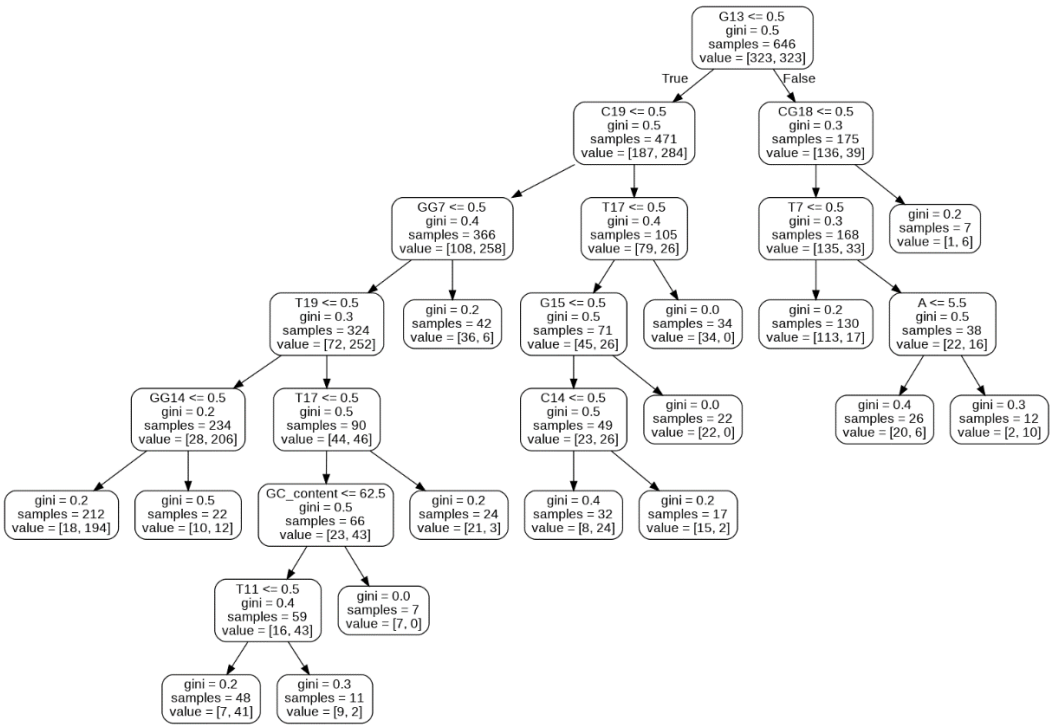
Els classificadors de *decision tree* són més ràpids d'entrenar que el *random forest*. Així doncs, si la capacitat de predicció del model és semblant a *random forest*, podria ser que *decision tree* fos preferible. Per saber si hauria funcionat, o no, el millor és provar-ho. Per tant, s'ha decidit valorar el funcionament de *decision tree* juntament amb els altres algoritmes descrits al a memòria. Com que Google Colab permet afegir nous algoritmes de forma senzilla, s'ha afegit una secció per Decision tree a la llibreta d'entrenament de models, disponible [aquí](https://github.com/marcexpositg/CRISPRed/blob/master/02.Model/2.5.2.EffModel.ipynb).

L'algoritme s'ha optimitzat provant diversos nodes, i s'ha identificat que 15 és el nombre amb més exactitud. Com que tan sols hi ha un decision tree, es pot visualitzar de forma senzilla. Aquest algoritme utilitza tan sols 13 de les features disponibles. És curiós observar similituds en les importàncies dels descriptors entre Random Forest i Decision trees. Al comparar el model amb la resta a través d'una ROC curve es pot veure que millora kNN però es queda curt per arribar a la capacitat de predicció de la resta. Cal mencionar que segurament es podrien optimitzar més paràmetres per millorar el rendiment. Les imatges de l'entrenament a partir de les que s'obtenen aquestes conclusions s'adjunten al comentari.









La conclusió que n'extrec és que el Decision Tree entrenat tan sols té en compte 13 descriptors, mentre que Random Forest, al utilitzar la combinació de varis arbres, té en compte un major nombre de descriptors. L'eficiència d'un gRNA no es pot predir a partir de la presència o absència de certs descriptors, sinó que es deu al conjunt global dels 20 nucleòtids que composen el gRNA. Per això, el model decision tree que té en compte tan sols pocs descriptors no és capaç de realitzar prediccions acurades.

– Quin serien els passos a seguir per noves dades?

Tal i com s'ha comentat a la memòria, s'està repetint l'experiment inicial procurant aconseguir una major representació de les regions d'interès per observar edicions genètiques. Una vegada es seqüenciïn les mostres, s'utilitzarà el procés de *trimming* descrit a la llibreta [1.1.SequencingDataProcessing.ipynb](https://github.com/marcexpositg/CRISPRed/blob/master/01.DescriptiveAnalysis/1.1.SequencingDataProcessing.ipynb) per eliminar els *reads* amb baixa qualitat. A continuació, es realitzarà el control de qualitat de les llibreries de gRNAs per assegurar que tots hi són representats en una proporció semblant, tal i com es descriu a [1.2.gRNALibDistribution.ipynb](https://github.com/marcexpositg/CRISPRed/blob/master/01.DescriptiveAnalysis/1.2.gRNALibDistribution.ipynb). Si els resultats són correctes, tal i com en el primer experiment que es mostra a la memòria, es procedirà a analitzar el *coverage* de les regions genòmiques d'interès, tal i com es descriu a [1.4.CoverageAnalysis.ipynb](https://github.com/marcexpositg/CRISPRed/blob/master/01.DescriptiveAnalysis/1.4.CoverageAnalysis.ipynb).

Si el coverage és superior a l'obtingut en la memòria, es quantificaran els resultats d'edició genètica tal i com es mostra a [2.2.LabelGen.ipynb](https://github.com/marcexpositg/CRISPRed/blob/master/02.Model/2.2.LabelGen.ipynb). Per fer-ne un anàlisi descriptiu, es recrearà l'anàlisi de les dades simulades tal i com es mostra a [2.3.OutcomesProfiling.ipynb](https://github.com/marcexpositg/CRISPRed/blob/master/02.Model/2.3.OutcomesProfiling.ipynb). El modelat es durà a terme amb dos models de predicció, un per l'eficiència i l'altre per la freqüència dels resultats. Per una banda, el model d'eficiència s'entrenarà i valorarà tal i com s'ha realitzat en la memòria i descrit a [2.5.EffModel.ipynb](https://github.com/marcexpositg/CRISPRed/blob/master/02.Model/2.5.EffModel.ipynb). Per l'altra banda, caldrà realitzar algunes modificacions en el modelat de la freqüència de cada resultat. Aquestes es basen en les conclusions extretes de les proves d'entrenament realitzades en la memòria per a aquesta part, descrites a [2.6.OutcomesModel.ipynb](https://github.com/marcexpositg/CRISPRed/tree/master/02.Model). Així, els models incorporaran descriptors de micro-homologia i s'utilitzarà *Keras* en comptes de *Scikit learn* per poder utilitzar *soft labels*. A més, es planeja incorporar altres models d'aprenentatge automàtic per comparar-los amb els descrits en la memòria.

Encara que les dades no permetessin observar edicions genòmiques, realitzar la memòria ha permès facilitar l'anàlisi i l'entrenament de models un cop es disposi de noves dades. Per exemple, la conversió de les coordenades de les regions genòmiques d'interès ([1.3.CoordiantesToC3H.ipynb](https://github.com/marcexpositg/CRISPRed/blob/master/01.DescriptiveAnalysis/1.3.CoordiantesToC3H.ipynb)) permetrà determinar el coverage en un menor període de temps. Tot i que el procés de simulació de dades no serà necessari, ha permès desenvolupar els *scripts* per entrenar models computacionals. Com que s'ha simulat les dades en un format que recrea l'experimental (obtenció de *reads*, i no de freqüències directament), els *scripts* es podran adaptar a les dades experimentals fàcilment.

– Què et sembla el TFM del teu company Alfonso Aguado que aborda un tema similar? Com valores els seus resultats?

Conec el treball de l'Alfonso, ja que els dos hem realitzat el TFM amb el mateix tutor extern i les seves dades havien de sortir del mateix experiment que he analitzat a la memòria i a causa del baix *coverage* també va simular les dades. Em sembla molt interessant observar les semblances/diferències entre els enfocs de cada un dels dos, que venen donades per la diferència en la tècnica d'edició. Tant la simulació de dades com l'entrenament dels models són diferents entre els treballs a causa de la diferència en els nostres objectius.

Per exemple, el seu model de càlcul d'eficiència de substitució és un de regressió mentre que el model d'eficiència d'edició de la meva memòria és un classificador binari. Mentre que en el seu model l'eficiència predita és el percentatge de nucleòtids de la seqüència modificats, en el meu l'eficiència és el percentatge de *reads* editats pel gRNA. Així, encara que els dos utilitzin el terme "eficiència", la defineixen diferent. Per edicions de *base editors*, la seva aproximació té sentit, ja que poden variar un o més nucleòtids de la seqüència, obtenint un valor entre 0 i 1. En canvi, per CRISPR-Cas9, com en el meu treball, no tindria sentit utilitzar la seva aproximació per predir els resultats, ja que els resultats d'edició no són substitucions, sinó delecions o insercions. Aquestes, donen productes d'edició de naturaleses diverses que no es poden representar en un sol valor continu entre 0 i 1, sinó que cal predir-les individualment en el model de freqüència dels productes d'edició com es fa en la meva memòria. Per això, el meu model d'eficiència prediu el percentatge de *reads* editats o no, sense entrar en detalls dels resultats editats. Aquesta aproximació es podria utilitzar també per *base editors*, però caldria generar unes dades diferents a les que ell ha generat. Tot i això, el seu treball no es centra en predir aquesta eficiència com a activitat del gRNA en si no en modelitzar els resultats obtinguts per *base editing* segons el gRNA, així que té sentit que no s'hagi inclòs.

En el segon apartat de modelatge, tots dos ens centrem en predir de forma més precisa els productes d'edició. La diferència en l'enfoc es deu a les diferències entre les productes d'edició de les tècniques. Per exemple, com que en el *base editing* hi ha substitucions, que no alteren la llargada del gRNA i que poden ser discretes i independents entre elles, l'Alfonso pot utilitzar un sistema per predir la probabilitat de cada nucleòtid segons posició de ser editat. En el meu model, en canvi, s'utilitzen categories diferents per cada producte. Això es deu a que CRISPR-Cas9 insereix insercions i delecions i no substitucions. Les insercions i delecions es realitzen sempre al voltant del punt de tall, i quan hi ha una deleció, les bases adjacents també han de ser eliminades, així que els nucleòtids no actuen de forma independent com en el cas de *base editing*. A més, seria impossible representar les insercions utilitzant la codificació de l'altre treball.

Així doncs, els resultats dels dos treballs difereixen a causa de la naturalesa de les tècniques d'edició genètiques utilitzades. Per tant, tot i que els descriptors són semblants (seqüència del gRNA), la tasca a realitzar és diferent, i és previsible que s'observin diferències en la capacitat de predicció dels algoritmes emprats per modelitzar-ho.

Trobo interessant que hagi comparat un nombre major de models i incorporat models que utilitzen xarxes neuronals. A més, el fet que el millor model que ha obtingut sigui el d'arbre de decisions em crida l'atenció, ja que en el meu treball he triat *Random Forests* sense passar per provar amb un sol arbre de decisió. Això fa que em decideixi a provar si un *decision tree* ofereix un valor de predicció comparable als models que he entrenat en la memòria.

Una segona lectura amb calma de certes parts del text hauria millorat l'acabat global del mateix (3r paràgraf pel final de la intro) o corregit certs petits errors (per exemple al començament pàgina 12 o a la pàgina 47 «a sigut difiícil» en comptes de: ha sigut difícil) que desvirtuen el gran treball fet.

Agraeixo que mencionis aquests errors, ja que així aprofito per compartir una versió revisada del treball que vaig editar passada l'entrega amb la intenció de corregir errors menors i tenir una memòria més polida per si la necessitava més endavant. Si t'interessa, l'he afegit al repositori GitHub del [treball](https://github.com/marcexpositg/CRISPRed/blob/master/ExpositGoy_Marc_Memoria_Revisada.pdf). Tot i això, he de comentar que no he pogut identificar els errors que menciones, ja que ni "sigut" ni "difícil" apareixen a la memòria.

– Quines assignatures i coneixements apresos durant el màster t'han servit per realitzar aquest TFM?

L'assignatura d'Eines de bioinformàtica em va ser útil per repassar els coneixements de programació en bash que tenia i per acostumar-me a treballar en un entorn Linux, tal i com he fet pel treball. L'assignatura de Genòmica Computacional em va ensenyar a utilitzar el navegador genòmic de UCSC, que vaig utilitzar en el procés de conversió de coordenades i em va facilitar molt la tasca. Les assignatures de Machine Learning i Análisis de Datos Omicos també m'han estat útils per el treball, tot i que al haver de realitzar-les en paral·lel al TFM vaig haver de formar-me prèviament en aquests àmbits fora del màster. Finalment, destacaria que el conjunt de la resta d'assignatures m'ha permès practicar programació (encara que fos en R). Això m'ha ajudat a acostumar-me a treballar de la forma en la que he realitzat el TFM, és a dir, en un entorn computacional i no al laboratori en bata.

– Quin seria el cost d'aquest projecte si tenim en compte les hores dedicades?

El projecte té una part experimental que incrementa molt la despesa, tant en material fungible com equipament necessari. Per exemple, la síntesi de la llibreria de gRNAs té un cost d'uns 2000€, tot i que es pot utilitzar en varis experiments i no només en aquest. El procés de clonació requereix múltiples intents, reactius i maquinària especialitzada i moltes hores, així que és complex estimar-ne el cost. El manteniment i manipulació de cèl·lules animals és també molt car. L'electroporació de tantes mostres pot arribar a costar uns 100€. Finalment, la preparació i seqüenciació de les mostres per Illumina arriba a costar uns 1000€, però cal tenir en compte que no només es seqüencien les mostres d'aquest projecte.

Per altra banda, la part computacional també requereix recursos. L'anàlisi de dades genòmiques és molt més senzill si s'utilitza un servidor extern com el clúster Marvin del PRBB. Tenir accés a aquest servei també és car. També cal tenir en compte que l'entrenament dels models i altres scripts requereixen un ordinador personal. Com que Google Colab utilitza la computació a distància i és gratis, no caldria que fos un ordinador gaire potent. Finalment, cal tenir en compte les despeses de personal. Considerant les hores dedicades amb un contracte de pràctiques de 5h/dia, sense tenir en compte les hores extres que s'hi acaben dedicant, caldria afegir-hi una despesa de 3000€.

Així, suposant que el treball es realitza en un laboratori amb la maquinaria requerida, i realitzant una estimació, es podria calcular que el cost total del projecte oscil·la entre 15,000-40,000€, segurament entre la franja dels 25,000€ i 30,000€.

– Quines són les parts que t'han costat més i menys del TFM?

Les parts que més m'han costat són també aquelles amb les que més he gaudit i après, i que descric a continuació. La quantificació dels resultats d'edició i el càlcul de la profunditat de seqüenciació van ser reptes a nivell de programació. La simulació de dades un repte a nivell d'integrar biologia i computació (trobar un enfoc computacional que recreï l'activitat biològica) i també d'adaptar eines ja existents per solucionar un problema nou. Identificar les aproximacions per utilitzar models de predicció de la freqüència dels resultats, i els paràmetres a provar per cada model, van posar a prova els meus coneixements sobre *machine learning* i obligar-me a investigar més sobre aquest àmbit en el qual era principiant. Finalment, convertir les coordenades de mm10 a C3H va ser un procés tediós i manual que va requerir moltes hores.

La part que menys m'ha costat ha estat el pre-processat i l'alineament de dades amb el genoma de referència, ja que es van utilitzar programes com samtools i trimmomatic que permeten automatitzar el procés i són simples d'utilitzar.

– Quin serà el següent pas en la teva carrera professional?

Després d'haver estat al MIT fent la tesi del màster en Bioenginyeria d'IQS, aquest any he rebut la beca de la caixa per fer el doctorat a Amèrica del nord. Així, l'any que ve començo un doctorat a la University of Washington, a Seattle. El màster m'ha permès formar-me com a bioinformàtic i complementar els coneixements d'estadística que em faltaven del grau de biotecnologia. Així, espero poder fer el doctorat en algun dels laboratoris que m'interessen de la University of Washington. En un es dediquen a dissenyar proteïnes *de novo* utilitzant Rosetta, i en l'altre realitzen anàlisis de dades òmiques provinents de tecnologies de nova seqüenciació. Tots combinen part experimental amb bioinformàtica, i gràcies al màster de la UOC espero poder tocar una mica de cada part.