Universitat Autònoma de Barcelona

Caracterització bioinformàtica de la relació entre l'impacte molecular de les variants patogèniques i el fenotip clínic

Òscar Marín Sala

Caracterització bioinformàtica de la relació entre l'impacte molecular de les variants patogèniques i el fenotip clínic

Òscar Marín Sala

Tesi doctoral

Doctorat en Bioquímica, Biologia Molecular i Biomedicina

Institut de Recerca de la Vall d'Hebrón (VHIR)

Dpt. de Bioquímica i Biologia Molecular

Universitat Autònoma de Barcelona

Juny 2017

Director:

Dr. Xavier de la Cruz i Montserrat

Tutor Acadèmic:

Dr. Enrique Querol i Murillo

«No tot el que és or és cosa que lluu,

ni tot muntaner ha d'anar perdut»

Agraïments

Un llarg viatge com aquest no hauria pogut ser possible sense l'ajuda i la companyia de tots vosaltres. Sou els que heu fet que els dies més foscos i els entrebancs no fossin més que anècdotes; però també amb els que he compartit els millors dies. La persona que sóc no s'entén sense la vostra part. Amb vosaltres he après que l'important no és el destí, sinó com –i , sobretot, amb *qui*– fas el viatge.

En primer lloc vull donar les gràcies al Dr. Xavier de la Cruz, director d'aquest treball, per mostrar-me com recórrer aquest camí, per la paciència davant les meves idees i per ensenyar a fixar-me en els detalls que importen.

Tot seguit vull agrair al Dr. Enrique Querol, qui em va introduir en el món de la ciència; i al Dr. Modesto Orozco, amb qui encara ara estic aprenent dia rere dia, Gràcies per les oportunitats que m'heu donat.

A la gent del grup al VHIR. Al Sergi, la Casandra, l'Elena, la Natàlia i el Josu; amb els que hem disseccionat, analitzat i interpretat qualsevol dada que se'ns posava davant... i a la resta de gent que heu passat més o menys temps pel grup. També a tota la resta de col·laboradors i companys del VHIR.

A la gent del MMB. Potser encara ha estat breu, però amb vosaltres cada dia aprenc quelcom nou. Al Ricard, una constant en aquest viatge, a l'Antonio, la Diana, el Diego i el Jurgen pels grans cafès i seus debats. A l'Adam i al Genís, pels bons consells. Al Pedro, el Víctor i el Juan: un plaer treballar amb vosaltres. A la resta del grup, per acollir-me tan bé.

Als que hi sou i als que hi fóreu. Als de Castellar. Al primer de tots, al Toni. Jo escric una tesi, ben aviat espero tenir a les mans el teu llibre! Als 4F: David, Lucas i Marc; per les infinites hores de diversió. Al Pau i l'Aïda. Al Borja i la

Desi, que per molt lluny que esteu us sentim molt aprop. A la resta d'amics del poble. Als de l'equip del Puig. També a tots de la carrera i del màster: hem après i estimat la biologia junts.

Perdoneu-me si no us he mencionat a tots, però sou *tants* els que formeu part d'aquesta història que no acabaria mai. Gràcies.

Als meus pares. Heu estat en tot moment allà, ajudant-me i donant-me suport en tot. Des del principi, i sempre. Marta, avui és la meva tesi. Qui sap si, d'aquí quatre anys, serà la teva. Pots fer el que et proposis. A la resta de la família. Als que ja no hi sou. Gràcies. Tot el que m'heu ensenyat forma part de cada lletra d'aquest treball.

I finalment per tu, Cris. Per estar amb mi en tot: en les coses bones i en les dolentes. Per la teva bondat, la teva paciència i el teu amor. Per haver fet tot aquest camí al meu costat. I per tot el camí que encara ens queda per fer.

Resum

L'adveniment de la seqüenciació de nova generació (NGS) promet canviar el paradigma de la medicina, però les dades provinents de la seqüenciació porten amb elles un conjunt de reptes tècnics i metodològics importants, i que dificulten la seva integració en la medicina de precisió. L'aprenentatge automàtic apareix com una possible solució a diversos d'aquests problemes, ja que és una eina molt potent capaç de processar i analitzar dades de gran complexitat. Aquesta tesi estudia temes clau per a la possible aplicació en clínica de les tècniques de NGS mitjançant eines bioinformàtiques i d'aprenentatge automàtic.

En primer lloc, s'estudien les característiques moleculars i evolutives de les variants patogèniques compensades en altres espècies (CPD), i quin rang d'impacte fenotípic poden produir els CPD.

En segon lloc, s'apliquen mètodes de xarxes neurals per la predicció de l'efecte de les variants puntuals patogèniques en la severitat de la malaltia, a partir d'atributs fisicoquímics i evolutius associats al canvi d'aminoàcid. S'usen les proteïnes F8 i F9 com a model. També s'analitzen les característiques de les variants que produeixen els efectes lleus i els severs de les malalties.

Finalment, s'apliquen mètodes basats en arbres de decisió per crear una metodologia de predicció de les CPD a partir de variables que descriuen el canvi molecular i la relació evolutiva d'una posició amb les altres de la proteïna. Després s'usen aquests mètodes per buscar la presència de variants CPD en humans amb l'estudi dels individus seqüenciats a 1000G, i s'analitza si aquestes variants poden ser una fracció de l'incidentaloma.

Abstract

The advent of Next Generation Sequencing (NGS) carries the promise to change medicine's paradigm, but sequencing data comes with a myriad of noticeable technical and methodological challenges. Those hurdles difficult the integration of NGS technologies in precision medicine. Machine Learning is a possible solution to some of those problems, as it is a powerful toolbox with algorithms capable of processing big and complex data. This thesis deals with key topics in the clinical application of NGS techniques using bioinformatics and machine learning methods.

First, we study the molecular and evolutionary characteristics of variants known as compensated pathogenic deviations (CPD), which are pathological variants appearing as wild type in other organisms, and its associated phenotype impact.

Second, we apply neural network models to predict the phenotype severity of pathological variants. We use physico-chemical and evolutionary attributes that describe the amino-acid change, using proteins F8 as F9 as our models. We also analyze the characteristics of variants associated to mild and severe versions of disease.

Last, we apply methods based on decision trees to create a CPD prediction methodology from descriptors of the molecular change and the evolutionary relationship between positions in the protein sequence. We use those predictors to search for CPD variants within humans, studying the sequenced individuals from the 1000G project. We study the likelihood that those variants are a fraction of the incidentalome.

Taula de Continguts

1.Introducció

La recerca descrita en aquesta tesis, així com la seva aplicabilitat, es situa en la intersecció entre tres àrees del saber: l'aplicació de la Next Generation Sequencing (NGS) a la medicina; l'estudi i predicció del fenotip de les variants patogèniques; i l'ús d'eines d'aprenentatge automàtic per a la construcció de models quantitatius de fenòmens biològics. La introducció cobreix els aspectes més rellevants d'aquestes tres àrees, de forma que el lector pugui comprendre la motivació de la recerca i pugui llegir de forma crítica els resultats que es presenten en els capítols següents.

1.1 De l'estructura de l'ADN a la seqüenciació de nova generació

1.1.1. La doble hèlix: estructura i seqüència

El 1953, Watson i Crick van publicar el descobriment de l'estructura molecular de l'àcid desoxiribonucleic (ADN) en forma de **doble hèlix** (Watson i Crick 1953). Dos cadenes de polinucleòtids en direccions oposades unides per ponts d'hidrògens entre els parells de nucleòtids d'Adenosina-Timina i Citosina-Guanina, que des de llavors es coneixen com a parells de Watson i Crick. Aquesta troballa va marcar un abans i un després en la biologia, i més en concret en la genètica: donava una resposta elegant a com els sers vius podien emmagatzemar molecularment tota la informació del genoma d'una forma comprimida en cada cèl·lula, aportant una peça molt important en el trencaclosques de l'herència i la reproducció.

Múltiples preguntes orfes de resposta podien ser investigades des d'aquell moment. Els enllaços fosfodiéster, que uneixen els nucleòtids d'una mateixa cadena, són més forts que els ponts d'hidrogen que permeten la doble hèlix. És

més fàcil, llavors, obrir la doble hèlix per llegir la informació que trencar el patró de seqüència. Es pot, per exemple, usar cadascuna de les cadenes, que són redundants, per a generar dos noves hèlix idèntiques, que són les que tindria la nova generació de cèl·lules –procés anomenat **replicació** (Meselson i Stahl 1958)–. També es pot usar una cadena per passar de la mínima unitat d'informació genètica, el gen, a àcid ribonucleic (ARN) missatger –procés anomenat **transcripció** (Crick 1958)–, que anirà a un ribosoma on, traduint cada codó, que és una agrupació de tres nucleòtids que correspon a un dels 20 aminoàcids, es sintetitzarà la proteïna –el procés anomenat **traducció** (Khorana et al. 1966)–. Es passa de la unitat mínima d'informació genètica, el gen, a la unitat funcional bàsica de la cèl·lula, la proteïna. S'establia, llavors, el dogma central de la biologia molecular (Crick 1958): La informació pot transmetre's d'ADN a ADN, d'ADN a ARN, i d'ARN a proteïna –i, en casos especials, d'ARN a ADN (Temin i Mizutani 1970; Baltimore 1970) o entre ARN–, però en cap cas en altres direccions (Figura 1.1). La informació que conté una proteïna ja no torna cap a l'ARN o l'ADN. Així doncs, qualsevol canvi en el patró de seqüència que es transmeti a la descendència ha de sorgir, forçosament, per **mutacions** en l'ADN.

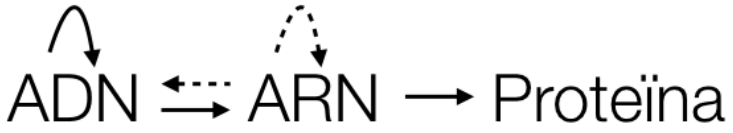

***Figura 1.1**: Dogma Central de la Biologia Molecular. Les fletxes contínues indiquen el flux principal d'informació genètica; les fletxes discontinues indiquen alternatives menys freqüents.*

Tot i això, per estudiar més detalladament la rellevància biològica de les molècules d'ADN, bé en humans o bé en altres organismes, faltava un pas molt

important: saber **quina és la seqüència** que tenen i, per extensió, quines proteïnes hi estan codificades. **Sanger** i col·laboradors van ser els que van introduir el primer mètode de seqüenciació d'ADN (Sanger et al. 1977), que va permetre anar estudiant sistemàticament la seqüència de diversos gens i, fins i tot, el genoma complet d'alguns virus (Sanger et al. 1977). La tècnica consisteix (Figura 1.2A) en que, a partir d'una de les cadenes d'ADN i d'una seqüència encebador –tros curt d'ADN que serveix com a inici de la síntesis–, una polimerasa va afegint nucleòtids a partir d'aquest encebador. La reacció es reprodueix per separat en quatre casos diferents, amb tots els desoxinucleòtids i un dels di-desoxinucleòtid (ddNTP), que és diferent a cada reacció, a una concentració aproximadament 100 vegades inferior a la del desoxinucleòtid. Aquest ddNTP, a l'afegir-se a la cadena d'ADN, atura la síntesi, de forma que en cada reacció hi haurà cadenes de la longitud corresponent a les posicions del nucleòtid. Un gel electroforètic farà córrer les seqüències a diferents posicions i, com que els ddNTP estan marcats radioactivament, cada reacció donarà un patró de bandes referent a les posicions del nucleòtid. Actualment n'hi ha prou amb una sola reacció, ja que s'usa fluorescència i un color diferent per a cada ddNTP, així que poden afegir-se els quatre a la mateixa reacció (Smith et al. 1986).

1.1.2 NGS: La seqüenciació d'ADN esdevé massiva

El mètode de Sanger, malgrat estar parcialment automatitzat, segueix sent massa lent i massa car per a seqüenciar genomes sencers. El **genoma humà**, que va començar a ser seqüenciat inicialment mitjançant Sanger, va ser un projecte d'un consorci públic de milers de milions de dòlars que va durar una mica més de deu anys (*International Human Genome Sequencing Consortium*

2001), un temps i un cost inassumible per qualsevol estudi més detallat sobre el genoma humà. Paral·lelament al projecte públic, una iniciativa privada encapçalada per l'empresa *Celera* va emprendre el mateix objectiu mitjançant seqüenciació *shotgun* (Figura 1.2B)*:* una metodologia que també fa ús de Sanger però, enlloc d'anar creant encebadors a mesura que es va seqüenciant, trenca a l'atzar les cadenes del cromosoma sencer en nombrosos petits trossos, els seqüencia tots, i els assembla progressivament en una seqüència més llarga mitjançant alineaments per semblança. Finalment el 2001, els dos projectes, privat i públic, van aconseguir una altra fita històrica en la biologia: el primer esborrany de la seqüència del genoma humà (*International Human Genome Sequencing Consortium* 2001; Venter et al. 2001), posteriorment, el 2004 s'assolí la seqüència complerta (*International Human Genome Sequencing Consortium* 2004).

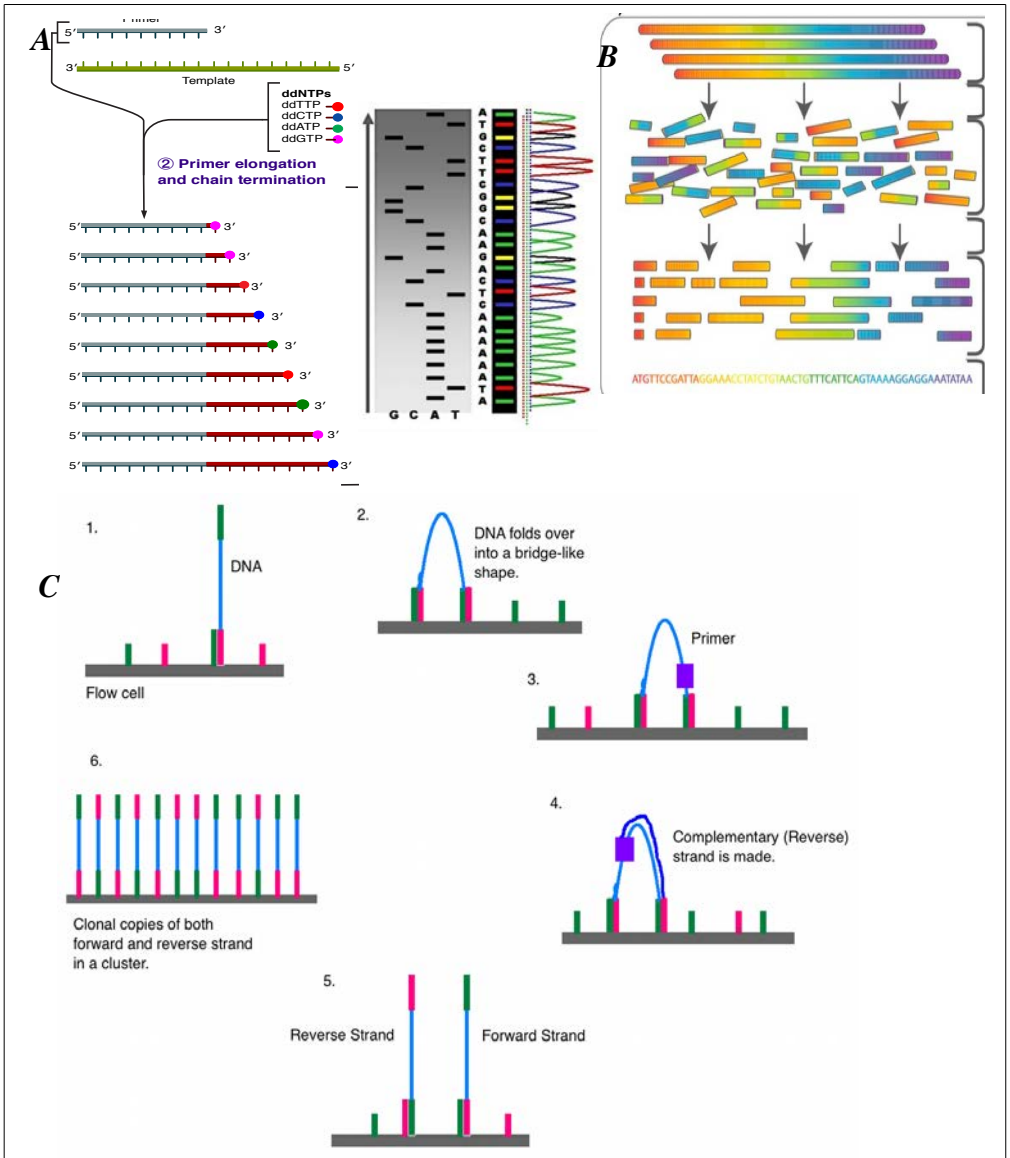

***Figura 1.2**: **A)** Seqüenciació Sanger (Adaptat de Abizar i Estevezj). **B)** Seqüenciació shotgun (Adaptat de Commins et al.) **C)** Seqüenciació de nova generació (Adaptat de DMLapato).*

El projecte del genoma humà va portar amb ell nombrosos avanços en l'entesa de la genètica humana, com aproximacions al nombre de gens que té el

ser humà –al voltant de 22300– (Pertea i Salzberg 2010), que va deixar palès que les diferències entre l'humà i altres espècies són molt menors del que inicialment s'esperava. Diversos grans projectes han continuat aprofundint el coneixement del genoma tot estudiant les variacions puntuals (SNP, per coherència amb l'anglès, Single Nucleotide Polymorphisms) entre individus (*International HapMap Consortium* 2003), observant les diferències en el genoma entre individus de la mateixa població i de poblacions diferents (*The 1000 Genomes Project Consortium* 2015); o estudiant genotip i fenotip de diverses tipologies de tumors cancerígens (*The Cancer Genome Atlas (TCGA) Research Network 2008*).

Però els avenços del projecte del genoma humà no només van ser purament científics, sinó que metodològicament es van assentar les bases per al que es va conèixer com a **Next Generation Sequencing** (NGS) o seqüenciació massiva en paral·lel. Es tracta d'un conjunt de tècniques, comercialment diferents però amb unes bases molt similars, que evolucionen el concepte de la tècnica de *shotgun* seqüenciant alhora una gran quantitat de fragments d'ADN (Figura 1.2C), fent el que s'hauria tardat setmanes en només un dia. En primer lloc, es trenca la cadena a seqüenciar en nombrosos fragments curts –la longitud dels quals dependrà de la tècnica comercial concreta–, que s'uneixen mitjançant un procés amb adaptadors a unes plaques o bombolles –també segons la tècnica–. Llavors, amb l'ajuda d'una polimerasa, es van llegint de forma iterativa els nucleòtids, que són identificats per la corresponent màquina de seqüenciació paral·lelament en els múltiples fragments que han quedat. L'evolució d'aquestes tècniques de nova generació ha reduït moltíssim el cost tan monetari com de temps per a seqüenciar un genoma sencer (Figura 1.3), i ha permès en els últims

anys disparar el nombre d'espècies que tenen el genoma seqüenciat –de 70 espècies eucariotes a finals de l'any 2004 a 4306 a principis de maig del 2017 (servei ncbi: genome[Internet] 2017)– i, fins i tot, portar la seqüenciació de genomes a la clínica (Chen i Snyder 2013).

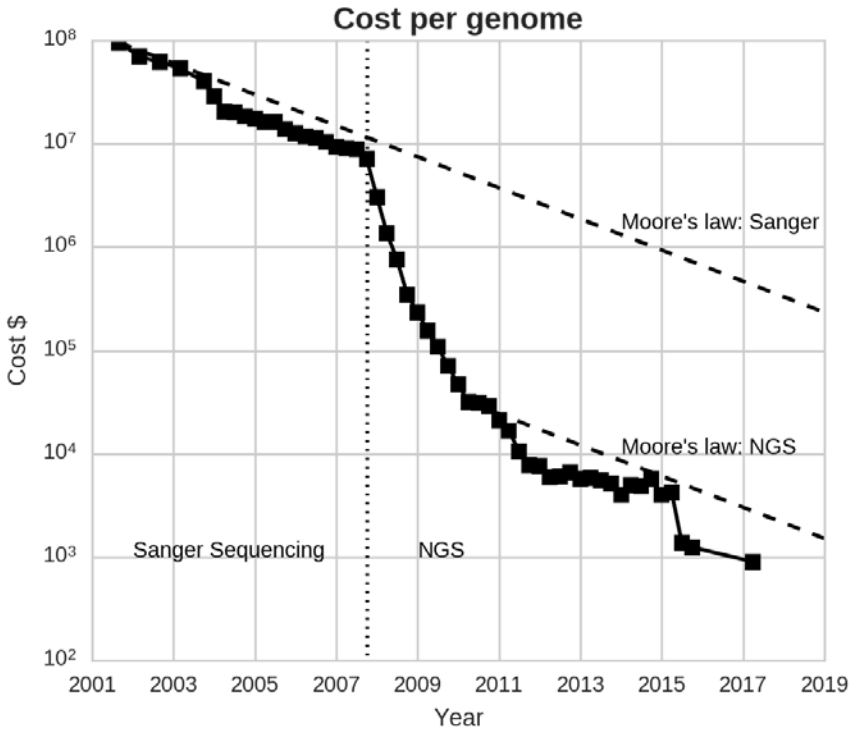

Figura 1.3: *Evolució del cost de la seqüenciació del genoma al llarg del temps. Fonts: www.genome.gov/sequencingcostsdata/ consultat el 09-05-2017 i www.scienceexchange.com/services/whole-genome-seq consultat el 09-05-2017*

El desenvolupament d'aquestes tècniques no s'ha aturat, ja que la **tercera generació** de seqüenciadors promet, en un futur no gaire llunyà, portar millor qualitat i més rapidesa amb un cost encara més baix i amb maquinàries molt més petites (Oxford Nanopore Technologies, consultat el 2017; Pacific

Biosciencies, consultat el 2017), que permetria l'adopció de tecnologies derivades de l'anàlisi del genoma en el dia a dia tan en recerca, com en clínica com, fins i tot, als domicilis.

1.2 NGS: creant el somni de la medicina personalitzada

El ràpid desenvolupament de les tècniques de seqüenciació ha permès la seva aplicació cada cop més generalitzada en certes àrees de la pràctica mèdica; com són el diagnosi, la prognosi o la teràpia. La medicina personalitzada, o medicina de precisió, resumeix aquests avenços. Medicina de precisió és el terme que s'usa per descriure els objectius de la medicina del futur: una medicina que tingui en compte les necessitats específiques i les particularitats de cada pacient; és a dir, capaç d'adaptar el tractament individualment, tenint en compte el metabolisme i la fisiologia del pacient; i no només de tractar, sinó d'**anticipar i prevenir** les malalties (Roden i Tyndale 2013). Molts d'aquests objectius passen per tenir el genoma del pacient, i, encara més important, **entendre**'l.

La conclusió del projecte del genoma humà, juntament amb l'adveniment del NGS, provocà una onada d'optimisme sobre la medicina de precisió (Daman i Weber 2012). L'èxit de casos com el del Trastuzumab, un anticòs monoclonal que actua sobre HER i que millora la supervivència en els càncer de mama que sobreexpressen el gen HER2 (Slamon et al. 2001; Roukos 2011) augurava un ràpid trasllat dels avenços en recerca bàsica cap a la pràctica clínica i a la millora dels tractaments actual. Altres casos paradigmàtics han estat els tractaments preventius als individus que porten mutacions als gens de BRCA1 i BRCA2 (Narod i Foulkes 2004), associats amb càncer de mama; o CDH1

(Ziogas i Roukos 2009), associat amb càncers gàstrics, que redueixen la prevalença en aquests individus sota risc. Portar a la pràctica l'estudi del genoma humà com a eina diagnòstica i de prevenció no ha estat, però, tan fàcil com es pensava: bé la falta de coneixements sobre quins són exactament els factors moleculars causants de les malalties (Hall et al. 2016) o sobre els factors ambientals (Delaney et al. 2016) i epigenètics (Han i He 2016) que modulen l'expressió del genotip en el fenotip; o bé la manca de marcs matemàtics adequats per modelar la complexitat genòmica (Colin et al. 2017); i d'altres raons (Laksman i Detski 2011); eviten que encara no es pugui aplicar amb propietat el terme *precisió* a la medicina de precisió, malgrat s'estigui avançant molt en totes les àrees esmentades.

1.2.1 Aplicació en diagnosis de la NGS

Encara que hi ha algunes dificultats tècniques i de coneixements que dificulten el trasllat a la clínica, el diagnòstic a partir de versions reduïdes del genoma ha anat obrint-se terreny en els centres més pioners del món: tècniques com el ***Whole Exome Sequencing*** **(WES)**, seqüenciació de tot l'exoma (Tetreault et al. 2015), que a partir d'una llibreria que marca les zones amb exomes, seqüencia només els fragments corresponents a les zones codificants de les proteïnes; o els **panells de seqüenciació**, que tenen com a objectiu només un conjunt de gens que ja se sap que són els potencialment més perillosos (Hall et al. 2014, Thakral et al 2016); són una via barata i més simple d'analitzar que no tot el genoma, conegut com a ***Whole Genome Sequencing (WGS)*** (Chang 2015; Oliver et al. 2015).

Hi ha diversos casos d'ús dels exomes i els panells de gens en l'entorn clínic, essent dos els més habituals: **i)** el diagnòstic genòmic en els casos de càncer,

mitjançant comparacions de l'exoma (o genoma) del tumor amb el del pacient, per observar-ne les variacions somàtiques (Liu et al. 2013; Roberts et al. 2013; Garofalo et al. 2016), i **ii)** La caracterització genètica de malalties rares –sovint mendelianes– per a les que se'n desconeix el gen causant (Bamshad et al. 2011; Shen et al. 2015) o per el diagnòstic concret del gen responsable en els casos de famílies de malalties rares fenotípicament molt semblants en les que el tractament depèn del gen implicat (Fuchs et al. 2012; Stark et al. 2016).

1.2.2. Els límits actuals de l'aplicació del NGS en la medicina

Els avenços esmentats són prometedors, i els casos d'èxit en els problemes anteriors poden facilitar la integració al món mèdic de les tecnologies derivades de la seqüenciació, i apropar-les posteriorment a la societat en altres aspectes, com són la investigació dels perfils metagenòmics del microbioma intestinal (Wang et al. 2015), o dels aspectes nutrigenòmics (Fallaize et al. 2013) i farmacogenòmics (Relling i Evans 2015) individuals; tot millorant els estudis personalitzats del genoma (Song et al. 2014).

Aquests avenços, però, han de venir precedits, però, d'una millora en alguns dels problemes que encara limiten la seqüenciació tan a nivell tècnic com a nivell del coneixement (He et al. 2017).

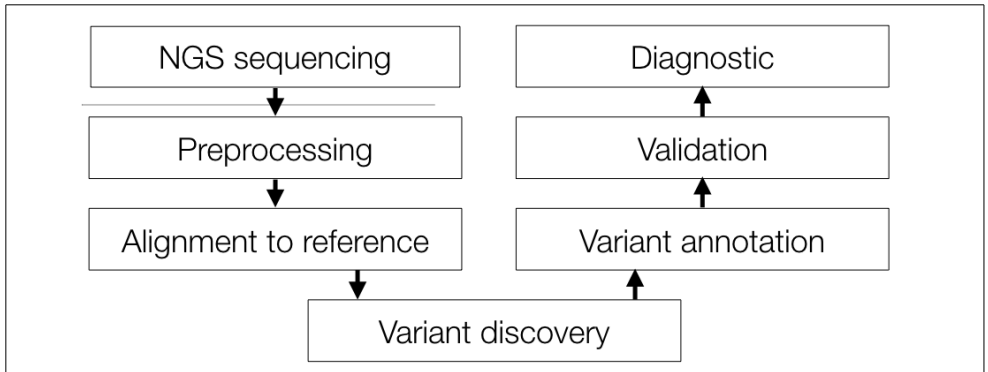

Figura 1.4: *Procés seguit habitualment en un estudi de NGS.*

Per una banda, la qualitat tècnica de la seqüenciació ha de millorar (Stapleton 2014; Zhou i Rockas 2014). Tot i que hi hagin diverses metodologies per netejar les lectures de cadena –o *reads*–, per netejar d'artefactes o per eliminar lectures duplicades (Li et al. 2009; McKenna et al. 2010; Bolger et al. 2014; Pandey et al. 2016); hi ha regions que encara no tenen prou qualitat ni profunditat –entès com a nombre de lectures d'una qualitat mínima que donen suport a una posició– com per a fer diagnòstics o prediccions amb un percentatge prou fiable. La millora ha de venir en tots els passos del processament de dades de NGS (figura 1.4): lectures més llargues –i per tan millor alineades–, de millor qualitat, amb menys artefactes i amb millor cobertura de totes les regions; acompanyat d'eines de processament més potents i més ràpides –actualment el postprocessament pot durar fins a 16 hores en entorns d'alt rendiment (Moncunill et al. 2014), que no estan disponibles arreu– faran més fàcil l'adopció de l'estudi del genoma en més camps.

Per altra banda, també s'ha de millorar el coneixement i les tècniques de la part d'anotació mèdica i funcional de les variants resultants, pas conegut com el «problema de la interpretació» (Schrijver 2012), i que és el coll d'ampolla més important en l'ús d'eines NGS en clínica. En un genoma podran aparèixer variants puntuals –tan en regions codificants, com en regions reguladores, com en regions intròniques per les que no se'n coneix funció–, indels, duplicacions, rearranjaments... Identificar dins del mar de variació quins són els canvis genètics que causen els canvis fenotípics, i separar correctament quins són els importants per al cas d'interès estudiat és una tasca per a la que encara no hi ha una eina o metodologia definitiva (Dand et al. 2015; Jalali Sefid Dashti i Gamieldien 2017). S'han dedicat nombrosos esforços per a l'anotació de canvis

puntuals en regions codificants, sobretot per discriminar variants patogèniques d'aquelles que són neutres (Adzhubei et al. 2010; Sim et al. 2012; Riera et al. 2014; Niroula et al. 2015; Riera et al. 2016; López-Ferrando et al. 2017), fins i tot en l'anotació d'altres característiques d'aquestes variants –canvis de solubilitat (Yang et al. 2016), severitat (Niroula i Vihinen 2017) o altres (Li et al. 2014; Anoosha et al. 2015)–, però no es tracta ni molt menys d'un camp resolt: encara s'ha de millorar en les prediccions (Riera et al. 2016) i s'ha de profunditzar en l'efecte real que tenen aquests canvis, sobretot en malalties no monogèniques (Keyes et al. 2015). A més a més, no hi ha encara mètodes per predir l'efecte fenotípic que tenen les variants que, malgrat no ser patogèniques modulen altres parts del fenotip: variants que fan que un enzim processi millor el substrat (Simpson et al. 2014) i quines conseqüències poden tenir, variants que modulen l'afinitat d'un receptor cap a un lligand (Feinberg et al. 2013), etc.

Per altra banda, un factor que també és important, que és d'origen conceptual i que s'haurà de tenir en compte en estudis individualitzats, és l'epigenètica. Estudis del posicionament i empaquetament dels nucleosomes (Flores i Orozco 2011) o els estudis de les marques químiques, com metilacions o acetilacions, que modulen l'expressió del genoma, són importants de cara a l'expressió o no del gens (Castillo-Fernandez et al. 2014). També s'haurà d'integrar a l'anàlisi, en la mesura del possible, part dels factors ambientals que afecten al genoma. Sigui en la forma dels registres electrònics (Wu et al. 2016), o mitjançant mesures preses amb *wearables* –sensors integrats en aparells electrònics comuns, com rellotges o telèfons mòbils– (Gao et al. 2016), o d'informació, directa o indirecta, de la dieta (Odriozola i Corrales 2015); la informació de l'ambient s'ha d'acabar integrant per personalitzar encara més el tractament

(Ordovás 2009).

Finalment, també s'haurà d'investigar l'efecte dels canvis en les potencials interaccions epistàtiques d'una proteïna amb l'altra (Hopf et al. 2017). L'efecte que produeix un canvi no és només local –entès com efectes directes en la mateixa proteïna–, sinó que pot transmetre's a través de la xarxa de proteïnes amb les que interactua, tan directa com indirectament (Lehner 2011; Wilkins et al. 2013; Moraru et al. 2015; Pedros et al. 2015).

A nivell computacional, no hi ha encara models matemàtics que integrin la totalitat dels factors anteriors. El nombre de graus de llibertat d'aquest problema és molt alt, però també ho és el nombre de dades. Els avenços en el camp de la genòmica hauran d'anar acompanyats, forçosament, per models més complexos i que usin més poder computacional per processar i analitzar les dades que genera la NGS.

1.3 Models matemàtics: «big data» i aprenentatge automàtic en l'estudi del genoma

En els últims anys han sorgit diverses paraules de moda com «Big Data» –terme referent a treballar amb dades a partir de GB, amb els canvis que implica a nivell computacional– i «machine learning» –aprenentatge automàtic– (Obermeyer i Emanuel 2016). Aquests camps es dediquen a l'estudi, desenvolupament i aplicació de tècniques i mètodes per treballar amb grans volums de dades i per crear models que n'aprenguin els patrons que es desprenen d'aquestes. L'abundància de dades que supera la capacitat de processament manual és ja un problema comú en totes les àrees de la ciència; concretament en el camp de la biologia molecular, on la seqüenciació de nova

generació crea gigabytes de dades, i cada cop s'estan seqüenciant més individus. Per poder aprofitar el potencial científic/mèdic de totes aquestes dades, cal integrar-les en models formals que permetin predir els fenòmens biològics. Avui en dia, aquesta tasca, enormement cara des del punt de vista computacional, es pot fer gràcies a les eines d'aprenentatge automàtic.

L'aprenentatge automàtic no és quelcom nou. Fou a mitjans del segle XX , amb l'inici de la computació, quan els primers mètodes de xarxes neurals, el perceptró (Rosenblatt 1957) i millores del perceptró amb major complexitat, van començar a sorgir, precisament inspirats en l'arquitectura de les neurones en el cervell. El perceptró, que originalment era una maquinaria física i que posteriorment va ser implementat com a algoritme, consistia de diverses cèl·lules fotovoltaiques que rebien la informació «sensorial» connectades a l'atzar amb unes «neurones», que tenien uns pesos –que s'anaven modificant en un procés d'entrenament per ajustar-se a reconèixer una figura concreta–, i que s'unien per a formar una sortida binària: el reconeixement correcte o no de la figura. Al final de l'entrenament el sistema era capaç de reconèixer imatges. Els perceptrons amb diverses capes –coneguts com a *multilayer perceptrons*–, en canvi, funcionaven amb diverses capes de neurones entre els receptors, o variables descriptives, i la sortida final; i eren capaços d'aprendre patrons més complexos. Dit d'altra forma: no era necessari que fossin patrons separables de forma linear, com passava amb el perceptró d'una sola capa (Grossberg 1973).

Les expectatives que van generar les xarxes neurals van ser, però, exagerades i impossibles de complir, fet que va conduir al que es coneix com a hivern de la intel·ligència artificial (Hendler 2008): un període amb una declivi en l'interès i el finançament en intel·ligència artificial. A partir de llavors el focus es va

centrar en models matemàtics més «simples» que, malgrat tot, també tenien poder predictiu, i que podien aplicar-se millor amb la quantitat de dades i de poder computacional que hi havia. Part de la recerca es va centrar en mètodes de classificació supervisats com les Support Vector Machines (SVM) (Boser et al. 1992), d'altra banda es van ampliar els coneixements dels mètodes sorgits a partir dels arbres de decisió, com el Random Forest (Ho 1995), el *boosting* (Breiman 1998*)*, o l'AdaBoost (Freund i Schapire 1997), i també es van desenvolupar més els mètodes no supervisats, donant lloc a tècniques com el BIRCH (Zhang et al. 1996) i DBSCAN (Ester et al. 1996).

Aquesta situació canvià amb l'adveniment d'internet i el creixement en poder computacional –tan en emmagatzematge com en capacitat de càlcul–, ja que en molts problemes el nombre de dades disponibles va augmentar molt, donant lloc al que es coneix com a *Big data*. Aquesta nova situació ha permès que les xarxes neurals siguin més sofisticades, és el camp conegut com a *deep learning,* on s'empren xarxes neurals amb un nombre de capes ocultes molt alt i una arquitectura molt més complexa, ja que fan servir mètodes com els *autoencoders* (Bengio 2009), les Màquines de Boltzmann Restringides (Smolensk 1986; Larochelle i Bengio 2008), o les xarxes neurals convolucionals (Matsugu et al. 2003).

Big data i aprenentatge automàtic són, doncs, dos termes que s'expliquen junts: fan falta els models desenvolupats per aprenentatge automàtic per a poder analitzar adequadament els grans volums de dades, i les xarxes neurals, una de les tècniques més importants d'aprenentatge automàtic, tenen millor rendiment quan es disposa d'una gran quantitat d'informació amb la que entrenar-se.

1.3.1 Aprenentatge automàtic en la biologia: modelant fenòmens moleculars

Les tècniques d'aprenentatge automàtic s'estan traslladant satisfactòriament a l'estudi de la biologia. Camps com la predicció de l'estructura secundària en proteïnes (Feng et al. 2014; Drozdetskiy et al. 2015) o en RNA (Achawanantakun i Sun 2013; Hamada et al. 2015), la predicció d'interaccions entre i amb proteïnes (Fernández-Recio 2012; Wong et al. 2013; Zhang et al. 2014; Xue et al. 2015) o la predicció de l'impacte funcional de mutacions (Adzhubei et al. 2010; Sim et al. 2012; Niroula et al. 2015; Riera et al. 2016; López-Ferrando et al. 2017) han utilitzat l'aprenentatge automàtic per predir característiques biològiques amb un encert acceptable. I aquests casos només són alguns dels exemples més clàssics, ja que s'ha usat en multitud de problemes més tan en biologia com en clínica, com per exemple en predicció d'interaccions entre fàrmacs (Cheng i Zhao 2014), en predicció d'epítops (Wang i Pai 2014) o per predir les possibilitats de supervivència en algunes malalties (Montazeri et al. 2016). En aquesta tesi (Capítol II) es veu com es pot aplicar l'aprenentatge automàtic a la predicció de la severitat de la malaltia associada a una mutació.

Les tècniques més sofisticades –xarxes neurals profundes, que necessiten una quantitat de dades ordres de magnitud superiors–, però, no s'han traslladat encara a la biologia, ja que encara s'estan usant molt esporàdicament (Zhou i Troyanskaya 2015; Tian et al. 2016, Vang i Xie 2017). Aplicar les tècniques més noves d'aprenentatge automàtic en problemes biològics i fer ús de la gran quantitat de dades que s'estan generant amb la NGS permetrà analitzar més ràpid i amb més exactitud el genoma, i permetrà comprendre'l millor (Miotto et al. 2017; Schmidt i Hildebrandt 2017).

1.4 La variació en el genoma: anotació i classificació de variants

L'últim pas computacional en un anàlisis tradicional de NGS és l'anotació i la classificació de les variants resultants. Es tracta de comparar la seqüència obtinguda en l'anàlisi amb la «canònica», que depèn del genoma de referència, anotar en quina regió està el canvi de nucleòtid –exó, intró, en quin gen es troba...–, observar quin canvi provocaria –canvi sinònim, substitució d'aminoàcid, canvi en splicing, canvi a codó stop...– i, finalment, predir el seu impacte funcional, que és especialment important en les substitucions d'aminoàcid, ja que el rang d'efectes possible va des d'essencialment neutre fins a molt greu o deleteri (Chakravorty i Hegde 2017).

Amb la constant millora de la velocitat i del preu de les tècniques de seqüenciació, l'anàlisi computacional de les dades generades pot convertir-se en el coll d'ampolla més important (Mardis 2010). Desenvolupar eines ràpides i, sobretot, precises, esdevé clau per reduir el sobrecost dels anàlisis bioinformàtics. Hi ha dos fases bàsiques: **i)** el processament tècnic i l'alineament de les lectures al genoma de referència i **ii)** la detecció i anotació dels canvis de nucleòtid. És en aquest segon punt on hi ha més variabilitat depenent del tipus d'anàlisi que es realitza, i hi ha múltiples programes i algoritmes que han sorgit per adreçar-los.

Per exemple, en un context oncològic en el que es fa l'anàlisi per buscar mutacions somàtiques, es creua el genoma de les cèl·lules normals –habitualment cèl·lules sanguínies– amb el genoma de les cèl·lules cancerígenes, i les diferències resultants s'anoten com a canvis somàtics (De Mattos-Arruda et al. 2015). Mètodes destacats per a buscar mutacions somàtiques són MuTect

(Cibulskis et al. 2013), que utilitza un classificador Bayesià per a filtrar variants amb baixa qualitat i per a descobrir els millors candidats somàtics; VarScan2 (Koboldt et al. 2012), que fa ús del test exacte de Fisher tenint en compte el nombre de lectures que donen suport a un o altre nucleòtid, i si el valor *p* supera un punt de tall, considera el canvi com a somàtic; o Smufin (Moncunill et al. 2014), que es salta el pas d'alinear a un genoma de referència i compara directament les lectures del cas normal amb les del cas tumoral, i només mapeja al genoma de referència les zones amb canvis potencials. Tots aquests mètodes reporten un bon funcionament, però encara hi ha fonts d'error que condueixen a possibles falsos positius, deguts per exemple a mals alineaments en les zones amb indels. Es pot usar una combinació de diversos mètodes (Fang et al. 2015) per millorar tan la precisió com la sensibilitat. És un camp en el que encara hi ha marge de millora (Hofmann et al. 2017), i que ha d'anar evolucionant paral·lelament als avenços en les tècniques de seqüenciació.

En el cas de les malalties mendelianes, e.g. malalties rares, els anàlisis per a trobar els SNPs hereditaris són diferents, ja que la comparació es fa respecte el genoma de referència (Warden et al. 2014). El mètode que més s'utilitza és el paquet de GATK (De Pristo et al. 2011), que a través de realinear les zones sospitoses de contenir variació extreu uns valors de versemblança per a cada possible variant, i finalment aplica la regla de Bayes per a decidir si marcar-lo com a SNP.

Un cop identificades les variants potencials, tots els tipus d'anàlisis acaben en el mateix punt: l'anotació de quin canvi es produeix i l'anotació del seu efecte funcional. Els programes més utilitzats són ANNOVAR (Wang et al. 2010) i SnpEff (Cingolani et al. 2012). Els dos s'encarreguen d'anotar cada

canvi de nucleòtid amb el canvi pràctic que provoca –en quina regió gènica està, si es dóna en un intró, si es dóna en una regió codificant si el canvi és sinònim o no, si està en una zona reguladora o d'splicing...–; les freqüències poblacionals, si és un canvi ja reportat, també poden filtrar segons el tipus de canvi o segons la freqüència; i finalment també poden anotar tipus de canvis específics, i. e. substitucions no sinònimes, amb el seu potencial efecte funcional. Els anàlisis de NGS acaben produint una llista de potencials variants causals (Jalali Sefid Dashti i Gamieldien 2017). El pas següent, que precedeix la decisió mèdica acaba depenent de la combinació de les prediccions dels efectes funcionals amb la informació que es té del pacient. És molt important, llavors, que les prediccions emprades tinguin la màxima exactitud possible, i donar informació complementaria per a que les decisions finals estiguin el millor informades.

Els mètodes més utilitzats en la predicció de l'efecte funcional són SIFT, Polyphen2 i CADD. SIFT (Sim et al. 2012) utilitza una mètrica derivada de la composició en la posició de l'alineament múltiple, la normalitza, i hi aplica un punt de tall per decidir la patogenicitat; Polyphen2 (Adzhubei et al. 2010) utilitza diversos descriptors heterogenis per a predir l'efecte mitjançant un classificador bayesià ingenu; i CADD (Kircher et al. 2014), que prediu en totes les posicions del genoma –no només en canvis no sinònims com els dos anteriors–, mesura el possible impacte amb multitud de descriptors mitjançant una Support Vector Machine.

Més recentment, s'han publicat uns mètodes amb un bon potencial de predicció de variants són: DEOGEN2 (Raimondi et al. 2017), que fa ús d'una dotzena de descriptors de diversos orígens per a predir la patogenicitat dels

canvis d'aminoàcid amb un Random Forest de 200 arbres; PMUT2017 (López-Ferrando et al. 2017), que utilitza dotze descriptors obtinguts en una *feature selection* sobre més de 100 descriptors, prediu substitucions no sinònimes també amb un Random Forest; EVMutation (Hopf et al. 2017), un mètode no supervisat que utilitza informació de totes les posicions d'un alineament múltiple per calcular l'efecte quantitatiu d'un canvi d'aminoàcid en la funció del gen. Finalment hi ha PhD-SNPg (Capriotti i Fariselli 2017), que utilitza un algoritme de gradient boosting, amb informació extreta de dos alineaments, per a predir l'impacte en qualsevol posició del genoma.

El panorama de la predicció de l'impacte funcional segueix evolucionant. Encara no hi ha un algoritme que sigui netament superior a la resta, malgrat que any rere any apareixen noves iteracions millorant algorismes anteriors amb els nous coneixements –PMUT2017 o PhD-SNPg el 2017, per exemple–, i segueixen creant-se nous mètodes aprofitant paradigmes diferents –EVMutation–, tot buscant percepcions o idees noves per a dur la predicció i la comprensió de les variants un pas més enllà, de forma que els anàlisis de NGS puguin obtenir millor informació.

Part de l'èxit de la medicina personalitzada depèn de l'evolució d'aquests predictors, i de la capacitat d'estendre la seva potència a nous problemes mèdics. Concretament, cal crear mètodes que ampliïn i especifiquin més detalls dels efectes funcionals del canvi, com la possible edat d'aparició de la malaltia, la gravetat, els teixits i/o xarxes gèniques més afectats, etc. ; tot aplicant els millors mètodes matemàtics i els algoritmes més eficaços; permetrà anar un pas més enllà i apropar definitivament l'anàlisi del genoma a la clínica, tot revertint en una millora per als pacients.

En aquesta tesi s'estudia la relació entre un tipus especial de variant patogènica, els CPD, i dos problemes de rellevància mèdica: la predicció de la severitat en patologies mendelianes (Capítol II) i l'impacte de l'incidentaloma (Capítol III). Aquests temes s'expliquen amb més detall en els seus capítols corresponents per evitar una heterogeneïtat excessiva en aquesta introducció.

Els CPDs són unes mutacions patogèniques especials, ja que es troben per defecte en altres organismes. Estudis previs (Ferrer-Costa 2007) suggereixen que el seu impacte molecular és menor que la mitjana de les variants patogèniques. Aquesta propietat suggereix que puguin estar associades a versions més tènues de la malaltia, fet que conferiria una capacitat predictiva en el problema de la severitat. Això s'estudia en el capítol I d'aquesta tesi. En aquest capítol es deriva, també, que hi ha una possible relació de les variants que produeixen poc impacte molecular, i la gravetat de les malalties resultants; que s'estudia en el capítol II. Finalment, i arrel de la participació en anàlisis d'exomes de glioblastoma, s'observa que la presència de variants tipus CPD podria tenir un impacte negatiu en els processos de diagnòstic. El capítol III tracta de l'estudi d'aquest problema, és a dir, de la quantitat de variants amb característiques de CPD que apareixen en experiments de seqüenciació.

1.5 Objectius de la Tesi

- Explorar la relació entre la severitat en l'impacte molecular d'una variant i que una variant estigui compensada.

- Caracteritzar els paràmetres moleculars de les variants compensades.

- Analitzar la predicció de la severitat del fenotip en el cas de les hemofílies.

- Caracteritzar les diferències dels paràmetres d'impacte moleculars entre les variants que causen fenotips lleus i les variants que causen fenotips greus.

- Analitzar la predictibilitat de les variants amb possible compensació.

- Explorar la possible fracció de variants compensades que apareixen en l'incidentaloma humà.

Capítol I

2. Relació entre l'impacte molecular i el fenotip: El cas dels CPD

2.1 Introducció

2.1.1 Les desviacions patogèniques compensades

La predicció de l'efecte que produeix un canvi d'aminoàcid és un dels elements clau en el diagnòstic clínic (MacArthur et al. 2014), i és una eina important en la inferència del fenotip d'un individu a partir del seu genotip. Els mètodes que tradicionalment funcionen millor per a predir variants patogèniques, com Polyphen2 (Adzhubei et al. 2010), que tot i tenir més de 7 anys segueix acumulant moltíssim ús –més de 1000 cites el 2016– o PON-P2 (Niroula et al. 2015), que és el predictor que reporta millor funcionament en revisions recents (Riera et al. 2016), utilitzen diversos descriptors amb característiques extretes d'alineaments múltiples (**M**ultiple **S**equence Alignments, MSA). Aquests descriptors proporcionen diferents mesures del grau de conservació entre espècies de la seqüència proteica al «locus» de la variant, e.g. la conservació en aquella posició o l'aparició o no de l'aminoàcid del que s'intenta determinar l'efecte. L'ús combinat d'aquests predictors, ens permet estimar la importància d'aquella posició en la integritat i el funcionament de la proteïna (Riera et al. 2014). Efectivament, els alineaments múltiples, i els patrons de conservació que s'hi poden trobar, reflecteixen milions d'anys de restriccions en els possibles canvis d'aminoàcid, i el seu ús en predicció ens permet emprar tota la informació acumulada al llarg de l'evolució per veure si una substitució concreta és permesa en aquella posició. De fet, s'ha observat que existeix una correlació significativa entre el grau de conservació d'un aminoàcid al MSA i la variació en energia lliure associada a la seva mutació (Riera et al. 2014). Malauradament, assignar tant pes als descriptors derivats del MSA pot donar lloc a que petits detalls, com un mal

alineament o un error de seqüenciació, puguin canviar per complet la predicció d'una variant (Colobran et al. 2016). Per exemple, el resultat de Polyphen2 per l'efecte de la substitució patogènica F376V en la proteïna FOXP3 és neutre per culpa de la seqüència de cavall, que té errors clars que introdueixen l'aminoàcid Valina a la posició 376, donant com a resultat una predicció errònia de neutre, quan el mateix alineament sense la seqüència defectuosa dóna la predicció correcta. Casos com l'anterior són deguts a mals alineaments en llocs puntuals, però la natura té un reservori encara més gran de casos així, que en aquest cas són correctes: al menys el 10% de totes les variants patogèniques en humà que s'han descrit es donen de forma natural en altres organismes (Jordan et al. 2015). Gràcies al temps que han tingut els processos evolutius per explorar possibles camins en l'espai de les seqüències, en algunes branques de l'arbre filogenètic la proteïna ha anat canviant la estructura de tal manera que, en algun moment en el temps, les característiques de la molècula fan que una substitució, que en una altra espècie seria deletèria, sigui neutra.

Aquest tipus de canvis (Figura 2.1) són anomenats Desviacions Patogèniques Compensades –**C**ompensated **P**athogenical **D**eviations a la literatura, o CPD– per coherència amb els primers treballs exhaustius dedicats al tema (Kondrashov et al. 2002), on també es va hipotetitzar que la compensació hauria d'estar dins de la mateixa molècula i estructuralment a prop del canvi que modula. Cal dir, però, que el concepte de compensació és encara més antic: Zuckerkandl i Pauling havien trobat un residu patogènic en humans a la seqüència d'hemoglobina d'orangutan (Zuckerkandl i Pauling 1962), i Motoo Kimura, el pare de la teoria neutra de l'evolució, ja havia estudiat el potencial rol evolutiu d'aquestes variants el 1985 (Kimura 1985).

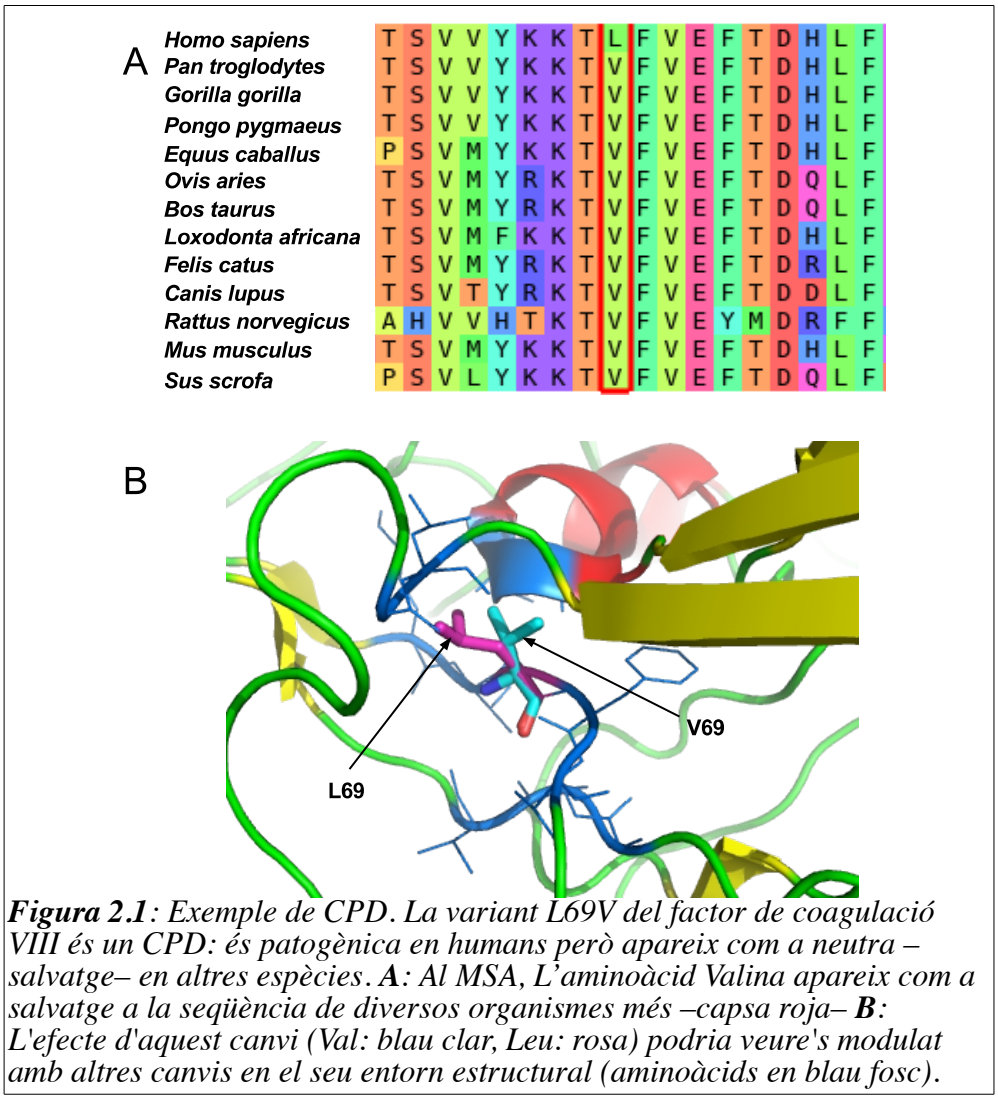

***Figura 2.1**: Exemple de CPD. La variant L69V del factor de coagulació VIII és un CPD: és patogènica en humans però apareix com a neutra –salvatge– en altres espècies. **A**: Al MSA, L'aminoàcid Valina apareix com a salvatge a la seqüència de diversos organismes més –capsa roja– **B**: L'efecte d'aquest canvi (Val: blau clar, Leu: rosa) podria veure's modulat amb altres canvis en el seu entorn estructural (aminoàcids en blau fosc).*

Més enllà dels primers estudis, s'han anat descrivint possibles mecanismes de compensació pels quals l'efecte potencialment deleteri en humans de certes substitucions podria suprimir-se. Aquests mecanismes poden ser tan canvis en una posició concreta de la proteïna (Barešić et al. 2010); com una acumulació de substitucions que, de forma additiva, han anat fent més robusta la proteïna i que, en un moment donat, li permeten absorbir substitucions que canvien molt

l'estabilitat de la proteïna (Xu i Zhang 2014; Starr i Thornton 2016); fins i tot, es poden trobar canvis compensatoris en altres proteïnes amb les que la proteïna patogènica interacciona o que formen part de la mateixa xarxa metabòlica (Lehner 2011).

Així doncs, la compensació també es pot definir com un fenomen epistàtic (Lehner 2011) que amplia els possibles camins evolutius d'una proteïna, tot obrint-li noves vies que, d'altra forma, tindria vetades (De Pristo et al. 2005). La hipòtesi de les compensacions ha estat fins i tot estudiada a través de models in vitro (Jordan et al. 2015), on es demostra experimentalment que canvis en la mateixa proteïna poden rescatar el fenotip patogènic. A més, hi ha diversos estudis in silico confirmant que, si es fa servir l'energia lliure de l'estructura com a representació de la fitness, un residu o una combinació de residus compensen l'efecte molecular del canvi (Rockah-Shmuel et al. 2015).

En paral·lel, alguns treballs han explorat si la probabilitat de compensació és la mateixa per a totes les variants patogèniques o si, en canvi, la seva natura modula la probabilitat de trobar-les compensades. Els primers resultats van indicar que és el segon cas (Ferrer-Costa et al. 2007): és difícil trobar com a CPD mutacions amb efectes moleculars molt disruptius, (e.g. a localitzacions enterrades que són fonamentals en l'estructura proteica, o que involucren canvis fisicoquímics molt severs). Així doncs, els CPD serien variants amb un efecte molecular més lleu que les variants patogèniques comunes (Barešić et al. 2010).

2.1.2 Severitat i fenotip dels CPD

Com s'ha vist, els coneixements sobre les propietats moleculars dels CPD augmenten ràpidament. Malauradament, encara se sap ben poc sobre el seu

impacte al fenotip, particularment en aspectes del fenotip «macroscòpics» com canvis morfològics, o fisiològics globals, severitat, etc. S'ha vist que les dades de patogenicitat de les malalties poden ser una primera font de coneixements per a resoldre aquest problema.

La severitat d'una variant, a més a més, és una propietat que, malgrat estar poc estudiada degut a la baixa quantitat de dades disponibles (Niroula i Vihinen 2017), pot resultar clau a l'hora d'oferir bons diagnòstics en estudis de NGS. Tenir informació no només del possible efecte patogènic d'una variant, sinó de si el seu possible impacte fenotípic és més greu o més lleu, serà un pas molt important en oferir proves de medicina personalitzada de millor qualitat (Niroula i Vihinen 2016) ja que permetrà estratificar les variants en dos grups. El primer corresponent a les variants d'efecte patogènic molt greu –gran impacte fenotip– que és més probable que donin lloc a malalties potencialment fatals, i que, per tant requereixen una actuació immediata. El segon grup correspon a les variants que, malgrat ser disruptives, donen lloc a una versió més lleu de la malaltia (impacte menor en el fenotip) en la que és menys urgent actuar (Green et al. 2013).

Lligant l'aspecte fenotípic amb l'existència dels CPD, llavors, la pregunta seria: Hi ha alguna relació entre la severitat d'una malaltia i que la variant associada pugui estar compensada?

Aquesta primera part del treball intenta esbrinar la relació entre la severitat reportada de les malalties associades a una variant patogènica –el subconjunt de canvis missense– amb que aquesta variant pugui ser un CPD. Es fa una caracterització de les variants amb un conjunt de descriptors que mostren l'impacte molecular i el canvi en el patró de conservació evolutiu que produeix

el canvi d'aminoàcid. Els canvis en aquests descriptors degut a les variants patogèniques es relacionen amb els corresponents fenotips clínics de la malaltia. Cal esmentar el calor d'aquesta relació més enllà de les aplicacions biomèdiques, ja que el fenotip clínic pot interpretar-se com una representació pràctica de l'efecte de la variant en la fitness d'un individu. Per tant, aquest estudi també té implicacions en teoria evolutiva.

Finalment, s'utilitzen les dades obtingudes per a comprovar si les representacions actuals de la fitness, com la $\Delta\Delta G$, poden extrapolar-se de forma general o si caldria tenir variables més precises.

2.2 Materials i mètodes

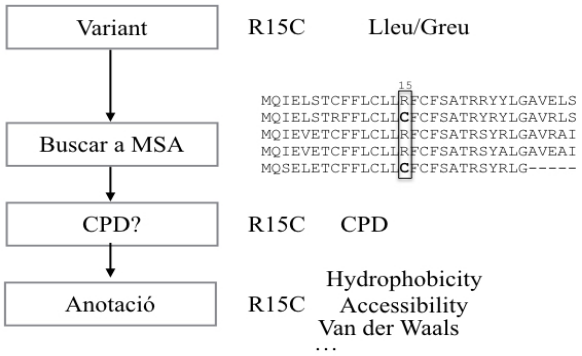

***Figura 2.2**: Resum del procés seguit amb cada variant patogènica.*

L'estudi s'estructura en quatre passos ben definits. **Primer**, l'obtenció de les dades de variants patogèniques amb notació de severitat (lleu o greu). **Segon**, la construcció d'alineaments múltiples d'ortòlegs dels quals poder-ne extreure les

variants que es consideren CPD. **Tercer**, la caracterització mitjançant descriptors a nivell molecular d'aquestes variants: propietats calculades a partir de l'estructura tridimensional de la proteïna, propietats evolutives que es deriven dels MSA i propietats fisicoquímiques. I **quart**, s'utilitza la informació recollida per a estudiar la relació entre impacte molecular i efecte fenotípic als CPD. El procés resumit (passos 1-3) es mostra a la Figura 2.2

2.2.1 Les variants patogèniques amb notació de fenotip

La base de dades de la que s'extreu la informació bàsica de patogenicitat és UniProt (Famiglietti et al. 2014, The UniProt Consortium 2017). Es cerca, mitjançant un programa escrit en Python (Van der Waalt et al. 2011, The Python Software Foundation), totes les variants patogèniques –anotades com a DISEASE a UniProt– que tenen anotació de severitat. Els mots clau que buscats són *mild*, que correspondrà a la categoria de lleu, i *lethal*, *severe* i *acute,* que es consideren com a categoria severa.

Per als casos del Factor de Coagulació VIII (F8 a partir d'ara) i el Factor de Coagulació IX (F9), s'obté també informació de les bases de dades de CHAMP i de CHBMP respectivament (CDC Hemophilia Mutation Project), i s'apliquen els mateixos criteris que a les variants d'UniProt. La severitat en aquestes variants s'obté a partir del camp de 'Reported Severity', que correspon a la presentació fenotípica de les seves malalties associades (Hemofílies A i B, respectivament).

Finalment, es seleccionen només aquelles proteïnes en les que tenim, com a mínim, cinc variants en cada categoria de severitat –per un total de, com a mínim 10 variants amb notació de severitat–.

2.2.2 L'obtenció de CPD

La classificació de variants en CPD o no CPD es fa en base a alineaments múltiples de seqüència (MSA) de proteïnes ortòlogues (Figures 2.1 i 2.2)

Les seqüències amb les que es construeixen els MSA s'obtenen cridant automàticament la interfície de programació d'aplicacions (API) d'ENSEMBL (release 80), (Aken et al. 2016) amb scripts de Python, i agafant les proteïnes ortòlogues amb humà a diversos organismes model. Es creen dos conjunts: un amb totes les seqüències i un altre restringint taxonòmicament als mamífers. Les seqüències de cada conjunt s'alineen mitjançant el programa Muscle (Edgar 2004) amb les opcions per defecte, i els MSA resultants s'utilitzen per a buscar possibles variants compensades. Com que el nombre d'organismes a ENSEMBL és baix, l'alineament d'ortòlegs tindrà bona qualitat i la probabilitat de tenir posicions mal alineades que puguin donar lloc a falsos positius és relativament baixa. Els resultats d'aquest capítol s'han generat amb els MSA de mamífers.

L'algoritme usat per a detectar CPD és el següent: Primer es busca la posició de la variant en la seqüència humana. Tot seguit es mira quin aminoàcid tenen la resta d'organismes en aquest posició de l'alineament. S'obté la conservació de la seqüència en aquesta posició –nombre de residus idèntics a l'humà dividit pel nombre de seqüències en el MSA– i la proporció de *gaps* –nombre d'espais buits en aquesta posició entre nombre de seqüències–; per a identificar CPD en posicions molt poc conservades o amb massa gaps, que corresponguessin a errors. Finalment, es considera que un residu és CPD si la variant patogènica en humà apareix al menys un cop com a 'wild-type' en una altra seqüència, de forma similar al que ja s'ha usat anteriorment per a descriure CPD (Kondrashov et al. 2002, Ferrer-Costa et al. 2007). La diferència és, però, que el pas actual

per assegurar la fiabilitat de la categorització del residu és en la construcció de l'alineament –MSA curts però només amb ortòlegs i de molta qualitat– enlloc d'aplicar nombrosos filtres de qualitat de la posició en un pas posterior.

Utilitzar poques seqüències al fer servir els ortòlegs d'ENSEMBL podria fer que es perdessin com a CPD variants que, potencialment, haurien aparegut usant un mètode amb el criteri més ampli. El problema de fons és quin equilibri es vol entre la quantitat de falsos positius i de falsos negatius obtinguts: amb un criteri més ampli s'obtindria més soroll –falsos positius– però també més possibles CPD. Seguint aquest protocol, s'obté millor qualitat malgrat la possibilitat d'haver deixat casos sense trobar. Per als objectius d'aquest estudi, que són de caire general i descriptiu, es va determinar que era més important tenir una caracterització acurada dels CPD, tot i deixar-ne alguns sense classificar, que agafar totes les variants que tenien certa opció de tenir compensació.

En total, s'obtenen 130 CPD associats a fenotips lleus i 86 CPD associats a fenotips greus. D'aquests, 124 i 50 corresponen als factors de coagulació VIII i IX, respectivament, i és en aquests on es centra la part més analítica d'aquesta primer capítol.

2.2.3 La caracterització molecular de les variants

Les propietats moleculars, que s'han escollit per a descriure l'impacte molecular dels CPD, es poden classificar en tres grans grups: i) descriptors de caràcter fisicoquímic, i intrínsecs al canvi de residu, ii) descriptors que provenen dels alineaments múltiples, i que per tant tenen un caire evolutiu i iii) descriptors derivats d'estructura –quan hi ha estructura tridimensional resolta

experimentalment–. Malgrat que hi ha una gran quantitat de possibles descriptors a escollir, ja que en estudis de predicció de patogenicitat s'han arribat a usar més de mil possibles característiques (Niroula et al. 2015), s'han triat set descriptors que conjuntament representen adequadament les diferències funcionals i fisicoquímiques de la proteïna variant envers la normal (Riera et al. 2014).

Representant els canvis en les **propietats fisicoquímiques** derivades de la substitució de l'aminoàcid tenim: el canvi d'hidrofobicitat que produeix la substitució (Fauchere i Pliska 1983), el canvi en el volum de Van der Waals (Bondi 1964), i el valor de la matriu de substitució de Blosum62 (Henikoff i Henikoff 1992).

Per a descriure la **conservació d'aminoàcid** de la posició de la variant en el MSA s'usa l'entropia de Shannon (Shannon 1948), que en el sentit matemàtic estricte és una representació de la informació continguda en aquella posició, i que en termes evolutius explica com de conservada està aquella posició de la seqüència. A més a més, s'usa el paràmetre $pssm_{nat}$ (Ng i Henikoff 2001), que representa el grau de conservació de l'aminoàcid natiu al lloc de la mutació, normalitzat per la seva freqüència en el MSA

L'entropia de Shannon es calcula com a $H(x) = -\sum p_i \log(p_i)$, on p_i és la freqüència de l'aminoàcid de tipus *i* –i pot ser Ala, Trp, Gly, etc– en la posició del MSA. El $pssm_{nat}$ es calcula com el $\log(p_{nat,j} / p_{nat,MSA})$, on $p_{nat,j}$ és la freqüència de l'aminoàcid natiu a la posició j, i $p_{nat,MSA}$ és la freqüència de l'aminoàcid en tot l'alineament.

L'impacte a nivell estructural es mesura usant els complexes

tridimensionals experimentals dipositats al Protein Data Bank –PDB– (Berman et al. 2000) mitjançant dos programes externs. Es calcula el canvi en l'estabilitat de la proteïna, entès com a canvi en l'energia lliure entre el mutant i l'estructura '*wild-type'* (Δ*ΔG),* mitjançant FoldX (Schymkowitz et al. 2011), i mesurada amb NACCESS (Hubbard et al. 1993) l'accessibilitat al solvent relativa d'aquella posició. Cal esmentar que l'accessibilitat relativa no sempre es mou en valors entre 0 i 1 –o 0 i 100, si es treballa amb percentatges–, ja que el càlcul de NACCESS es fa usant l'accessibilitat en el tri-pèptid Ala-X-Ala, on X és l'aminoàcid en qüestió usat com a referència. Com que en algunes proteïnes hi ha casos amb angles inusuals o que tenen una geometria distorsionada, poden aparèixer canvis amb l'accessibilitat relativa major a 1 (documentació a Hubbard et al. 1993).

El canvi d'estabilitat, o Δ*ΔG,* estimat mitjançant Fold-X, és una mesura que s'ha utilitzat habitualment com a representació de la fitness –una mesura relativa de la probabilitat de l'individu de deixar descendència– però també s' usa sovint a com a indicador de la viabilitat de la proteïna –manteniment de la integritat estructural, de la funció...–. Diverses teories evolutives biofísiques (Gong et al. 2013; Khatri i Goldstein 2015) utilitzen aquesta interpretació dels valors de Δ*ΔG* per inferir els possibles camins de seqüència que es podrien haver donat en la proteïna. En aquest sentit, es podria considerar que Δ*ΔG* és el descriptor més important per raons de la literatura: és el que més s'utilitza, fora d'aquest context de predicció, per mesurar *in silico* la funció i estabilitat de la proteïna.

El canvi a nivell estructural només es pot mesurar en aquells residus que apareixen en alguna estructura experimental per la proteïna d'interès. És a dir,

que els anàlisis fets amb aquests descriptors queden restringits al subconjunt de residus amb els que hi ha informació estructural. Per a la proteïna F8 s'usa l'estructura dipositada al PDB (Berman et al. 2000) 2R7E, i per a F9 es fan servir les estructures amb codis 1CFH, 1IXA i 3LC5, que cobreixen respectivament diverses parts de la proteïna. El mapatge entre la seqüència d'aminoàcids i la posició en l'estructura s'ha fet amb el servei d'API de pdbsws (Martin 2005) a través d'scripts de Python. Aquest mapatge és necessari degut a que la numeració del PDB pot no coincidir amb la d'UniProt, atès que a vegades comença a 1 independentment de la posició en la seqüència consens, i a vegades comença en altres punts per raons històriques.

2.2.4 Anàlisi comparatiu del fenotip de la malaltia, filogènia i mètodes computacionals

Addicionalment als descriptors calculats, es fa servir una estratègia més per a comprovar, en aquest cas, com afecta una CPD al fenotip de més alt nivell. En concret, s'empren anotacions del fenotip de la malaltia relacionada amb canvis a la proteïna; i també s'utilitzen arbres filogenètics per visualitzar en quin punt apareix un CPD en el procés evolutiu.

La **visualització de termes fenotípics d'alt nivell** es fa a partir dels termes estandarditzats de Human Phenotype Ontology (HPO) (Köhler et al. 2017).

Per a aquesta representació es fan servir totes les proteïnes amb qualsevol anotació de severitat (lleu o greu), sense cap filtre per al nombre de variants mínim. Es busca si alguna variant severa de la proteïna, si n'hi ha, és CPD –segons alineaments amb mamífers d'Ensembl, seguint el mateix protocol que al pas 2.2.2– , i es fa el mateix amb les variants lleus. Després, es filtra per a

conservar només els gens per als que hi ha algun CPD i se'n fan dos subconjunts: un per a gens amb CPD de variants d'efecte fenotípic greu, i un altre per a gens amb CPD lleus. Aquests són els casos per als que es farà l'anàlisi descrit a continuació..

Amb les proteïnes seleccionades, es busca primer tota l'ontologia HPO de cadascun dels gens. Després es seleccionen les malalties amb la paraula clau *anormalitat de* (literalment *'Abnormality of'*), tot excloent els termes massa generals –com *anormalitat de l'abdomen* (*'Abnormality of the abdomen'*)– i eliminant la redundància entre els termes, traient tots aquells que estan inclosos en algun altre, e.g. excloent *anormalitat de la fisiologia del sistema nerviós* ('*Abnormality of the nervous system physiology'*) si ja tenim inclosa *anormalitat del sistema nerviós* ('*Abnormality of the nervous system'*). Finalment, es recullen els termes resultants per a cada malaltia i es guarden tots aquells termes que apareixen sovint en totes les proteïnes d'interès. Al final hi haurà una selecció de termes d'alt nivell d'HPO, amb un recompte per als casos lleus, un altre per als severs, i un últim tenint en compte totes les proteïnes.

La **representació mitjançant un arbre filogenètic** –en forma de cladograma– de les seqüències dels diferents organismes es fa a partir dels MSA. La filogènia de les seqüències es computa mitjançant el mètode de màxima parsimònia, i els arbres es creen mitjançant el paquet de R (R Core Team 2013) de ggtree (Yu et al. 2017). A més a més, es representa gràficament en els arbres quins organismes tenen un CPD i quins no. Això pot permetre veure en quin punt de l'arbre evolutiu apareix el CPD.

Finalment, els gràfics que apareixen en aquest capítol han estat realitzats a través del jupyter notebook (Pérez et al. 2007), amb el paquet de Python

matplotlib (Hunter 2007), usant esporàdicament funcions dels paquets de *numpy* (van der Walt et al. 2011) i de *pandas* (McKinney 2010).

2.3 Resultats

Els resultats representen una aproximació esgraonada a l'estudi de la relació entre l'impacte molecular i el fenotip –la severitat– en el cas dels CPD. En primer lloc s'observa com els CPD poden estar associats tant a versions lleus com greus de l'hemofília. En segon lloc, es caracteritza l'impacte molecular dels CPD en relació a les variants no CPD. Finalment, es descriu la relació entre impacte molecular i severitat.

2.3.1 La severitat en CPD

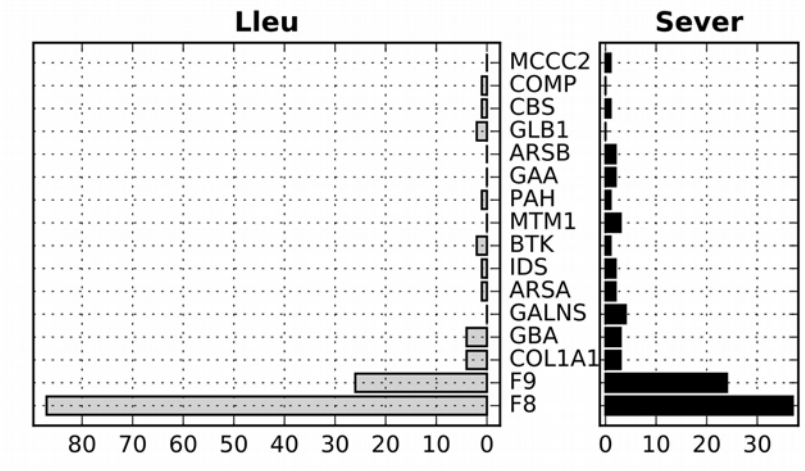

Figura 2.3: *Figura de piràmide mostrant el nombre de CPD trobats en cada proteïna candidata per les categories de lleu i de sever.*

Els CPD estan distribuïdes al llarg de tot el possible espectre de la severitat (Figura 2.3). És a dir, es troben CPD en ambdós casos possibles: en lleus i en severes. Tanmateix, en F8 i F9, hi ha una petita inclinació significativa a trobar més CPD en variants lleus (test estadístic χ^2, p-valor < 0.01 i p-valor < 0.01 per

F8 i F9 respectivament), que per raons de quantitat són les proteïnes que d'aquí en endavant s'estudiarà: fer servir de cara a les estadístiques els descriptors en proteïnes per a les que hi ha tan pocs CPD pot aportar soroll i variabilitat en forma d'*outliers*, o, en el millor dels casos, simplement no aportar res. Addicionalment, les dades estaran tan esbiaixades per les contribucions de F8 i F9 que el que s'observi serà molt semblant als resultats per a aquestes proteïnes. Per tant, es considera que és millor per aquest estudi acceptar el des-balanç de les dades i centrar l'anàlisi en dos casos concrets: es podrà fer una caracterització més exhaustiva i rigorosa en F8 i F9. Potser el resultat peca de ser menys generalitzable, però, sigui com sigui, a mida que vagin apareixent noves dades, qualsevol nova investigació podrà tenir un punt de partida sòlid basat en els casos estudiats aquí.

El percentatge de CPD trobat en aquesta població de variants està al voltant del 10%, xifra similar –tot i que lleugerament més alta en el cas de les variants lleus– al que s'ha descrit en treballs anteriors. (Ferrer-Costa et al. 2007, Barešić et al. 2010).

Proteïna	Lleu		Greu	
	CPD	*Total*	CPD	*Total*
F8	87	493	37	480
F9	26	119	24	275

Taula 2.1. *- Nombre de CPD per a F8 i per a F9 en les dos categories estudiades de severitat respecte el total de variants respectiu.*

Que no s'observi la tendència de que els CPD tendeixen a ser més lleus en la resta de casos podria venir donat per la ja comentada falta de dades o per una absència real de la tendència. De tota manera, manquen dades per a poder donar

suport a qualsevol de les teories.

En resum, els CPD poden estar associats tan a versions suaus com greus de les malalties, en totes les proteïnes considerades.

2.3.2 L'impacte molecular dels CPD

El resultat anterior suggereix que els CPD obtinguts poden no ser sempre molecularment suaus, en contra d'algunes observacions prèvies (Ferrer-Costa et al. 2007; Barešić et al 2010). Per estudiar aquest punt, es decideix estimar l'impacte molecular dels CPD en relació al de les mutacions que no ho són. De forma general, la comparació de les distribucions per les diferents propietats emprades mostra que els CPD tendeixen a tenir un impacte significativament més lleu que la resta de variants patogèniques (Test estadístic de Mann-Whitney-Wilcoxon (MWW), amb correcció de Bonferroni, significatiu per la majoria de casos, Taula Apèndix-1 A1.1). Malgrat això, en tots els descriptors hi ha casos repartits per tot l'espectre possible de l'impacte (Figura 2.4), fet que remarca que només es tracta d'una tendència estadística i no d'una característica intrínseca: Les mutacions amb un impacte molecular greu serien, generalment, més difícils de compensar, però això no vol dir que no es puguin compensar.

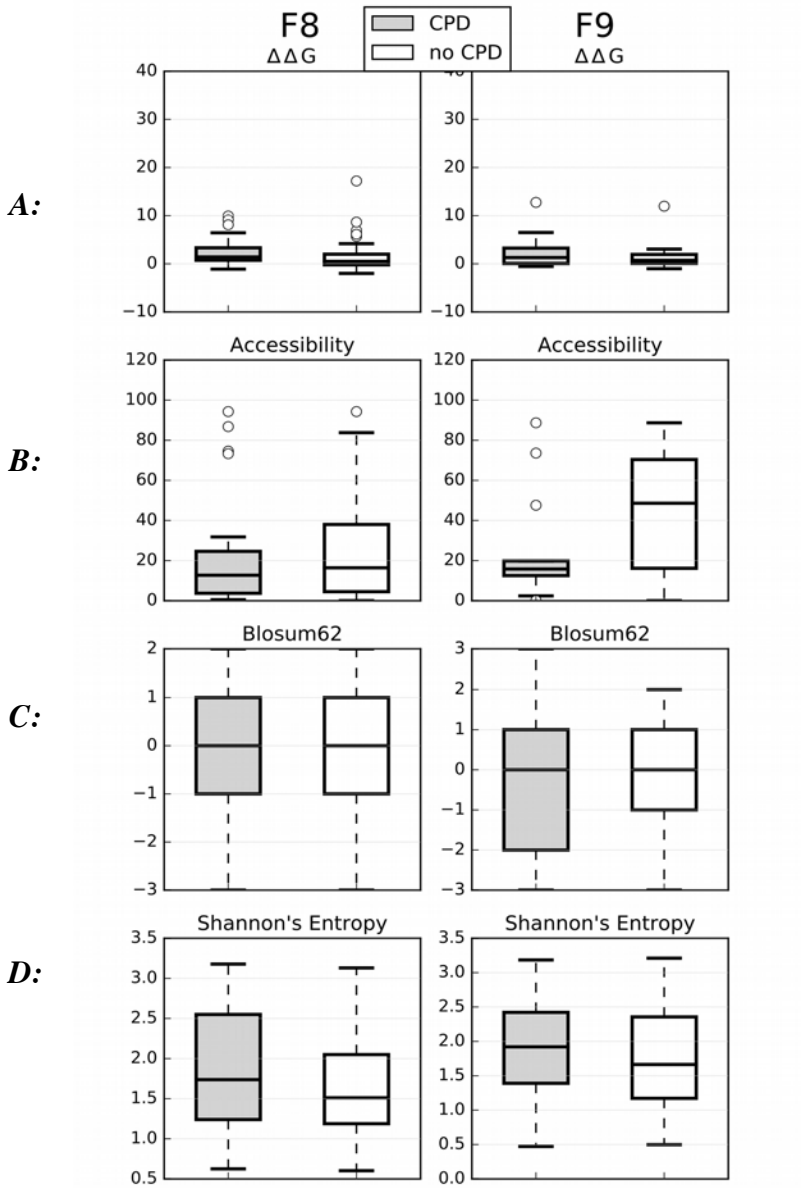

Figura 2.4: *Boxplots il·lustrant les diferències en la distribució d'alguns dels descriptors d'impacte molecular per a F8 i F9.* ***A****:* $\Delta\Delta G$ ***B****: Accessibilitat relativa.* ***C****: Blosum62.* ***D****: Entropia de Shannon.*

El cas de l'accessibilitat relativa –grau d'exposició al solvent d'un residu–, és interessant ja que aquesta propietat està relacionada amb l'efecte estructural de les mutacions (Matthews 1995). Es veu, però, que encara que els CPD tinguin una mitjana més elevada –que indica un efecte molecular més tolerable–, hi ha una superposició molt gran entre les dos distribucions. El mateix succeeix en la resta de paràmetres, tal i com s'il·lustra en la figura 2.4 amb els casos de la $\Delta\Delta G$, el blosum62 o l'entropia de Shannon. Mereix menció especia la $\Delta\Delta G$, ja que és un factor que s'usa en nombrosos estudis (Bershtein et al. 2006; Tokuriki i Tawfik 2009; McKeone et al. 2014; Murakami et al. 2015) com a representació indirecta de la fitness d'una proteïna. Es pot observar que, precisament en el cas de CPD vs no CPD, la diferència que s'observa no és significativa (MWW no significant per a F8 i al límit de no significació per a F9, Taula A1.1).

Les observacions anteriors confirmen un aspecte destacat a les dades: la tendència a la significança no és absoluta, ni entre els descriptors ni entre les proteïnes. Per una banda, l'entropia de Shannon i el valor de la matriu blosum62 tenen diferència significativa en ambdós casos. Per altra banda, la diferència de hidrofobicitat i la de volum de Van der Waals només són significatives en F8. A més a més, $Pssm_{nat}$ no apareix com a significant en cap dels casos, tot i ser una mesura que s'ha usat fructíferament en casos de predicció patogènica, sigui en combinació amb altres descriptors (Riera et al. 2015), o en solitari però amb una versió una mica més complexa, tal i com s'utilitza en el mètode de SIFT (Sim et al. 2012). Aquesta manca de predictibilitat és fins a cert punt sorprenent si es té en compte que un altre descriptor amb caràcter evolutiu –entropia de Shannon– si que ho és. Aquesta

variabilitat de les dades fa que no es pugui generalitzar completament la tendència que s'havia observat inicialment (Ferrer-Costa 2007) a que els CPD tinguin un impacte més lleu.

En resum, els CPD no tenen necessàriament un impacte lleu sobre la proteïna, i això pot explicar parcialment les observacions de les Figures 2.2 i 2.3.

2.3.3 El rang fenotípic dels CPD

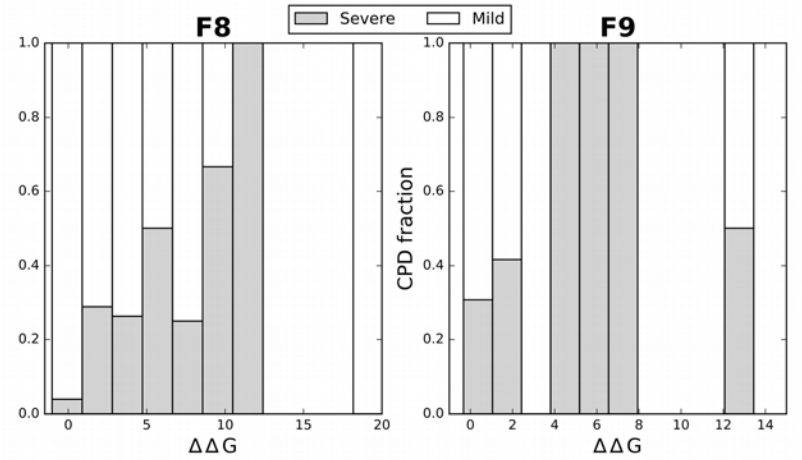

Figura 2.5: *Proporció de CPD amb fenotips lleus i greus en diferents fraccions del ΔΔG per a F8 i F9. La fracció sense pintar correspon a zones en les que no hi ha casos.*

En les seccions anteriors, s'ha observat que els CPD poden estar associats tant a versions greus com versions lleus de la malaltia, en el cas de F8 i F9 (Figura 2.3). També s'ha vist (Figura 2.4) que els CPD tendeixen a tenir un impacte molecular menys fort que els no CPD, malgrat que no és sistemàtic. No hi ha, llavors, una correspondència simple entre impacte molecular –propietats de la proteïna– i fenotip clínic –severitat–, com a mínim en F8 i F9. No es pot

descartar aquesta possibilitat, però, degut a la dispersió de valors observada en la Figura 2.4, en la que es veu que l'impacte molecular dels CPD pot variar substancialment. En aquesta secció s'estudia si aquest és el cas, i si hi ha realment una relació monòtona entre la variació dels paràmetres moleculars i el fenotip clínic; és a dir, si el primer explica el segon. Amb aquest objectiu, es divideixen les poblacions de CPD de F8 i F9 en dos grups: CPD associats a versions lleus i CPD associats a versions greus de la malaltia. L'anàlisi es centra en el paràmetre molecular del canvi d'estabilitat ($\Delta\Delta G$) que proporciona la interpretació més simple i integrada de l'impacte molecular de les variants, i que també està relacionat amb aspectes evolutius. Representar la fracció de CPD associats a versions severes de la malaltia com a funció de $\Delta\Delta G$ (Figura 2.5), s'observa una certa relació creixent entre les dos variables, però també es troba que, fins i tot en valors elevats del canvi d'estabilitat, hi ha CPD associats a casos lleus de la malaltia. La relació que s'aprecia entre l'impacte molecular –$\Delta\Delta G$ en aquest cas– i fenotip clínic no és unívoca.

En concordança amb el resultat anterior, s'observa que els CPD poden trobar-se en tot tipus de gen, independentment de la severitat de les malalties associades. Es pot observar que proteïnes amb CPD estan relacionades amb sistemes en els quals les fallades més greus podrien donar lloc a la mort (Figura 2.6)

Degut a les implicacions d'aquest resultat, es decideix explorar fins a quin punts es dóna aquesta relació tènue entre fenotip i genotip al prescindir de la condició de CPD, i generalitzar els resultats a la població completa de variants patogèniques amb anotació de severitat. Quan es representen aquestes dades per als paràmetres Blosum62, entropia de Shannon i $\Delta\Delta G$ (Figura 2.7 i tests

estadístics de MWW amb bonferroni significatius en la majora de descriptors, Apèndix 1, Taula A1.2), s'observa que hi ha un important grau de solapament entre les distribucions, que indica que, a nivell molecular, un cas greu d'hemofília pot estar associat tant a una variant greu com a una lleu.

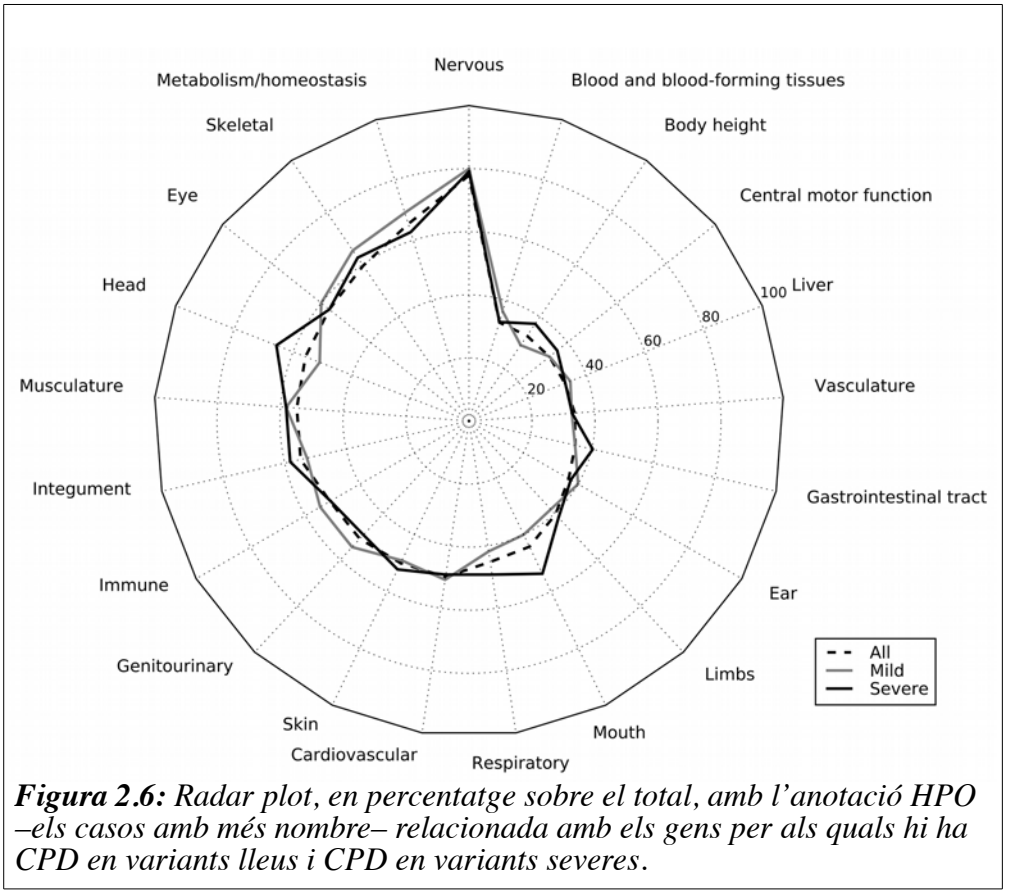

Figura 2.6: *Radar plot, en percentatge sobre el total, amb l'anotació HPO –els casos amb més nombre– relacionada amb els gens per als quals hi ha CPD en variants lleus i CPD en variants severes.*

En resum, s'observa (Figura 2.5, 2.6, 2.7) que no hi ha un mapatge simple de lleu a lleu o de greu a greu, entre impacte molecular i severitat, en els casos estudiats de F8 i de F9. Aquesta observació s'aplica tant als CPD com a les variants patogèniques en general.

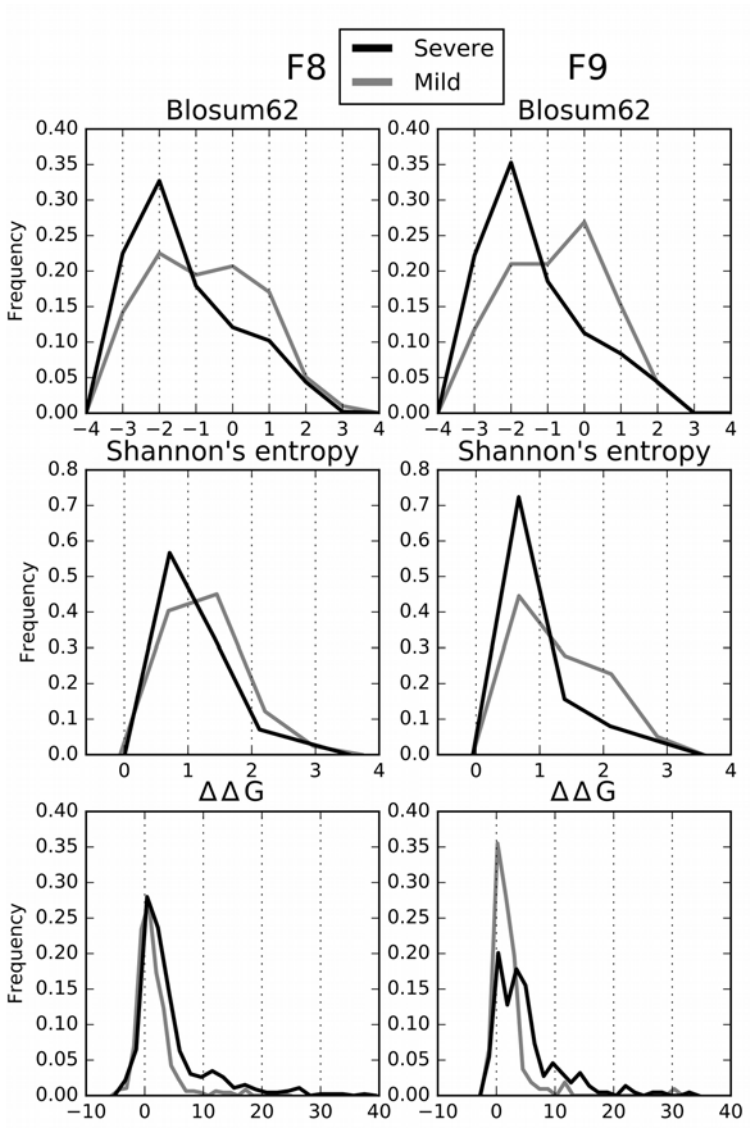

Figura 2.7: *Histogrames de línia mostrant les diferències entre lleus i greus.* ***A****: Blosum62.* ***B****: Entropia de Shannon.* ***C****: ΔΔG de FoldX.*

2.4 Discussió

Els resultats descrits en les seccions anteriors indiquen que el fet que una mutació causal sigui CPD aporta una informació limitada sobre la possible severitat de la malaltia associada. Tot i que aquest resultat redueix l'aplicabilitat clínica dels CPD, poden haver-hi unes inesperades repercussions en el camp de l'evolució de les proteïnes. Els estudis biofísics han donat explicacions al comportament dual, patogènic en alguns organismes i neutre en altres, dels CPD (De Pristo et al. 2005; Ferrer-Costa et al. 2007; Xu i Zhang 2014; Jordan et al. 2015). De fet, ara ja està demostrat tan *in silico* com *in vitro* que els canvis compensatoris són el principal responsable de la supressió dels possibles efectes deleteris (Xu i Zhang 2014; Jordan et al. 2015), i que aquests efectes, en termes de les seves propietats intrínseques moleculars, tenen una certa tendència a ser més lleus (Ferrer-Costa et al. 2007, Barešić et al. 2010). Aquesta relació, fins ara, s'havia estès per a explicar al llarg de l'evolució l'aparició dels CPD mitjançant el descriptor de ΔΔG com una representació de la fitness (De Pristo et al. 2005; Ferrer-Costa et al. 2007; Barešić et al. 2010; Sikosek i Chan 2014). De Pristo i col·laboradors, a més a més, utilitzen una relació determinista entre ΔΔG i la fitness de l'organisme, en la que el canvi de la fitness associat a una mutació varia exponencialment en funció del canvi que produeix en la ΔΔG. Els resultats obtinguts en aquesta secció indiquen que aquests models són inadequats per als CPD. De fet, la relació que s'observa entre la severitat de la malaltia, un fenotip directament relacionat amb la fitness i la ΔΔG –o altres descriptors similars– és feble (Figures 2.4 i 2.7). Aquesta manca de relació queda palesa en el fet que petits canvis en la ΔΔG poden estar associats tan a versions lleus com a versions greus de l'hemofília, la seva malaltia associada

(Figura 2.4 i 2.5).

No hi ha tampoc, tal i com s'observa en la figura 2.5, cap relació determinista entre la fracció de casos greus i la ΔΔG per les variants patogèniques amb notació de severitat.

Com es pot observar a la figura 2.5, hi ha tan casos de gran ΔΔG associats a efectes lleus, com casos amb un canvi mínim de ΔΔG associats a efectes greus. Hi ha CPD en tot el rang d'impacte molecular (Figura 2.4). Els sistemes en els que es manifesten les malalties en les que hi ha CPD per variants lleus, o CPD per variants greus; tenen en ambdós casos implicacions greus (Figura 2.6). És a dir, els resultats obtinguts indiquen una relació probabilística entre ΔΔG i la fitness més que una relació determinista com s'havia usat en alguns estudis.

Aquest factor aleatori molt probablement reflexa la contribució de factors externs al propi gen: sigui per interacció epistàtica amb la resta del fons genètic de l'individu o per qualsevol nombre de factors ambientals que podrien intervenir. És un cas en el que la clàssica relació de F=G+A+GA –Fenotip és igual a la genètica, més l'ambient, més la interacció entre la genètica i l'ambient–, guanya especial importància. Una bona opció per a modelar de forma quantitativa aquesta equació és combinar aproximacions actuals basades en l'estabilitat (De Pristo et al. 2005) amb informació de la resta del genoma de l'individu (Hopf et al. 2017). Això permetria millorar la representació de la relació genotip-fenotip, i ajudaria a desenvolupar un model que superés les limitacions dels basats exclusivament en la ΔΔG.

Focalitzant el cas dels CPD, els termes intergènics en els models de fitness ajudaran a representar millor els possibles efectes de la compensació molecular.

Per això, cal aprofundir en el coneixement dels orígens de la modulació del fenotip de les mutacions (genotip, ambient). Cada cop hi ha més dades que ajuden a entendre aquests fenòmens, sobretot la seva base genètica. Per exemple, Vu i col·laboradors (Vu et al. 2015) utilitzen ARNi per estudiar el fenotip de mutacions de pèrdua de funció en dos línies de *C. Elegans,* descobrint que, en el voltant del 20% dels gens estudiats, la severitat del fenotip resultant canvia. Aquestes diferències només poden ser degudes a la resta de genètica de l'individu i no a la pèrdua de funció en sí, ja que la pèrdua és igual en els dos casos. Per altra banda, hi ha estudis que ajuden a entendre millor sistemes concrets, com en la porfíria eritropoiètica congènita, una malaltia causada per mutacions en el gen UROS, on s'ha trobat que la severitat de la malaltia és deguda a les variants presents en el gen ALAS2, que és un enzim implicat en el procés (To-Figueras et al. 2011). O en l'hemofília, on s'ha vist que hi ha alguns altres factors que poden influir en la severitat, com altres polimorfismes en el mateix gen, canvis en els gens TNF-alfa o IL-10 que estan associats amb més severitat del fenotip clínic, o factors ambientals com l'índex de massa corporal (Pavlova i Oldenburg 2013).

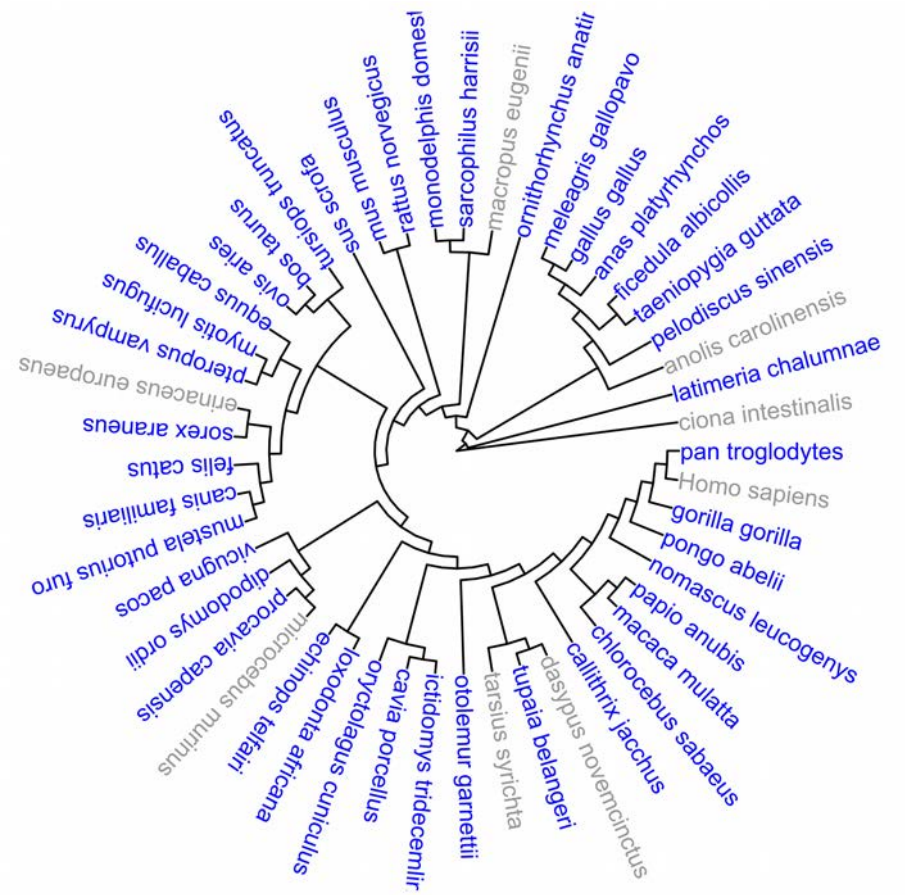

Figura 2.8: *Arbre evolutiu construït a partir de les seqüències de F8. En blau, els organismes que tenen a la posició 69 una Valina. En gris els que no. «Homo sapiens» es troba a la part del mig, a la dreta. Una Valina és patogènica en humà.*

Addicionalment, els estudis evolutius poden ajudar a localitzar efectes compensatoris intragènics, analitzant l'aparició dels CPD en diversos punts de l'arbre filogenètic. És important, ja que els CPD poden donar-se tant en branques concretes, com en múltiples punts, associats a diferents canvis de seqüència. No hi ha tampoc un enriquiment considerable a les espècies de la

part més llunyana d'humà de l'arbre evolutiu. Fins i tot, hi ha alguns casos on el que podria haver succeït és una possible pèrdua de la compensació en els humans, com el que es mostra a la Figura 2.8.

En l'exemple de la Figura 2.8, del cas de F8, a la posició 69 l'aminoàcid salvatge en la majoria d'organismes és una Valina. En humà, però, la Valina provoca hemofília A de caràcter lleu, sent una Leucina l'habitual. Podria haver sorgit algun altre canvi que fos compensat per la Leucina, i la pèrdua d'aquesta, llavors, dispararia l'efecte deleteri de l'hipotètic canvi compensat. O viceversa, ja que algun altre canvi hauria produït que la Valina, que apareix en tants organismes, deixés de ser viable. Hi ha múltiples vies i mecanismes per els quals poden donar-se les compensacions (Lehner 2011) que, de fet, podrien ser un dels motors de l'evolució (Breen et al. 2012). Els CPD poden ser un factor important alhora de descobrir noves vies en l'espai virtual de seqüència, obrint múltiples camins que, sense elles, quedaven ocults. Aquests camins podrien modular la funció i/o l'estructura de la proteïna, conduint als avantatges selectius necessaris per perviure.

Els futurs models hauran de tenir en compte la possible presència dels CPD per a fer una millor representació del fitness: hi ha variants amb possible efecte molt deleteri que pot atenuar-se amb un canvi compensatori.

Malgrat la complexitat tècnica del problema, la creixent quantitat de dades disponible permetrà acabar models raonables per als CPD. Aquests models serien de naturalesa empírica, ja que actualment no hi ha una teoria formal que integri els diversos nivells de complexitat biològica. Un aspecte interessant dels resultats d'aquest capítol (Figura 2.7) és que obren la porta a l'estudi de la severitat en un context de diagnòstic clínic, abordat en el següent capítol.

Capítol II

3. El component intrínsec i la predicció de la severitat en hemofílies A i B

3.1 El component intrínsec de la severitat

La possible severitat d'una malaltia és un factor que serà important en la futura era de la medicina personalitzada (MacArthur et al. 2014, Niroula et al. 2016). Per exemple, en tot estudi de seqüenciació, degut a l'incidentaloma (Berg et al. 2013; Jamuar et al. 2016), solen sortir moltes més variants potencialment patogèniques de les que realment es manifesten; en aquesta situació, un criteri clau en la priorització clínica de les variants podria ser la severitat associada al canvi. És a dir, a l'hora de triar una explicació molecular per als símptomes del pacient, les variants patogèniques associades a versions greus de la malaltia estaran per sobre d'aquelles associades a versions lleus de la malaltia. En estudis genòmics prospectius personalitzats, la severitat associada a la malaltia també podria ser un factor important a l'hora de determinar un protocol d'actuació mèdica (Green et al. 2013).

En el capítol anterior s'ha suggerit, en el context de l'estudi dels CPD, que hi podria haver una relació entre la severitat del fenotip i diferents mesures del grau d'impacte molecular d'una variant. De fet, les prediccions de patogenicitat –discriminació entre si una variant és neutra o és patogènica – (Adzhubei et al. 2011; Sia et al 2012; Riera et al. 2016; Niroula et al. 2016; López-Ferrando et al. 2017) fan ús de les mesures d'impacte molecular per a predir aquesta versió binaria i més extrema del pas de genotip a fenotip que és la patogenicitat.

Aquest capítol explora la possibilitat de crear predictors del fenotip d'una malaltia, centrant l'estudi en el cas de la severitat de les hemofílies A i B –causades per les proteïnes F8 i F9 respectivament–; i buscant l'existència d'un component intrínsec (associat al canvi d'aminoàcid) que contribueix a aquesta

severitat. El component intrínsec es defineix com a tota la contribució al fenotip fruït de l'impacte molecular al gen mutat, únicament –canvis d'estabilitat de la proteïna codificada, diferències en les propietats fisicoquímiques associades al canvi d'aminoàcid, etc–. És una forma d'analitzar quina importància té el genotip en una hipotètica equació en la que el fenotip, F, s'expressa com F=G+A+GA (Henry et al. 2011; Gjuvsland et al 2013), on G és el genotip, A l'ambient, i GA les interaccions de les dos primeres.

Alguns estudis generals ja han suggerit una possible relació entre l'impacte molecular i la gravetat del fenotip posterior (Ferrer-Costa et al. 2004; Stone i Sidow 2005, Hamasaki-Katagiri et al 2012; Sergeev et al. 2013). En aquest cas, s'estudiarà exhaustivament la relació entre el fenotip i els valors d'un conjunt de descriptors relacionats amb l'impacte molecular en els casos de l'hemofília A i B. En aquest context, s'explorarà la capacitat predictiva dels mètodes d'aprenentatge automàtic en el cas de la severitat.

3.2 Materials i Mètodes

L'objectiu final del treball presentat en aquest capítol és identificar i quantificar la relació entre la severitat clínica d'un fenotip i diversos indicadors de l'impacte de les mutacions patogèniques en la funció de la proteïna.

Per a fer això s'usen com a model els factors de coagulació 8 i 9, que estan relacionats respectivament amb les hemofílies A i B. Per aquestes dos proteïnes, tal i com s'ha vist en el capítol anterior, hi ha la major part de canvis d'aminoàcid disponibles amb notació de severitat; en gran part es deu a les bases de dades del Center for Disease Control and prevention (CDC). Aquestes bases de dades categoritzen la variació recollida en les dos malalties tan a la

literatura com en altres bases de dades públiques (Center for Disease Control and prevention, 2015).

La quantificació del factor intrínsec en la severitat es fa obtenint un model predictiu amb un software de xarxes neurals –usant els descriptors d'impacte com a paràmetres d'entrada–, i intentant predir amb el model la categoria a la que cada mutació pertany (Figura 3.1). L'aproximació a aquest problema es divideix en dos parts.

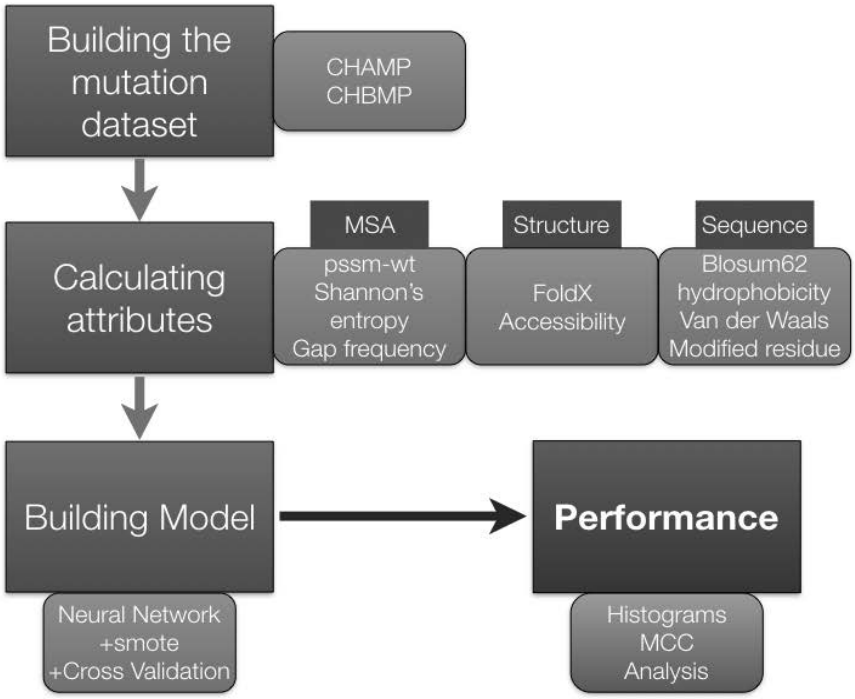

Figura 3.1: *Procés seguit per a crear un predictor.*

En primer lloc, s'estudia fins a quin punt es pot resoldre la versió binaria del problema de la traducció del genotip al fenotip; la predicció de la patogenicitat

de les mutacions. Aquest pas permetrà establir fins a quin punt els paràmetres que mesuren l'impacte molecular es relacionen amb els dos extrems del fenotip estudiat –individu afectat, que presenta els símptomes, vs individu no afectat, que no presenta els símptomes–. **Posteriorment,** es fan servir aquests paràmetres per a predir la severitat, emprant el mateix tipus d'aproximació: aprenentatge automàtic, supervisat, per classificar les mutacions en dos categories diferents.

Entre la primera i la segona part canvien només les categories per les que s'ha d'entrenar les xarxes: en un cas neutres vs patogèniques, en l'altre lleus vs greus.

Finalment, s'estableix la viabilitat de predir la severitat del fenotip esbrinant quins són els descriptors clau per a la predicció; d'aquesta forma es pot tenir una visió global dels paràmetres que afecten o determinen la severitat clínica.

3.2.1 Recol·lecció de dades.

Hi ha dos conjunts genèrics de variants que s'utilitzen en aquest capítol, que s'anomenaran a partir d'ara *pat* i *sev* –per les prediccions de patogenicitat i severitat, respectivament–. S'afegeix el nombre 8 o 9 quan fan referència al gen F8 o al gen F9, respectivament: e.g. pat8 es refereix al conjunt de mutacions neutres i patogèniques de F8.

Les dades emprades en el conjunt de *sev* en aquest capítol són les mateixes que les del capítol anterior: totes les variants patogèniques lleus i severes recollides d'UniProt (Famiglietti et al. 2014, The UniProt Consortium 2017) i de CHAMP i CHBMP. Corresponen a 490 variants lleus i 480 greus per a F8, i 119 variants lleus i 275 greus per a F9. En els dos casos, les variants estan

distribuïdes uniformement per tota la seqüència de la proteïna, tal i com s'observa en la Figura 3.2, exceptuant la regió central –cadena B – de F8, que és una regió molt glicosilada que és eliminada de la proteïna madura, pel que té baixa importància en la funció pro-coagulant de F8 (Oldenburg et al. 2004).

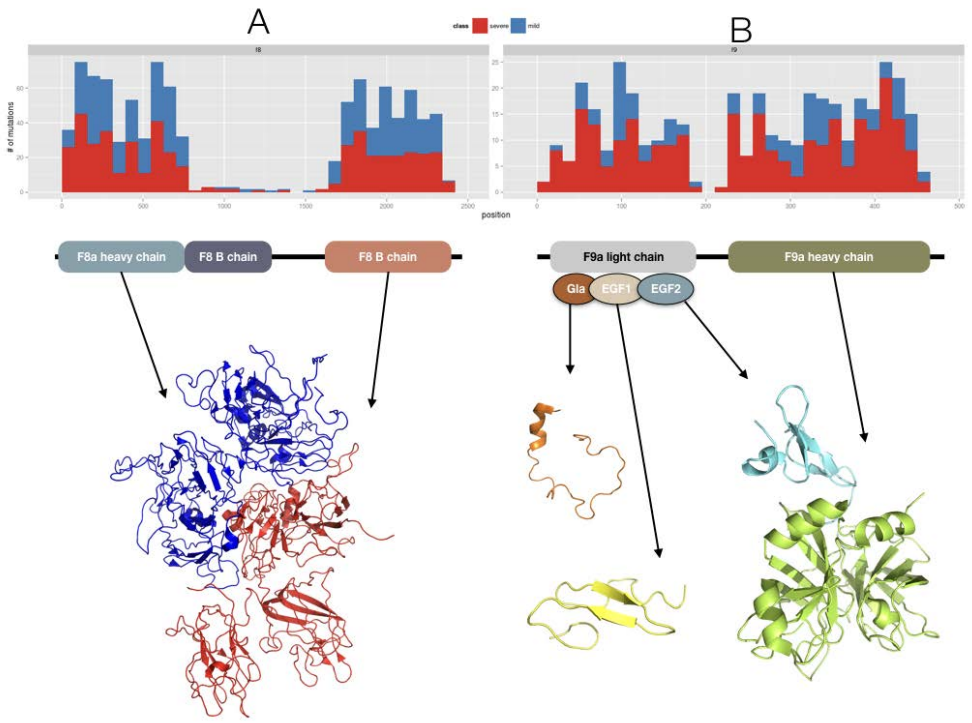

Figura 3.2: *Histogrames amb la distribució de mutacions al llarg de la proteïna –en vermell, variants severes, en blau, lleus–, el domini al que corresponen i la estructura PDB corresponent –si n'hi ha–.* ***A****: F8.* ***B****: F9.*

Per als conjunts de *pat* es recullen dos tipus de dades: variants patogèniques –en aquest cas, independentment de si tenen anotacions de severitat– i variants d'efecte neutre. Les variants patogèniques s'obtenen altre cop de CHAMP i CHBMP. L'obtenció de les variants neutres és més complexa.

Per a buscar variants neutres hi ha dos estratègies que s'han usat tradicionalment en el camp de la predicció de patogenicitat (Ramensky et al. 2002; Capriotti et al. 2006; Riera et al. 2014): **i)** Usar polimorfismes coneguts de la proteïna, o **ii)** Usar canvis en proteïnes properes a la humana com a potencials variants neutres.

En aquest cas, i degut al baix nombre de polimorfismes que hi ha descrits en F8 i en F9 comparats amb el nombre de variants patogèniques – més de 1000 –, s'escull la segona opció, que habitualment resulta en un nombre més elevat de variants. Això és important, ja que els mètodes usats d'aprenentatge artificial, a més a més, necessiten un nombre de casos el més equilibrat possible, pel que fa falta aconseguir tantes variants neutres com es pugui.

El model utilitzat s'anomena model per homologia, i parteix d'un alineament múltiple (MSA) de la família proteica per a obtenir les variants neutres. Es construeixen dos MSA, un per F8 i l'altre per a F9, emprant el programa Muscle (Edgar 2004), que alinea les seqüències obtingudes d'UniRef100 (Boutet et al. 2007) per a cada proteïna per separat. Les seqüències s'obtenen fent una busca d'homòlegs amb psi-blast (Altschul et al. 1990; Altschul et al. 1997) i la respectiva seqüència humana, amb paràmetres per defecte e-value=0.001 i nombre d'iteracions=2. S'eliminen les seqüències que tenen menys d'un 40% d'identitat amb la seqüència humana, ja que es consideren excessivament divergents (Ferrer-Costa et al. 2002) i poden introduir soroll en la construcció dels MSA. Un cop obtingut aquest, s'identifiquen les seqüències amb al menys un 95% d'identitat amb la seqüència humana, i s'extreuen tots els canvis de les seqüències properes, que seran les mutacions neutres d'aquest model. La idea darrere del mètode és que a identitats de seqüència molt elevades és molt

probable que la proteïna humana i els seus homòlegs mantinguin la mateixa funció i la mateixa estructura, pel que les diferències entre elles no correspondrien a variants funcionals, serien neutres.

Proteïna	Neutres	Patogèniques	Total *pat*	Lleus	Greus	Total *sev*
F8	196	1082	*1278*	474	461	*935*
F9	33	484	*517*	109	219	*328*

Taula 3.1: *El nombre de variants usats en cadascun dels casos*.

Aplicant aquest protocol s'obtenen 41 variants neutres per a pat8 i 22 per a pat9. Els nombres totals de variants utilitzats en els dos mètodes es resumeixen en la Taula 3.1.

3.2.2 Descriptors d'impacte molecular

Els paràmetres que es calculen per descriure el canvi d'aminoàcid en les molècules són pràcticament els mateixos que s'han descrit en l'apartat anterior, ja que a priori tenien un potencial predictiu en la severitat, tal com s'havia suggerit en alguns estudis anteriors (Ferrer-Costa et al. 2002). Cal esmentar que, en aquest cas, els paràmetres derivats del MSA –entropia de Shannon, $pssm_{nat}$ i un de nou, freqüència de gaps en la posició– no contenen únicament seqüències d'ortòlegs, provenen de la busca psi-blast descrita en l'apartat anterior, que també pot incloure paràlegs.

Els descriptors són: canvi en el volum de Van der Waals, diferència d'hidrofobicitat, valor de la matriu BLOSUM62, entropia de Shannon, freqüència de gaps, $pssm_{nat}$, $\Delta\Delta G$ (canvi d'estabilitat de la proteïna mutant respecta la nativa), accessibilitat relativa i importància del residu.

També s'ha estudiat la capacitat predictiva de dos nous descriptors: la freqüència d'espais buits en l'alineament –o gaps– en la posició de la mutació i la importància del residu. El primer es calcula com el nombre de gaps en la posició de la mutació dividit pel nombre total de seqüències en el MSA. És una mesura addicional del grau de divergència evolutiva i reflecteix la importància relativa de la posició mutada respecte la resta de posicions de la proteïna. El segon, la importància del residu, és un descriptor booleà –Cert o Fals–, que indica si el residu té alguna funció d'especial importància descrita a UniProt (The UniProt Consortium 2017). Els termes emprats corresponen a les funcions següents: unió a un substrat, formació de ponts disulfur, N-glicosilació o pertinença al centre actiu. Aquest paràmetre serveix per capturar funcions de l'aminoàcid que no queden ben descrites per la conservació del residu via entropia de Shannon i $pssm_{nat}$ (Ferrer-Costa et al. 2004), com poden ser la funció d'aquest residu en un centre actiu, o formant algun enllaç entre cadenes no adjacents.

3.2.3 El mètode de predicció: Xarxes Neurals.

L'algoritme usat per a la predicció és una xarxa neural del tipus *multilayer perceptron,* implementat a través del programari de WEKA (v3.7) (Hall et al. 2009), un paquet estàndard però molt flexible d'aprenentatge automàtic que permet que tots els resultats siguin totalment reproduïbles.

Per compensar els mals balanços en les dades –ja que hi ha moltes més variants patogèniques que variants neutres obtingudes per homologia, i els nombres de variants lleus i greus també difereix– s'usa l'algoritme d'SMOTE (Chawla et al. 2002), que realitza un sobre-mostreig de la població en minoria creant punts sintètics en un procés de pseudo-interpolació. Aquests punts

s'obtenen de l'espai dels descriptors a partir de mostrejar els segments que uneixen els k-veïns-propers de la classe minoritària, tants com siguin necessaris. L'SMOTE escollit varia en funció de la proporció entre classes (majoritària entre minoritària), a raó de 100% per duplicar la població amb menys casos, 200% per triplicar...

Un altre factor important a tenir en compte és el sobre-ajustament (Hawkins 1995). Per aquesta raó, s'han de dividir les dades en diferents conjunts:: un amb el que entrenar el mètode –conjunt d'entrenament– i un altre amb el que mesurar-ne el seu comportament –conjunt de test–. Sinó, la xarxa neural pot haver-se après molt bé els casos amb els que s'entrena, però generalitzar molt malament, donant unes mesures que sobreestimin el comportament del predictor. Per altra banda, com que la quantitat de dades és baixa, només dividir les dades en dos conjunts pot privar al mètode d'aprendre's alguna secció o interacció entre descriptors de les dades. Usar una validació creuada amb *k*-particions (Kohavi 1995) –en aquest cas 3– redueix aquest efecte: Es divideixen les dades en *k* parts, i s'usen 1- *k* parts per a l'entrenament i 1 per al test, i s'entrena el mètode; aquest procés es repeteix *k* vegades, per les *k* combinacions possibles de conjunts de test. El comportament del mètode, aleshores, s'avalua sobre la mitjana dels resultats als *k* conjunts de test.

Finalment, per atenuar l'impacte dels possibles biaixos en la partició de les dades, es repeteix 100 vegades el procés de k-particions, amb particions diferents cada cop –generades a l'atzar–. La capacitat predictiva proporcionada correspondrà a la mitjana sobre les cent particions.

Es valoren els resultats amb quatre mesures estàndard (Baldi et al. 2000; Vihinen 2013; Vihinen 2014): el Coeficient de Correlació de Matthews (MCC),

la Sensibilitat, l'Especificitat i l'Exactitud. Aquests paràmetres reflecteixen diverses vessants del procés predictiu, i s'obtenen a partir de diferents combinacions dels nombres de falsos positius (FP), falsos negatius (FN), vertaders positius (TP) i vertaders negatius (TN).

L'**exactitud** és la fracció de casos correctament identificats; va de 0 a 1: (TP+TN)/(TP+FP+TN+FN).

La **sensibilitat** és la fracció de variants positives (en aquest cas patogènics o greus) correctament identificats; va de 0 a 1 (Altman i Bland 1994): TP/ (TP+FN)

L'**especificitat** és la fracció de variants negatives (en aquest cas neutres o lleus) correctament identificades; va de 0 a 1 (Altman i Bland 1994): TN/ (TN+FP)

El **Coeficient de Correlació de Matthews** (MCC) és una mesura de la qualitat de la classificació que té en compte possibles des-balanços de les dades i és una mesura més completa que l'exactitud. Va de -1 a +1, on 0 correspon a un mètode totalment estocàstic, 1 a un mètode perfecte, i -1 a un mètode amb error sistemàtic (Matthews 1975): (TP*TN-FP*FN) /(TP+FN)(TP+FN) (TN+FP)(TN+FN)^(1/2)

3.2.4 Validació del mètode.

Per a avaluar els predictors entrenats i la seva funcionalitat es comparen amb altres predictors de la literatura, fent servir els mateixos conjunts de dades. Assolir una capacitat d'encert similar a la dels predictors estat de l'art significa que el model proposat funciona prou bé, validant el seu ús per a estudiar l'impacte a nivell molecular de les mutacions.

Els mètodes escollits per a realitzar la validació són SIFT (Ng i Henikoff 2001; Sim et al. 2012) i Polyphen2 (Adzhubei et al. 2010) per al cas de la predicció de la patogenicitat, i MAPP (Stone i Sidow 2005) per al cas de la predicció de severitat.

En el cas de Polyphen2 (Adzhubei et al. 2010), s'utilitzen les dos versions del predictor, la entrenada en HumDiv i la entrenada en HumVar. Polyphen2 és un mètode que utilitza l'aprenentatge automàtic, mitjançant un classificador bayesià ingenu, amb un conjunt de descriptors tan estructurals i de seqüència com derivats de MSA. S'escullen a partir d'una llista més gran de descriptors, seleccionats amb un algoritme voraç que, iterativament, va seleccionant els descriptors que donen millor resultat. En comparació amb l'aproximació d'aquest treball, els descriptors escollits no provenen de cap selecció feta sobre el conjunt estudiat de dades i no corren el risc de patir sobre-ajustament. Els dos mètodes resultants de Polyphen2 provenen d'usar dos conjunts d'entrenament diferents. HumDiv està entrenat amb les variants de UniProt anotades com a causants de malaltia mendeliana, i amb les variants neutres extretes per homologia. HumDiv, en canvi, usa totes les variants d'UniProt que tenen associació amb malaltia com a patogèniques, i tota la resta de variants com a neutres.

SIFT (Sim et al. 2012), d'altra banda, només utilitza l'homologia de seqüència per a predir el possible impacte d'una variant. Es pot interpretar com el poder predictiu total en aquella proteïna de la informació continguda en la conservació de la seqüència.

MAPP (Stone i Sidow 2005) és un predictor de patogenicitat que no només classifica en patogènic o neutre, sinó que ofereix un valor que, com més alt,

més fort és el potencial impacte. Segons els autors, aquest valor pot discriminar entre els casos més lleus i els casos més greus de severitat. Per a interpretar aquest valor de forma binària en la predicció de severitat els autors no donen cap possible punt de tall.

Per a obtenir el punt de tall en cada proteïna, es segueix una aproximació exhaustiva, generant les possibles prediccions a múltiples punts de tall (Figura 3.3), i triant el més òptim per a la classificació en aquella proteïna. Així doncs, la predicció de la severitat en MAPP serà la millor possible donat el mètode.

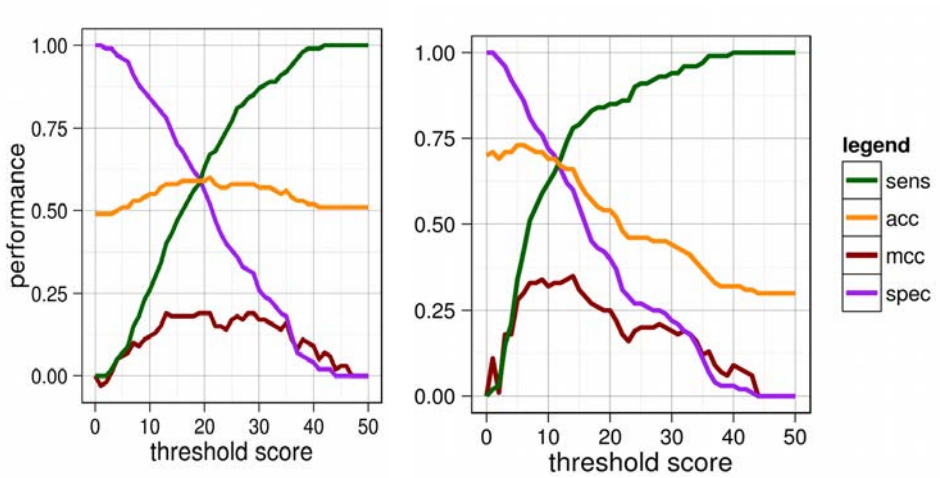

***Figura 3.3**: Mostrant el comportament de MAPP segons el punt de tall. El gràfic de l'esquerra correspon a F8. El de la dreta a F9. El valor de punt de tall escollit correspon a la intersecció entre la sensibilitat i l'especificitat.*

3.2.5. El component intrínsec de la severitat.

Finalment, per a estudiar més detalladament l'origen del component intrínsec de l'impacte molecular en la severitat, es mesura separadament el poder predictiu de cadascun dels paràmetres: d'estructura, del canvi de seqüència, i de conservació del residu natiu. A nivell tècnic, s'entrena un perceptró amb un sol

descriptor i una sola capa, construint-ho amb el programa WEKA (v3.7) (Hall et al. 2009). Aquesta aproximació es segueix tan en el problema de neutre vs patogènic com en el de lleu vs greu.

Els anàlisis realitzats en aquest capítol estan fets amb una combinació de scripts de Python (v2.7 i v3.4) (Python Software Foundation), i Bash (v3.2) (Free Software Foundation). Les figures resultants estan creades amb R (R-Core Team 2013), amb el paquet de ggplot2 (Wickham 2009).

3.3 Resultats i Discussió

Tal i com es suggereix al capítol anterior i el que la evidència prèvia indica (Stone i Sidow 2005), hi ha un cert grau de relació entre l'impacte molecular produït per una substitució d'aminoàcid la severitat de la malaltia resultant. Per a estudiar aquesta relació s'escullen com a models les malalties de l'hemofília A i B, relacionades amb mutacions a F8 i F9 respectivament. A partir d'aquí, es procedeix de forma esgraonada, basant-se en que el component intrínsec de la severitat és una versió quantitativament refinada del component intrínsec que determina la patogenicitat/neutralitat de les mutacions. **Primer** (secció 3.3.1), partint de les propietats específiques de proteïna descrites al capítol anterior, es verifica la seva capacitat predictiva en l'avaluació de la patogenicitat/neutralitat de les mutacions a F8 i F9. Aquest pas es fa emprant models d'aprenentatge automàtic, i les seves mètriques habituals de comportament –descrites a Materials i Mètodes–; aquestes s'utilitzaran com a mesures del component intrínsec. En el **segon** pas (secció 3.3.2), es segueix el mateix procediment per estudiar l'existència d'aquest component per la severitat de les malalties. **Finalment** (secció 3.3.3) s'integren els resultats anteriors per estudiar la

predicció de la severitat.

3.3.1 La predicció de la patogenicitat i el paper de les propietats fisicoquímiques i evolutives

En un estudi recent s'ha descobert que un seguit de propietats fisicoquímiques i del canvi genètic –conservació de seqüència en la posició, hidrofobicitat, ratio GC, Energia lliure del mRNA, potencial de fosforilació...– prediuen amb bona exactitud la patogenicitat de variants en hemofília (Hamasaki-Katagiri et al. 2013). Els autors també consideren les prediccions de Polyphen2 (Adzhubei et al. 2010), SIFT (Sim et al. 2012), CONDEL(González-Pérez i López-Bigas 2011), PROVEAN (Choi et al. 2012) i Mutation-Assesor (Reva et al. 2007), que funcionen moderadament bé – al voltant del 80% d'exactitud, Taula 2 de Hamasaki-Katagiri et al. 2013– malgrat les seves diverses aproximacions. CONDEL, per exemple, és un predictor consens entre PolyPhen2, SIFT i Mutation-Assessor; i PROVEAN, que també pot usar-se amb indels, només fa ús de la similitud entre la proteïna mutant i els homòlegs. Amb aquesta informació prèvia a hemofília, i el fet que usar mètodes de predictors específics d'una proteïna pot millorar el seu poder predictiu (Riera et al. 2015, Riera et al. 2016), es dissenya l'anàlisi específic per predictor de patogenicitat, és a dir, per al problema de discriminar variants neutres de patogèniques.

Les distribucions de valors dels descriptors d'impacte molecular

L'estudi de les distribucions de valors dels descriptors d'impacte molecular per als dos tipus de mutacions considerats –patogènics i neutres– és un pas important per entendre'n la capacitat predictiva. Els resultats obtinguts en estudiar els diferents descriptors (Figura 3.4) mostra que, tot i un cert poder

discriminatori, les seves distribucions per les variants neutres i patogèniques es superposen. Això significa que no hi ha cap propietat que en solitari pugui diferenciar totalment entre els dos tipus de variants, reproduint els resultats anteriors d'hemofília (Hamasaki-Katagiri et al. 2013) i d'altres proteïnes (Ferrer-Costa et al 2004; Riera et al. 2015). A part d'això, el comportament de cada descriptor té les seves característiques pròpies: zones de valors corresponents a una de les classes, similars a les dos proteïnes. Per exemple, la diferència de volum de Van der Waals (Figura 3.4A) té un pic proper a 0 per a variants neutres, mentre que les patogèniques són més freqüents als valors més grans; a Blosum62 (Figura 3.4A) hi ha un gran pic a -2 per a les variants patogèniques; o als canvis d'estabilitat, o ΔΔG –calculats amb FoldX– (Figura 3.4B), on la majoria de la població es concentra al voltant del 0, amb més presència de les patogèniques a valors més elevats. Aquesta tendència de la ΔΔG ja ha estat reportada per altres autors en les dos hemofílies (Rallapalli et al. 2014). Sorgeixen diferències entre F8 i F9 en el $pssm_{nat}$ (3.4B), ja que les mutacions neutres presenten un comportament de tipus bimodal a F9, mentre que a F8 la tendència és molt menys pronunciada. A l'entropia de Shannon també hi ha diferències entre proteïnes, encara que no tan pronunciades.

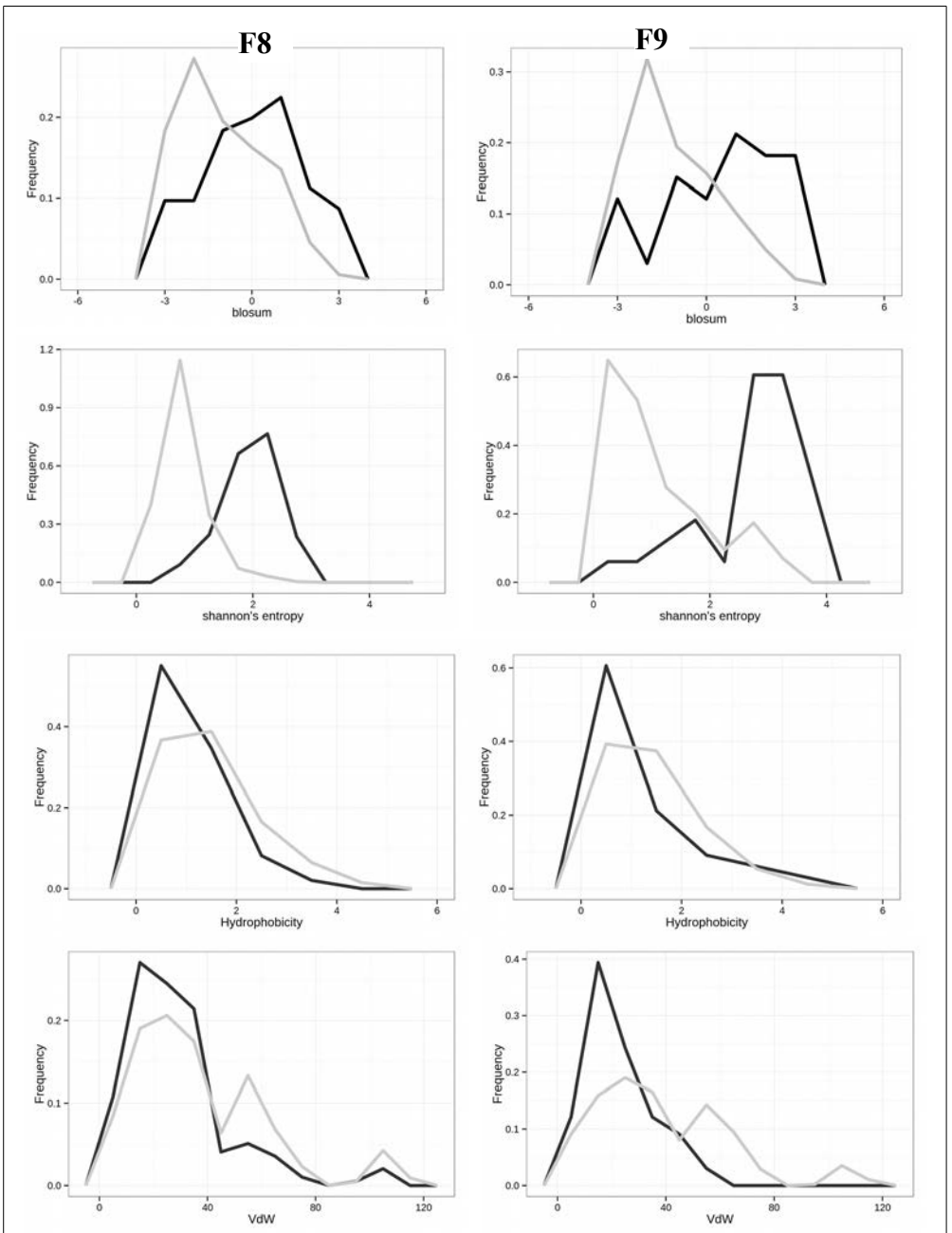

Figura 3.4A: *Histogrames amb la distribució de les variants patogèniques (en clar) respecte les neutres (en fosc) per a blosum62, Entropia de Shannon i diferències d'Hidrofobicitat i de Volum de Van der Waals.*

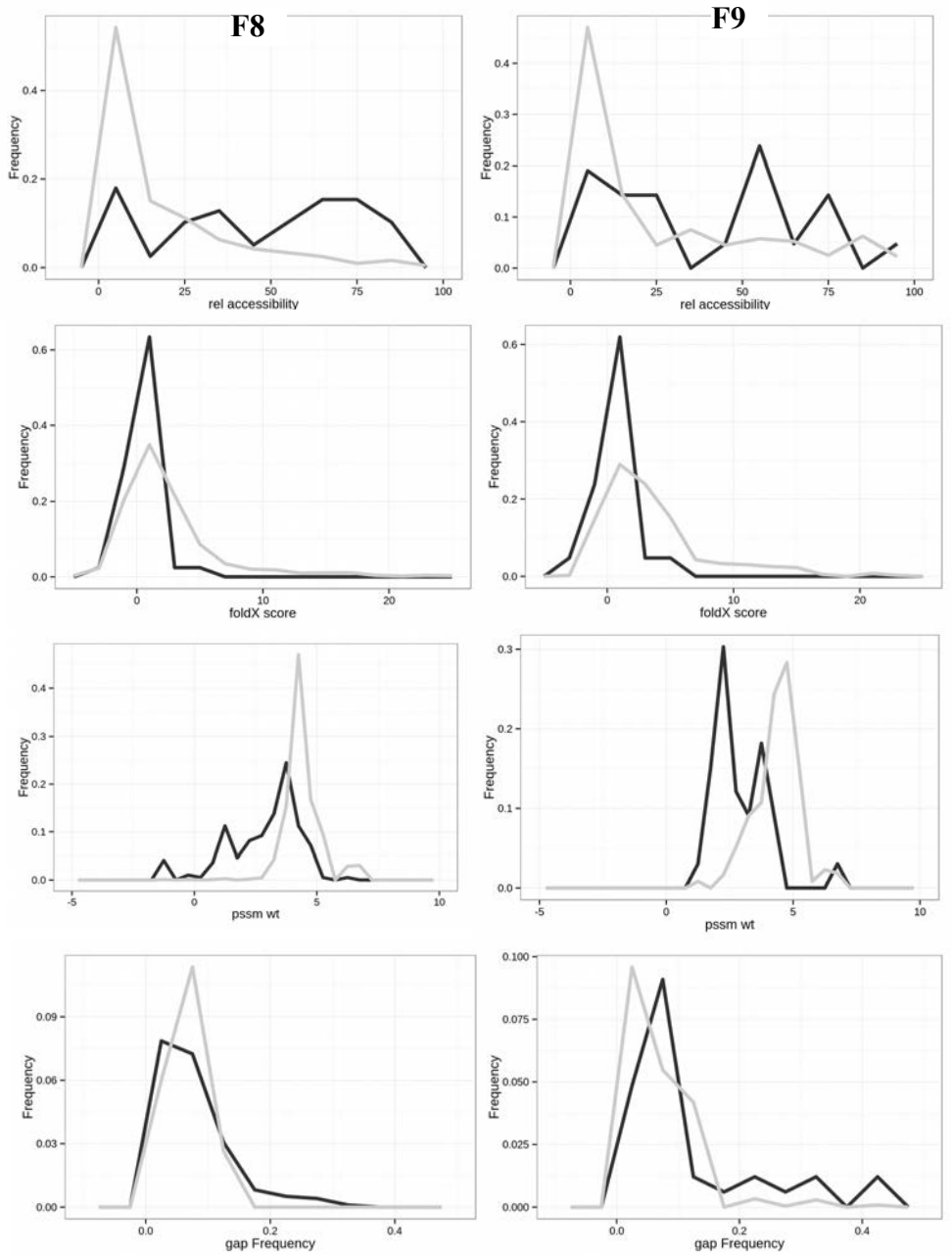

Figura 3.4B: *Histogrames amb la distribució de les variants patogèniques (en fosc) respecte les neutres (en clar) per a l'accessibilitat relativa, la ΔΔG de foldX, la freqüència de gaps i el* $pssm_{nat}$.

Aquesta diferència entre propietats evolutives és deguda a que el $pssm_{nat}$ és el descriptor que depèn en més mesura de l'alineament a nivell global, tenint en compte la composició d'aminoàcids en la posició estudiada i la composició d'aminoàcids de la resta de l'alineament. En canvi, l'entropia de Shannon, que també depèn del MSA, només depèn de la posició de la variant: això fa que que el comportament d'aquest paràmetre també variï segons el MSA de la proteïna, però de forma menys extrema que el $pssm_{nat}$.

La variabilitat observada en les mesures dependents dels MSA (Figura 3.4 A i 3.4B) suggereix que l'aproximació a fer predictors de patogenicitat específics de proteïna pot ser important en aquells casos on aquestes variables són les més discriminatòries. Malgrat que fins ara no s'ha trobat quins paràmetres ni quins efectes fan millorar alguns i empitjorar altres casos (Riera et al. 2016), les observacions d'aquest capítol en el cas de l'hemofília apunten a que hi ha una contribució de la qualitat i tipologia dels MSA utilitzats en la capacitat d'encert dels predictors específics a una proteïna.

La capacitat predictiva dels descriptors d'impacte molecular

Les diferències observades en les figures 3.4A i 3.4B, en les que poca superposició entre les categories indica un bon poder predictiu i molta superposició indica el contrari –ja que vol dir que les dos poblacions tenen una distribució de valors similar–, es poden quantificar fent predictors amb un sol descriptor (Taula 3.2).

F8				
Descriptor	**Exactitud**	**Especificitat**	**Sensibilitat**	**MCC**
$pssm_{nat}$	0.81	0.66	0.81	0.22
Entropia Sh.	0.86	0.78	0.87	0.34
Rel. Acc.	0.79	0.68	0.80	0.22
ΔΔG	0.57	0.78	0.56	0.13
F9				
Descriptor	**Exactitud**	**Especificitat**	**Sensibilitat**	**MCC**
$pssm_{nat}$	0.80	0.73	0.80	0.29
Entropia Sh.	0.83	0.76	0.83	0.33
Rel. Acc.	0.70	0.54	0.71	0.12
ΔΔG	0.68	0.60	0.68	0.14

Taula 3.2: *Predictibilitat dels descriptors per separat en neutre vs patogènic*.

Els resultats obtinguts (Taula 3.2) repliquen el que la literatura ja apuntava (Hicks et al. 2011; Castellano i Mazza 2013; Riera et al. 2014), que és que els descriptors derivats dels MSA tenen per ells mateixos un poder predictiu significatiu –MCC de >0.2 en $pssm_{nat}$ i MCC de >0.3 per a l'entropia–. Altres descriptors, com és el cas de ΔΔG, tenen un cert poder predictiu, però és menor que el dels descriptors provinents de MSA. Un factor que empitjora la capacitat predictiva d'algunes propietats és la falta d'estructura en algunes regions de la proteïnes. Per exemple, en el cas de F9, enlloc d'una estructura global per a tota la proteïna, hi ha diverses estructures que només corresponen a fragments

proteics (Figura 3.2). Per a certes mutacions en aquests fragments, el valor del càlcul d'accessibilitat o d'estabilitat és enganyós, ja que en les estructures disponibles els aminoàcids natius estan més exposats al solvent o tenen menys interaccions residu-residu que a la realitat. Aquest pot ser un dels motius que explica la diferència en el comportament de l'accessibilitat relativa entre F8 i F9. No hi ha cap altre manera de millorar aquests descriptors que esperar a que sorgeixin estructures resoltes que cobreixin més secció de la proteïna, ja que treballar amb models per homologia (Venselaar et al. 2010) o ab-initio (Spencer et al. 2015) introduiria un altre factor difícilment quantificable d'incertesa i de variabilitat.

La combinació de tots els descriptors és, però, el que realment aconsegueix obtenir uns resultats més bons, tan a la literatura (Ramensky et al. 2002; Ferrer-Costa et al. 2004; Niroula et al. 2015) com en aquest treball (Taula 3.3).

La Xarxa Neural construïda obté valors comparables als estudis anteriors (Hamasaki-Katagiri et al. 2013), i als obtinguts amb amb Polyphen2 i SIFT, tal i com es pot observar a la comparació amb les mateixes dades (Taula 3.3). El nombre de descriptors escollit, malgrat ser de només 8, aconsegueix un poder discriminatori molt alt – Exactitud >0.88–, comparable als predictors clàssics, validant la metodologia utilitzada. Les diferències en la predictibilitat entre el model construït amb la xarxa neural i els models anteriors rauen en la diferència entre els descriptors utilitzats, i en les diferències en la categorització de les dades, donades pel model neutre d'homologia, que és bastant conservador (Riera et al. 2014).

F8				
Mètode	**Exactitud**	**Sensibilitat**	**Especificitat**	**MCC**
MODELpat	0.92	0.93	0.85	0.5
PPH2-HD	0.95	0.95	0.86	0.58
PPH2-HV	0.93	0.93	0.88	0.51
SIFT	0.82	0.81	0.95	0.36
F9				
Mètode	**Exactitud**	**Sensibilitat**	**Especificitat**	**MCC**
MODELpat	0.88	0.89	0.77	0.42
PPH2-HD	0.85	0.85	0.91	0.47
PPH2-HV	0.85	0.82	0.91	0.42
SIFT	0.85	0.86	0.77	0.37

Taula 3.3: *Prediccions per a les dades de pat8 i pat9 del mètode entrenat (MODEL), Polyphen2 Hum-Div, Polyphen2 HumVar i SIFT.*

L'impacte del model de variants neutres sobre la capacitat dels predictors

Aquest és, potser, un dels factors amb més efecte en la capacitat predictiva de les xarxes neurals. El model d'homologia té els seus pros i les seves contres. Per un cantó, el nombre obtingut de variants neutres és molt més elevat que l'obtingut per Hamasaki-Katagiri (Hamasaki-Katagiri et al. 2013), cosa que fa més fàcil una generalització adequada del model, ja que les variants corresponen a un espectre més ampli dels possibles camins que podria prendre la proteïna en l'espai virtual de seqüència (de Pristo et al. 2005). Per altra banda, però, el model de neutralitat per homologia no té en compte la possible aparició de compensacions, fet precisament estudiat en el capítol anterior i que

ja s'ha vist que té una importància cabdal en l'obertura de nous camins evolutius (Breen et al. 2012). Aquest tipus de compensacions afecten al model neutre ja que poden donar lloc a falsos positius. De fet, algunes variants aparentment neutres són en realitat patogèniques degut a que a l'espècie on s'ha trobat van acompanyades d'un canvi, absent en humans, que suprimeix el seu efecte. Aquesta situació podria afectar a un 10% del total de neutres (Jordan et al. 2014).

Addicionalment, el model de variants neutres és incomplet en el sentit de que hi hauria encara una gran fracció de variants neutres que no s'han trobat. Un dels possibles factors és per què les seqüències del MSA no han arribat a explorar encara aquesta zona de l'espai de seqüència (Povolotskaya i Kondrashov 2010), ja que només són una fracció del total de seqüències de F8 o de F9 que s'han pogut donar o que es donen en altres espècies.

També hi hauria, com ja s'ha comentat, una fracció indeterminada de variants, neutres en humà, que no es podrien detectar amb el model d'homologia per què venen acompanyades a la seqüència humana de compensacions que la permeten –i que no es tenen en compte en el model d'homologia–. Buscar aquestes variants és, però, molt complicat: ja és difícil trobar una compensació quan se saps quina és la variant compensada (Ferrer-Costa et al. 2007; Jordan et al. 2014), i la dificultat augmenta enormement –tan per longitud de seqüència com per nombre de seqüències que s'analitzen– quan el que es busca és una potencial compensada. En resum, cal veure el model d'homologia com el que és, un compromís entre la necessitat d'augmentar el nombre de variants neutres que es poden analitzar malgrat l'existència d'alguns problemes associats.

Resumint aquesta secció, s'ha demostrat que hi ha un component intrínsec a

l'impacte fenotípic en el cas de la patogenicitat, i que aquest component intrínsec queda recollit pels descriptors escollits. Aquest resultat permet emprar aquests descriptors com a punt de partida per estudiar el cas de la severitat de la malaltia, un problema directament relacionat amb la predicció de la patogenicitat.

3.3.2 El component intrínsec en la severitat de l'hemofília

Basat en els resultats anteriors (secció 3.3.1), s'utilitza la mateixa aproximació per buscar el component intrínsec de la severitat. S'entrena una xarxa neural amb els mateixos descriptors, però en aquest cas el conjunt d'entrenament està constituït únicament per variants patogèniques per les que hi ha informació de la severitat de les malalties de l'hemofília A i B –F8 i F9 respectivament–. Lleu o greu són les dos classes a discriminar pel predictor. S'accepta l'existència d'un component molecular intrínsec a la severitat quan el predictor té una habilitat predictiva millor que l'atzar, que a nivell pràctic vol dir que els valors de la sensibilitat, l'especificitat i l'exactitud són majors de 0.5, i que el MCC és major que 0.

Predicció de mutacions lleus i greus i el component intrínsec de la severitat

A la Taula 3.4 s'observa que, de fet, els paràmetres acompleixen les condicions esmentades, el que permet concloure que, dins de les condicions d'aquest experiment, aquest component existeix i és detectable, malgrat ser menor que el component intrínsec en la patogenicitat. Aquest menor efecte és consistent amb el que suggerien estudis anteriors: quan hi havia algun anàlisi, aquest apuntava a una petita, però existent, contribució (Ferrer-Costa et al. 2002; Stone i Sidow 2005).

F8				
Mètode	**Exactitud**	**Sensibilitat**	**Especificitat**	**MCC**
MODELsev	0.63	0.59	0.66	0.27
MAPP	0.58	0.57	0.59	0.16
F9				
Mètode	**Exactitud**	**Sensibilitat**	**Especificitat**	**MCC**
MODELsev	0.72	0.71	0.75	0.44
MAPP	0.68	0.65	0.70	0.34

Taula 3.4: *Mètriques de la predicció per a lleu i greu, els conjunts de sev8 i sev9 amb MODELsev, que és la xarxa neural entrenada, i amb MAPP amb el punt de tall amb millor comportament*.

La conclusió no depèn del mètode emprat, ja que tant la xarxa neural pròpia com el software de MAPP (Stone i Sidow 2005) aconsegueixen trobar una senyal moderadament positiva. De fet, ambdós predictors fins i tot identifiquen una senyal respectable a F9: el percentatge d'encert està per sobre de 0.7 i el MCC és de 0.44, un valor que no sempre s'aconsegueix en els estudis de patogenicitat (Hicks et al. 2011; Castellano i Mazza 2013; Riera et al. 2016;

López-Ferrando et al. 2017).

Els resultats obtinguts (Taula 3.4) mostren clarament que, al menys en les hemofílies A i B, hi ha un efecte **intrínsec** al gen –que només depèn d'aquest i no de la resta de fons gènic o de l'ambient– en la severitat. Això es replica tan amb MAPP, com amb els predictors entrenats específicament per a F8 i F9, que veuen que hi ha una fracció que correspon al ~70% de les variants, que pot explicar-se només amb els descriptors intrínsecs. Això tampoc significa que la contribució de la severitat sigui del 70%, ja que no és una quantificació exacte, però sí que permet dir que hi ha un efecte important del genotip en la severitat.

Un anàlisi fi de les dades mostra un aspecte interessant relacionat amb l'ús de les mètriques d'avaluació de les eines bioinformàtiques. A F9 S'observa que, tot i que el MCC té valors comparables entre *MODELsev* i *MODELpat*, aquesta similitud és en part deguda a l'estructura de les dades. *Pat9* té un nombre molt baix de lleus, que prediu moderadament bé –0.77 d'especificitat–, però pitjor que les patogèniques –0.89 de sensibilitat–; i aquesta heterogeneïtat en la predicció –millor en patogèniques que en neutres– es reflexa en el MCC –0.42–, més baix del que el valor global de l'exactitud –0.89– per el des-balanç. En canvi, en *MODELsev*, l'exactitud –>0.7– és bona però no és excel·lent, tot i tenir un MCC similar –0.44–. Aquest cas mostra la importància de no fiar-se exclusivament d'una sola mesura a l'hora de valorar els predictors. N'hi ha que són molt bones –com l'MCC–, però cap valor dóna una representació del que realment està passant com una combinació de les mesures, exemplificant la importància de l'esforç de Vihinen per crear un estàndard en les mesures usades a l'hora de descriure el funcionament de predictors (Vihinen 2012; Vihinen 2013, Vihinen 2014).

Propietats clau en la predicció de la severitat a F8 i F9

Un cop obtinguts els predictors, una pregunta natural és demanar quins són els paràmetres clau o que tenen una contribució més important. No és fàcil donar una resposta, ja que normalment els predictors funcionen com a capses negres de les que no és simple interpretar-ne els resultats. Per afavorir la comprensió dels resultats i augmentar-ne l'aplicabilitat, es construeix la distribució dels diferents paràmetres emprats (Figura 3.5), amb l'objectiu d'identificar l'origen de la capacitat predictiva en les xarxes neurals.

El primer que es veu és que les diferències entre poblacions de variants s'observen molt més tènuement que amb la patogenicitat, tal i com podia suposar-se per les mètriques dels predictors. Nogensmenys, segueix havent-hi diferències, encara que observables en categories diferents. La matriu de Blosum62 manté a F8 i F9 una diferència apreciable entre els casos lleus i els greus, indicant que les diferències fisicoquímiques entre aminoàcids natiu i mutant són un factor que afecta la severitat de la malaltia. Però no és l'únic que queda palès. Hi ha una tendència important de les mutacions amb valors elevats de ΔΔG a tenir un fenotip més sever. Aquesta tendència s'observa clarament en observar les prediccions fetes només amb un sol descriptor (Taula 3.5): ΔΔG és molt –de fet, el descriptor que és més– predictiu en els casos en que hi ha estructura. Cal remarcar que a F9 també Blosum62, l'entropia de Shannon o fins i tot l'accessibilitat relativa, prediuen moderadament bé per separat –exactitud >60%–; fet que no es replica a F8, on tenen menor potencial predictiu.

F8				
Descriptor	**Exactitud**	**Especificitat**	**Sensibilitat**	**MCC**
Blosum62	0.58	0.58	0.57	0.16
Entropia Sh.	0.54	0.55	0.54	0.10
Rel. Acc.	0.58	0.54	0.60	0.14
ΔΔG	0.61	0.73	0.49	0.24
F9				
Descriptor	**Exactitud**	**Especificitat**	**Sensibilitat**	**MCC**
Blosum62	0.60	0.63	0.60	0.22
Entropia Sh.	0.67	0.62	0.70	0.30
Rel. Acc.	0.64	0.57	0.68	0.25
ΔΔG	0.64	0.76	0.58	0.33

Taula 3.5: *Capacitat predictiva dels descriptors per separat en lleu-greu.*

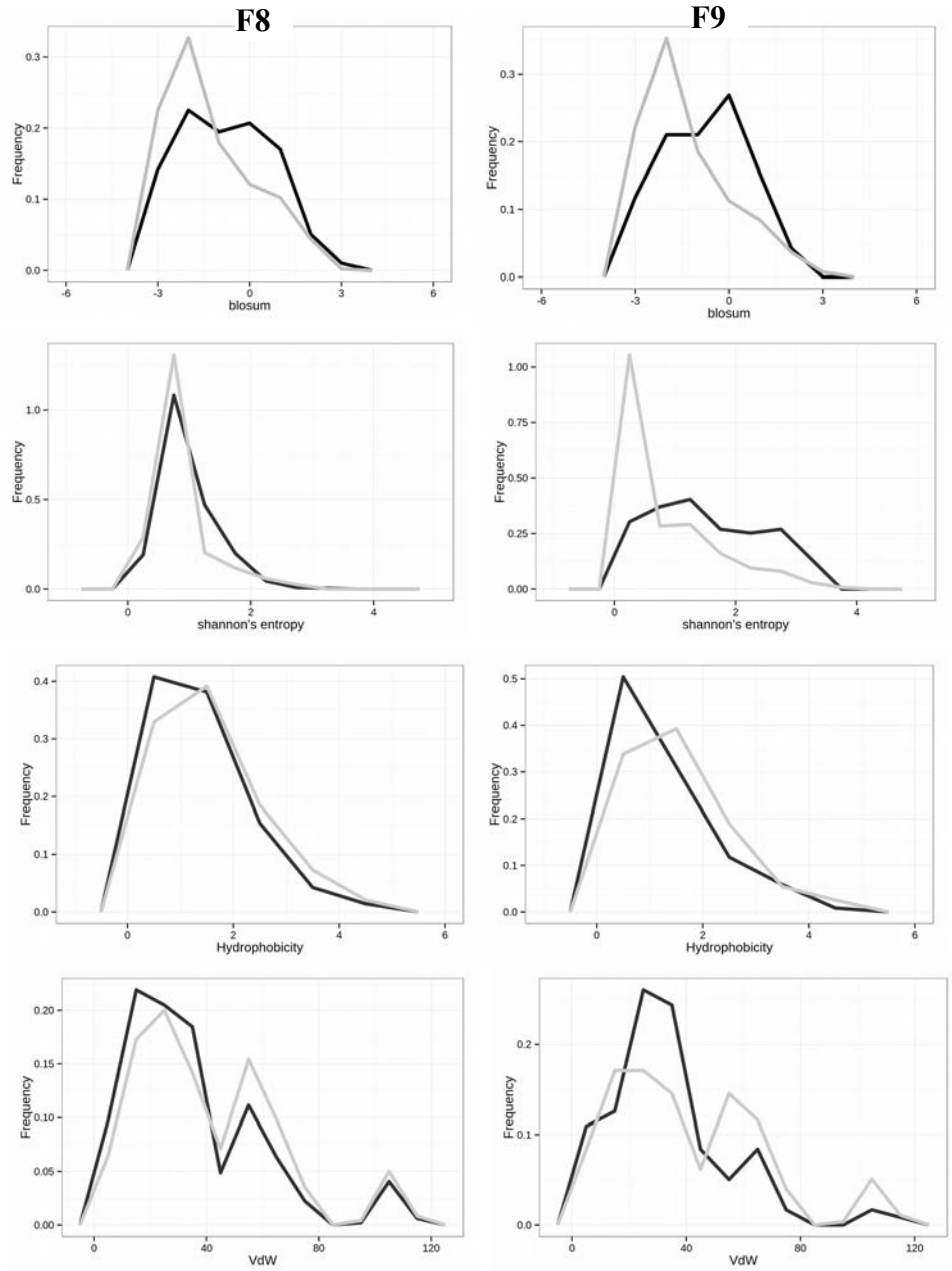

Figura 3.5A: *Histogrames amb la distribució de les variants greus (en fosc) respecte les lleus (en clar) per a blosum62, Entropia de Shannon i Deltes d'Hidrofobicitat i de Volum de Van der Waals.*

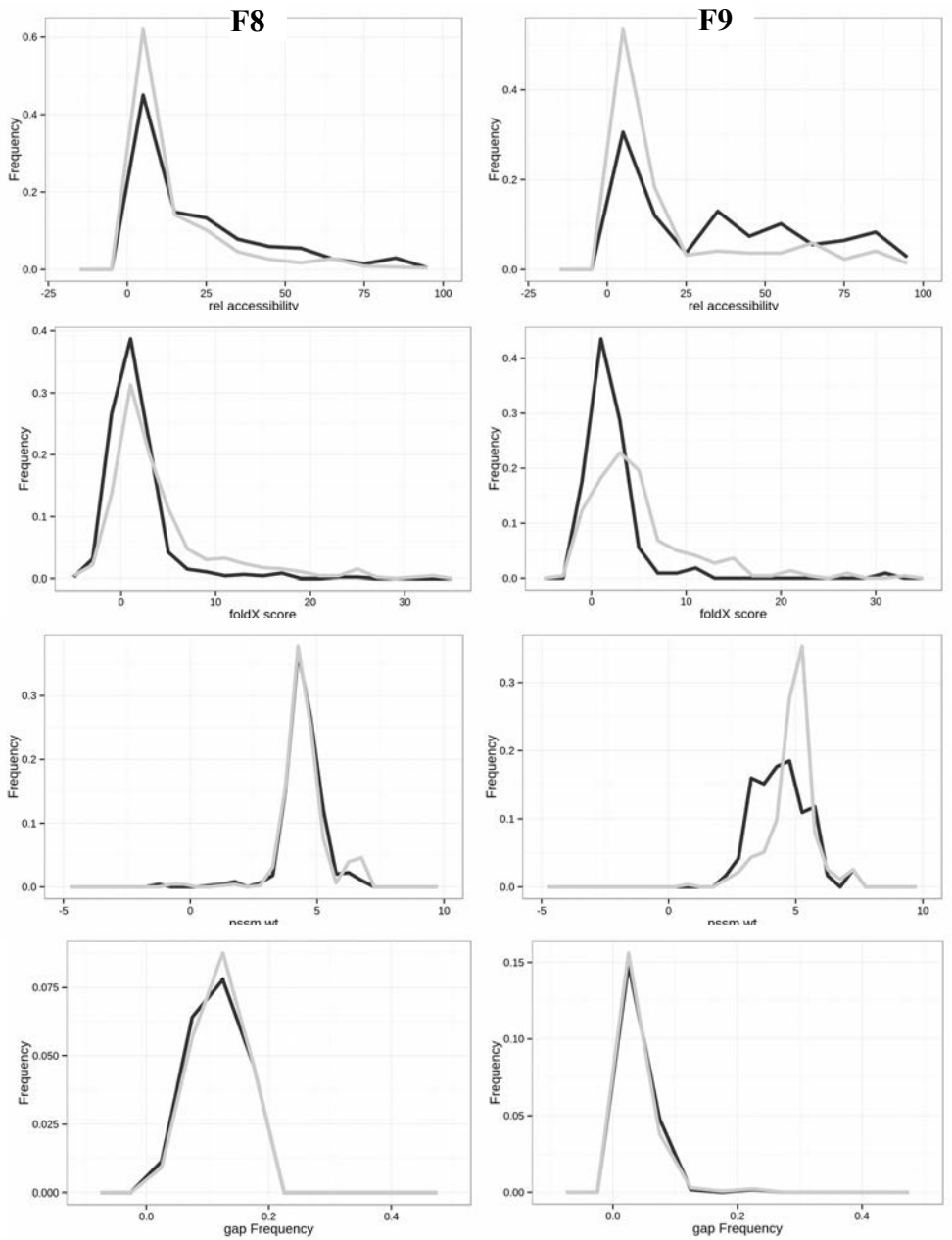

Figura 3.5B: *Histogrames amb la distribució de les variants lleus (en fosc) respecte les greus (en clar) per a l'accessibilitat relativa, la ΔΔG de foldX, la freqüència de gaps i el* $pssm_{nat}$.

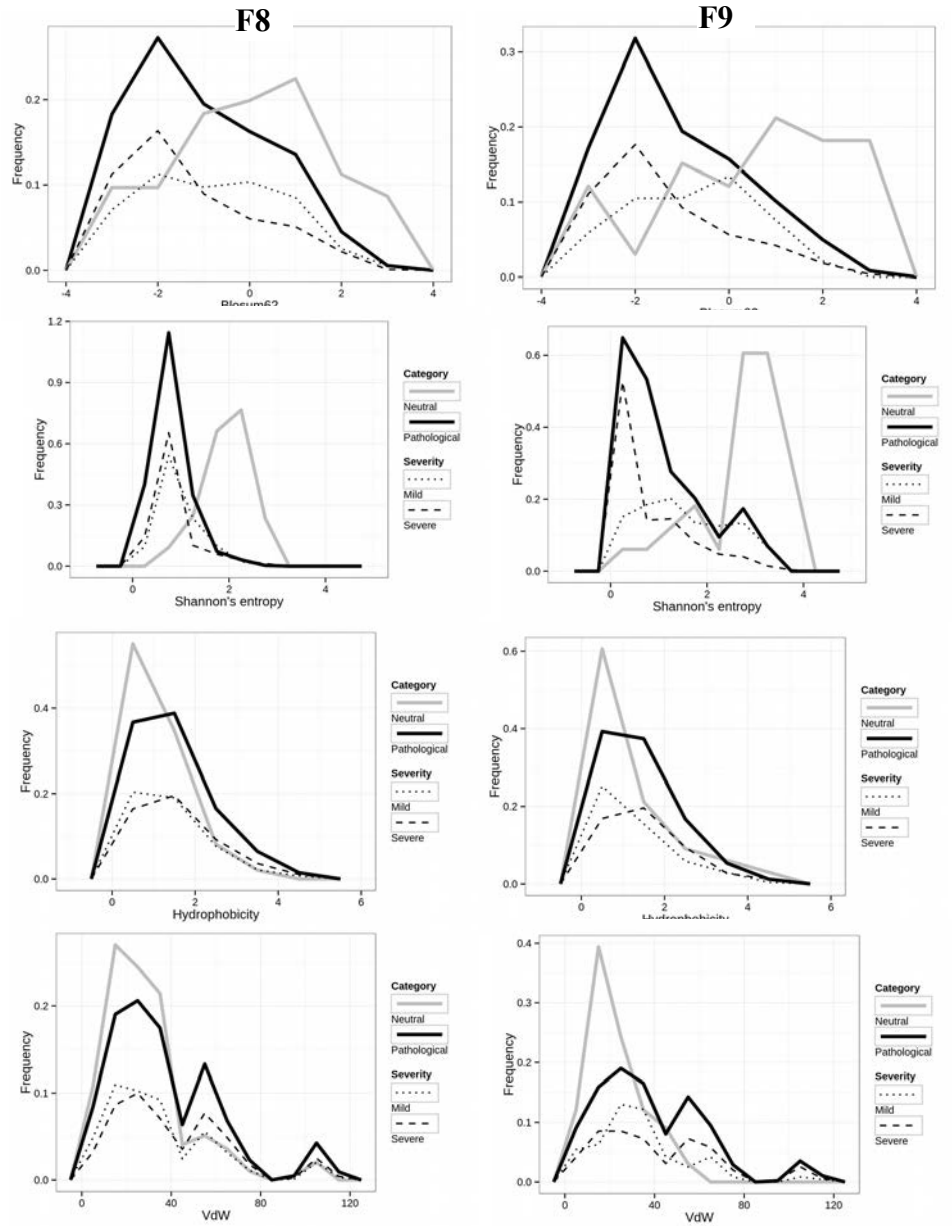

Figura 3.6A: *Histogrames amb la distribució de les variants greus (en fosc, a ratlles) i les lleus (en clar, puntejat) respecte les neutres (clar) i les patogèniques(fosc) per a blosum62, Entropia de Shannon i Deltes d'Hidrofobicitat i de Volum de Van der Waals.*

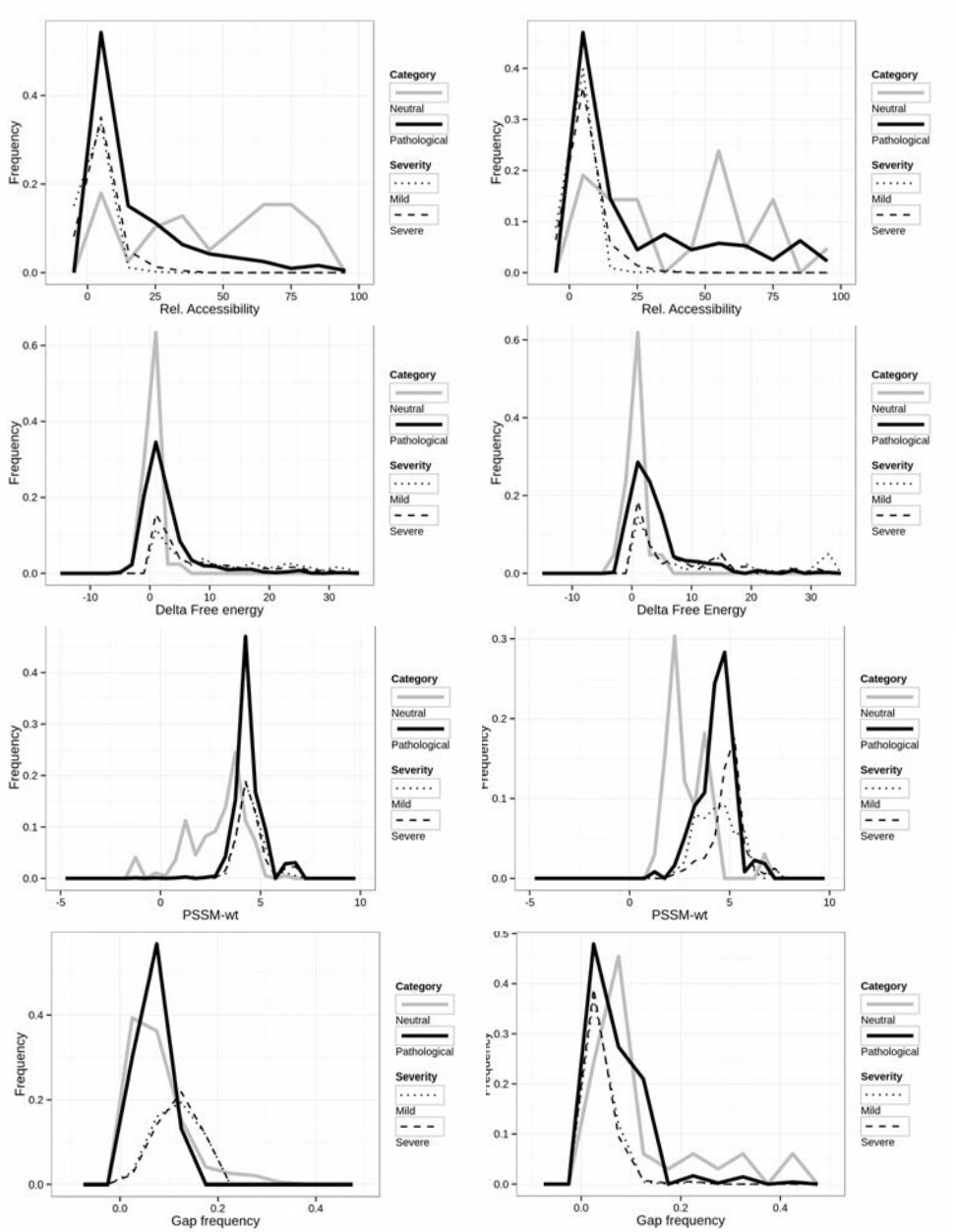

Figura 3.6B: *Histogrames amb la distribució de les variants greus (en fosc, a ratlles) i les lleus (en clar, puntejat) respecte les neutres (clar) i les patogèniques (fosc) per a l'accessibilitat relativa, la ΔΔG de foldX, la freqüència de gaps i el* $pssm_{nat}$.

Comparant al detall els casos de neutre-patogènic i de lleu-greu, es pot enriquir l'observació anterior.

En primer lloc, els descriptors fisicoquímics del canvi –blosum62, canvi d'hidrofobicitat i canvi en el volum de Van der Waals– generalment aporten més o menys el mateix: són bastant discriminatoris en el cas de neutre-patogènic, i menys en lleu-greu.

Els descriptors derivats de l'estructura tridimensional –ΔΔG i accessibilitat relativa– aparenten tenir més poder predictiu en el cas de la severitat (Taula 3.5) que en el cas de la patogenicitat (Taula 3.2) Cal assenyalar, però, que la diferència en el MCC també té un component tècnic d'origen mostral: hi ha poques variants neutres en comparació a les patogèniques, creant un des-balanceig molt gran (Baldi et al. 2000). Hi ha certa gradació de l'impacte molecular segons com canvia l'estabilitat de la proteïna, i també amb l'exposició al solvent de l'aminoàcid: com menor impacte o més exposat està el residu, més tendència hi ha a un efecte neutre; i com més extrem és el canvi o més enterrat, més probable és que l'impacte sigui més greu.

Finalment, pels descriptors evolutius –derivats de l'alineament–, es veu que el $pssm_{nat}$ és el descriptor que més varia: és molt predictiu en el cas de neutre-patogènic, pràcticament no discrimina en lleu-greu en F8, i manté molt poca discriminació en F9. L'entropia de Shannon, en canvi, que també és molt predictiva en neutre vs patogènic, manté una capacitat predictiva apreciable en la severitat, sobretot en el cas de F9. És a dir, els descriptors evolutius, que són els millors en el cas de la patogenicitat (Adzhubei et al. 2012; Riera et al. 2014; Niroula et al. 2015; López-Ferrando et al. 2017), no tenen tanta importància en la severitat, malgrat que els fisicoquímics mantenen predictibilitat. Aquestes

observacions podrien explicar-se pel fet que els MSA, que únicament tenen una seqüència per espècie, tot i reflectir bé quins possibles canvis tenen impacte sobre la fitness –relacionat amb la predicció neutre/patogènica–, no representen adequadament la magnitud del canvi –relacionat amb la predicció lleu/greu–: tota senyal que redueixi la fitness és seleccionada negativament independentment de la magnitud de la reducció. En e cas dels canvis fisicoquímics, que no depenen d'una senyal indirecta com la del MSA, la situació és diferent: és una senyal intrínseca computable directament i que queda mostrada amb els descriptors.

En resum, part de la disminució del poder predictiu en la severitat prové dels paràmetres evolutius, que són els més discriminants en el cas de la patogenicitat.

3.3.3 La predicció de la severitat de les variants requereix més informació que la predicció de la patogenicitat

En les seccions anteriors s'ha vist que la capacitat predictiva de les propietats emprades és menor per al problema de la severitat (Taula 3.2) que per al problema de la patogenicitat (Taula 3.5). Per entendre millor les raons d'aquest fet, es representen en una mateixa figura les distribucions per les diferents poblacions de mutacions (Figura 3.6). Les variants lleus i les variants severs són els dos casos extrems, quan es subdivideixen, de la patogenicitat. Les variants lleus tendeixen a valors més propers als de les variants neutres, i les greus a valors més allunyats. Això fa que les categories de lleu i sever tinguin una distinció molt més subtil que les de neutre i patogènic. Aquest fet s'explica tenint en compte que les dos primeres són una subdivisió del cas patogènic –i que aquesta és una distribució unimodal–. És a dir, que hi ha una reducció en la

informació disponible per al problema de discriminació, que hauria de ser compensada mitjançant la inclusió de més informació, de fonts diferents, per recuperar la capacitat predictiva perduda. I ni així és segur que pogués arribar-s'hi, ja que la informació que resta per incloure és molt complicada de representar amb un nombre petit de variables. Això es deu a que en el fenotip intervenen més factors, com l'ambient (Jelier et al. 2011), la resta de fons gènic (Girirajan i Elsea 2009), o fins i tot factors ambientals heretats (Burga i Lehner 2012). Tenir en compte aquests factors en un procés com el de construir un predictor és molt complex, ja que no n'hi ha mesures acurades, tenen una gran variabilitat, sovint pateixen de cert grau d'indefinició i no hi ha prou dades de qualitat per a parametritzar els models numèrics.

En resum, es veu que les prediccions de la severitat són un cas en el que la utilitat dels predictors no específics de proteïna (Niroula et al. 2017) és molt més limitada que en el de la predicció de la patogenicitat. Afegir informació específica de cada proteïna, quan existeix, hauria de dur els algoritmes a capacitats d'encert més properes a l'ús clínic, com ja les tenen en el cas de la patogenicitat. El cas de l'hemofília il·lustra el valor de la informació específica, ja que els extrems lleu-greu no són iguals en les dues proteïnes estudiades, F8 i F9: són dos casos molt similars funcionalment –estan en la mateixa via metabòlica, la cascada de la coagulació–, però tenen una gradació diferent del fenotip lleu, sent sent l'impacte funcional de les variants a F8 més greu molecular i fenotípicament (Tagariello et al. 2009) que a F9.

Finalment, cal remarcar que hi ha uns pocs casos més, a part de F8 i F9, en els quals s'ha observat que la relació entre el grau de l'efecte molecular i el fenotip és parcial. Per exemple, en la proteïna RS1, causant de la malaltia de

XLRS, Sergeev i col·legues (Sergeev et al. 2013) han trobat que la severitat de l'efecte molecular causat per la variant, mesurat mitjançant variables estructurals –via ΔΔG i accessibilitat relativa–, correlaciona amb el fenotip clínic en la malaltia .

No n'hi ha prou amb aquests resultats per poder generalitzat aquestes troballes a la resta de proteïnes com en el cas de la patogenicitat: la manca de mutacions amb notació de severitat n'és la barrera principal. Hi ha un altre factor, però, que també és molt limitant en aquest aspecte: la falta d'estàndards en l'anotació del fenotip. No sempre hi ha escales de severitat ben establertes, i decidir què és lleu i què és greu ha quedat, sovint, a la lliure interpretació dels autors de cada article. Aquests factors, específics per al problema de predicció de la severitat, són quelcom que podria resoldre's en un futur no gaire llunyà gràcies als esforços d'alguns projectes com el Variome Project (Smith i Vihinen 2015), que pretén estandarditzar la col·lecció de dades de variants obtingudes dels genomes. O bé mitjançant l'aplicació més generalitzada en l'anotació del fenotip de variants amb l'ontologia d'HPO (Köhler et al. 2014); que dóna uns termes comuns i unes directrius amb les que aplicar-los.

Solucionar aquests problemes obrirà les portes a l'aplicació de la predicció del fenotip en diagnòstic clínic o en estudis del genoma.

Capítol III

4. La fracció de CPD en l'incidentaloma

4.1 L'incidentaloma en la medicina de precisió

En el capítol I d'aquesta tesi s'ha vist com l'ús d'informació biomèdica permet estudiar la relació genotip-fenotip en el cas dels CPD i utilitzar la informació resultant per contrastar la validesa dels models evolutius en aquestes mutacions. En aquest capítol s'estudia una extensió del concepte de CPD al cas que tant la mutació patogènica com la compensatòria estiguin presents en la seqüència humana, no en una altra espècie. Aquest cas s'anomenarà hCPD (human-Compensated Pathogenic Deviations), i es veurà com el seu estudi té una inesperada rellevància mèdica en l'aplicació dels experiments de NGS al diagnòstic molecular. L'origen d'aquesta aplicabilitat rau en el que es coneix com incidentaloma.

El concepte d'**incidentaloma** apareix amb les primeres aplicacions de NGS en l'entorn clínic, i es defineix com el conjunt de canvis potencialment malignes en el genoma que es troben en un experiment de seqüenciació però que són fenotípicament asimptomàtics: no expliquen els símptomes de la malaltia del pacient, ni van acompanyats de símptomes propis (identificables) de la malaltia associada al gen. Per exemple, un pacient expressa símptomes de la malaltia de Fabry, i en l'experiment de seqüenciació presenta una mutació al gen de l'alfa-galactosidasa, que confirma el diagnòstic de Fabry. En el mateix experiment, s'observa que el pacient també és portador d'una variant potencialment patogènica a TP53, proteïna associada al càncer: Aquesta darrera mutació seria part de l'incidentaloma, juntament amb d'altres mutacions semblants que es poguessin trobar. En estudis generals de l'incidentaloma (Kohane et al. 2012) s'ha descrit que es poden arribar a identificar, de mitjana, fins a 100 variants amb potencial patogènic, per genoma estudiat. Això crea un

problema mèdic immediat: com identificar quina és la variant causal de la malaltia del pacient. També crea, però, un problema moral greu a l'hora d'informar els pacients sobre els resultats de la seqüenciació i la seva interpretació: cal reportar totes les variants que s'han trobat en el seu cas, encara que no tinguin res a veure amb la malaltia que es vol diagnosticar? Incloent-hi les que corresponguin a malalties sense tractament conegut? Aquesta situació ha generat un intens debat, tant en entorns clínics com en els de recerca, sobre quin és el protocol que s'hauria de seguir en aquest casos (Berg et al. 2013; Akle et al. 2015). Fins i tot hi s'han desenvolupat alguns projectes dedicats a catalogar millor els tipus de variació (Burn i Watson 2016), per ajudar a decidir en quins casos és millor informar-ne als pacients i ens quins casos no és necessari. Malgrat tot, no s'ha estudiat fins ara quines són les característiques de tanta variació a priori patogènica, ni si les variants indiquen el mateix o bé reflexen diferents situacions. De fet, hi ha diverses possibilitats: les variants corresponen a malalties que no s'han manifestat encara, a malalties amb baixa penetrància, a variants en gens que simplement produeixen canvis fenotípics no directament associats amb malaltia o a variants erròniament categoritzades a la literatura com a patològiques. Part de l'incidentaloma, fins i tot, podria correspondre a variants que en un individu en concret no són patogèniques perquè en una altra regió del genoma, per epístasis (Lehner 2011) –en el mateix gen, en un altre gen, o via una combinació de canvis – hi hagi un altre canvi que compensi el potencial efecte patogènic de la primera variant. Aquest darrer cas és particularment interessant per aquest treball, ja que està directament relacionat amb els CPD. L'única diferència és que, en el cas que s'estudia en aquest capítol, tant la mutació patogènica com la mutació compensatòria es trobarien en el genoma humà

La presència de compensacions moleculars a canvis que d'altra manera serien patogènics és un fet que s'ha descrit *in vivo*, en la literatura (Brandis i Hughes 2013; Xu i Xhang 2014; Jordan et al. 2015; Moura de Sousa 2017), però que, fins ara, pràcticament no s'ha trobat en humans, malgrat hi hagi algun cas concret estudiat (Mankad et al. 2006). És, però, una categoria que no hauria de passar desapercebuda ja que el fet que no puguin existir-ne casos dins del mateix ser humà seria de gran importància en el context de la medicina de precisió. Concretament, ajudaria a millorar el rendiment diagnòstic a l'aplicar experiments de NGS en entorns clínics.

En aquest capítol es presenta una primera aproximació a la quantificació del nombre de hCPD, en individus sans. El punt de partida en aquest capítol serà l'estudi del genotip dels individus seqüenciats en l'experiment de seqüenciació massiu de 1000G (The 1000 Genome Project Consortium 2015). Es caracteritzaran totes les variants patogèniques fent servir un mètode de predicció de CPD, entrenat a partir de les característiques fisicoquímiques i evolutives d'aquestes mutacions, molt properes als hCPD. Al final d'aquest procés, s'obté una estimació del nombre de hCPD en la població estudiada.

4.2 Materials i Mètodes

Buscar variants compensades és una tasca difícil (Kondrashov et al. 2002). Això es deu a que no hi ha una metodologia ben definida i simple per trobar en estudis particulars d'una sola proteïna les compensacions associades a una variant: es requereixen extensos estudis de mutagènesis i provés bioquímiques que serveixin de representació de la fitness (Jordan et al. 2014). Si, a més a més, l'estudi de les possibles compensacions té un àmbit global, no restringit a una sola proteïna, el nombre de possibilitats augmenta enormement: les

potencials compensacions intergèniques poden estar en més de vint-mil proteïnes, i, fins i tot quan es redueixen candidats als del possible interactoma de la proteïna –quan està descrit i si és fiable (Alvarez-Ponce 2016)– el nombre de graus de llibertat segueix sent molt alt. A més a més, les possibles representacions de la viabilitat d'una proteïna depenen molt del context, tal i com s'ha vist en els capítols 1 i 2, i no hi ha encara un paràmetre prou representatiu per utilitzar globalment. De fet, en el primer capítol s'ha vist que ni la ΔΔG no dóna una representació prou acurada de la fitness del mutant més que en alguns casos particulars i aïllats: altres representacions suggerides –predictors com EV-mutation (Hopf et al. 2017) o CADD (Kircher et al. 2014), per exemple–, no donen cap criteri per a interpretar els seus resultats des d'un punt de vista de la viabilitat.

Així doncs, el procediment seguit per buscar hCPD tindrà un caire probabilístic i descriptiu. La metodologia seguida consta de, **primer**, entrenar un predictor amb les dades de compensació més àmplies que es poden aconseguir: la compensació que s'ha anat adquirint al llarg dels milions d'anys d'evolució. Són les CPD «clàssiques» analitzades al Capítol 1, però en aquest cas per a tot el genoma. Tot i que algunes restriccions poden ser diferents entre CPD i hCPD, un predictor que generalitzi sobre els patrons habituals de compensació té utilitat descriptiva (Cheng et al 2015). **Segon**, s'utilitza aquest predictor de CPD per a buscar, en individus prèviament seqüenciats, variants patogèniques candidates a tenir compensacions. **Finalment**, s'usa un tractament estadístic del resultat per veure quina probabilitat hi ha que existeixin hCPD en aquestes condicions; tot comparant les dades d'individus reals amb dades sintètiques.

4.2.1 El predictor de CPD

Obtenció de CPD: Els alineaments

Tal i com s'ha vist anteriorment, l'obtenció de predictors requereix, en primer lloc, obtenir un conjunt de variants que serveixi per al procés d'aprenentatge. En aquest cas, la font de variants escollida per a servir com a punt de partida de la predicció és la base de dades de *humsavar* d'Uniprot (Apweiler et al. 2004), on hi ha recollits canvis de residu en proteïnes humanes que apareixen a la literatura. Es seleccionen totes aquelles mutacions amb anotació de *DISEASE* - malaltia-, que es fan servir com a patogèniques.

Per a cadascuna de les proteïnes corresponents a les variants obtingudes es creen dos MSA de seqüències d'ortòlegs.

El primer MSA –a partir d'ara anomenat Ensembl– es crea amb seqüències provinents d'ortòlegs de la base de dades Ensembl (release 86) (Vilella et al. 2009) utilitzant Muscle (Edgar 2004) com a alineador múltiple, com al capítol 1. Les seqüències d'Ensembl provenen d'estudis de seqüenciació massiva de tot el genoma de diversos organismes de referència, i l'ortologia de seqüències s'obté mitjançant un estudi de comparació entre totes les proteïnes de la referència –*Homo sapiens*– respecte totes les proteïnes de l'organisme en qüestió. Aquests alineaments, llavors, són curts, ja que el nombre de seqüències és menor de 100, però és d'alta confiança: hi ha molts menys espais buits –gaps– i el el patró de conservació serà més apreciable visualment.

El segon MSA –d'ara en endavant anomenat euNOG– s'obté a partir de les dades d'eucariotes d'eggNOG v4.5 (Huerta-Cepas et al 2016), una base de dades que conté grups d'ortòlegs predits automàticament. Aquests MSA són més grans

i tenen molta més varietat, podent arribar a tenir més de 1000 organismes cadascun, però també són menys precisos i contenen molts més possibles errors d'alineament. La idea a l'usar euNOG és poder captar patrons evolutius més àmpliament, i tenir una representació més acurada dels camins evolutius que pot seguir la proteïna per l'espai de seqüència (Povolotskaya i Kondrashov 2010).

La combinació dels dos mètodes dóna dos visions complementàries sobre els CPD: una més selectiva però més fiable, i una de més àmplia i amb més varietat però amb més possibles fonts d'error. Els següents passos descrits es repeteixen per a cadascun dels alineaments.

Obtenció de CPD: l'algoritme per buscar CPD

L'obtenció de CPD es fa seguint criteris molt similars al capítol I, però ajustant lleugerament en euNOG el nombre de seqüències: la variant patogènica ha d'aparèixer al menys en seqüències de 5 organismes diferents. Cada variant d'UniProt queda anotada com a CPD o no CPD, i totes aquestes seran anotades mitjançant un conjunt de descriptors que caracteritzen la seqüència en termes evolutius i fisicoquímics (Figura 4.1).

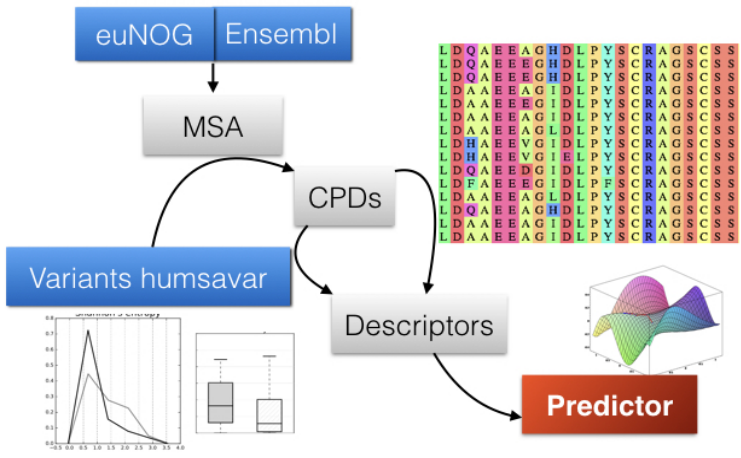

Figura 4.1: *Diagrama del procés seguit per obtenir un predictor de CPDs.* ***i)*** *Es creen els MSA,* ***ii)*** *Amb les variants, es busquen als MSA si són CPD,* ***iii)*** *Es caracteritzen els CPD i els no CPD;* ***iv)*** *S'entrena l'algoritme.*

Els descriptors dels CPD

S'usen tres tipus de paràmetres, descrits a continuació: Derivats de MSA (i per tant molt dependents de la font d'ortòlegs escollida), derivats del canvi de residu i anotacions de bases de dades externes.

Els descriptors calculats a partir dels MSA han estat: entropia de Shannon (descrita en el capítol I); entropia de Shannon en finestres de 5 i 10 aminoàcids; informació mútua màxima de la posició vs la resta de posicions de l'alineament, informació mútua mitjana de la posició respecte la resta, informació mútua ajustada màxima, informació mútua ajustada mitjana i identitat de seqüència a la posició i en finestres de 5 i 10 aminoàcids.

L'entropia de Shannon en finestres és una mètrica que fa la mitjana del valor de l'entropia de Shannon (Shannon 1948) (capítol I) en els residus que estan seqüencialment a prop de l'aminoàcid, en aquest cas 5 o 10 residus endavant i enrere, més l'entropia de l'aminoàcid en qüestió. Són mesures correlacionades però que donen informació no només de les restriccions evolutives de l'aminoàcid, sinó dels residus que estan prop d'aquest.

***A**: Donades les posicions en l'alineament* U *i* V*, en els que* i *són els aminoàcids en* U*, i* j *són els aminoàcids en* V*,* a_i *el nombre d'elements a* U*, i* b_i *el nombre d'elements a* V*:*

$$MI(U,V)=\sum_{i=1}^{a_i}\sum_{j=1}^{b_i} p(i,j)\log\left(\frac{p(i,j)}{p(i)\,p(j)}\right)$$

on p(i,j) és la fracció de casos on i *i* j es donen alhora*, p(i) és la fracció de l'aminoàcid* i *a* U*, i p(j) la fracció de* j *a* V.

***B**: Donada l'entropia de Shannon a* U *com a* H(U)*, i a* V *com a* H(V)*; i donat* n_{ij} *= max(1,* a_i+b_i-N*); on* N *és 20, el nombre de possibles aminoàcids, la MI esperada adoptant un model hipergeomètric és:*

$$E[MI(U,V)]=\sum_{i=1}^{a_i}\sum_{j=1}^{b_i}\sum_{n_{ij}}^{min(a_i,b_i)}\frac{n_{ij}}{N}\cdot\log\left(\frac{N\cdot n_{ij}}{a_i b_j}\right)\cdot\frac{a_i!\,b_j!\,(N-a_i)!\,(N-b_j)!}{N!\,n_{ij}!\,(a_i-n_{ij})!\,(b_j-n_{ij})!\,(N-a_i-b_j+n_{ij})!}$$

i, llavors:

$$AMI(U,V)=\frac{MI(U,V)-E[MI(U,V)]}{max(H(U),H(V))-E[MI(U,V)]}$$

***Figura 4.2**: Equacions per calcular **(A)**: la Informació Mútua (MI) i **(B)**: la Informació Mútua Ajustada (AMI).*

Els paràmetres d'informació mútua (Everett 1956), que es calculen comparant la posició de l'aminoàcid en l'alineament respecte totes les altres, dóna informació

de possibles restriccions coevolutives i de potencial epístasi intragènica que han tingut aquelles posicions de forma conjunta. És a dir, tenen el potencial d'identificar possibles tendències compensatòries dins la mateixa proteïna. L'equació per al càlcul de la informació mútua es troba en la Figura 4.2A. La informació mútua ajustada (Vinh et al. 2010) és una versió que corregeix per la possibilitat de que part de la senyal pot ser deguda a l'atzar, per el que dóna una visió més completa però considerablement més costosa de calcular (Figura 4.2B). En cadascuna, es calcula tant una mitjana amb totes les posicions com la màxima entre totes les posicions. S'ha usat la implementació del paquet de python scikit-learn (versió 0.18.1) (Pedregosa et al. 2011) per a calcular aquests paràmetres.

La identitat de seqüència en la posició es calcula com el nombre de d'aminoàcids idèntics en una posició entre el nombre de seqüències, i la identitat en finestres és la mitjana de la identitat en totes les posicions escollides. La formula és: identitat de posició = nombre d'aminoàcids idèntics (a referència) / nombre d'aminoàcids total.

Els descriptors relacionats amb l'impacte molecular del canvi han sigut la diferència d'hidrofobicitat (Fauchere i Pliska 2983), la diferencia en el radi de Van der Waals (Bondi 1964), el valor de la matriu Blosum62 (Henikoff i Henikoff 1992) –descrits en el capítol I– i el valor de la matriu de Pam250 (Dayhoff et al. 1978). Aquestes variables descriuen explícita o implícitament l'impacte fisicoquímic del canvi de residu.

Els descriptors extrets de bases de dades externes han sigut: i) l'anotació de la funció del residu en UniProt –tal i com està descrit al capítol II– i ii) l'anotació de l'expressió monoal·lèlica del gen, tal i com està descrit en la base

de dades de dbMAE (Savova et al 2016). De dbMAE s'extreuen dos descriptors: un amb la informació experimental (monoal·lèlic o no) i un altre amb les prediccions.

El model de predicció: boosting de pendent extrem

Una vegada calculats tots els valors per al conjunt de variants, s'entrena un model d'aprenentatge automàtic per separat per a cadascun dels conjunts de CPD –euNOG i Ensembl–. S'utilitza com a mètode d'aprenentatge el boosting de pendent extrem –*Extreme gradient boosting*– amb el paquet de xgboost (v0.6) (Chen i Guestrin 2016), amb modificació d'algun híper-paràmetre: n_estimators=500 i learning_rate=0.12 en ambdós; i, en Ensembl, scale_pos_weight= nombre de no CPD dividit pel nombre de CPD–.

El boosting de pendent extrem (Chen i Guestrin 2016) és un algoritme que modifica l'algoritme de boosting mitjançant l'ús d'una funció de pèrdua diferenciable en l'optimització, que és la que s'utilitza per actualitzar els pesos dels descriptors. El boosting (Freund i Schapire 1995) és un algoritme consistent en combinar diversos classificadors dèbils, en aquest cas arbres de decisió, en un classificador més robust. El boosting està relacionat amb els coneguts dels Random Forests (Ho 1995), però es diferència en el tipus d'arbres utilitzats: Random Forest usa diversos arbres de decisió complerts (baix biaix però alta variància) i els ajunta amb agregació per boostrap, reduint la variància al fer la mitjana dels resultats. En canvi, el boosting usa classificadors febles (molt biaix però poca variància), i s'optimitza mitjançant la creació de grups –*ensembles*– que redueixen el biaix global dels classificador (Freund i Schapire 1995). El boosting de pendent va actualitzant iterativament la classificació amb nous arbres febles que corregeixen l'error dels classificadors anteriors

mitjançant la funció de pèrdua, juntament amb un element de regulació que controlar la complexitat del model –que es tradueix en un control del sobre-ajust–. En resum, el boosting de pendent redueix la variància i el biaix, i n'evita el sobre-ajust. El paquet escollit, xgboost, és una implementació molt ràpida, acurada i flexible del boosting de pendent.

El model de predicció: Les mesures d'encert

L'obtenció dels valors de predictibilitat del model es fa mitjançant una validació creuada amb 10-particions i fent la mitjana dels resultats en 100 llavors diferents –com al capítol 2–, tot evitant donar uns valors afectats amb circularitats (Vihinen 2012; Grimm et al. 2015) i evitant que estiguem valorant un comportament sobre-ajustat a les dades d'entrenament. Les mètriques usades són el coeficient de correlació de Matthews, la sensibilitat, l'especificitat, l'exactitud –descrits en el capítol II–, l'àrea sota la corba ROC (AUC) (Fawcett 2006) i el valor-F (F1s) (Powers 2011).

L'AUC és la probabilitat que un classificador assigni una probabilitat més alta a un cas aleatori positiu que a un cas aleatori negatiu. És l'àrea que formen les corbes formada per els casos positius verdaders (TPR) en un eix, amb el ràtio de casos de falsos positius (FPR) – també definit com a 1-especificitat– en l'altre eix..

F1s és, formalment, la mitjana harmònica de la precisió i l'exhaustivitat. F1s = 2 * (precisió*exhaustivitat) / (precisió+exhaustivitat). És una mesura robusta davant classificacions amb un nombre mal balancejat de casos – molts casos negatius i molt pocs de positius, per exemple, on una estratègia que donaria bons resultats en altres mesures seria simplement classificar-ho tot com a

negatiu –.

El conjunt d'aquests valors dóna una visió general de la predicció en global, de la predicció malgrat possibles biaixos en la composició de les dades, de com de bé es prediuen els casos positius i de com de bé els negatius. Tots ells es calculen mitjançant el paquet de python de scikit-learn (v.0.18.1) (Pedregosa et al. 2011).

4.2.2 L'incidentaloma a 1000 genomes

Els predictors obtinguts s'han aplicat per a buscar potencials hCPD i estimar el seu nombre per individu. Aquest anàlisis a nivell individual, es va fer amb les dades públiques del projecte de 1000 genomes (The 1000 Genome Project Consortium 2015), que conté dades genotípiques de les 2504 persones de diverses poblacions. Aquestes dades es poden obtenir del repositori de ftp públic, on hi ha informació individualitzada de tots els canvis de nucleòtid que es troben, a més d'algunes anotacions a nivell global –com freqüències poblacionals dels SNP–, i a nivell puntual –com prediccions de l'efecte funcional del canvi de nucleòtid–.

El protocol d'identificació dels hCPD es descriu a la Figura 4.3. En primer lloc es recullen els genotips de tots els individus i es processen per obtindre'n només aquells canvis que modifiquen la seqüència de la proteïna, juntament amb la descripció de la substitució d'aminoàcid corresponent. Posteriorment s'anoten totes les variants obtingudes amb els descriptors dels predictors euNOG i Ensembl (Figura 4.3). Es prediu el potencial com a CPD per a totes les variants amb les quatre combinacions possibles de predictors i MSA: euNOG amb descriptors euNOG, euNOG amb descriptors Ensembl, Ensembl

amb descriptors Ensembl i Ensembl amb descriptors euNOG. Juntament amb la predicció binària, es guarda la probabilitat que el mètode usa per a classificar la variant: internament assigna un valor entre 0 i 1, si és major de 0.5 l'assigna com a categoria CPD, si és menor de 0.5 l'assigna com a no CPD. Aquest valor és proporcional a la confiança que té el predictor amb l'assignació de la categoria, així que, en un cas teòric com aquest, és útil fer-lo servir en càlculs posteriors.

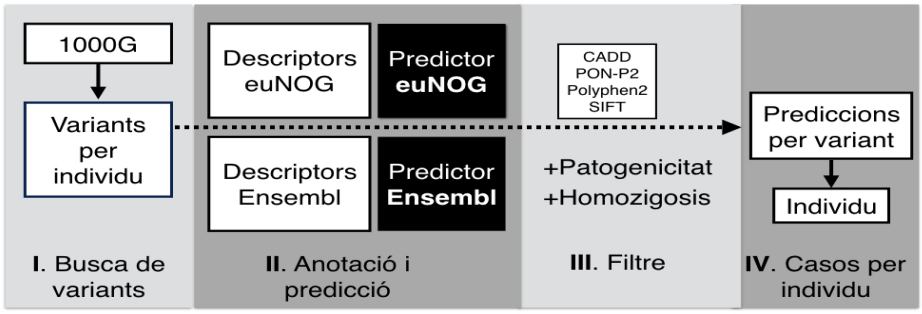

Figura 4.3: *Diagrama del procés seguit per analitzar les dades de 1000G.* ***i)*** *Processar les variants,* ***ii)*** *Calcular els descriptors en ambdós conjunts, i calcular les probabilitats dels predictors,* ***iii)*** *Selecció de les variants patogèniques i les variants per les que cada individu és homozigot,* ***iv)*** *Obtenció del genotip de cada individu.*

Una vegada recollides les probabilitats de CPD per a tots els casos, es mira com estan distribuïts en cada individu, seleccionant les variants que són possibles canvis patogènics (Figura 4.3). El criteri usat és que al menys dos de quatre predictors usats –Polyphen2 (Adzhubei et al. 2010), CADD (punt de tall 15) (Kircher et al. 2014), SIFT(Sim et al. 2012) i PONP2 (Niroula et al. 2015)– el prediguin com a patogènic.

Finalment es seleccionen només les mutacions per a les que l'individu és homozigot.

Les variants obtingudes al final d'aquest protocol –predites com a patogèniques i com a CPD – són les que s'usaran en els anàlisis posteriors. Constitueixen una representació de l'incidentaloma de cada individu.

4.2.3 La probabilitat de compensació en l'incidentaloma

La probabilitat de hCPD

Per a tenir una primera mesura estadística de si realment és possible tenir hCPD, o, dit d'altra manera, per establir si part de l'incidentaloma s'explicar per la presència de mecanismes compensatoris, es parteix d'un model simple de la probabilitat de tenir hCPD. Es mira per a cada possible hCPD que té l'individu quina és la seva probabilitat de ser-ho mitjançant la fórmula $p(hCPD)=p(hCPD\text{-}euNOG)*p(hCPD\text{-}Ensembl)*\eta$. En aquest cas, el valor η és un factor corrector empíric emprat per incloure l'efecte de possibles factors externs: els errors en els predictors de patogenicitat o el fet que no tots els possibles hCPD estaran realment compensats, és a dir, factors que no hi ha forma de comprovar formalment ni experimentalment. Com a valors possibles per a η hi ha dues opcions extremes: La primera, 0.7, correspondria a un valor molt conservador de la confiança dels predictors de patogenicitat, suposant que funcionen bé en al menys el 0.7 dels casos (Riera et al. 2014). La segona opció per a η, 0.07, multiplicaria el valor anterior per 0.1, la fracció que treballs anteriors en CPD (Kondrashov et al. 2002; Ferrer-Costa et al. 2007; Barešić et al. 2010) han definit com a consens en la fracció de variants patogèniques que a priori són CPD. Es decideix treballar amb la versió més conservadora, $\eta=0.07$.

L'expressió triada per p(hCPD), addicionalment, fa la suposició conservadora que els factors p(hCPD-euNOG) i p(hCPD-Ensembl) són independents, quan és probable que no sigui així. Així doncs, el valor obtingut per la probabilitat de que una variant sigui hCPD serà més baix que amb altres conjectures, però aplicant paràmetres més restrictius es reforçen les possibles conclusions positives posteriors.

Amb les p(hCPD) per a cada variant, es pot calcular una probabilitat de que un individu tingui al menys una hCPD. p(individu té algun hCPD) = 1-p(individu no té cap hCPD), i aquesta p(individu no té cap hCPD) és el producte de la probabilitat de cada variant de no ser hCPD, o 1 – p(hCPD).

La població sintètica

Finalment, s'ha introduït una última comprovació, per controlar la possibilitat de que el resultat obtingut tingui un origen tècnic i no biològic. Amb aquest objectiu, s'obté un model aleatori per la probabilitat que un individu no tingui hCPD (sota qualsevol η escollida); aquest model s'ha construït a partir d'una població en la que els individus no tenen hCPD. Obtenir una població natural així és impossible, pel que en el seu lloc es crea una població sintètica agafant com a punt de partida totes les possibles variants patogèniques que hi pot haver en un genoma segons el criteri anterior – al menys dos dels quatre predictors diuen que són patogèniques– de la base de dades de dbNSFP (v3.4a) (Liu et al. 2016). Per cada individu d'aquesta població, es crea un genotip, en el que el nombre de variants s'obté mitjançant el següent protocol.

(i) Obtenir un nombre a l'atzar, entre 0 i 100, que correspondrà al percentatge sobre el quantil (de 0 a 100) de la població.

(ii) Es recullen a quina població correspon cada individu de 1000G. A l'atzar s'escull una d'aquestes, sense reemplaç entre individus.

(iii) A continuació, s'extreu el nombre de variants de l'individu. Aquest correspon al valor –obtingut al pas (i) – del quantil de la distribució gaussiana de variants patogèniques de la població de 1000 genomes obtinguda al pas (ii).

(iv) S'extreuen a l'atzar de dbNSFP tantes variants com s'ha obtingut del pas anterior i s'assignen a l'individu. En aquest punt s'obté el genotip de l'individu sintètic.

(v) S'aplica tot el procés descrit en l'apartat 4.2.2 a les variants obtingudes del pas (iv), calculant-ne els descriptors i les probabilitats de ser CPD.

(vi) S'aplica la fórmula de probabilitat d'hCPD en cada variant, i amb elles s'obté la probabilitat per individu.

(vii) Ajuntant tots els individus, s'obté la distribució de probabilitat, que correspon al model d'atzar.

Les dos poblacions, 1000G i sintètica, es comparen en euNOG i Ensembl mitjançant el test estadístic de Kolmogorov-Smirnoff (Kolmogorov 1933).

4.2.4. Programari utilitzat

Tots els gràfics d'aquest capítol s'han realitzat amb els paquets de python de pandas (v0.18.1) (McKinney 2010), matplotlib (v1.5.1) (Hunter 2007) i seaborn (v0.7.1) (Wascom et al. 2016).

La busca de CPD i el càlcul d'alguns dels descriptors s'ha fet amb un programa propi amb el lenguatge Python (v2.7.11) fent ús del paquet numpy (v1.11.3) (van der Walt 2011).

4.3. Resultats i discussió

4.3.1 Construcció del predictor de CPDs

Els conjunts de CPD

El primer pas va ser la construcció dels conjunts de CPD, seguint el procés de la secció 4.2.1. EL nombre de CPD obtinguts varia en funció dels alineaments múltiples usats. En aquest cas, s'obté un total de 4177 i 13915 CPD per a Ensembl i euNOG respectivament. El percentatge de CPD recollits és major que altres aproximacions usades a la literatura (Kondrashov et al. 2002; Ferrer-Costa et al. 2007; Jordan et al. 2015), que consideraven el percentatge de CPD sobre les variants al voltant del 10%; en aquest cas s'ha recollit al voltant d'un 15% amb Ensembl, i al voltant d'un 50% amb euNOG. Aquestes diferències tenen dos origens. Primer, per l'aplicació d'unes restriccions menys exigents en els passos de filtrat i de nombre de seqüències d'altres espècies necessàries per a definir la variant com a CPD. I, segon, per la diferència en el tipus d'alineament utilitzat, un aspecte important en el conjunt euNOG.. La gran quantitat de seqüències que formen part d'aquest conjunt permet un mostreig més gran de l'espai de seqüència, oferint una font de noves CPD legítimes, encara que aquesta major quantitat de seqüències pugui portar associats més errors d'alineament, traduïts en un cert percentatge de falsos positius. En aquest cas, la presència d'errors en la categorització no seria, tampoc, un factor extremadament dolent. Els algoritmes d'aprenentatge automàtic basats en arbres de decisió –entre altres–, amb una quantitat de dades mitjanament gran com aquesta, són robustos a dades amb soroll en la categorització dels casos (Ghosh et al. 2017).

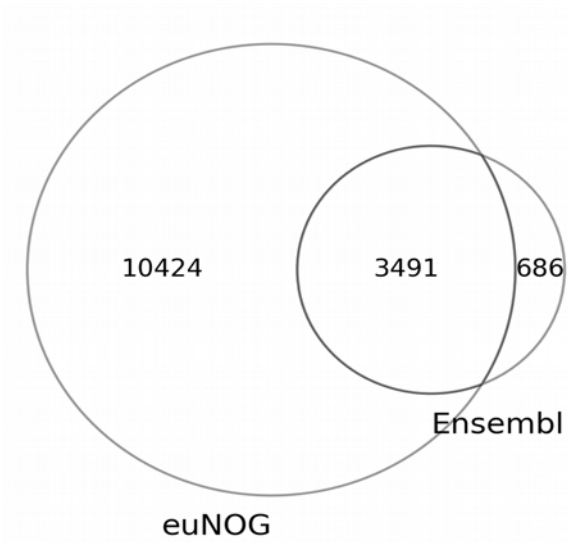

***Figura 4.4**: Diagrama de Venn dels CPD trobats amb els MSA d'euNOG i amb els d'Ensembl.*

Si es comparen els dos conjunts de CPD, s'observa que hi ha un considerable grau de solapament (Figura 4.4), tot i que el 15% dels CPD d'Ensembl no són trobats per euNOG, fet sorprenent per la major magnitud dels MSA d'euNOG. Aquest efecte pot ser degut a: **i**) Errors d'alineament en algun dels dos MSA, tan en Ensembl per trobar el CPD, o **ii)** com en euNOG per a no trobar-lo; **iii**) que el possible CPD en Ensembl estigués en un nombre baix de seqüències i, que malgrat aparèixer en un nombre també baix en euNOG, no són seqüències suficients per al filtre més estricte en euNOG.

Els atributs dels CPD

Malgrat la diferència tant en nombre de CPD que s'han trobat –molt més alta en euNOG–, com en la composició i llargada dels MSA –molt més gran en euNOG–, els CPD trobats tenen un comportament molt semblant en relació als paràmetres emprats per mesurar-ne l'impacte molecular. En la Figura 4.5 s'aprecia que, en general, els CPD presenten una tendència a ser més «lleus» que els no CPD en el seu impacte molecular intrínsec –Blosum62 i

hidrofobicitat–, d'acord amb el que s'havia observat al capítol I.

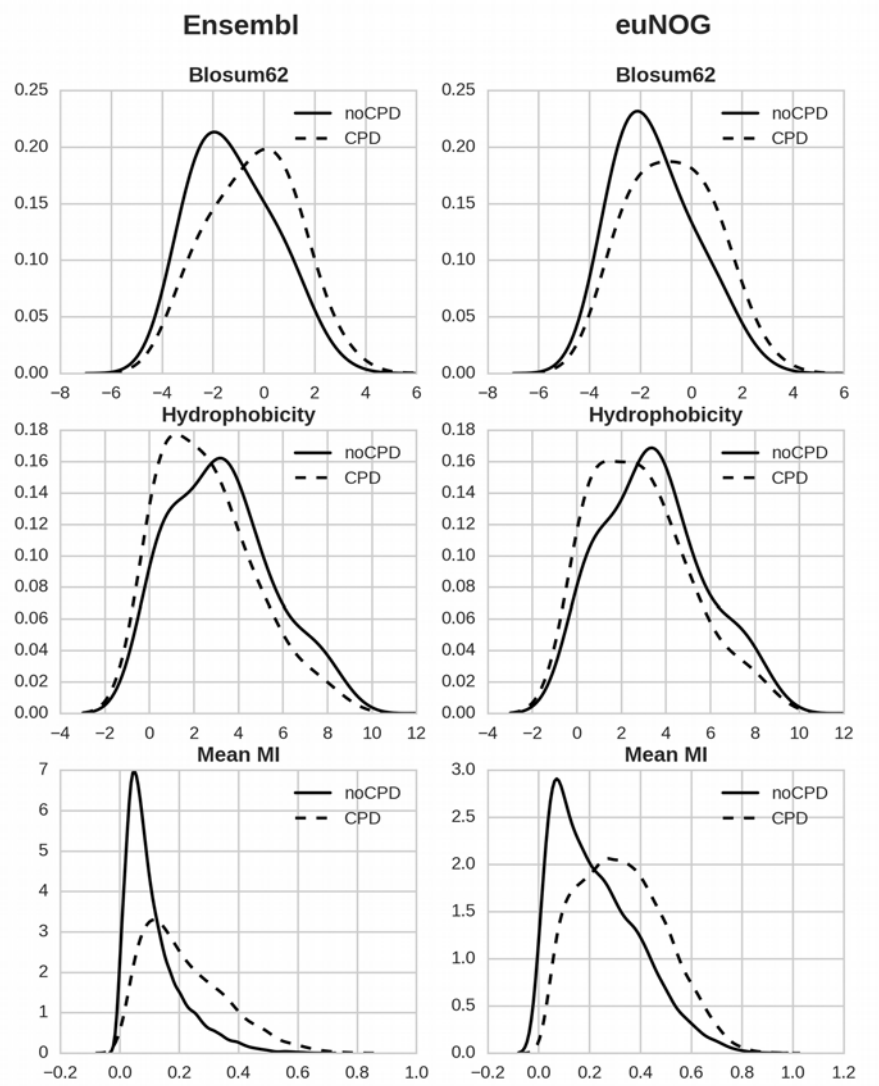

Figura 4.5: *Kernel density estimation (KDE)* de diversos descriptors – Blosum62, Canvi d'hidrofobicitat i Informació Mútua mitja – per la població de CPD i per la població de no CPDs en els dos conjunts. *l'eix d'ordenades reflexa un valor adimensional, el valor del qual depèn dels valors a l'eix d'abscises, de forma que l'àrea de la corba resultant sigui 1.*

Els valors del paràmetre de la «Informació Mútua mitjana» mostren, a més a més, que les posicions amb CPD estan més correlacionades amb canvis simultanis d'aminoàcid en altres posicions, que les posicions amb variants no CPD. Aquesta observació és important, ja que torna a deixar palesa la importància evolutiva dels CPD, i mostra que canvis en aquestes posicions concretes estan més **correlacionats** amb canvis a altres posicions del que s'esperaria si només afectés l'atzar. Això és consistent amb els coneixements previs de les CPD, a les zones candidates a la compensació.

La creació dels predictors de CPD

Els predictors entrenats amb les dades i paràmetres descrits, aconsegueixen diferenciar els CPD dels no CPD, tal i com es pot observar en la taula 4.1. LA capacitat predictiva és superior a la de l'atzar. Per exemple, els valors de MCC en ambdós conjunts de dades són comparables a la capacitat predictiva que tenen alguns mètodes de predicció de patogenicitat (Riera et al. 2016).

Font de dades	*Acc*	*Sns*	*Spc*	*Mcc*	*Auc*	*F1s*
Ensembl	0.74	0.76	0.71	0.48	0.74	0.75
euNOG	0.72	0.70	0.72	0.32	0.71	0.44

Taula 4.1*: Mètriques de predicció dels models de classificació de variants en CPD/no CPD. Provenen de la mitjana de la cross-validació 10-fold en 100 llavors diferents.*

És important assenyalar que els resultats també són consistents entre les dos classes: s'aprèn a diferenciar igualment bé entre els casos de CPD i els casos de no CPD, i l'algoritme no ha quedat esbiaixat en cap direcció, fins i tot havent-hi, en el cas d'Ensembl, un desproporció gran en les dades –nombre de CPD <<

nombre de no CPD–.

Per analitzar la capacitat de generalització dels models generats, també es calculen les prediccions creuades, en les que un predictor entrenat amb un conjunt de dades s'aplica a les dades de l'altre conjunt (Taula 4.2). Els resultats són pitjors que a la taula 4.1, però apunten a una certa capacitat de reconeixement. Si es va al detall, es veu que, per una banda, el predictor d'Ensembl validat amb les dades d'euNOG manté unes mètriques millors que l'atzar en tots els camps, sense cap biaix cap a CPD o cap a no CPD. Per altra banda, el predictor d'euNOG validat amb les dades d'Ensembl, malgrat mantenir unes mètriques globals bones –MCC, AUC, F1S i exactitud– , té un biaix considerable cap a predir qualsevol variant patogènica com a no CPD: encerta en el 96% dels casos dels que no són CPD per a Ensembl, però només prediu correctament com a CPDs el 27% d'aquests. Aquest efecte queda parcialment esmorteït en el cas dels paràmetres globals –MCC, AUC, F1s i exactitud– degut a que al conjunt d'Ensembl hi ha més no CPD –~23000– que CPD –~4000–.

Font predictor vs dades	*Acc*	*Sns*	*Spc*	*Mcc*	*Auc*	*F1s*
Ensembl vs euNOG	0.62	0.61	0.63	0.24	0.62	0.62
EuNOG vs Ensembl	0.85	0.27	0.96	0.33	0.62	0.37

Taula 4.2*: Mètriques de predicció dels models de classificació de variants en CPD/no CPD, creuant el predictor entrenat amb uns descriptors amb els descriptors obtinguts de l'altre conjunt de MSA.*

Les diferències de comportament a nivell de especificitat i sensibilitat possiblement rauen en el comportament de paràmetres evolutius com l'

«informació mútua mitjana». A la Figura 4.5 es veu que el punt de tall entre les corbes de CPD i de no CPD canvia: és de ~0 .14 a Ensembl i de ~0.23 per euNOG. Aquesta diferència, tot i ser petita, té un efecte considerable al predir les dades Ensembl amb el model euNOG, ja que la distribució de valors per als no CPD està molt més concentrada en el cas d'Ensembl. Aquest canvi en els punts de tall entre les distribucions CPD i no CPD prové de les diferències de composició dels MSA: euNOG té més seqüències més diferents entre elles, que comporta una pèrdua de correlació entre posicions. En resum, l'experiment d'informació creuada suggereix que el model més fiable és l'obtingut a partir d'Ensembl, i que el model d'euNOG és més limitat.

Finalment, l'altre factor que la predictibilitat creuada aporta és la reducció en la sospita que part de la capacitat predictiva provingui d'una fuga de dades (Kaufman et al. 2011). Dit d'altra manera: que el que realment veuria l'algoritme, de forma indirecta, fos la classificació de la variant en CPD o no CPD; i per tant els resultats posteriors serien artefactes. La predictibilitat creuada mostra que sí que hi ha component predictiu en els paràmetres escollits, i, encara que la fuga de dades no es pot descartar totalment, valida la metodologia.

4.3.2 La variació a 1000 genomes

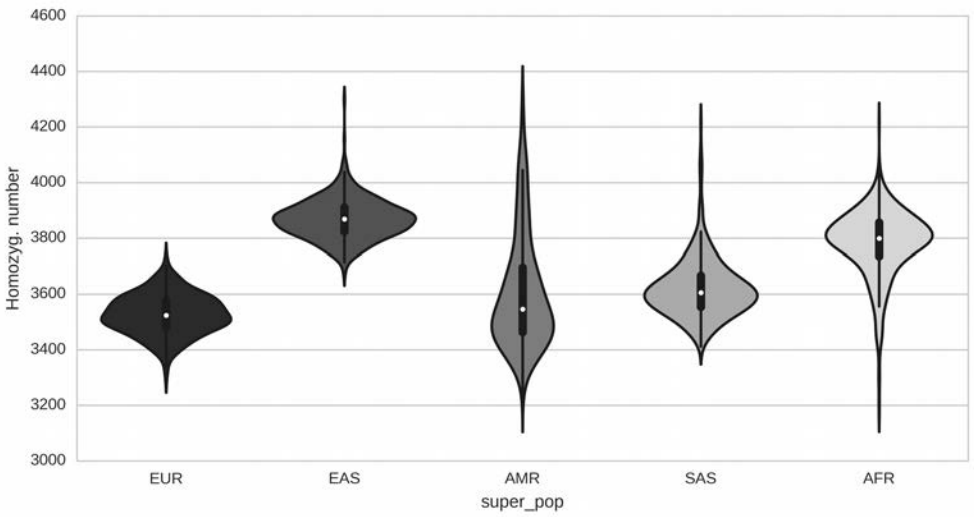

***Figura 4.6**: Violin-plots del nombre de variants homozigòtiques per individu en cadascuna de les superpoblacions. Correspondències amb superpoblació EUR: Europea; EAS: Àsia Oriental, AMR: Americana, SAS: Àsia Meridional, AFR: Africana.*

Abans d'aplicar els predictors obtinguts en la secció 4.3.1 a les dades del projecte 1000 genomes humans (*The 1000 Genomes Project Consortium* 2015), es fa una descripció de la natura d'aquestes dades, centrant l'estudi en la variabilitat genètica observada. Això és rellevant per construir la població sintètica descrita en la secció 4.2.3

El projecte internacional de seqüenciar 1000 genomes humans (*The 1000 Genomes Project Consortium* 2015) va concloure el 2015 amb les dades de 2504 individus del que es consideren 5 super-poblacions diferents –africana, europea, americana, asiàtica meridional i asiàtica oriental–, contenint cadascuna diverses divisions més segons la procedència, per un total de 26 poblacions –o

ètnies– diferents. El projecte de 1000 genomes obté, llavors, les freqüències poblacionals i totals de les variants trobades en l'estudi. Aquestes s'han usat posteriorment com a mètode de cribratge en estudis de NGS (Kim et al. 2013; Smedley et al. 2014; James et al. 2016), considerant com a polimorfismes, i per tant considerant no patogèniques, les variants que en alguna de les poblacions apareixen amb més de cert percentatge –sovint usant com a punt de tall l'1% de freqüència en la població (Kim et al. 2013; Smedley et al. 2014; James et al. 2016)–. El nombre de SNP amb canvi missense i que estan en homozigosi en els individus usats en aquest capítol està al voltant els 3000-4000 per cas, com s'aprecia en la figura 4.6.

Hi ha diferències en els nombres de variants per a cadascuna de les superpoblacions. En el cas de les superpoblacions americana i africana, la distribució és molt més àmplia: hi ha individus amb nombres més diferent que en altres. Això es deu a que algunes de les poblacions han tingut episodis de mescla ètnica recentment (*The 1000 Genomes Project Consortium 2015*). Els individus mostrejats de les poblacions europees, en canvi, mostren un nombre mitjà de variants més baix i més homogeni degut a que es tracta de poblacions amb menys mescla ètnica (*The 1000 Genomes Project Consortium 2015*) i que, a més, s'han fet servir més en els estudis genòmics i, per tant, hi ha cert biaix en la composició de les referències. Aquestes diferències en la composició són un factor que, llavors, s'haurà de tenir en compte en la creació de la població sintètica per reduir aquest possible biaix..

No totes les variants obtingudes, però, són potencialment patogèniques o poden tenir un efecte en la funcionalitat de la proteïna. En la figura 4.7A es pot apreciar la coincidència entre els diversos predictors de patogenicitat usats.

PON-P2 (Niroula et al. 2015) és el predictor més restrictiu dels quatre que s'han tingut en compte; i a més a més no dóna una resposta concreta en una fracció important de les variants, ja que només classifica com a patogènic o neutre en aquelles variants on resulta amb un valor predictiu extrem: dóna poques prediccions, però més fiables. Per altra banda, SIFT (Sim et al. 2012) és el predictor que més casos dóna com a patogènics. Aquesta tendència a predir la majoria com a possiblement danyí no és compartida ni amb PON-P2, ni amb Polyphen2 (Adzhubei et al. 2010), ni amb CADD (Kircher et al. 2014). Això du a pensar que SIFT tendeix a donar un excés de Falsos Positius, quelcom que es confirma quan s'observa que moltes d'aquestes variants es troben habitualment a la població –la seva freqüència és elevada– i per tant és possible que siguin polimorfismes –és a dir, variants funcionalment neutres–.

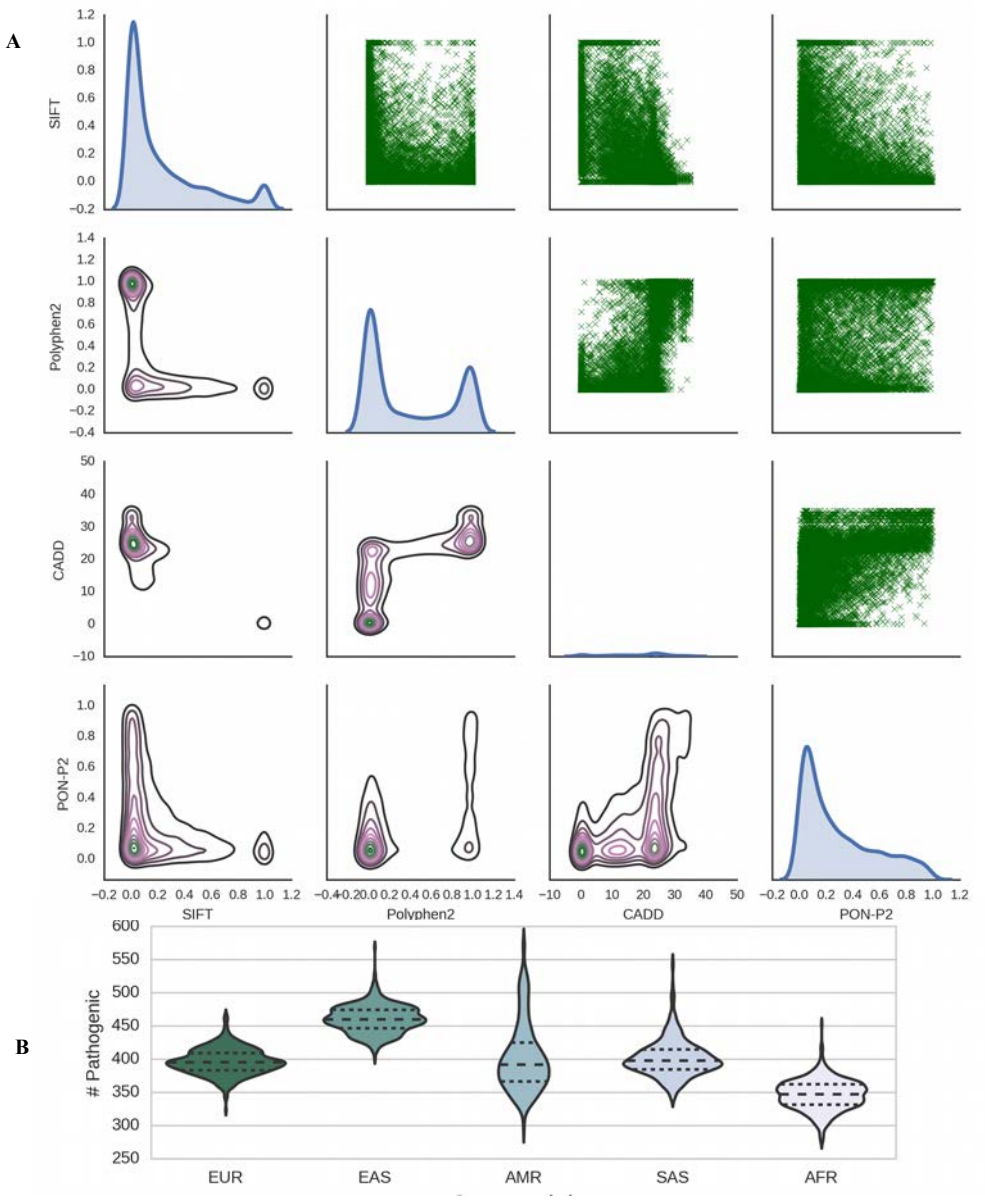

***Figura 4.7**: Patogenicitat en les variants de 1000G. **A:** Distribucions bivariants –part esquerra– i núvols de punts –part dreta– de les prediccions en un mostreig de 10000 casos de 1000G; i les seves distribucions de probabilitat –diagonal– **B:** Gràfics violí del nombre de variants patogèniques en les super-poblacions de 1000G.*

A partir de les dades dels predictors de patogenicitat es decideix quines són les variants patogèniques amb les que treballar –patogenicitat entesa, en aquest cas, com a obtenir valors de patogenicitat en al menys dos dels quatre predictors, secció 4.2.2–. El nombre varia segons la super-població, tal i com s'aprecia a la Figura 4.7B. El conjunt de la superpoblació africana conté menys possibles variants patogèniques que la resta. Aquesta observació és consistent amb les observacions sorgides de l'anàlisi exhaustiu de 1000G (*The 1000 Genomes Project Consortium* 2015) estenent el que altres estudis ja havien afirmat (Lohmueller et al. 2008): que les poblacions africanes tenen menys variants patogèniques malgrat el major nombre de variació total, ja que la resta de superpoblacions haurien patit un coll d'ampolla al voltant del moment que van migrar d'Àfrica –segons la hipòtesi de l'únic origen recent des d'Àfrica (Posth et al. 2016)–. Un coll d'ampolla és una reducció de població tan gran que limita molt la variació genètica, i com a conseqüència els individus posteriors presenten una variabilitat genètica molt reduïda, podent canviar significativament la proporció d'al·lels. Aquest coll d'ampolla observat produeix una major aparició de variants patogèniques, gràcies a una reducció en la pressió selectiva en la diversitat genètica de la població. És una situació anàloga al que s'ha observat més recentment –des d'un punt de vista històric– en les poblacions de jueus Asquenazita (Slatkin 2004) o la població Islandesa (Helgason 2000); que tenen un enriquiment en certes variants patogèniques degut al reduït nombre d'individus del que descendeixen. De nou, el conjunt americà és el que té un rang possible més elevat degut a la mescla ètnica. Finalment, les petites diferències en els nombres de variants potencialment patogèniques entre les poblacions europees i les poblacions asiàtiques podrien explicar-se per un biaix d'origen tècnic, intrínsec a l'ús de GRCh37 com a

genoma de referència, que està fet a partir del genoma de només 13 individus. Alguns dels problemes del conjunt de genoma GRCh37 han estat arreglats en actualitzacions, i s'han establert definitivament en la nova versió, GRCh38 (Schneider 2017), ja que s'inclou la possibilitat de diversos al·lels en algunes posicions.

4.3.3 Estimació de la presència de CPDs a 1000 genomes.

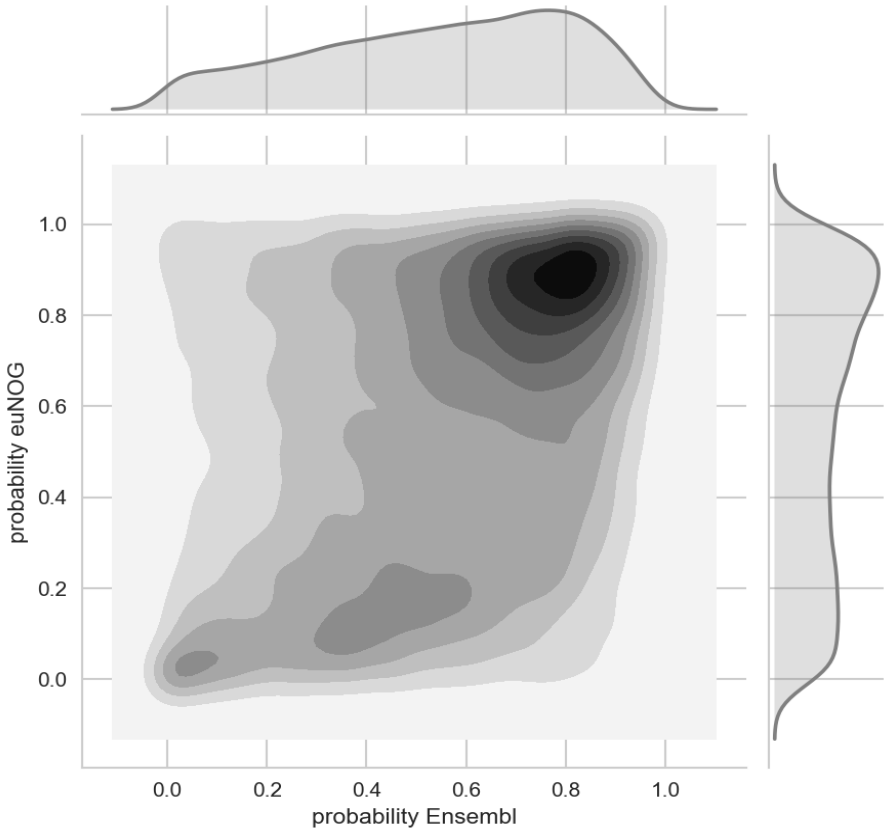

***Figura 4.8**: Distribució bivariant de les prediccions d'euNOG i Ensembl sobre un subset de 25000 casos de variants patogèniques.*

Aplicar directament la predicció binaria de CPD com a estimació dels hCPD

no seria correcte. Hi ha massa factors no controlats que intervenen en la possibilitat de que una variant patogènica en humà sigui un hCPD respecte les restriccions imposades en els predictors de CPD. Malgrat aquestes diferències, però, es poden usar els predictors per a fer una estimació estadística: inferir el potencial de cada variant per a tenir compensacions i, a partir d'aquí, amb els conjunts obtinguts, realitzar aproximacions de la potencialitat de compensació dels individus.

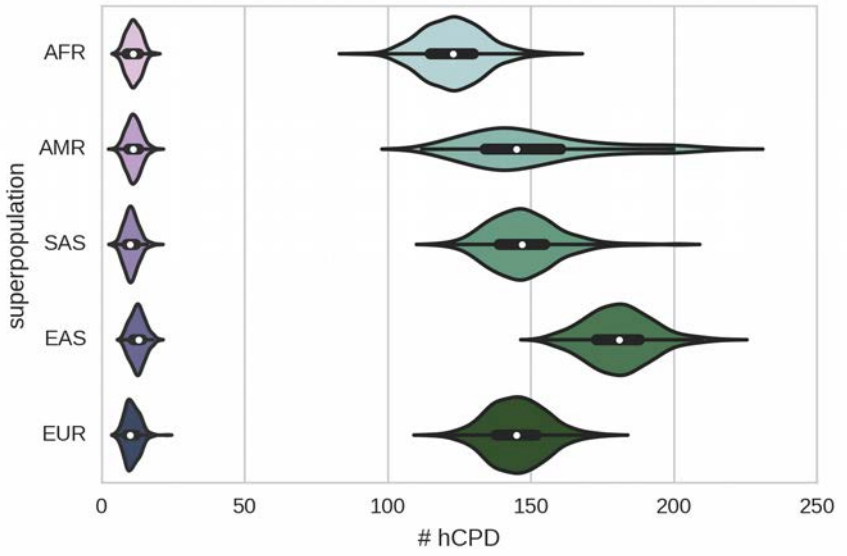

***Figura 4.9**: Prediccions hCPD. Nombre d'hCPD segons filtres: a l'esquerra més de 0.9 segons euNOG i Ensembl, a la dreta més de 0.5 en ambdós.*

Els predictors coincideixen bastant en els valors que assignen a les variants (coeficient de correlació de pearson=0.32, p-valor=0), tal i com es pot observar a la figura 4.8 en una mostra representativa del total de variants patogèniques predites en les dades de 1000G. La regió superior dreta de la figura, que és la part on coincideixen els dos en afirmar que són variants amb potencial d'ésser

compensades, **són els casos que es consideren com a possibles hCPD**.

Si s'apliquen els predictors a totes les variants obtingudes dels 1000G, el nombre de variants patogèniques considerades amb potencial hCPD resulta, però, més alt del que s'esperava –estudis anteriors xifraven en més o menys 10% el nombre de CPD en altres espècies (Kondrashov et al. 2002; Ferrer-Costa et al. 2007; Jordan et al. 2015)–. Efectivament, si es posa el punt de tall de hCPD a 0.5, els predictor euNOG dóna 135097 de 266256 variants amb potencial de hCPD, més d'un 50%; i el predictor Ensembl dóna 156318 de 235013; més d'un 60%. D'altra banda, posant el punt de tall a 0.9 s'obtenen 34486 variants com a possibles hCPD amb euNOG –aproximadament un 20%–, i 13795 amb Ensembl –aproximadament un 10%–; valors més propers a l'esperat. En aquest context, s'observa que hi ha, segons les prediccions, un cert nombre de hCPD en cadascun dels individus, que varia segons els punts de tall que es decideixi posar. Si el punt de tall és de 0.9 –el més restrictiu– tant en euNOG com en Ensembl, s'obtindrà un valor d'entre 5 i 10 hCPD per individu, sense gaire variació segons la població (Figura 4.9, gradient lila). Si s'agafa un valor de 0.5 com a punt de tall, el nombre de hCPD creix variant entre 100 i 200, amb fluctuacions segons la població (Figura 4.9, gradient verd), que són proporcionals al nombre de variants patogèniques per individu (Figura 4.7B).La distribució dels valors predictius obtinguts, diferenciant abans i després de posar el filtre de patogenicitat són diferents. A les representacions de la figura 4.10 es pot apreciar, en les regions superior dreta –corresponents a major potencial hCPD segons un predictor– és més densa abans dels filtres (Figura 4.10A i 4.10C) que després, amb només les predites com a patogèniques (Figura 4.10B i 4.10D). Això mostra quelcom interessant: que els polimorfismes, que són les variants que el filtre fa desaparèixer, són més fàcils

de compensar que les variants patogèniques. Els valors de predicció d'hCPD més alts dels polimorfismes son consistents tant amb el que diu la teoria neutra de l'evolució (Kimura 1968), que no assigna canvi de funció als polimorfismes; com amb la teoria quasi neutra de l'evolució (Ohta 1973), que assigna canvis de funció més moderats o molt poc deleteris als polimorfismes.

En la secció 4.3.1 s'ha comprovat que Ensembl és millor representació gràcies a que els MSA mantenen millor les senyals evolutives, que no queden diluïdes com en euNOG. En la figura 4.10A i 4.10B, que mostren el model euNOG, s'aprecia com predir amb els valor d'Ensembl afecta als descriptors evolutius. EuNOG, que està entrenat amb uns valors sorgits d'uns MSA molt més grans, no aporta el mateix tipus d'informació, i al predir amb Ensembl té més tendència a predir no hCPD que la resta. Aquest efecte no s'aprecia en l'altra predicció creuada (figura 4.10C i 4.10B), ja que els resultats del predictor Ensembl mostren uns valors més coincidents entre els conjunts.

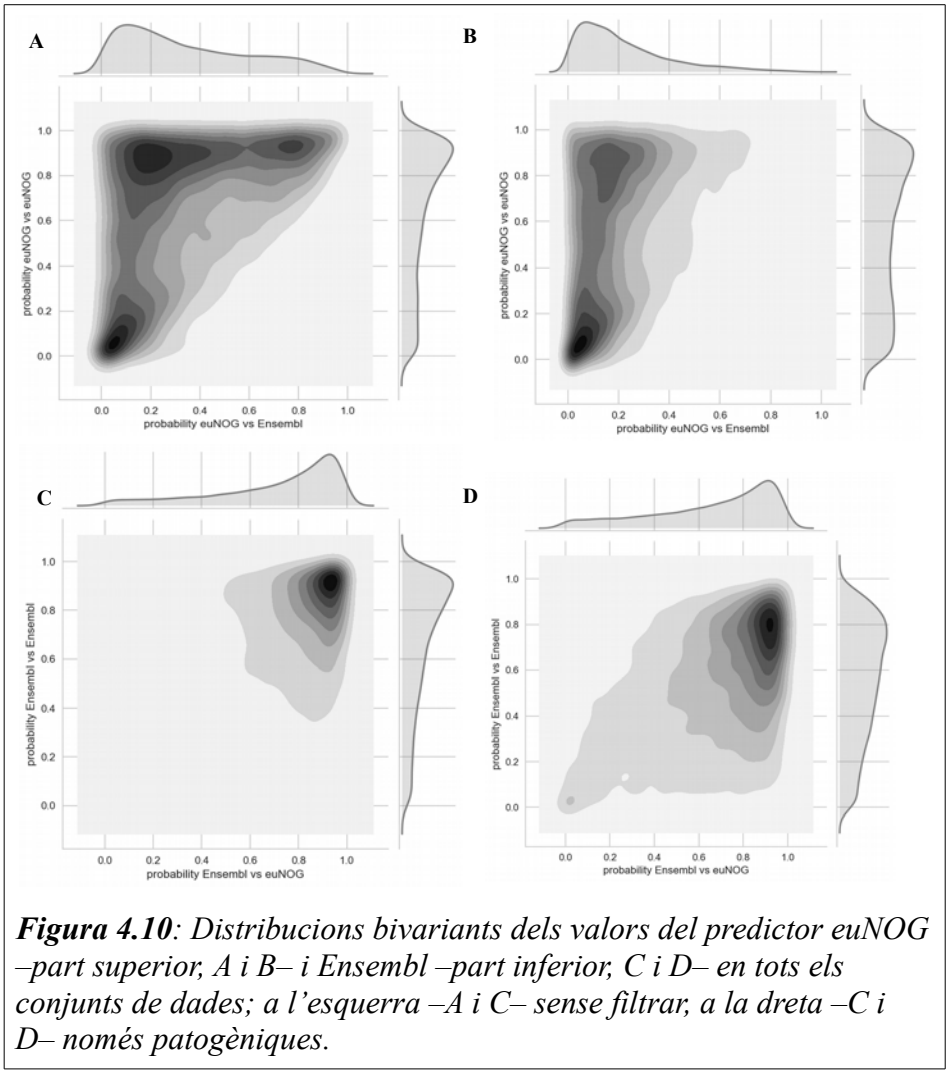

***Figura 4.10**: Distribucions bivariants dels valors del predictor euNOG –part superior, A i B– i Ensembl –part inferior, C i D– en tots els conjunts de dades; a l'esquerra –A i C– sense filtrar, a la dreta –C i D– només patogèniques.*

4.3.4 Establint la presència d' hCPD en l'incidentaloma

En la secció anterior s'han obtingut les prediccions de hCPD de totes les variants patogèniques dels individus de 1000G. A partir d'aquestes s'obté quina probabilitat té cada individu de **no tenir hCPD**. Aquestes probabilitats, tal i com es veu a la figura 4.11 –línia blava–, són molt properes a 0. És a dir, que la majoria d'individus tindrà hCPD.

Cal recordar, però, que els càlculs de probabilitat per a cada individu s'han fet sota les restriccions i els predictors utilitzats. Per si aquestes circumstàncies han introduït un artefacte tècnic respecte a un comportament a l'atzar, s'ha emprat com a referència la població sintètica. Tal i com s'ha descrit a l'apartat 4.2.3, es genera una població artificial de 1000 nous individus totalment aleatòria, amb la qual, després d'aplicar el mateix procediment que a 1000G i els mateixos predictors, se'n pot obtenir quina probabilitat de no tenir hCPD dels individus sintètics. Aquesta població artificial, al provindre d'un mostreig aleatori de les possibles variants patogèniques en el ser humà, no tindrà les restriccions que la presència de compensació en la mateixa espècie imposaria: un potencial enriquiment de variants que, d'altra forma, serien més deletèries.

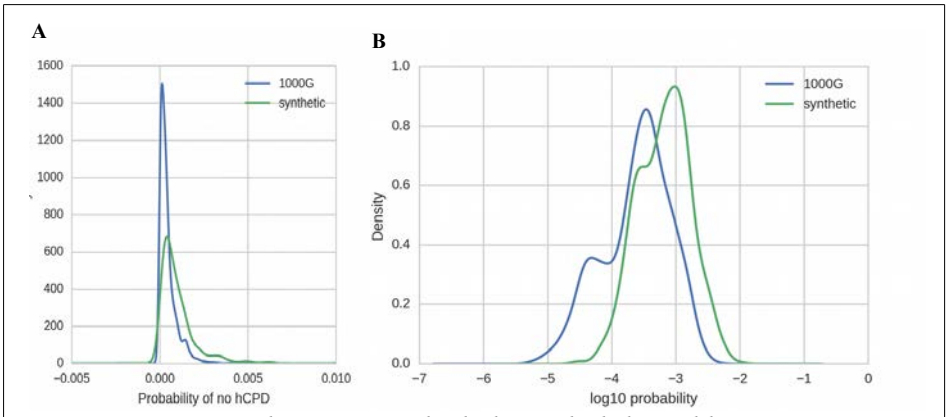

***Figura 4.11**: Distribucions totals de la probabilitat d'hCPD per individu a 1000G –blau– i a la població sintètica –verd– amb η=0.07. **A)** KDE de la probabilitat de tots els individus **B)** A; però amb l'eix d' abscisses log10.*

Una vegada creada la població sintètica, es calculen les seves probabilitats –Figura 4.11, línia verda–, i es comparen amb una població real, la de 1000G. La distribució de probabilitats de no tenir hCPD de 1000G és **més baixa** que la

sintètica, (Figura 4.11A i 4.11B, test Kolmogorov-Smirnoff per 2 mostres amb p-valor=$2.66 \cdot 10^{-62}$), tenint en compte una η=0.07. La diferència, a més a més, no varia canviant la η, doncs les probabilitats de les dos poblacions canviarien proporcionalment. D'aquí es pot concloure, doncs, que **existeix un enriquiment** en la població humana de variants patogèniques amb potencial compensable, o **hCPD**. Malgrat això, en aquest punt encara no es pot afirmar amb total seguretat que aquesta major probabilitat de hCPD –o menor probabilitat de no tenir-ne– sigui la causant de la diferència amb l'atzar o si és merament una conseqüència d'un altre efecte intrínsec al mostreig.

L'enriquiment de les variants patogèniques amb un potencial a ser compensades en el patrimoni genètic humà, com es veu a la Figura 4.11, si s'ajunta amb el fet que realment poden existir compensacions a la variant patogènica humana dins la mateixa espècie (Brandis i Hughes 2013; Xu i Zhang 2014, Jordan et al. 2015; Moura de Sousa 2017), com en l'humà (Mankad et al. 2006), és una explicació a una fracció de l'incidentaloma.

Les troballes d'aquest capítol no són aïllades, sinó que concorden amb altres treballs recents. Un estudi de quasi 300 nous exomes de la població espanyola (Dopazo et al. 2016) troba que, en comparació amb les altres poblacions europees de 1000G, les variants relacionades amb la susceptibilitat a malalties complexes apareixen en la mateixa freqüència. Es veu, en canvi, que hi ha un enriquiment en la freqüència de determinades variants associades a malalties rares o mendelianes, algunes dels quals només s'han trobat en la població espanyola. També descriu l'aparició de nombrosos polimorfismes privats –exclusius– de la població. La co-ocurrència de variants rares i polimorfismes en la població indica potencials compensacions.

Però, encara que part de l'incidentaloma pugui ésser explicat per les hCPD, resten encara més variants amb potencial patogènic sense aclarir. Els errors en l'anotació i/o la predicció de la patogenicitat (Tang i Thomas 2016), els mateixos errors de seqüenciació (MacArthur i Tyler-Smith 2010) o una penetrança incompleta d'una variant causal (Giudicessi i Ackerman 2013) són altres possibles causes de les variants incidentals (Jamuar et al. 2016). De fet, els predictors de patogenicitat prediuen el canvi de funció com a causant de malaltia, però la modulació de la funció no sempre està associada a la patogenicitat, sinó que el canvi pot ser beneficiós (MacArthur i Tyler-Smith 2010); o aquesta patogenicitat pot ésser depenent de restriccions evolutives d'origen ambiental que ha sofert la població (Oh et al. 2015). Exemples d'aquests casos es troben en canvis en el metabolisme de toxines o nutrients, on l'efecte del canvi depèn de la disponibilitat ambiental del substrat, i on el mateix canvi pot arribar, fins i tot, a conferir resistència contra determinats patògens. Tampoc tota la variació deletèria condueix a malaltia: És el cas de CASP12 (Xue et al. 2006), que modula la resposta immune davant algunes endotoxines, i que té una variant sense sentit molt freqüent en poblacions no africanes, on l'exposició a les endotoxines pot ser menor. La manca de CASP12, llavors, produeix una resposta immune més lleu en cas de sèpsia, i per tant facilita la supervivència en infeccions sistèmiques, tot i la pitjor resposta a endotoxines, que en aquestes poblacions no és tan necessària. Un altre cas d'exemple es dóna en el gen ACTN3, que és el productor d'un component estructural d'alguns músculs, i en el que alguns variants sense sentit s'han relacionat amb alta resistència física en esportistes d'elit (Yang et al. 2003). L'absència d'aquestes variants, però, s'ha relacionat amb esportistes d'elit en disciplines que requereixen força i explosivitat (Roth et al. 2008).

Observar des d'un punt de vista poblacional l'incidentaloma, tenint en compte els factors anteriors, explica evolutivament la presència d'algunes d'aquestes variants que puntualment són patogèniques, però que no sempre redueixen la fitness de l'individu: són variants que poden formar un reservori de variació capaç de donar respostes poblacionals a canvis en l'ambient –com en l'exemple de CASP12–, o poden presentar un enriquiment actual degut a necessitats adaptatives anteriors –com el cas de les variants que trunquen ACTN3, que confereix més resistència física, que s'ha teoritzat que haurien estat favorables per algunes poblacions antigues d'humans (Ruxton i Wilkinson 2013)–. El que puntualment és desfavorable no sempre redueix la fitness, i viceversa, el que puntualment és favorable no sempre l'augmenta.

En aquest capítol s'ha estudiat una categoria de variants patogèniques nova, les hCPD, que poden tenir importància en el context dels estudis clínics mitjançant seqüenciació. En un anàlisis de NGS poden trobar-se diverses variants amb possible efecte funcional no associades a cap malaltia que el pacient expressi. Els futurs mètodes de priorització de variants hauran de tenir en compte, a més de la penetrància incompleta, la variació no deletèria, i la variació en gens sense interès clínic; que alguns dels candidats poden tenir compensacions que en redueixin o n'atenuïn la patogenicitat. Això, llavors, en reduiria la seva importància en aquell cas. Les hCPD són, doncs, un cas d'especial importància en un marc d'accionabilitat davant de troballes incidentals amb interès clínic. Tenir noves pistes sobre l'efecte fenotípic dels canvis genètics, com els hCPD o la severitat comentada en altres capítols, són factors molt important de cara a la medicina de precisió.

5. Conclusions

- En el cas de les hemofílies A i B, hi ha una relació probabilística entre les variants CPD i el seu impacte molecular: les CPD tendeixen a generar un impacte més lleu en la proteïna, malgrat que la relació no és determinista.

- Per els factors de coagulació F8 i F9, el valor de $\Delta\Delta G$ per si mateix no és una representació prou bona de la fitness d'una proteïna, sinó que s'hauria de combinar amb mesures que considerin les interaccions epistàtiques i la coevolució.

- Hi ha un component molecular intrínsec a la severitat d'una malaltia: La gravetat de l'impacte molecular, descrita amb atributs fisicoquímics i evolutius, mostra una relació amb la gravetat de la malaltia

- Es pot predir amb un encert moderat la gravetat d'una malaltia causada per un canvi d'aminoàcid usant els mètodes emprats en la predicció de la patogenicitat. Tot i això, l'encert és significativament menor que en el problema de la patogenicitat, i actualment no té qualitat suficient per ser útil en clínica.

- Es pot predir si una variant té potencial de ser CPD amb un encert moderat. Hi ha atributs fisicoquímics i evolutius característics de les CPD compartits independents de les dades d'origen.

- Hi ha variants amb potencial de CPD en el ser humà. Hi ha una fracció de l'incidentaloma que pot ser explicada amb aquestes variants, que té importància en els estudis clínics de NGS.

- Els mètodes d'aprenentatge automàtic són una eina que pot donar resposta a dificultats tècniques i conceptuals dels estudis òmics i que poden accelerar-ne el trasllat a l'ambient clínic.

6. Apèndix

6.1 Apèndix: Taules

Descriptor	Proteïna	p-valor
Entropia de Shannon	F8	7.109e-23*
	F9	1.714e-15*
Accessibilitat Relativa	F8	0.00233
	F9	0.00124
Delta-Delta G	F8	0.00379
	F9	0.0003
$Pssm_{nat}$	F8	0.0030
	F9	1.488e-07*
Delta hidrofobicitat	F8	2.603e-06*
	F9	0.167
Delta Van der Waals	F8	0.0002*
	F9	0.0345
Blosum62	F8	9.342e-12 *
	F9	9.995e-06*

Taula A1.1: Correlacions Mann-Whitney Wilcoxon dels descriptors per a CPD-noCPD (Capítol 1, Apartat 2.3.2). Comparació amb el valor 0.0003 per a una significació de 99% de probabilitat.

Descriptor	Proteïna	p-valor
Entropia de Shannon	F8	1.882e-08*
	F9	1.095e-08*
Accessibilitat Relativa	F8	7.053e-10*
	F9	1.309e-05 *
Delta-Delta G	F8	3.822e-18*
	F9	6.572e-10*
$Pssm_{nat}$	F8	0.912
	F9	0.0067
Delta hidrofobicitat	F8	0.0006
	F9	0.0029
Delta Van der Waals	F8	8.442e-05*
	F9	0.0418
Blosum62	F8	2.821e-09*
	F9	5.751e-06*

Taula A1.2: Correlacions Mann-Whitney Wilcoxon dels descriptors per a CPD lleu vs CPD sever. (Capítol 1, Apartat 2.3.3) Comparació amb el valor 0.0003 per a una significació de 99% de probabilitat.

7. Abreviacions

ACC: Exactitud.

ADN: Àcid desoxiribonucleic.

ARN: Àcid ribonucleic.

API: Interfície de Programació d'Aplicacions.

AUC: Àrea sota la corba ROC.

CDC: Center for Disease Control and prevention.

CHAMP: CDC Hemophilia A Mutation Project.

CHBMP: CDC Hemophilia B Mutation Project.

CPD: Desviació patogènica compensada.

F1s: Mètrica del Valor F1.

F8: Factor de Coagulació 8.

F9: Factor de Coagulació 9.

FN: Falsos Negatius.

FP: Falsos Positius.

hCPD: Desviació patogènica amb potencial de compensació en humà.

HPO: Human Phenotype Ontology.

KDE: Kernel Density Estimation.

MCC: Coeficient de Correlació de Matthews.

MSA: Alineament Múltiple de Seqüència

MWW: Test estadístic no paramètric Mann-Whitney-Wilcoxon

NGS: Seqüenciació de nova generació.

NoCPD: variant no CPD

PDB: Protein Data Bank.

SNP: Single Nucleotide Polymorphism

SNS: Sensibilitat.

SPC: Especificitat.

SVM: Support Vector Machines

TN: Vertaders Negatius.

TP: Vertaders Positius.

8. Bibliografia

Abizar. Figure. CC BY-SA 3.0, https://commons.wikimedia.org/w/index.php?curid=3800855 .

Achawanantakun R, Sun Y. Shape and secondary structure prediction for ncRNAs including pseudoknots based on linear SVM. *BMC Bioinformatics*. 2013;14(Suppl 2):S1. doi:10.1186/1471-2105-14-S2-S1.

Adzhubei IA, Schmidt S, Peshkin L, et al. A method and server for predicting damaging missense mutations. *Nature methods*. 2010;7(4):248-249. doi:10.1038/nmeth0410-248.

Aken BL, Ayling S, Barrell D et al. The Ensembl gene annotation system. *Database (Oxford)* 2016; 2016 baw093. doi:10.1093/database/baw093.

Akle S, Chun S, Jordan DM, Cassa CA. Mitigating false-positive associations in rare disease gene discovery. *Human mutation*. 2015;36(10):998-1003. doi:10.1002/humu.22847.

Altman DG, Bland JM. Diagnostic tests. 1: Sensitivity and specificity. *BMJ : British Medical Journal*. 1994;308(6943):1552.

Altschul SG, Gish W, Miller W, et al. Basic local alignment search tool. *Journal in Molecular Biology*. 1990; 215(3): 403-10. doi: 10.1016/S0022-2836(05)80360-2.

Altschul SF, Madden TL, Schäffer AA, et al. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Research.* 1997; 25(17):3389-3402. doi: 10.1093/nar/25.17.3389.

Alvarez-Ponce D. Recording negative results of protein-protein interaction assays: an easy way to deal with the biases and errors of

interactomic data sets. *Briefings in Bioinformatics*. 2016. [Epub]. doi:10.1093/bib/bbw075.

Anoosha P, Sakthivel R, Gromiha MM. Prediction of protein disorder on amino acid substitutions. *Anals in Biochemistry.* 2015; 491:18-22. doi:10.1016/j.ab.2015.08.028.

Apweiler R, Bairoch A, Wu CH, et al. UniProt: the Universal Protein knowledgebase. *Nucleic Acids Research.* 2004;32(Database issue):D115-D119. doi:10.1093/nar/gkh131.

Baldi P, Brunak S, Chauvin Y et al. Assessing the accuracy of prediction algorithms for classification: an overview. *Bioinformatics*. 2000; 16(5): 412-24. doi: 10.1093/bioinformatics/16.5.412.

Baltimore D. RNA-dependent DNA polymerase in virions of RNA tumour viruses. *Nature.* 1970; 226(5252): 1209-1211. doi:10.1038/2261209a0.

Bamshad MJ, Ng SB, Bigham AW, et al. Exome sequencing as a tool for Mendelian disease gene discovery. *Nature Reviews in Genetics.* 2011; 12(11):745-755. doi:10.1038/nrg3031.

Barešić A, Hopcroft LEM, Rogers HH et al. Compensated Pathogenic Deviations: analysis of structural effects, *Journal of Molecular Biology.* 2010;396(1):19-30. doi:10.1016/j.jmb.2009.11.002.

Bengio Y. Learning deep architectures for AI. *Foundations and Trends in Machine Learning.* 2009; 2(1):1-127. doi:10.1561/2200000006.

Berg JS, Adams M, Nassar N, et al. An informatics approach to analyzing the incidentalome. *Genetics in medicine : official journal of the American College of Medical Genetics*. 2013;15(1):36-44. doi:10.1038/gim.2012.112.

Berman HM, Westbrook J, Feng Z, et al. The Protein Data Bank. *Nucleic Acids Research*. 2000;28(1):235-242.

Bershtein S, Segal M, Bekermar R et al. Robustness-epistasis link shapes the fitness landscape of a randomly drifting protein. *Nature*. 2006;444:929-932. doi: 10.1038/nature05385.

Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics*. 2014;30(15):2114-2120. doi:10.1093/bioinformatics/btu170.

Bondi A. van der Waals volumes and radii. *Journal of Physical Chemistry*. 1964; 68(3): 441-451. doi:10.1021/j100785a001.

Boser BE, Guyon IM, Vapnik VN. A training algorithm for optimal margin classifiers. *Proceedings of the fifth annual workshop on Computational learning theory.* 1992; COLT '92:144. doi:10.1145/130385.130401.

Boutet E, Lieberherr D, Tognolli M et al. UniprotKB/Swiss-Prot. *Methods in Molecular Biology.* 2007; 406:89-112.

Brandis G, Hughes D. Genetic characterization of compensatory evolution in strains carrying rpoB Ser531Leu, the rifampicin resistance mutation most frequently found in clinical isolates. *The Journal of Antimicrobial Chemotherapy.*2013; 68(11):2493-2497. doi:10.1093/jac/dkt224.

Breen MS, Kemena C, Vlasov PK et al. Epistasis as the primary factor in molecular evolution. *Nature*. 2012; 490(7421):535-8. doi: 10.1038/nature11510.

Breiman L. Arcing classifier (with discussion and a rejoinder by the author). *Annals in Statistics*. 1998;26(3):801-849. doi:10.1214/aos/1024691079.

Burga A, Lehner B. Beyond genotype to phenotype: why the phenotype of an individual cannot always be predicted from their

genome sequence and the environment that they experience. *FEBS J.* 2012; 279(20):3765-75. doi: 10.1111/j.1742-4658.2012.08810.x.

Burn J, Watson M. The Human Variome Project. *Human Mutation.* 2016. 37(6):505-507. doi:10.1002/humu.22986.

Capriotti E, Calabrese R, Casadio R. Predicting the insurgence of human genetic diseases associated to single point protein mutations with support vector machines and evolutionaty information. *Bioinformatics*. 2006; 22(22):2729-34. doi: 10.1093/bioinformatics/btl423.

Capriotti E, Fariselli P. PhD-SNPg: a webserver and lightweight tool for scoring single nucleotide variants. *Nucleic Acids Research.* 2017. [Epub]. doi:10.1093/nar/gkx369.

Castellano S, Mazza T. Congruency in the prediction of pathogenic missense mutations: state-of-the-art web-based tools. *Briefings in bioinformatics*. 2013; 14(4): 448-459. doi: 10.1093/bib/bbt013.

Castillo-Fernandez JE, Spector TD, Bell JT. Epigenetics of discordant monozygotic twins: implications for disease. *Genome Medicine*. 2014;6(7):60. doi:10.1186/s13073-014-0060-z.

Centers for Disease Control and Prevention. CDC Hemophilia A Mutation Project (CHAMP). https://www.cdc.gov/ncbddd/hemophilia/champs.html. 2015. Consultat el 20-4-2017.

Centers for Disease Control and Prevention. CDC Hemophilia B Mutation Project (CHBMP). https://www.cdc.gov/ncbddd/hemophilia/champs.html. 2015. Consultat el 20-4-2017.

Chakravorty S, Hegde M. Gene and variant annotation for mendelian disorders in the era of advanced sequencing technologies. *Annual Review of Genomics and Human Genetics.* 2017. [epub]. doi:10.1146/annurev-genom-083115-022545.

Chang J. Core services: Reward bioinformaticians [Internet]. *Nature.* Disponible a http://www.nature.com/news/core-services-reward-bioinformaticians-1.17251 Consultat el 05-05-2017.

Chawla NV, Bowyer KW, Hall LO et al. SMOTE: Synthetic Minority Over-sampling Technique. *Journal of Artificial Intelligence Research*. 2002; 16:321-357. doi: 10.1613/jair.953.

Chen R, Snyder M. Promise of personalized omics to precision medicine. *Wiley interdisciplinary reviews Systems biology and medicine*. 2013;5(1):73-82. doi:10.1002/wsbm.1198.

Chen T, Guestrin C. XGBoost: A Scalable Tree Boosting System. *In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '16)*. 2016 785-794. doi:10.1145/2939672.2939785.

Cheng F, Zhao Z. Machine learning-based prediction of drug–drug interactions by integrating drug phenotypic, therapeutic, chemical, and genomic properties. *Journal of the American Medical Informatics Association : JAMIA*. 2014;21(e2):e278-e286. doi:10.1136/amiajnl-2013-002512.

Cheng Z, Zhou S, Guan J. Computationally predicting protein-RNA interactions using only positive and unlabeled examples. *Journal of Bioinformatics and Computational Biology.* 2015. 13(3):1541005. doi:10.1142/S021972001541005X.

Choi Y, Sims GE, Murphy S et al. predicting the functional effect of amino acid substitutions and indels. de Brevern AG, ed. *PLoS ONE*. 2012; 7(10): e46688. doi:10.1371/journal.pone.0046688.

Cibulskis K, Lawrence MS, Carter SL, et al. Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. *Nature biotechnology*. 2013; 31(3):213-219. doi:10.1038/nbt.2514.

Cingolani P, Platts A, Wang LL, et al. A program for annotating and

predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3. *Fly*. 2012;6(2):80-92. doi:10.4161/fly.19695.

Colijn C, Jones N, Johnston IG, Yaliraki S, Barahona M. Toward Precision Healthcare: context and mathematical challenges. *Frontiers in Physiology*. 2017;8:136. doi:10.3389/fphys.2017.00136.

Colobran R, Álvarez de la Campa E, Soler-Palacín P et al. Clinical and structural impact of mutations affecting the residue Phe367 of FOXP3 in patients with IPEX syndrome. *Clinical Immunology.* 2016;163:60-5. doi:10.1016/j.clim.2015.12.014.

Commins J, Toft C, Fares MA. Computational biology methods and their application to the comparative genomics of endocellular symbiotic bacteria of instects. *Biology Procedures Online.* 2009. Accessed via SpringerImages. Figure. CC BY-SA 2.5, https://commons.wikimedia.org/w/index.php?curid=17509619

Crick FH. On protein synthesis. *Symposia of the Society of Experimental Biology.* 1958; 12:138-63.

Dammann M, Weber F. Personalized medicine: caught between hope, hype and the real world. *Clinics*. 2012;67(Suppl 1):91-97. doi:10.6061/clinics/2012(Sup01)16.

Dand N, Schulz R, Weale ME, et al. Network−informed gene ranking tackles genetic heterogeneity in exome−sequencing studies of monogenic disease. *Human Mutation*. 2015;36(12):1135-1144. doi:10.1002/humu.22906.

Dayhoff MO, Schwartz RM, Orcutt BC. A model of evolutionary change in proteins. *Atlas of Protein Sequence and Structure.* 1978; 5 suppl 3:353-352.

De Mattos-Arruda L, Mayor R, Ng CK et al. Cerebrospinal fluid-derived circulating tumour DNA better represents the genomics

alterations of brain tumours than plasma. *Nature Communications*. 2015;6:8839. doi:10.1038/ncomms9839.

DePristo MA, Weinreich DM, Hartl DL. Missense meanderings in sequence space: a biophysical view of protein evolution. *Nature Reviews in genetics*. 2005; 6(9):678-87. doi: 10.1038/nrg1672.

DePristo MA, Banks E, Poplin RE, et al. A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nature genetics*. 2011;43(5):491-498. doi:10.1038/ng.806.

Delaney SK, Hultner ML, Jacob HJ, et al. Toward clinical genomics in everyday medicine: perspectives and recommendations. *Expert Review of Molecular Diagnostics*. 2016;16(5):521-532. doi:10.1586/14737159.2016.1146593.

DMLapato. Figure. CC BY-SA 4.0, https://commons.wikimedia.org/w/index.php?curid=43777596.

Dopazo J, Amadoz A, Bleda M, et al. 267 Spanish exomes reveal population-specific differences in disease-related genetic variation. *Molecular Biology and Evolution*. 2016;33(5):1205-1218. doi:10.1093/molbev/msw005.

Drozdetskiy A, Cole C, Procter J, Barton GJ. JPred4: a protein secondary structure prediction server. *Nucleic Acids Research*. 2015;43(Web Server issue):W389-W394. doi:10.1093/nar/gkv332.

Edgar, R.C. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Research.* 2004;32(5):1792-179. doi:10.1093/nar/gkh340.

Ester M, Kriegel HP, Sander J, Xu X. A density-based algorithm for discovering clusters in large spatial databases with noise. *Proceedings of the 2nd International conference on Knowledge Discovery and Data Mining.* 1996; KDD-96:226-231.

Estevezj. Figure. CC BY-SA 3.0,

https://commons.wikimedia.org/w/index.php?curid=23264166

Everett H. Theory of the universal wavefunction. *Thesis, Princeton University*. 1956.

Fallaize R, Macready AL, Butler LT, et al. An insight into the public acceptance of nutrigenomic-based personalized nutrition. *Nutrition Research Reviews.*2013; 26(1):39-48. doi:10.1017/S0954422413000024.

Famiglietti ML, Estreicher A, Gos A, et al. Genetic variations and diseases in UniProtKB/Swiss-Prot: the ins and outs of expert manual curation. *Human Mutation*. 2014;35(8):927-935. doi:10.1002/humu.22594.

Fang LT, Afshar PT, Chhibber A, et al. An ensemble approach to accurately detect somatic mutations using SomaticSeq. *Genome Biology*. 2015; 16(1):197. doi:10.1186/s13059-015-0758-2.

Fauchere JL, Pliska V. Hydrophobic parameters pi of amino-acid side chains from the partitioning of N-acetyl-amino-acid amides. *European Journal of Medical Chemistry.* 1983; 18(3):369-375.

Fawcett T. An introduction to ROC analysis. *Pattern Recognition Letters.* 2006; 27(8):861-874. doi:10.1016/j.patrec.2005.10.010.

Feinberg H, Rowntree TJW, Tan SLW et al. Common polymorphisms in human langerin change specificity for glycan ligands. *The Journal of Biological Chemistry*. 2013;288(52):36762-36771. doi:10.1074/jbc.M113.528000.

Feng Y, Lin H, Luo L. Prediction of protein secondary structure using feature selection and analysis approach. *Acta biotheoretica.* 2014;62(1):1-14. doi:10.1007/s10441-013-9203-7.

Fernández-Recio J. Prediction of protein binding sites and hot spots. *WIREs Computational Molecular Science.* 2011; 1:680-698. doi:10.1002/wcms.45.

Ferrer-Costa C, Orozco M, de la Cruz X. Characterization of disease-associated single amino acid polymorphisms in terms of sequence and structure properties. *Journal of Molecular Biology.* 2002; 315(4):771-86. doi: 10.1006/jmbi.2001.5255.

Ferrer-Costa C, Orozco M, de la Cruz X. Sequence-based prediction of pathological mutations. *Proteins*. 2004; 57(4):811-9. doi: 10.1002/prot.20252.

Ferrer-Costa C, Orozco M, de la Cruz X. Characterization of compensated mutations in terms of structural and physico-chemical properties, *Journal of Molecular Biology.* 2007; 365(1):249-256. doi:10.1016/j.jmb.2006.09.053.

Flores O, Orozco M. nucleR: a package for non-parametric nucleosome positioning. *Bioinformatics*. 2011; 27(15):2149-2150. doi:10.1092/bioinformatics/btr345.

Free Software Foundation. Bash [Unix Shell Program]. v3.2. www.gnu.org/software/bash/

Freund Y, Schapire R. A desicion-theoretic generalization of on-line learning and an application to boosting. *Computationa Learning Theory. EuroCOLT.* 1995; 23-37. doi:10.1007/3-540-59119-2_166.

Fuchs SA, Harakalova M, van Haaften G et al. Application of exome sequencing in the search for genetic causes of rare disorders of copper metabolism. *Metallomics.* 2012; 4(7):606-613. doi:10.1039/c2mt20034a.

Gao W, Emaminejad S, Nyein HYY, et al. Fully integrated wearable sensor arrays for multiplexed *in situ* perspiration analysis. *Nature*. 2016;529(7587):509-514. doi:10.1038/nature16521.

Garofalo A, Sholl L, Reardon B, et al. The impact of tumor profiling approaches and genomic data strategies for cancer precision medicine. *Genome Medicine*. 2016; 8:79. doi:10.1186/s13073-016-0333-9.

Genome [Internet]. Bethesda (MD): National Library of Medicine (US), National Center for Biotechnology Informaiton; 2004 – [cita 03/05/2017]. Disponible: https://www.ncbi.nlm.nih.gov/genome/

Ghosh A, Manwani N, Sastry PN. On the robustness of decision tree learning under label noise. *PAKDD 2017: Advances in Knowledge Discovery and Data Mining.* 2017; 685-697. doi:10.1007/978-3-319-57454-7_53.

Girirajan S, Elsea SH. Distorted Mendelian transmission as a function of genetic background in Rai1-haploinsufficient mice. *European Journal of Medical Genetics.* 2009; 52(4):224-8. doi:10.1016/j.ejmg.2008.12.002.

Giudicessi JR, Ackerman MJ. Determinants of incomplete penetrance and variable expressivity in heritable cardiac arrhythmia syndromes. *Translational researc: the journal of laboratory and clinical medicine*. 2013;161(1):1-14. doi: 10.1016/j.trsl.2012.08.005.

Gjuvsland AB, Vik JO, Beard DA et al. Bridging the genotype–phenotype gap: what does it take? *The Journal of Physiology*. 2013;591(8):2055-2066. doi:10.1113/jphysiol.2012.248864.

Gong LI, Suchard MA, Bloom JD. Stability-mediated epistasis constrains the evolution of an influenza protein. Pascual M, ed. *eLife*. 2013;2:e00631. doi:10.7554/eLife.00631.

González-Pérez A, López-Bigas N. Improving the Assessment of the Outcome of Nonsynonymous SNVs with a Consensus Deleteriousness Score, Condel. *American Journal of Human Genetics*. 2011; 88(4): 440-449. doi:10.1016/j.ajhg.2011.03.004.

Green RC, Berg JS, Grody WW, et al. ACMG Recommendations for reporting of incidental findings in clinical exome and genome sequencing. *Genetics in medicine : official journal of the American College of Medical Genetics.* 2013;15(7):565-574. doi:10.1038/gim.2013.73.

Grimm DG, Azencott C, Aicheler F, et al. The evaluation of tools used to predict the impact of missense variants is hindered by two types of circularity. *Human Mutation*. 2015;36(5):513-523. doi:10.1002/humu.22768.

Grossberg S. Contour enhancement, short-term memory and constancies in reverbrating neural networks. *Studies in Applied Mathematics.* 1973. 52: 213-257.

Hall M, Frank E, Holmes G, et al. The WEKA data mining software: An update. *SIGKDD Explorations*. 2009;11(1):10-18. doi:10.1145/1656274.1656278.

Hall MA, Moore JH, Ritchie MD. Embracing complex associations in common traits: critical considerations for precision medicine. *Trends in Genetics*. 2016; 32(8):470-484. doi:10.1016/j.tig.2016.06.001.

Hall MJ, Forman AD, Pilarski R, et al. Gene panel testing for inherited cancer risk. *Journal of the National Comprehensive Cancer Network.* 2014. 12(9):1339-13346.

Hamada M 2015. RNA secondary structure prediction from multi-aligned sequences. *Methods in Molecular Biology.* 2015;1269:17-38. doi:10.1007/978-1-4939-2291-8_2.

Hamasaki-Katagiri N, Salari R, Simhadri VL et al. Analysis of F9 point mutations and their correlation to severity of haemophilia B disease. *Haemophilia*. 2012; 18(6):933-40. doi:10.1111/j.1365-2516.2012.02848.x.

Hamasaki-Katagiri N, Salari R, Wu A, et al. A Gene-Specific Method for Predicting Hemophilia-Causing Point Mutations. *Journal of molecular biology*. 2013; 425(21): 4023-4033. doi:10.1016/j.jmb.2013.07.037.

Han Y, He X. Integrating Epigenomics into the Understanding of Biomedical Insight. *Bioinformatics and Biology Insights*.

2016;10:267-289. doi:10.4137/BBI.S38427.

Hawkins DM. The problem of overfitting. *Journal of chemical information and computer sciences.* 2004; 44:1-12. doi:10.1021/ci0342472.

He KY, Ge D, He MM. Big Data Analytics for Genomic Medicine. Cho WC, ed. *International Journal of Molecular Sciences*. 2017;18(2):412. doi:10.3390/ijms18020412.

Helgason A, Sigurðardóttir S, Gulcher JR, Ward R, Stefánsson K. mtDNA and the origin of the icelanders: deciphering signals of recent population history. *American Journal of Human Genetics*. 2000;66(3):999-1016.

Hendler J. Avoiding another AI winter. *IEEE Intelligent Systems* 23.2 2008; 2-4.

Henikoff S, Henikoff JG. Amino acid substitution matrices from protein blocks. *Proceedings of the National Academy of Sciences of the United States of America*. 1992;89(22):10915-10919.

Henry CS, Overbeek R, Xia F et al. Connecting genotype to phenotype in the era of high-throughput sequencing. *Biochimica et Biophysica Acta*. 2011; 1810(10):967-77. doi:10.1016/j.bbagen.2011.03.010.

Hicks S, Wheeler DA, Plon SE, Kimmel M. Prediction of Missense Mutation Functionality Depends on both the Algorithm and Sequence Alignment Employed. *Human mutation*. 2011; 32(6):661-668. doi:10.1002/humu.21490.

Ho TK. Random Decision Forests. *Proceedings of the 3rd International conference on Document Analysis and Recognition.* 1995: 278-282.

Hofmann AL, Behr J, Singer J, et al. Detailed simulation of cancer exome sequencing data reveals differences and common limitations

of variant callers. *BMC Bioinformatics*. 2017; 18:8. doi:10.1186/s12859-016-1417-7.

Hopf TA, Ingraham JB, Poelwijk FJ, et al. Mutation effects predicted from sequence co-variation. *Nature Biotechnology*. 2017; 35(2):128-135. doi:10.1038/nbt.3769.

Hubbard S, Thornton J. NACCESS, Computer Program. Department of Biochemistry Molecular Biology, University College London. 1993.

Huerta-Cepas J, Szklarczyk D, Forslund K, et al. eggNOG 4.5: a hierarchical orthology framework with improved functional annotations for eukaryotic, prokaryotic and viral sequences. *Nucleic Acids Research.* 2016;44(Database issue):D286-D293. doi:10.1093/nar/gkv1248.

Hunter JD. Matplotlib: A 2D graphics environment. *Computing in Science and Engineering.* 2007; 9(3):90-95. doi:10.1109/MCSE.2007.55.

International HapMap Consortium. The International HapMap Project. *Nature.* 2003; 426(6968): 789-796. doi:10.1038/nature02168.

International Human Genome Sequencing Consortium. Initial sequencing and analysis of the human genome. *Nature*. 2001; 409(6822):860-921. doi:10.1038/35057062.

International Human Genome Sequencing Consortium. Finishing the euchromatic sequence of the human genome. *Nature*. 2004; 431(7011):931-945. doi:10.1038/nature03001.

Jalali Sefid Dashti M, Gamieldien J. A practical guide to filtering and prioritizing genetic variants. *Biotechniques*. 2017; 62(1):18-30. doi:10.2144/000114492.

James RA, Campbell IM, Chen ES, et al. A visual and curatorial

approach to clinical variant prioritization and disease gene discovery in genome-wide diagnostics. *Genome Medicine*. 2016;8:13. doi:10.1186/s13073-016-0261-8.

Jamuar SS, Kuan JL, Brett M, et al. Incidentalome from Genomic Sequencing: A Barrier to Personalized Medicine? *EBioMedicine*. 2016;5:211-216. doi:10.1016/j.ebiom.2016.01.030.

Jelier R, Semple JI, Garcia-Verdugo R, Lehner B. Predicting phenotypic variation in yeast from individual genome sequences. *Nature Genetics*. 2011; 43(12):1270-1274. doi:10.1038/ng.1007.

Jordan DM, Frangakis SG, Golzio C, et al. Identification of *cis*-suppression of human disease mutations by comparative genomics. *Nature*. 2015;524(7564):225-229. doi:10.1038/nature14497.

Kaufman S, Rosset S, Perlich C. Leakage in data mining: formulation, detection and avoidance. *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*. 2011; KDD'11:556-563. Doi:10.1145/2020408.2020496.

Keyes KM, Davey Smith G, Koenen KC, Galea S. The mathematical limits of genetic prediction for complex chronic disease. *Journal of epidemiology and community health*. 2015;69(6):574-579. doi:10.1136/jech-2014-204983.

Khatri BS, Goldstein RA. Simple Biophysical Model Predicts Faster Accumulation of Hybrid Incompatibilities in Small Populations Under Stabilizing Selection. *Genetics*. 2015;201(4):1525-1537. doi:10.1534/genetics.115.181685.

Kim JH, Jarvik GP, Browning BL, et al. Exome sequencing reveals novel rare variants in the ryanodine receptor and calcium channel genes in malignant. Hyperthermia Families. *Anesthesiology*. 2013;119(5):1054-1065. doi:10.1097/ALN.0b013e3182a8a998.

Kimura M. Evolutionary rate at the molecular level. *Nature*. 1968;

217(5129):624-626. doi:

Kimura M. The role of compensatory neutral mutations in molecular evolution. J*ournal of Genetics*. 1985;64:7–19.

Kircher M, Witten DM, Jain P, O'Roak BJ, Cooper GM, Shendure J. A general framework for estimating the relative pathogenicity of human genetic variants. *Nature genetics*. 2014;46(3):310-315. doi:10.1038/ng.2892.

Koboldt DC, Zhang Q, Larson DE, et al. VarScan 2: Somatic mutation and copy number alteration discovery in cancer by exome sequencing. *Genome Research*. 2012; 22(3):568-576. doi:10.1101/gr.129684.111.

Kohane IS, Hsing M, Kong SW. Taxonomizing, sizing, and overcoming the incidentalome. *Genetics in medicine : official journal of the American College of Medical Genetics*. 2012;14(4): 10.1038/gim.2011.68. doi:10.1038/gim.2011.68.

Kohavi R. A study of cross-validation and bootstrap for accuracy estimation and model selection. *Proceedings of the Fourteenth International Joint Conference on Artificial Intelligence*. 1995; 2(12):1137-43. doi:10.1.1.133.9187.

Köhler S, Doelken SC, Mungall CJ, et al. The Human Phenotype Ontology project: linking molecular biology and disease through phenotype data. *Nucleic Acids Research*. 2014;42(Database issue):D966-D974. doi:10.1093/nar/gkt1026.

Köhler S, Vasilevsky NA, Engelstad M, et al. The Human Phenotype Ontology in 2017. *Nucleic Acids Research*. 2017;45(Database issue):D865-D876. doi:10.1093/nar/gkw1039.

Kolmogorov A. Sulla determinazione empirica di una legge di distribuzione. *Giornale dell'Istituto Italiano degli Attuari.*1933; 4.83-91.

Kondrashov AS, Sunyaev S, Kondrashov FA. Dobzhansky–Muller incompatibilities in protein evolution. *Proceedings of the National Academy of Sciences of the United States of America.* 2002; 99(23):14878-14883. doi:10.1073/pnas.232565499.

Kyte J, Doolittle RF. A simple method for displaying the hydropathic character of a protein. *Journal of Molecular Biology.* 1982;157(1):105-132. doi:10.1016/0022-2836(82)90515-0.

Laksman Z, Detsky AS. Personalized Medicine: Understanding Probabilities and Managing Expectations. *Journal of General Internal Medicine*. 2011;26(2):204-206. doi:10.1007/s11606-010-1515-6.

Larochelle H, Bengio Y. Classification using discriminative restricted Boltzmann machines. *Proceedings of the 25th international conference on Machine learning.* 2008; ICML '08:536. doi:10.1145/1390156.1390224.

Lehner B. Molecular mechanisms of epistasis within and between genes. *Trends in Genetics*. 2011; 27(8):323-31. doi:10.1016/j.tig.2011.05.007.

Li H, Handsaker B, Wysoker A, et al. The Sequence Alignment/Map format and SAMtools. *Bioinformatics*. 2009;25(16):2078-2079. doi:10.1093/bioinformatics/btp352.

Li M, Petukh M, Alexov E, Panchenko AR. Predicting the Impact of Missense Mutations on Protein–Protein Binding Affinity. *Journal of Chemical Theory and Computation*. 2014;10(4):1770-1780. doi:10.1021/ct401022c.

Liu X, Wang J, Chen L. Whole-exome sequencing reveals recurrent somatic mutation networks in cancer. *Cancer letters.* 2013; 340(2): 270-276. doi:10.1016/j.canlet.2012.11.002.

Liu X, Wu C, Li C, Boerwinkle E. dbNSFP v3.0: a one-stop database of functional predictions and annotations for human non-

synonymous and splice site SNVs. *Human Mutation*. 2016;37(3):235-241. doi:10.1002/humu.22932.

Lohmueller KE, Indap AR, Schmidt S, et al. Proportionally more deleterious genetic variation in european than in african populations. *Nature*. 2008; 451(7181):994-997. doi:10.1038/nature06611.

López-Ferrando V, Gazzo A, de la Cruz X et al. Pmut: a web-based tool for the annotation of pathological variants on proteins, 2017 update. *Nucleic Acids Research.* 2017. gkx313 doi:10.1093/nar/gkx313.

MacArthur DG, Manolio TA, Dimmock DP, et al. Guidelines for investigating causality of sequence variants in human disease. *Nature*. 2014;508(7497):469-476. doi:10.1038/nature13127.

MacArthur DG, Tyler-Smith C. Loss-of-function variants in the genomes of healthy humans. *Human Molecular Genetics*. 2010;19(R2):R125-R130. doi:10.1093/hmg/ddq365.

Mankad A, Taniguchi T, Cox B, et al. Natural gene therapy in monozygotic twins with Fanconi anemia. *Blood*. 2006;107(8):3084-3090. doi:10.1182/blood-2005-07-2638.

Mardis ER. The $1,000 genome, the $100,000 analysis? *Genome Medicine*. 2010; 2(11):84. doi:10.1186/gm205.

Martin AC. Mapping PDB chains to UniProtKB entries. *Bioinformatics*. 2005; 21(23):4297-301. doi:10.1093/bioinformatics/bti694.

Matsugu M, Mori K, Mitari Y, Kaneda Y. Subject independent facial expression recognition with robust face detection using a convolutional neural network. *Neural Networks.* 2003; 16(Special Issue 5-6):555-559. doi:10.1016/S0893-6080(03)00115-1.

Matthews BW. Comparison of the predicted and observed

secondary structure of T4 phage lysozyme. *Biochimica et Biophysica Acta (BBA) - Protein Structure.* 1975; 405 (2): 442–451. doi:10.1016/0005-2795(75)90109-9.

Matthews BW. Studies on protein stability with T4 Lysozime. *Advances in Protein Chemistry.* 1995. 46:249-78. doi:10.1016/S0065-3233(08)60337-X.

McKenna A, Hanna M, Banks E, et al. The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Research*. 2010;20(9):1297-1303. doi:10.1101/gr.107524.110.

McKeone R, Wikstrom M, Kiel C, Rakoczy PE. Assessing the correlation between mutant rhodopsin stability and the severity of retinitis pigmentosa. *Molecular Vision*. 2014;20:183-199.

McKinney W. Data Structures for Statistical Computing in Python, *Proceedings of the 9th Python in Science Conference.* 2010;51-56.

Meselson M, Stahl FW. The replication of DNA in Escherichia coli. *Proceedings of the National Academy of Sciences of the United States of America*. 1958;44(7):671-682.

Miotto R, Wang F, Wang S, et al. Deep learning for healthcare: review, opportunities and challenges. *Briefings in Bioinformatics.* [epub]. doi:10.1093/bib/bbx044.

Moncunill V, Gonzalez S, Beà S, et al. Comprehensive characterization of complex structural variations in cancer by directly comparing genome sequence reads. *Nature Biotechnology*. 2014; 32(11):1106-1112. doi:10.1038/nbt.3027.

Montazeri M, Montazeri M, Montazeri M, Beigzadeh A. Machine learning models in breast cancer survival prediction. *Technology and Healthcare.* 2016;24(1):31-42. doi:10.3233/THC-151071.

Moraru M, Black LE, Muntasell A, et al. NK cell and Ig interplay in

defense against Herpes Simplex Virus type 1: epistatic interaction of CD16A and IgG1 allotypes of variable affinities modulates antibody-dependent cellular cytotoxicity and susceptibility to clinical reactivation. *Journal of Immunology.* 2015;195(4):1676-1684. doi:10.4049/jimmunol.1500872.

Moura de Sousa J, Balbontín R, Durão P, Gordo I. Multidrug-resistant bacteria compensate for the epistasis between resistances. de Visser A, ed. *PLoS Biology*. 2017;15(4):e2001741. doi:10.1371/journal.pbio.2001741.

Murakami S, Oshima H, Hayashi T, Kinoshita M. On the physics of thermal-stability change upon mutation of a protein. *The Journal of Chemical Physics*. 2015; 143(12): 125102. doi:10.1063/1.4931814.

Narod SA, Foulkes WD. BRCA1 and BRCA2: 1994 and beyond. *Nature Reviews Cancer.* 2004; 4(9):665-676. doi:10.1038/nrc1431

Ng PC, Henikoff S. Predicting Deleterious Amino Acid Substitutions. *Genome Research*. 2001;11(5):863-874. doi:10.1101/gr.176601.

Niroula A, Urolagin S, Vihinen M. PON-P2: Prediction Method for Fast and Reliable Identification of Harmful Variants. Tosatto SCE, ed. *PLoS ONE*. 2015;10(2):e0117380. doi:10.1371/journal.pone.0117380.

Niroula A, Vihinen M. Variation interpretation predictors: principles, types, performance and choice. *Human Mutation.* 2016; 37(6):579-97. doi:10.1002/humu.22987.

Niroula A, Vihinen M. Predicting severity in disease causing variants. *Human Mutation.* 2017; 38(4):357-364. doi:10.1002/humu.23173.

Obermeyer Z, Emanuel EJ. Predicting the Future — Big Data, Machine Learning, and Clinical Medicine. *The New England journal of medicine*. 2016;375(13):1216-1219.

doi:10.1056/NEJMp1606181.

Odriozola L, Corrales FJ. Discovery of nutritional biomarkers: future directions based on omics technologies. *International Journal of Food Sciences and Nutrition.* 2015; 66 Suppl 1:S31-40.

Oh HJ, Choi D, Goh CJ, Hahn Y. Loss of gene function and evolution of human phenotypes. *BMB Reports*. 2015;48(7):373-379. doi:10.5483/BMBRep.2015.48.7.073.

Ohta T. Slightly deleterious mutant substitutions in evolution. *Nature.* 1973; 246(5428): 96-98.

Oldenburg J, Ananyeva NM, Saenko EL. Molecular basis of haemophilia A. *Haemophilia. 2004;* 10(4):133-139. doi:10.1111/j.1365-2516.2004.01005.x

Oliver GR, Hart SN, Klee EW. Bioinformatics for clinical next generation sequencing. *Clinical Chemistry.* 2015; 61(1):124-135. doi:10.1373/clinchem.2014.224360.

Online Mendelian Inheritance in Man, OMIM®. Johns Hopkins University, Baltimore, MD. MIM Number: 306700 and 306900. Consultat el 2017. https://omim.org/

Ordovás JM. Integración del medio ambiente en el análisis «ómico». *Revista española de Cardiología.* 2009; 62 Suppl 2:17-22. doi:10.1016/S0300-8932(09)72118-9

Oxford Nanopore Technologies. Our goal: to enable the analysis of any living thing, by any person, in any environment [Internet]. https://nanoporetech.com/about-us Consultat el 03/05/2017.

Pacific Biosciences. Advance genomics with single molecule, real-time (SMRT) sequencing [Internet]. www.pacb.com/smrt-science/smrt-sequencing/ Consultat el 03/05/2017.

Pandey RV, Pabinger S, Kriegner A, Weinhäusel A. ClinQC: a tool

for quality control and cleaning of Sanger and NGS data in clinical research. *BMC Bioinformatics*. 2016;17:56. doi:10.1186/s12859-016-0915-y.

Pavlova A, Oldenburg J. Defining severity in hemophilia: more than factor levels. *Seminars in Thrombosis and Hemostasis*. 2013; 39(7):702-710. doi:10.1055/s-0033-1354426.

Pedregosa F, Varoquax G, Gramfort A et al. Scikit-learn: Machine Learning in Python. *The Journal of Machine Learning Research.* 2011; 12:2825-2830.

Pedros C, Gaud G, Bernard I, et al. An epistatic interaction between Themis1 and Vav1 modulates regulatory T cell function and Inflamatory Bowel Disease development. *Journal of Immunology*. 2015; 195(4):1608-1616. doi:10.4049/jimmunol.1402562.

Pérez F, Granger EB. IPython: A System for Interactive Scientific Computing. *Computing in Science and Engineering.* 2007;9(3):21-29. doi:10.1109/MCSE.2007.53.

Pertea M, Salzberg S. Between a chicken and a grape: estimating the number of human genes. *Genome Biology*. 2010; 11(5): 206. doi:10.1186/gb-2010-11-5-206.

Posth C, Renaud G, Mittnik A, et al. Pleistocene mitochondrial genomes suggest a single major dispersal of non-africans and a late glacial population turnover in europe. *Current Biology*. 2016; 26(6):827-833. doi:10.1016/j.cub.2016.01.037.

Povolotskaya IS, Kondrashov FA. Sequence space and the ongoing expansion of the protein universe. *Nature*. 2010; 465(7300):922-6. doi:10.1038/nature09105.

Powers DMW. Evaluation: from precision, recall and f-measure to ROC, informedness, markedness and correlation. *Journal of Machine Learning Technologies.* 2011; 2(1):37-63. doi:10.9735/2229-3981.

Python Software Foundation. Python Language Reference, versió 2.7. Disponible http://www.python.org

R Core Team (2013). R: A language and environment for statistical computing. R. *Foundation for Statistical Computing*. 2013. http://www.R-project.org/.

Raimondi D, Tanyalcin I, Ferté J et al. DEOGEN2: prediction and interactive visualization of a single amino acid variant deleteriousness in human proteins. *Nucleic Acids Research.* 2017. [epub]. doi:10.1093/nar/gkx390.

Ramensky V, Bork P, Sunyaev S. Human non-synonymous SNPs: server and survey. *Nucleic Acids Research*. 2002;30(17):3894-3900.

Relling MV, Evans WE. Pharmacogenomics in the clinic. *Nature*. 2015; 526(7573):343-350. doi:10.1038/nature15817.

Reva B, Antipin Y, Sander C. Determinants of protein function revealed by combinatorial entropy optimization. *Genome Biology*. 2007;8(11):R232. doi:10.1186/gb-2007-8-11-r232.

Riera C, Lois S, de la Cruz X. Prediction of pathological mutations in proteins: the challenge of integrating sequence conservation and structure stability principles. *WIREs Computational Molecular Science.* 2014; 4:249–268. doi:10.1002/wcms.1170.

Riera C, Lois S, Domínguez C et al. Molecular damage in Fabry disease: characterization and prediction of alpha-galactosidase A pathological mutations. *Proteins*. 2015;83(1):91-104. doi:10.1002/prot.24708.

Riera C, Padilla N, de la Cruz X. The complementarity between protein-specific and general pathogenicity predictors for amino acid substitutions. *Human Mutation*. 2016;37(10):1013-24. doi:10.1002/humu.23048.

Roberts ND, Kortschak RD, Parker WT, et al. A comparative

analysis of algorithms for somatic SNV detection in cancer. *Bioinformatics*. 2013; 29(18):2223-2230. doi:10.1093/bioinformatics/btt375.

Rockah-Shmuel L, Tóth-Petróczy Á, Tawfik DS. Systematic mapping of protein mutational space by prolonged drift reveals the deleterious effects of seemingly neutral mutations. Orengo CA, ed. *PLoS Computational Biology.* 2015;11(8):e1004421. doi:10.1371/journal.pcbi.1004421.

Roden D, Tyndale R. Genomic Medicine, Precision Medicine, Personalized Medicine: What's in a Name? *Clinical pharmacology and therapeutics*. 2013; 94(2):169-172. doi:10.1038/clpt.2013.101.

Rosenblatt F. The Perceptron -- a perceiving and recognizing automaton. *Cornell Aeronautical Laboratory.* 1957. Report 85-460-1.

Roth SM, Walsh S, Liu D, et al. The *ACTN3* R577X nonsense allele is under-represented in elite-level strength athletes. *European journal of human genetics : EJHG*. 2008;16(3):391-394. doi:10.1038/sj.ejhg.5201964.

Roukos DH. Trastuzumab and beyond: sequencing cancer genomes and predicting molecular networks. *Pharmacogenomics Journal*. 2011; 11(2):81-92. doi:10.1038/tpj.2010.81.

Ruxton GD, Wilkinson DM. Endurance running and its relevance to scavenging by early hominins. *Evolution*. 2013; 67(3):861-867. doi:10.1111/j.1558-5646.2012.01815.x

Sanger F, Nicklen S, Coulson AR. DNA sequencing with chain-terminating inhibitors. *Proceedings of the National Academy of Sciences of the United States of America*. 1977; 74(12):5463-5467.

Savova V, Patsenker J, Vigneau S et al. dbMAE: the database of autosomal monoallelic expression. Nucleic Acids Research. 2016;44(Database issue):D753-D756. doi:10.1093/nar/gkv1106.

Schmidt B, Hildebrandt A. Next-generation sequencing: big data meets high performance computing. *Drug Discovery Today.* 2017;22(4):712-717. doi:10.1016/j.drudis.2017.01.014.

Schneider VA, Graves-Lindsay T, Howe K, et al. Evaluation of GRCh38 and de novo haploid genome assemblies demonstrates the enduring quality of the reference assembly. *Genome Research*. 2017;27(5):849-864. doi:10.1101/gr.213611.116.

Schymkowitz J, Borg J, Stricher F et al. The FoldX web server: an online force field. *Nucleic Acids Research*. 2005;33(Web Server issue):W382-W388. doi:10.1093/nar/gki387.

Sergeev YV, Vitale S, Sieving PA, et al. Molecular modeling indicates distinct classes of missense variants with mild and severe XLRS phenotypes. *Human Molecular Genetics*. 2013;22(23):4756-4767. doi:10.1093/hmg/ddt329.

Sikosek T, Chan HS. Biophysics of protein evolution and evolutionary protein biophysics. *Journal of the Royal Society Interface.* 2014;11(100):20140419. doi:10.1098/rsif.2014.0419.

Sim N-L, Kumar P, Hu J, Henikoff S, Schneider G, Ng PC. SIFT web server: predicting effects of amino acid substitutions on proteins. *Nucleic Acids Research*. 2012;40(Web Server issue):W452-W457. doi:10.1093/nar/gks539.

Simpson EH, Morud J, Winiger V, et al. Genetic variation in COMT activity impacts learning and dopamine release capacity in the striatum. *Learning & Memory*. 2014;21(4):205-214. doi:10.1101/lm.032094.113.

Shannon, CE. A Mathematical Theory of Communication. *Bell System Technical Journal.* 1948; 27(3): 379–423. doi:10.1002/j.1538-7305.1948.tb01338.x

Shen T, Lee A, Shen C, Lin CJ. The long tail and rare disease research: the impact of next-generation sequencing for rare

Mendelian disorders. *Genetics Research.* 2015; 97:e17. Doi:10.1017/S0016672315000166.

Shrijver I, Aziz N, Farkas DH et al. Opportunities and challenges associated with clinical diagnostic genome sequencing: a report of the association for molecular pathology. *Journal of Molecular Diagnosis.* 2012; 14(6):525-540. doi:10.1016/j.jmoldx.2012.04.006.

Slamon DJ, Leyland-Jones B, Shak S et al. Use of chemotherapy plus a monoclonal antibody against HER2 for metastatic breast cancer that overexpresses HER2. *New England Journal of Medicine*. 2001; 344(11):783-792. doi:10.1056/NEJM200103153441101.

Slatkin M. A population-genetic test of founder effects and implications for Ashkenazi jewish diseases. *American Journal of Human Genetics*. 2004;75(2):282-293.

Smedley D, Köhler S, Czeschik JC, et al. Walking the interactome for candidate prioritization in exome sequencing studies of Mendelian diseases. *Bioinformatics*. 2014; 30(22):3215-3222. doi:10.1093/bioinformatics/btu508.

Smith LM, Sanders JZ, Kaiser RJ, et al. Fluorescence detection in automated DNA sequence analysis. *Nature.* 1986; 321(6071):674-679. doi: 10.1038/321674a0.

Smolensk P. Chapter 6: Information processing in dynamical systems: foundations of harmony theory. *Parallel Distributed Processing: Explorations in the Microstructure of Cognition, Volume 1: Foundations. 1986;* 194-281.

Song M, Lee HW, Kang D. The potential application of personalized preventive research. *Japanese Journal of Clinical Oncology.* 2014; 44(11):1017-1024. doi:10.1093/jjco/hyu135.

Spencer M, Eickholt J, Cheng J. A Deep Learning Network Approach to *ab initio* Protein Secondary Structure Prediction.

IEEE/ACM transactions on computational biology and bioinformatics / IEEE, ACM. 2015;1 2(1):103-112. doi:10.1109/TCBB.2014.2343960.

Stapleton AE. A biologist, a statistician, and a bioinformatician walk into a conference room... and walk out with a great metagenomics project plan. *Frontiers in Plant Science.* 2014; 5:250. doi:10.3389/fpls.2014.00250.

Stark Z, Tan TY, Chong B, et al. A prospective evaluation of whole-exome sequencing as a first-tier molecular test in infants with suspected monogenic disorders. *Genetics in medicine.* 2016; 18(11):1090-1096. doi:10.1038/gim.2016.1.

Starr TN, Thornton JW. Epistasis in protein evolution. *Protein Science.* 2016; 25(7):1204-18. doi:10.1002/pro.2897.

Stone EA, Sidow A. Physicochemical constraint violation by missense substitutions mediates impairment of protein function and disease severity. *Genome Research.* 2005;15(7):978-986. doi:10.1101/gr.3804205.

Tagariello G, Iorio A, Santagostino E, et al. Comparison of the rates of joint arthroplasty in patients with severe factor VIII and IX deficiency: an index of different clinical severity of the 2 coagulation disorders. *Blood*. 2009; 114(4):779-784. doi:10.1182/blood-2009-01-195313.

Tang H, Thomas PD. Tools for predicting the functional impact of nonsynonymous genetic variation. *Genetics*. 2016;203(2):635-647. doi:10.1534/genetics.116.190033.

Temin HM, Mizutani S. RNA-dependent DNA polymerase in virions of Rous sarcoma virus. *Nature*. 1970; 226(5252): 1211-1213. doi:10.1038/2261211a0.

Tetreault M, Bareke E, Nadaf J, et al. Whole-exome sequencing as a diagnostic tool: current challenges and future opportunities. *Expert*

Review of Molecular Diagnostics. 2015; 15(6): 749-760. doi:10.1586/14737159.2015.1039516.

Thakral G, Vierkoetter K, Namiki S, et al. AML multi-gene panel testing: A review and comparison of two gene panels. *Pathology – Research and Practice.* 2016; 212(5); 372-380. doi:10.1016/j.prp.2016.02.004.

The 1000 Genomes Project Consortium. A global reference for human genetic variation. *Nature*. 2015; 526(7571):68-74. doi:10.1038/nature15393.

The Cancer Genome Atlas (TCGA) Research Network. Comprehensive genomic characterization defines human glioblastoma genes and core pathways. *Nature*. 2008; 455(7216):1061-1068. doi:10.1038/nature07385.

The UniProt Consortium. UniProt: the universal protein knowledgebase. *Nucleic Acids Research.* 2017; 45 (D1): D158-D169. doi:10.1093/nar/gkw1099.

Tian K, Shao M, Wang Y, et al. Boosting compound-protein interaction prediction by deep learning. *Methods.* 2016;110:64-72. doi:10.1016/j.ymeth.2016.06.024.

To-Figueras J, Ducamp S, Clayton J et al. ALAS2 acts as a modifier gene in patients with congenital erithropoyetic porphyria. *Blood*. 2011; 118(6):1443-51. doi:10.1182/blood-2011-03-342873.

Tokuriki N, Tawfik DS. Stability effects of mutations and protein evolvability. *Current Opinion in Structural Biology.* 2009;19(5):596-604. doi:10.1016/j.sbi.2009.08.003.

Vang YS, Xie X. HLA class I binding prediction via convolutional neural networks. *Bioinformatics.* 2017. [epub]. doi: 10.1093/bioinformatics/btx264

Venselaar H, Joosten RP, Vroling B, et al. Homology modelling and

spectroscopy, a never-ending love story. *European Biophysics Journal*. 2010; 39(4):551-563. doi:10.1007/s00249-009-0531-0.

Venter JC, Adams MD, Myers EW, et al. The sequence of the human genome. *Science.* 2001; 291(5507):1304-1351. doi:10.1126/science.1058040.

Vihinen M. How to evaluate performance of prediction methods? Measures and their interpretation in variation effect analysis. *BMC Genomics*. 2012;13(Suppl 4):S2. doi:10.1186/1471-2164-13-S4-S2.

Vihinen M. Guidelines for reporting and using prediction tools for genetic variation analysis. *Human Mutation.* 2013; 34(2): 275-82. doi:10.1002/humu.22253.

Vihinen M. Proper reporting of predictor performance. *Nature Methods.* 2014; 11(8):781. doi:10.1038/nmeth.3032.

Vilella AJ, Severin J, Ureta-Vidal A et al. EnsemblCompara GeneTrees: Complete, duplication-aware phylogenetic trees in vertebrates. *Genome Research.* 2009;19(2):327-335. doi:10.1101/gr.073585.107.

Vinh NX, Epps J, Bailey J. Information theoretic measures for clusterings comparison. *Proceedings of the 26th Annual Conference on Machine Learning.* 2009. ICML'09:1. Doi:10.1145/1553374.1553511.

Vu V, Verster AJ, Shertzberg M et al. Natural variation in gene expression modulates the severity of mutant phenotypes. *Cell*. 2015; 162(2):391-402. doi:10.1016/j.cell.2015.06.037.

van der Walt S, Colbert SC and Varoquaux G. The NumPy Array: A Structure for Efficient Numerical Computation. *Computing in Science & Engineering.* 2011;13:22-30. doi:10.1109/MCSE.2011.37.

Wang K, Li M, Hakonarson H. ANNOVAR: functional annotation

of genetic variants from high-throughput sequencing data. *Nucleic Acids Research*. 2010; 38(16):e164. doi:10.1093/nar/gkq603.

Wang HW, Pai TW. Machine learning-based methods for prediction of linear B-cell epitopes. *Methods in Molecular Biology.* 2014;1184:217-236. doi:10.1007/978-1-4939-1115-8_12.

Wang W-L, Xu S-Y, Ren Z-G et al.. Application of metagenomics in the human gut microbiome. *World Journal of Gastroenterology : WJG*. 2015;21(3):803-814. doi:10.3748/wjg.v21.i3.803.

Warden CD, Adamson AW, Neuhausen SL, Wu X. Detailed comparison of two popular variant calling packages for exome and targeted exon studies. Emes R, ed. *PeerJ*. 2014; 2:e600. doi:10.7717/peerj.600.

Wascom M, Botvinnik O, drewokane et al. Seaborn v0.7.1. 2016. doi:10.5281/zenodo.54844.

Watson JD, Crick FH. Molecular structure of nucleic acids; a structure for deoxyribose nucleic acid. *Nature.* 1953*;* 171(4356): 737-738. doi:10.1038/171737a0.

Wickham H. ggplot2: Elegant Graphics for Data Analysis. *Springer-Verlag New York*. 2009.

Wilkins AD, Venner E, Marciano DC, et al. Accounting for epistatic interactions improves the functional analysis of protein structures. *Bioinformatics*. 2013;29(21):2714-2721. doi:10.1093/bioinformatics/btt489.

Wong GY, Leung FH, Ling SH. Predicting protein-ligand binding site using support vector machine with protein properties. *IEEE/ACM Transactions on Computational Biology and Bioinformatics.* 2013;10(6):1517-1529. doi:10.1109/TCBB.2013.126.

Wu P-Y, Cheng C-W, Kaddi CD, et al. -Omic and electronic health

record Big Data analytics for precision medicine. *IEEE Transactions on Biomedical Engineering.* 2016; 64(2):262-273. doi:10.1109/TBME.2016.2573285.

Xu J, Zhang J. Why human disease-associated residues appear as the wild-type in other species: genome-scale structural evidence for the compensation hypothesis. *Molecular Biology and Evolution*. 2014;31(7):1787-1792. doi:10.1093/molbev/msu130.

Xue Y, Daly A, Yngvadottir B, et al. Spread of an inactive form of Caspase-12 in humans is due to recent positive selection. *American Journal of Human Genetics*. 2006;78(4):659-670.

Xue LC, Dobbs D, Bonvin AMJJ, Honavar V. Protein-Protein interface predictions by data-driven methods: a review. *FEBS letters*. 2015;589(23):3516-3526. doi:10.1016/j.febslet.2015.10.003.

Yang N, MacArthur DG, Gulbin JP, et al. *ACTN3* Genotype Is Associated with Human Elite Athletic Performance. *American Journal of Human Genetics*. 2003;73(3):627-631.

Yang Y, Niroula A, Shen B, Vihinen M. PON-Sol: prediction of effects of amino acid substitutions on protein solubility. *Bioinformatics.* 2016;32(13):2032-2034. doi:10.1093/bioinformatics/btw066

Yu G, Smith DK, Zhu H, et al. ggtree: an R package for visualization and annotation of phylogenetic trees with their covariates and other associated data. *Methods in Ecology and Evolution*. 2017, 8(1):28-36. doi:10.1111/2041-210X.12628.

Zhang T, Ramakrishnan R, Livny M. BIRCH: an efficient data clustering method for very large databases. *Proceedings of the 1997 ACM SIGMOD international conference on Management of data.* 1996; SIGMOD '96:103-114. doi:10.1145/233269.233324.

Zhang S-W, Hao L-Y, Zhang T-H. Prediction of protein–protein interaction with pairwise Kernel Support Vector Machine.

International Journal of Molecular Sciences. 2014;15(2):3220-3233. doi:10.3390/ijms15023220.

Zhou X, Rokas A. Prevention, diagnosis and treatment of high-throughput sequencing data pathologies. *Molecular Ecology.* 2014; 23(7):1679-1700. doi:10.1111/mec.12680.

Zhou J, Troyanskaya OG. Predicting effects of noncoding variants with deep learning-based sequence model. *Nature Methods.* 2015;12(10):931-934. doi:10.1038/nmeth.3547.

Ziogas D, Roukos DH. *CDH1* Testing: can it predict the prophylactic or therapeutic nature of total gastrectomy in hereditary diffuse gastric cancer? *Annals of Surgical Oncology*. 2009;16(10):2678-2681. doi:10.1245/s10434-009-0598-y.

Zuckerkandl E and Pauling L Molecular disease, evolution and genic heterogeneity. *Horizons in Biochemistry: Albert Szent−Györgyi Dedicatory Volume*. 1962:189–225.