UNIVERSITAT ROVIRA i VIRGILI

IDENTIFICACIÓ D'ACTIVITY CLIFFS EN COMPOSTOS QUE S'UNEIXEN A LA PROTEASA M^{pro} del SARS-CoV-2

TREBALL DE FI DE GRAU
Grau de Biotecnologia

Òscar Albert Cañamero
Tutor: Gerard Pujadas

Tarragona
Novembre 2020

Índex

Dades del centre

Aquest treball de fi de grau s'ha dut a terme gràcies a l'ajut del grup de recerca de Quimioinformàtica i Nutrició de la Universitat Rovira i Virgili, que pertany al departament de Bioquímica i Biotecnologia de la Universitat Rovira i Virgili. El grup de recerca centra la seva activitat en la utilització d'eines computacionals per a trobar nous compostos bioactius i nous usos de determinades molècules de pocs centenars de Da.

Resum i paraules claus

Els *activity cliffs* són parelles de compostos que presenten estructures molt similars però que tenen una diferència d'activitat molt alta. Aquests permeten capturar modificacions químiques que influeixen en l'activitat biològica, pel que són de gran utilitat i interès en l'anàlisi de la relació estructura-activitat dels compostos. En aquest projecte s'han buscat *activity cliffs* entre molècules que inhibeixen la M^{pro} del SARS-CoV-2 i molècules que no l'inhibeixen. Per a fer-ho s'ha desenvolupat una aplicació que permet trobar *activity cliffs* mitjançant la representació en un dendrograma de les molècules de dos arxius segons la seva similitud (un arxiu per a les molècules que inhibeixen la M^{pro} i un altre per a les que no l'inhibeixen). Per a calcular la similitud entre molècules s'utilitzen fingerprints i el càlcul de l'índex de Tanimoto entre els fingerprints de les diferents molècules. L'aplicació permet diferenciar quins compostos pertanyen a cada arxiu i mostra l'estructura de les molècules quan l'usuari interacciona amb el nom d'aquestes al dendrograma. En total s'han trobat 18 *activity cliffs* entre totes les molècules analitzades.

Paraules clau: *Activity cliff*, SARS-CoV-2, M^{pro}, Tanimoto, Fingerprint, COVID-19, Dendrograma.

1. Introducció

El SARS-CoV-2 o coronavirus 2 de la síndrome respiratòria aguda greu és un coronavirus d'origen animal causant de la malaltia de la COVID-19 (Coronavirus Disease-2019). Els coronavirus són virus ARN monocatenaris positius (+ssRNA) de la família dels *Coronaviridae* que es poden classificar en quatre gèneres principals: el gènere *Alphacoronavirus*, el gènere *Betacoronavirus*, el gènere *Gammacoronavirus* i el gènere *Deltacoronavirus*. Generalment els Alphacoronavirus i Betacoronavirus infecten a mamífers, mentre que els Gammacoronavirus i Deltacoronavirus infecten a ocells (1). Abans de l'arribada del SARS-CoV-2 només es coneixien sis coronavirus que podien causar malalties als humans: el HCoV-229E (229E), el HCoV-OC43 (OC43), el coronavirus de la síndrome respiratoria aguda greu (SARS-CoV), el HCoV-NL63 (NL63), el HCoV-HKU1 (HKU1) i el coronavirus de la síndrome respiratòria de l'Orient Mitjà (MERS-CoV) (2). Generalment els coronavirus 229E, OC43, HKU1 i NL63 provoquen refredats comuns o malalties respiratòries lleus, mentre que el SARS-CoV, el MERS-CoV i el SARS-CoV-2 provoquen símptomes més severs, podent arribar a causar la mort (2). El primer brot de SARS-CoV va sorgir de la província de Guandong el Novembre de 2002 i va afectar a 8.098 persones, 774 de les quals van morir (3). El MERS-CoV va ser detectat per primer cop a Aràbia Saudita el 2012 i aproximadament un 35% dels casos de MERS-CoV que han estat diagnosticats han resultat fatals (4). Tant aquests dos virus com el SARS-CoV-2 són betacoronavirus. Els primers casos de pneumònia causada per COVID-19 van aparèixer a la ciutat de Wuhan el Desembre de 2019, i d'ençà la malaltia s'ha estès de manera global, provocant una pandèmia mundial que ha afectat a més de 42 milions de persones i provocat més d'1 milió de morts (5). Els símptomes de la COVID-19 poden variar, essent els lleus la febre, la tos seca, el malestar general, la diarrea, la pèrdua de l'olfacte o del gust, el mal de cap, la conjuntivitis o el mal de gola. Els símptomes més greus inclouen la dificultat per a respirar, la falta d'alè, el mal de pit o la pressió en aquest (6).

El genoma del SARS-CoV-2 està format per una cadena d'ARN de 29.903 nucleòtids (7), presentant el següent ordre de gens de 5' a 3': *ORF1a* i *ORF1b*, espícula (*Spike*, S), embolcall (*Envelope*, E), membrana (*Membrane*, M) i nucleocàpsida (*Nucleocapsid*, N). El SARS-CoV-2 presenta les seqüencies terminals característiques dels betacoronavirus, tenint 265 nucleòtids a l'extrem 5′ i 229 nucleòtids a l'extrem 3′. Els gens *ORF1a* i *ORF1b*

estan format per 21.291 nucleòtids i codifiquen 16 proteïnes no estructurals . A continuació d'aquests es situen 13 ORFs, entre els quals es troben els gens que codifiquen les proteïnes estructurals prèviament mencionades i les proteïnes accessòries 3a, 3b, 6, 7a, 7b, 9a, 9b i 10 (7). Tot i que el genoma del SARS-CoV-2 és similar al del SARS-CoV, existeixen diferencies entre les proteïnes accessòries d'aquests, com ara la manca de proteïna accessòria 8b en el SARS-CoV-2 i diferencies significatives en el número d'aminoàcids en les proteïnes 8b i 3b (8). A la Figura 1 es poden observar aquestes diferencies.

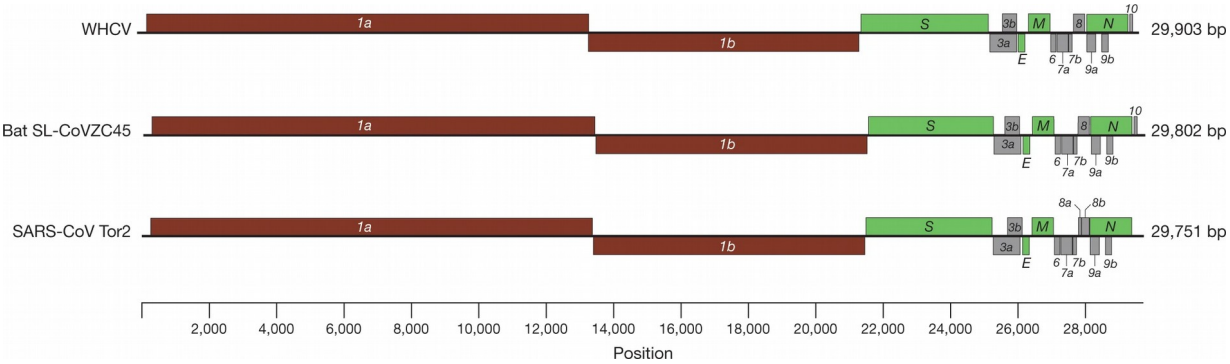


Figura 1. En aquesta imatge es mostra la organització dels gens del SARS-CoV-2 (WHCV), un coronavirus associat als ratpenats (bat SL-CoVZC45) i un coronavirus associat a humans (SARS-CoV Tor2). Figura extreta de Wu F et al, 2020 (7).

La proteïna S o de l'espícula, que es codifica a partir del gen S, és una glicoproteïna de 1273 aminoàcids que està formada per un pèptid senyal trobat a l'extrem N-terminal (aminoàcids 1-13), la subunitat S1 (residus 14-185) i la subunitat S2 (residus 686-1273). La subunitat S1 es la responsable de la unió al receptor i la subunitat S2 és responsable de la fusió amb la membrana del virus (9). La proteïna forma homotrímers que sobresurten a la membrana del virus per a facilitar la unió amb la cèl·lula hoste. La proteïna E o embolcall, que és la proteïna més petita de l'estructura del virus, es troba a la membrana lipídica i forma part del procés de producció i maduració del virus. La proteïna M o de membrana és la proteïna més abundant en l'embolcall lipídic i té un rol important en la determinació de la forma d'aquest. La unió de la proteïna N amb la proteïna M ajuda a l'estabilització de la nucleocàpsida i promou la unió entre l'ARN i aquesta, estabilitzant el complexe ARN-proteïna N (10). La proteïna N, que es codifica a partir del gen N, s'uneix a l'ARN viral per a formar la nucleocàpsida, estabilitzant l'ARN viral. Aquesta està involucrada en processos relacionats amb el cicle de replicació viral i en la resposta de la cèl·lula hoste a la infecció viral (10). La membrana del virus conté proteïnes S glicosilades que s'uneixen a l'enzim convertidor d'angiotensina 2 o ACE2 (Angiotensin-converting

enzyme 2), donant lloc a l'entrada i unió mitjançant la fusió de la membrana viral i la cèl·lula hoste provocant que la proteasa serina transmembranal 2 o TMPRSS2 (Transmembrane protease serine 2), trobada a la membrana cel·lular de la cèl·lula hoste, promogui l'entrada del virus activant la proteïna S. Quan el virus entra a la cèl·lula s'allibera l'ARN viral, les poliproteïnes es tradueixen i comença la replicació i transcripció de l'ARN viral un cop es talla la poliproteïna i es duu a terme l'acoblament del complexe replicasa-transcriptasa (9).

Tot i que actualment s'estan desenvolupant diferents vacunes, els canvis que poden patir les diferents soques dels coronavirus fan que la troballa d'opcions terapèutiques comuns contra aquests sigui una necessitat.

1.1. La SARS-CoV-2 M^{pro}

La proteasa principal del SARS-CoV-2 o M^{pro} (també anomenada 3CL-pro o nsp5) ha esdevingut una de les dianes principals en estudis per al desenvolupament de tractaments antivirals per a la COVID-19.

La M^{pro} es troba a les poliproteïnes pp1a i pp1ab, que s'obtenen a partir de la traducció dels ORF1a i ORF1b. La traducció de l'ORF1b es duu a terme a partir d'un canvi en la pauta de lectura per part de la RNA-polimerasa a la posició -1 del solapament dels gens ORF1a i ORF1b (11). La poliproteïna pp1ab conté les proteïnes no estructurals necessàries per a la replicació i transcripció del virus, que es mostren a la Taula 1. La M^{pro} és un dels enzims claus en el cicle viral: aquesta té un paper essencial en la replicació de l'ARN viral ja que s'encarrega de processar les proteïnes pp1a i pp1ab junt amb la proteïna PL^{pro} (papain-like protease o nsp3). La M^{pro} es separa de la poliproteïna pp1ab mitjançant una autoproteòlisi i a continuació s'encarrega de digerir la poliproteïna per 11 llocs per a obtindre les proteïnes no estructurals (12,13). La Figura 2 mostra la distribució de totes les proteïnes no estructurals que contenen les poliproteïnes pp1a i pp1ab i la Taula 1 llista les funcions de cadascuna d'aquestes (14).

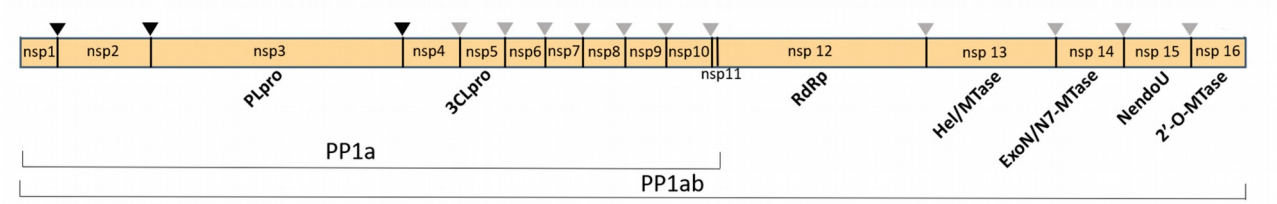


Figura 2. Distribució de les diferents proteïnes no estructurals en les poliproteïnes pp1a i pp1ab. Les fletxes negres són els punts per on talla la PL^{pro}, mentre que les fletxes de color gris són els punts per on talla la M^{pro}. Figura extreta de Romano M. et al., 2020 (14).

Proteïna	Funció
nsp1	Degradació i inhibició de la traducció de l'ARN-m de l'hoste.
nsp2	Desconeguda.
nsp3	Processament de la poliporteïna, desubiquitinació, formació de les vesícules de doble membrana.
nsp4	Formació de les vesícules de doble membrana.
nsp5	Processament de les poliproteïnes.
nsp6	Formació de les vesícules de doble membrana.
nsp7	Co-factor de la RNA Polimerasa RNA-dependent.
nsp8	Primasa, co-factor de la RNA Polimerasa RNA-dependent.
nsp9	Unió de l'ARN de cadena simple.
nsp10	Cofactor per a la nsp14 i 16.
nsp11	Desconeguda.
nsp12	RNA Polimerasa RNA-dependent, nucleotidiltransferasa.
nsp13	Helicasa, 5' RNA trifosfatasa.
nsp14	Exoribonucleasa 3' a 5', correcció d'errors, formació de la caputxa 5'.
nsp15	Endoribonucleasa.
nsp16	Formació de la caputxa 5', ribosa 2'-O-metiltransferasa.

Taula 1. Funcions de les diferents proteïnes no estructurals que formen part de les poliproteïnes pp1a i pp1ab (11).

Les M^{pro} del SARS-CoV-2 i del SARS-CoV tenen una semblança del 96% en la seva seqüencia d'aminoàcids (15), presentant diferències en 12 residus i només un d'ells, el residu Ser46 (Ala46 al SARS-CoV), es troba a prop del centre catalític. La Figura 3 mostra l'alineament de les seqüencies d'aminoàcids de l'M^{pro} del SARS-CoV i del SARS-CoV-2, on es poden veure les diferencies entre aquestes. Al SARS-CoV-2 el residu Thr285 del SARS-CoV és una Ala, i el residu Ile286 del SARS-CoV és una Leu. Aquests canvis tenen una gran importància en l'activitat de l'M^{pro}, especialment el canvi T285A, que permet que els dos dominis III de la M^{pro} s'apropin (12,16). El fet de que hi hagi una conservació alta entre les M^{pro} d'ambdues soques fa que la cerca d'un fàrmac que tingui com a diana aquesta proteïna sigui molt interessant per al tractament del SARS-CoV-2 i possibles futures soques de coronavirus. A més a més, aquesta proteïna no està present en els humans, cosa que fa que sigui una diana relativament segura.

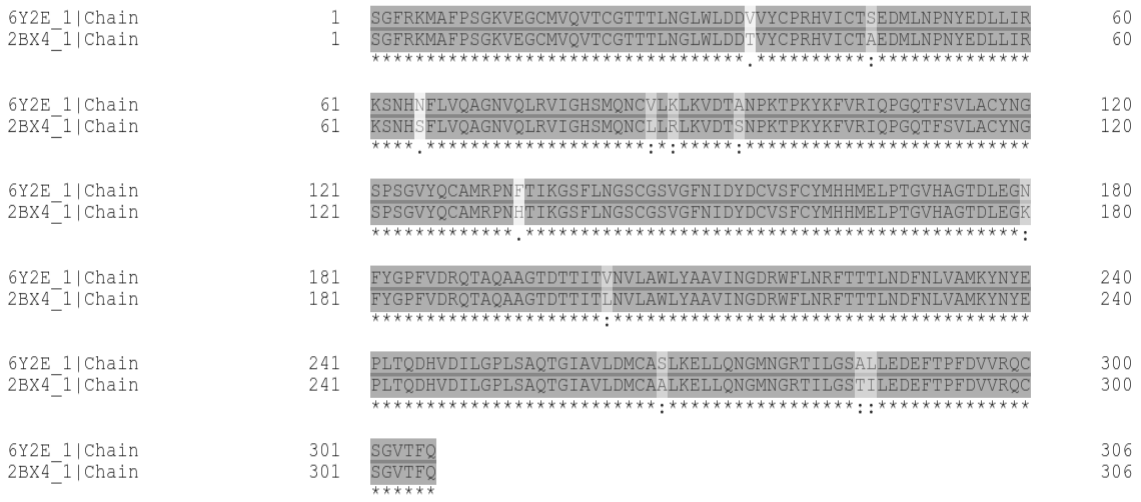


Figura 3. Alineament de les seqüencies d'aminoàcids de les proteases del SARS-CoV (PDB: 2BX4) i SARS-CoV-2 (PDB: 6Y2E). Figura obtinguda utilitzant uniprot.

La M^{pro} és un homodímer amb dues subunitats idèntiques. Cada subunitat té una llargada de 306 residus i està formada per tres dominis. El domini I està format pels residus del 8 al 100, el domini II està format pels residus del 102 al 184 i el domini III està format pels residus del 201 a 301. Els dominis I i II presenten el mateix plegament, amb una estructura de sis tires làmina-β antiparal·leles, mentre que el domini III està format per cinc α-hèlix ordenades en un gran clúster globular antiparal·lel. Els dominis II i III estan connectats per un llaç que està format pels residus 185 a 198 (13). El lloc d'unió al substrat de la M^{pro} es troba en un espai entre els dominis I i II. El domini III s'encarrega de regular la dimerització a través d'un pont salí entre el residu Glu290 d'una subunitat i el residu Arg4 de l'altre subunitat. La formació del dímer és necessària per a que la proteïna sigui activa, ja que el residu N-terminal d'una subunitat interactua amb el residu Glu166 de l'altre, formant la butxaca S_1 del lloc d'unió al substrat (17).

La zona activa té una díada catalítica formada pels residus His41 i Cys145 i les butxaques S_1, S_2, S_3 i S_1', que estan formades pels següents residus: la butxaca S_1 està formada pels residus Phe140, Leu141, Asn142, His163, Glu166 i el residu Ser1 de l'altre protòmer; la butxaca S_2 està formada pels residus Met49, Tyr54, His164, Asp187 i Arg188; la butxaca S_3 està formada pels residus Met165, Leu167, Gln189, Thr190 i Gln192; la butxaca S_1' està formada pels residus His41, Gly143, Ser144 i Cys145 (18). La Figura 4 mostra l'estructura de la M^{pro} i diferencia els dominis i les butxaques d'aquesta per colors.

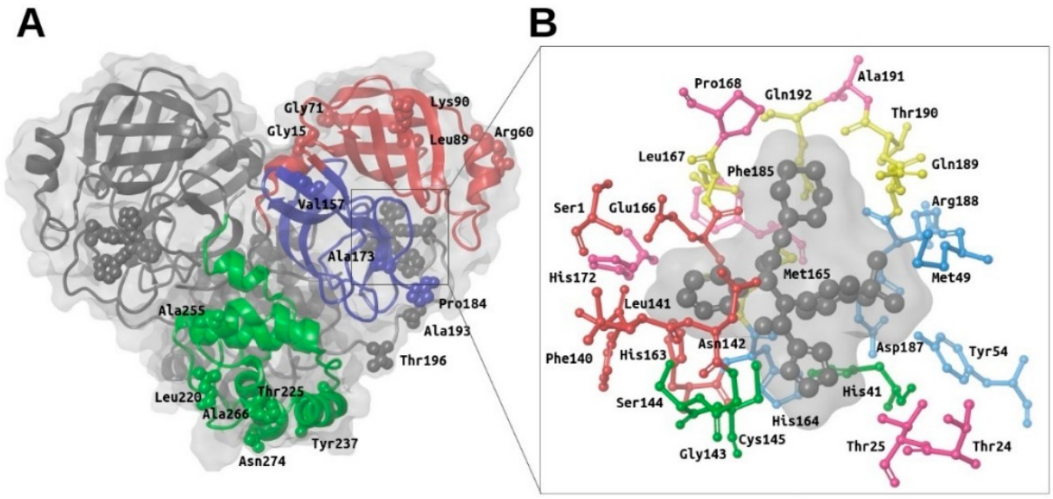


Figura 4. La imatge A mostra l'estructura general de la M^{pro} i la seva estructura homodimèrica. Es representa el domini I de color vermell, el domini II de color blau i el domini III de color Verd. La imatge B mostra els residus més importants dels diferents subespais, on es representen els d'S1 de vermell, els d'S2 de blau, els d'S3 de groc, els d'S1´ de verd i els demès residus importants de rosa. Figura extreta de Gimeno et al., 2020 (17).

La unió d'un lligand no covalent a la M^{pro} provoca els següents canvis al lloc d'unió: els residus Met49/Arg188 (butxaca S_2) i Met165/Gln189 (Butxaca S) pateixen canvis conformacionals a les seves cadenes laterals; tots els residus de la butxaca S_3, el residu Pro168 i Ala191 tenen la seva cadena principal desplaçada lleugerament; els residus Ser1 i Asn142 veuen petits canvis al final de les seves cadenes laterals (17,19). Les tres principals regions del lloc d'unió de la M^{pro} a partir de l'anàlisi de complexos proteïna-lligand cristal·litzats (lligands 13^a, 13b, N3 i X77) són els següents: una butxaca hidrofòbica formada per les butxaques S_3 i S_2, on el lligand s'uneix mitjançant interaccions hidrofòbiques amb els residus Met165, Gln189, Met49, Asp187 i His41; La butxaca S_1, on els àtoms d'hidrogen i nitrogen de la cadena principal del residu Glu166 formen enllaços pont d'hidrogen amb tots els lligands observats; la butxaca S_1', on el lligand s'uneix mitjançant interaccions covalents i no-covalents amb els residus Cys145 i His41 de la díada catalítica i es produeix un enllaç pont d'hidrogen amb el nitrogen de la cadena principal del residu Gly143 (17).

1.2. *Activity cliffs*

Els *activity cliffs* són parelles de compostos o grups de compostos que presenten estructures molt similars i que són actius contra la mateixa diana, però que tenen una diferència en la seva activitat vers aquesta diana molt alta. Així doncs, els *activity cliffs*

capturen modificacions químiques que influeixen en l'activitat biològica, pel que són de gran utilitat i interès en l'anàlisi de la relació estructura-activitat dels compostos (20). Per a poder trobar *activity cliffs* s'ha de poder descriure i representar la similitud entre compostos d'alguna manera. Una de les tècniques més utilitzada al llarg dels anys per a representar la similitud entre compostos ha estat la utilització de fingerprints moleculars junt amb el càlcul de l'índex de Tanimoto entre diferents molècules (20). A la Figura 5 es mostra un exemple d'*activity cliff* (21).

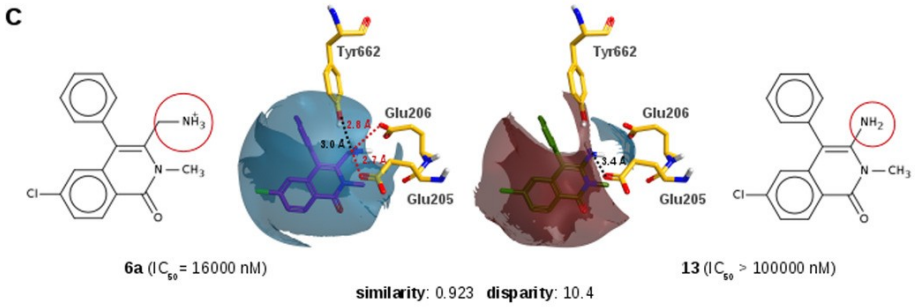


Figura 5. Exemple d'activity cliff. Les dues molècules tenen un índex de similitud de 0.923 i una disparitat de 10,4 entre elles, que es va calcular amb la següent formula: Disparitat=Δ activitat / (1-similitud). Figura extreta de Ojeda-Montes et al., 2018 (21).

1.3. Fingerprints moleculars

Els fingerprints moleculars són una manera de mapejar l'estructura d'una molècula a un vector (veure Figures 6 i 7), on els components d'aquest seran bits (0 o 1), per a que es pugui dur a terme una comparació amb vectors de bits d'altres molècules. La comparació entre vectors de bits de diferents molècules ha de poder ser expressada d'una manera que permeti quantificar la similitud entre els dos vectors. Hi han moltes maneres de trobar la similitud entre dos vectors, però el paràmetre més utilitzat per a fer-ho quan s'utilitzen fingerprints moleculars és el coeficient de Tanimoto (22), que es calcula mitjançant la següent formula:

$$S(A,B)=\frac{c}{a+b-c}$$

On, donats els fingerprints moleculars de dos compostos A i B amb similitud S, a és igual al nombre de bits amb valor 1 del vector A, b és igual al nombre de bits amb valor 1 del vector B i c és igual al nombre de bits que han estat 1 en ambdós vectors.

Hi han diferent tipus de fingerprints segons la manera que es té d'obtenir les cadenes de bits a partir de la representació de l'estructura de la molècula. Els principals tipus de fingerprints que existeixen són els següents (22):

- Fingerprints de subestructures o fingerprints estructurals. Aquests fingerprints activen un bit del vector de bits depenent de si hi ha presència d'una subestructura o característica a la molècula o no. Aquest tipus de fingerprints són útils quan sabem que la majoria de les molècules que analitzem presentaran alguna de les subestructures o característiques que codifiquen el fingerprint en qüestió. Cada subestructura que es pugui trobar amb el fingerprint tindrà un bit únic. Els fingerprints MACCS són fingerprints estructurals, i normalment tenen el seu nom conté el número de subestructures que poden trobar, tenint el fingerprint MACCS166 166 claus de subestructures identificables. A la figura 6 es mostra un exemple del funcionament d'un fingerprint estructural.

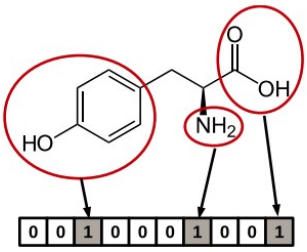


Figura 6. Exemple d'assignació de bits a un fingerprint estructural de 10 bits. Figura extreta de Cereto-Massagué et al., 2015 (22).

- Fingerprints topològics. Aquests fingerprints analitzen tots els fragments d'una molècula utilitzant camins lineals de diferents número d'enllaços, i apliquen una funció de hash a cadascun d'aquests camins. La funció de hash transforma els camins en cadenes de caràcters d'una llargada determinada i mapeja aquesta cadena a un element del vector. L'utilització d'una funció de hash fa que un bit pugui ser activat per més d'un camí, pel que es poden causar col·lisions de bits, que es tradueixen en una pèrdua d'informació. Com més llarg sigui el vector de bits que descrigui el fingerprint menys col·lisions hi hauran i més característiques diferents es podran mapejar. Aquests fingerprints s'utilitzen majoritàriament per a fer filtratges ràpids de subestructures. A la figura 7 es mostra un exemple de com s'assignen els bits en un fingerprint topològic. En aquesta s'observen els diferents camins generats, a partir dels quals es crea el vector de bits.

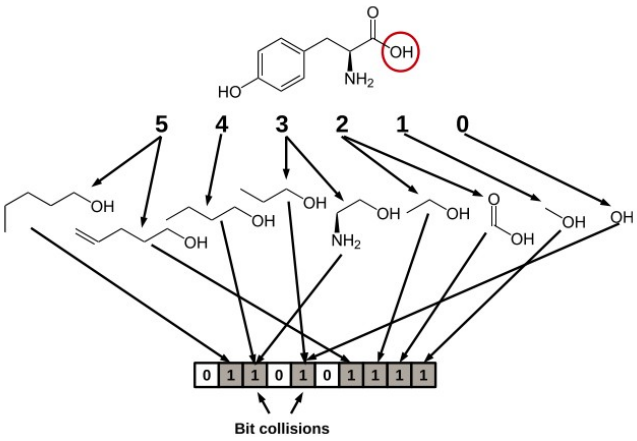


Figura 7. Exemple de l'assignació de bits en un fingerprint topològic de 10 bits i camins de fins a 5 passos. Figura extreta de Cereto-Massagué et al., 2015 (22).

- Fingerprints circulars. Aquests fingerprints es diferencien dels fingerprints topològics en què en comptes de crear camins lineals de la molècula s'analitzen els entorns de cada àtom. Els entorns venen delimitats per un radi. Aquests fingerprints s'utilitzen majoritàriament per a fer cerques de similitud d'estructures completes. El fingerprint ECFP o Extended-Conectivity Fingerprints són els fingerprints circulars més utilitzats, i el fingerprint Morgan d'RDKit és una variació d'aquest fingerprint. A la Figura 8 es poden observar els recorreguts que es fan en un fingerprint circular de radi 2. En total es fan tres iteracions per cada àtom que no sigui un hidrogen de la molècula, una amb radi 0, una amb radi 1 i l'última amb radi 2 (23). Per cada iteració de cada àtom es crea una clau de hash i s'assignen bits com a la Figura 7.

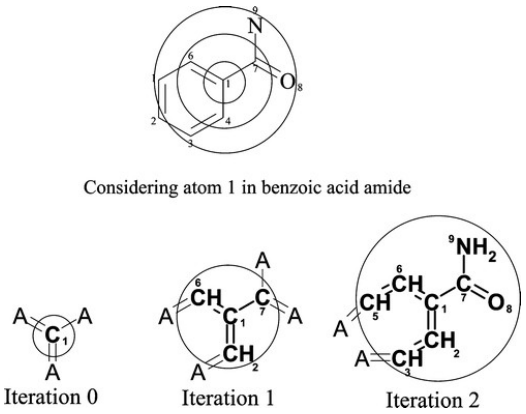


Figura 8. Exemple dels camins recorreguts en un fingerprint circular per cada iteració. Figura extreta de Rogers & Hahn, 2010 (23).

1.4. Clústering jeràrquic aglomeratiu

El clústering jeràrquic aglomeratiu (Hierarchal aglomerative clustering o HAC) és un mètode d'agrupació d'observacions en grups basada en la distància que hi ha entre els diferents grups. L'algorisme genèric de HAC es coneix com algorisme de Lance-Williams, i aquest permet utilitzar el mateix algorisme per a dur a terme el HAC amb diferents diferents mètodes de càlcul de distancia (24). Suposant que dos clústers Ci i Cj pròxims que s'han unit formant un clúster Cij en un pas previ i un altre clúster Ck, es defineixen les següents distàncies:

- d(ij)k és la distancia entre el clúster Cij i Ck
- dij, dik i djkz són les distancies entre els clústers Ci-Cj, Ci-Ck i Cj-Ck respectivament.

Amb les que es pot escriure la formula de Lance-Williams:

$$d_{(ij)k} = \alpha_i d_{hi} + \alpha_j d_{hj} + \beta d_{ij} + \gamma |d_{ik} - d_{ij}|$$

On αi, αj, β i γ són paràmetres que depenen del mètode de càlcul de distància utilitzat.
L'estratègia per a dur a terme un HAC és, primer de tot, considerar tots els elements com a un clúster. A partir dels clústers inicials es duen a terme iteracions recursives, on a cadascuna d'aquestes es van actualitzant les distàncies entre clústers i formant-ne de nous aplicant la formula de Lance-Williams, fins que finalment s'obté un únic clúster i s'acaba l'algorisme amb tots els elements classificats en diferents clústers.

1.5. Format SDF i SMILES

Els arxius SDF (Structure-Data File) i SMILES (Simplified Molecular Input Line Specification) són arxius que contenen informació estructural d'una o més molècules.

Un arxiu SDF conté informació estructural i no estructural de les molècules que el formen. La informació d'una molècula està definida per les parts que es mostren a la Figura 9. L'encapçalament conté la identificació de la molècula i altre informació diversa d'aquesta. La taula de connexions conté tota la informació de l'estructura de la molècula. La primera línia de la taula de connexions conté informació del nombre d'àtoms del compost, nombre d'enllaços d'aquest, una serie de números binaris que indiquen diferents característiques i per últim la versió de la taula de connexions. A continuació es troba el bloc d'àtoms, que especifica els símbols atòmics, massa, carrega, estereoquímica i el número d'hidrògens associats de cada àtom. L'última part de la taula de connexions descriu tots els enllaços

de la molècula. La segona part de la descripció de la molècula conté informació no estructural d'aquesta com pot ser el punt de fusió d'aquesta. La separació entre molècules es defineix amb amb quatre «$» (25).

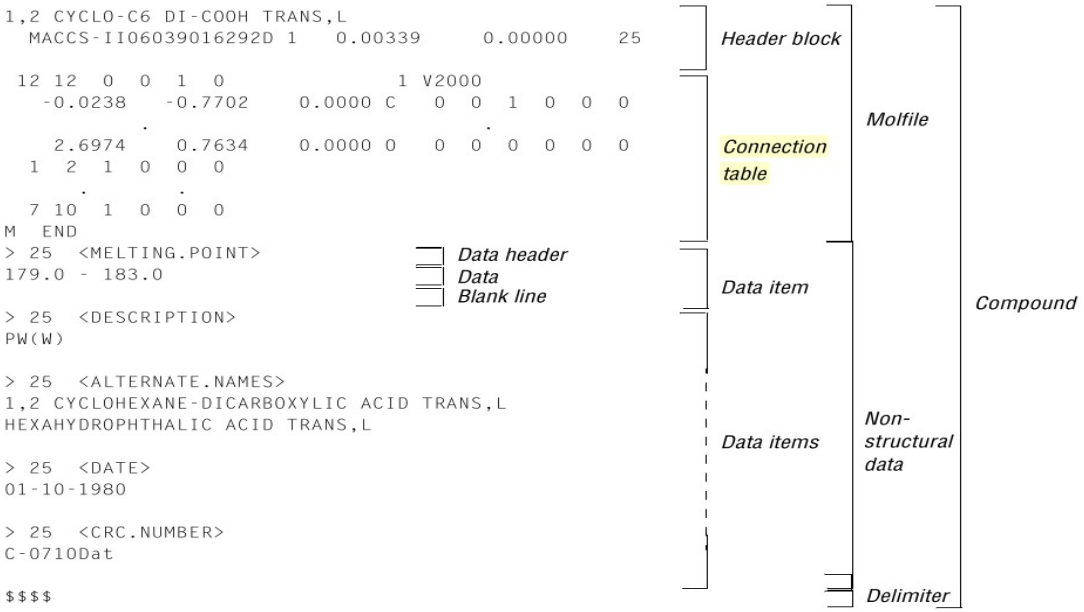


Figura 9. Estructura de la descripció d'una molècula en un arxiu SDF. Figura extreta de Mauthe & Thomas, 2005 (25).

El format SMILES descriu molècules utilitzant una sèrie de caràcters que acaba en un espai. Aquesta representació permet representar els àtoms, enllaços, les ramificacions, les estructures cícliques i l'aromaticitat d'una molècula (26). A la Figura 10 es mostra una molècula en format SMILES i la seva representació.

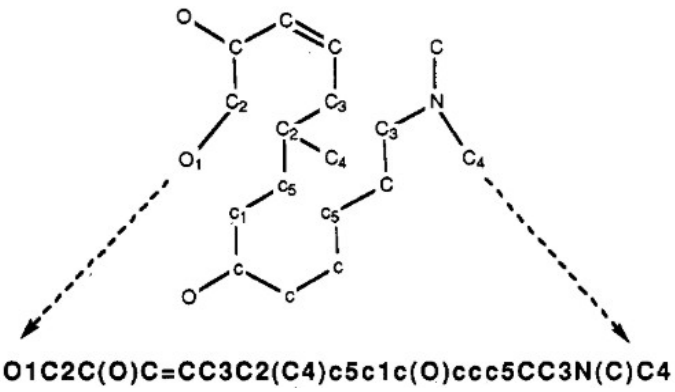


Figure 10. Imatge que mostra la representació d'una molècula a partir de l'SMILES d'aquesta. Figura extreta de Weininger D., 1988 (26).

2. Hipòtesi I objectius

Petits canvis en l'estructura d'un compost poden provocar grans canvis en la seva activitat.

L'objectiu principal és trobar *activity cliffs* entre compostos que s'uneixen a la M^{pro} del SARS-CoV-2. Específicament es volen trobar *activity cliffs* entre un conjunt de compostos que s'uneixen a la proteasa M^{pro} del SARS-CoV-2 i l'inhibeixen i un conjunt de compostos que no l'inhibeixen.

L'objectiu secundari és el disseny d'una eina per a la identificació d'*activity cliffs* que permeti els següents punts:

- Representar els compostos de dos arxius, un amb molècules que inhibeixen la M^{pro} i l'altre amb molècules que no, segons la seva similitud estructural calculada mitjançant l'ús de fingerprints.
- Diferenciar els compostos d'un arxiu dels compostos de l'altre.
- Visualitzar les estructures 2D de tots els compostos.

3. Metodologia

3.1. Disseny de l'aplicació

El primer pas va ser determinar les necessitats d'usabilitat de l'aplicació, que es llisten a continuació:

- S'han de poder pujar un o dos arxius de molècules en format SDF o SMILES.
- S'ha de poder seleccionar un fingerprint d'entre un llistat de disponibles.
- S'ha de crear un dendrograma o arbre filogenètic on es representin les molècules distribuïdes segons la seva similitud.
- S'ha de poder fer zoom i canviar la mida del dendrograma o arbre.
- S'han de poder crear clústers i canviar el color del dendrograma o arbre filogenètic per a diferenciar-los.
- S'ha de poder identificar quines molècules pertanyen a cada arxiu.
- S'ha de poder descarregar una imatge del dendrograma o arbre.

L'aplicació va ser dissenyada com una aplicació web basada en Python 2.7 i està implementada utilitzant un microframework de Python anomenat Flask. Flask és un microframework perquè el seu nucli és molt lleuger, es a dir que té poques funcionalitats per defecte. Aquestes funcionalitats es poden anar ampliant a mesura que el desenvolupament de l'aplicació avança i augmenten les necessitats d'aquesta. La modularitat que ofereix fa que el desenvolupament sigui més còmode i que només es tingui el que es necessita. La part de visualització de les pàgines (front-end) s'ha implementat utilitzant HTML (27), JavaScript (28), CSS (29) i jQuery (30). A mesura que s'expliquin les funcions més importants de l'aplicació s'aniran mencionant les llibreries que les implementen.

L'aplicació consta de dues pàgines: la pàgina inicial, on s'introdueixen els inputs necessaris per a crear l'arbre i la pàgina del gràfic, on es pot observar el dendrograma o arbre filogenètic i es poden aplicar les diferents opcions de visualització. A la pàgina inicial hi ha un formulari (Figura 11) que conté dos camps on es poden seleccionar arxius, un camp de selecció de fingeprints a partir d'una llista d'aquests (Figura 12), un camp de

selecció de mètode de càlcul de distàncies entre clústers per a dur a terme el clústering jeràrquic (Figura 12) i finalment un botó per a crear un arbre a partir dels arxius i paràmetres prèviament seleccionats.

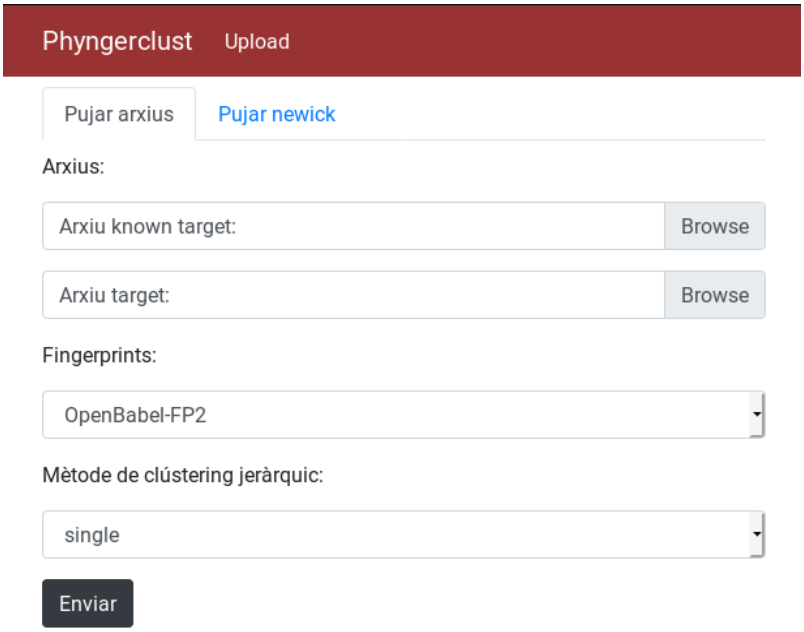


Figura 11. Formulari de la pàgina principal. Els dos primers camps accepten arxius SDF o SMILES, i es poden pujar arxius als dos camps o solament al primer. El tercer i quart camp contenen una llista de fingerpints i una llista de mètodes de clústering jeràrquics respectivament.

Les 15 opcions de fingerprints que es mostren a la Figura 12 requereixen 3 llibreries de Python diferents per a ser utilitzades. La primera de totes és ChemFP (31), una llibreria quimioinformàtica de codi obert (versió 1.5) que permet trobar la similitud de parelles de molècules a una alta velocitat i utilitzant pocs recursos de memòria. Aquesta eina pot utilitzar les funcions de fingerprints d'altres llibreries per als seus càlculs. Les altres dos llibreries són RDKIT (32) i OpenBabel (33). Aquestes llibreries també son de codi obert i tenen les seves implementacions de fingerprints pròpies que ChemFP utilitza per a efectuar el càlcul de similitud entre compostos.

Els mètodes de clústering jeràrquic que es llisten a la Figura 12 són implementats per la llibreria de codi obert SciPy (34), que utilitza la llibreria NumPy de Python per a implementar diferents algoritmes matemàtics.

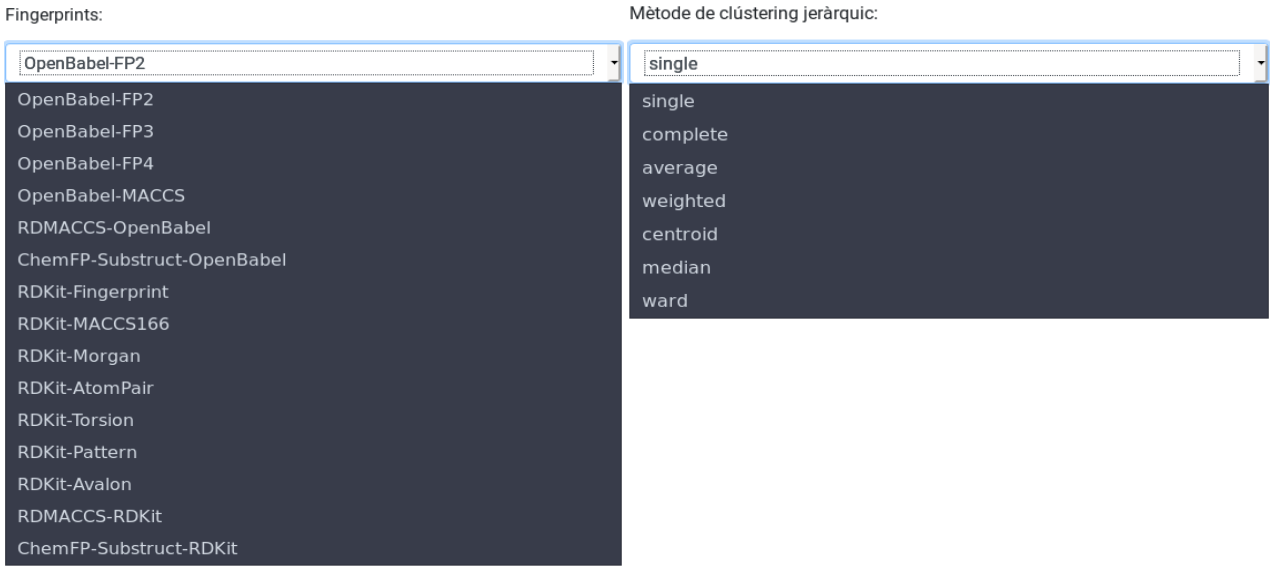


Figura 12. Llistat dels diferents fingerprints i mètodes de càlcul de distàncies per al clústering jeràrquic que pot utilitzar l'aplicació.

Quan l'usuari envia el formulari de la pàgina inicial l'aplicació duu a terme una serie d'accions per a crear un dendrograma que l'usuari pugui visualitzar. Aquestes accions es llisten a continuació:

1. Es carreguen totes les molècules dels arxius i es transformen en cadenes de bits utilitzant el fingerprint seleccionat per l'usuari utilitzant la llibreria ChemFP. Cada vector de bits generat es guarda a una llista per al següent pas.

2. Es calculen els índexos de Tanimoto de totes les parelles de compostos utilitzant dos nivells d'iteració sobre la llista obtinguda al punt 1. S'obté la distància entre els compostos de cada parella (1-similitud) i es guarden en una matriu simètrica.

3. A partir de la matriu de distància es duu a terme un clústering jeràrquic amb l'eina SciPy utilitzant el mètode que l'usuari ha seleccionat al formulari, obtenint una matriu d'enllaços.

4. Es transforma la matriu d'enllaços a un arbre en format newick amb el que finalment es crea el dendrograma o arbre filogenètic. La transformació es duu a terme en dos passos. Primer es passa la matriu d'enllaços a un objecte TreeNode de la llibreria scikit-bio utilitzant la funció from_linkage (35), que accepta com a paràmetres la matriu d'enllaços obtinguda al pas 3 i una llista amb els noms dels objectes de la matriu (s'utilitza l'índex de la matriu de distàncies). L'objecte TreeNode és una estructura de dades que representa un arbre, on cada objecte

TreeNode té una relació amb els seus nodes pares i fills. Finalment es transforma l'objecte TreeNode en un objecte Tree de la llibreria ETE3 (36), amb el que es pot obtindre un arbre en format newick fàcilment.

Un cop l'aplicació obté l'arbre en format newick, es redirigeix la pàgina per a mostrar el dendrograma o arbre filogenètic a l'usuari. L'arbre es renderitza utilitzant la llibreria de JavaScript Phylotree (37), que permet crear arbres filogenètics interactius. A la Figura 13 es poden veure les diferents opcions que es poden aplicar a l'arbre, que es mostren en una barra vertical a l'esquerra de la pàgina del gràfic. D'entre aquestes destaquen les opcions per a seleccionar si es vol representar el gràfic com un dendrograma o com un arbre filogenètic (Layout linear o radial), l'opció d'ampliar o disminuir l'amplada i l'alçada del gràfic, l'opció d'activar la vista d'imatges de les molècules (generades amb RDKit) quan es passi el ratolí per damunt del nom d'aquestes, les opcions per a crear clústers i pintar-los, les opcions per a canviar el color del nom de les molècules d'un arxiu i les opcions per a descarregar una imatge del gràfic. A la figura 17 de l'annexe 2 es pot observar la pantalla que es veu quan es crea un gràfic, sense cap tipus de modificació. El tipus de gràfic per defecte és el dendrograma. A la figura 18 de l'annexe 2 es pot observar un gràfic al que s'han creat clústers.

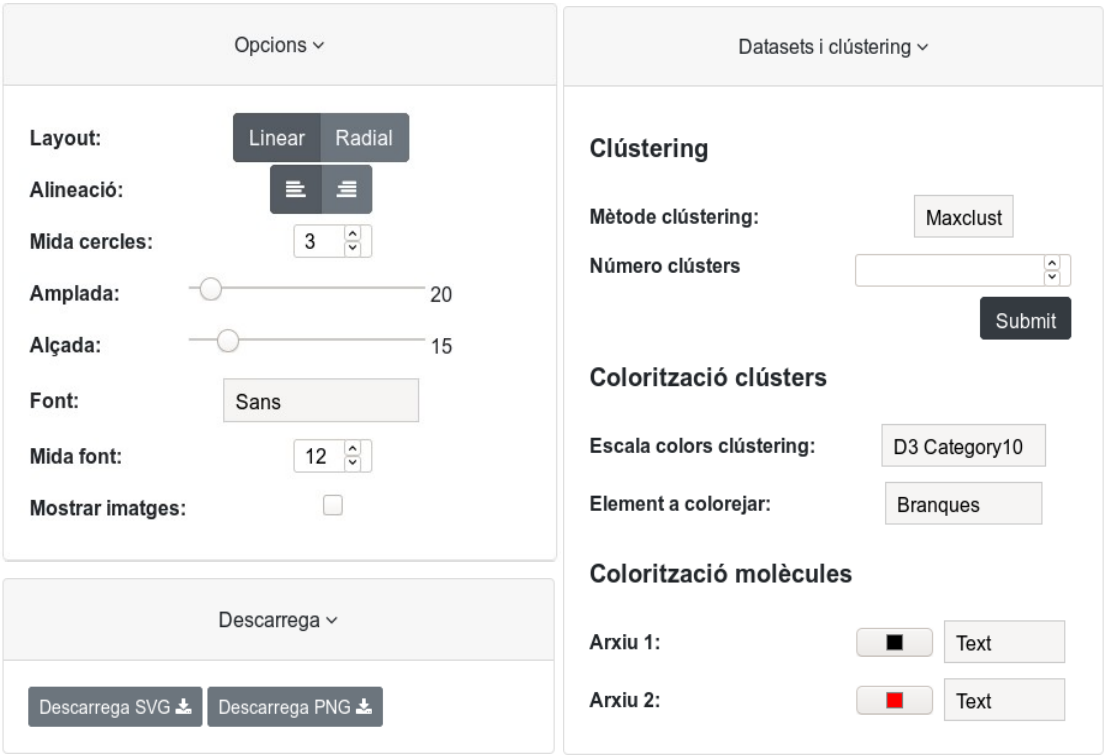


Figura 13. Figura que mostra les diferents opcions de l'arbre. Els apartats es troben en un desplegable vertical que només pot tenir un apartat visible, essent l'ordre d'aquest el següent: Opcions, Datasets i clústering i Descarrega.

3.2. Identificació d'*activity cliffs*

Per a identificar *activity cliffs* entre compostos que s'uneixen a la M^{pro} del SARS-CoV-2 es van utilitzar dos arxius proporcionats pel grup de recerca de quimioinformàtica i nutrició de l'URV.

El primer arxiu conté molècules conegudes que inhibeixen la M^{pro} del SARS-CoV-2, i el segon conté molècules conegudes que no l'inhibeixen. Aquestes molècules s'han extret de la pàgina de COVID Moonshot (38). COVID Moonshot és una iniciativa de col·laboració oberta per a la troballa d'un antiviral lliure de patents, on diferents grups d'investigació i empreses aporten propostes de compostos que podrien ser candidats a inhibidors de la M^{pro}, a partir de les quals es fan diferents fases de selecció de compostos fins a obtenir compostos amb els que es fan assajos de cribratge antiviral en cèl·lules. Les diferents propostes de compostos s'identifiquen per ID's que s'assignen seguint el següent format: "<PRIMERES 3 LLETRES DEL NOM>-<PRIMERES 3 LLETRES DE LA INSTITUCIÓ>-<cadena de 8 caràcters aleatoris>". Un exemple d'ID seria GER-UNI-caecb3b0, on la persona que ha fet la proposta és el Gerard Pujadas i la institució la Universitat Rovira i Virgili. Dins de cada proposta les molècules s'identifiquen per un número que s'afegeix al final de l'ID de la proposta (p. ex. GER-UNI-caecb3b0-1).

A partir dels dos arxius es van crear dos arbres, un utilitzant el fingerprint molecular MACCS166 d'RDKit i el mètode d'agrupament jeràrquic average i l'altre utilitzant el fingerprint molecular Morgan d'RDKit i el mètode d'agrupament jeràrquic average. Es van crear dos arbres amb diferents fingerprints per assegurar que es trobarien tots els *activity cliffs* d'entre totes les molècules, ja que els dos arbres presentaven diferents distribucions de molècules, una basada en la similitud de presència de subestructures (MACCS166) i l'altre basada en l'anàlisi dels àtoms veïns de cada àtom d'una molècula. Es va afegir l'activitat de les molècules que inhibien la M^{pro} al nom d'aquestes per a que es pogués obtindre aquesta fàcilment al observar el gràfic.

A la Figura 14 es pot observar l'arbre que s'obté utilitzant el fingerprint molecular Morgan d'RDKit. Les molècules amb el nom en negre pertanyen al grup de compostos que inhibeixen la M^{pro} i les molècules amb el nom en vermell pertanyen al grup de molècules que no l'inhibeixen. La figura 16 de l'annexe 1 conté la imatge de l'arbre que s'obté utilitzant el fingerprint molecular MACCS166 i el mètode de clústering jeràrquic average. La cerca de fingerprints s'ha dut a terme buscant molècules dels dos arxius que es

trobessin a prop l'una de l'altre. El fet de que s'utilitzessin dos colors per a diferenciar les molècules d'un arxiu de l'altre va facilitar la troballa de parelles actives-inactives. A la Figura 15 es pot observar un exemple de troballa d'un *activity cliff*. Quan es trobava una molècula d'un arxiu envoltada de molècules d'un altre arxiu s'observaven les estructures de les molècules passant el ratolí per damunt del nom d'aquestes i es decidia si les parelles que s'observaven eren *activity cliffs* o no.

Figura 14. Arbre filogenètic obtingut a partir dels dos arxius utilitzant el fingerprint molecular Morgan d'RDKit i el mètode de clústering aglomeratiu average de la llibreria SciPy.

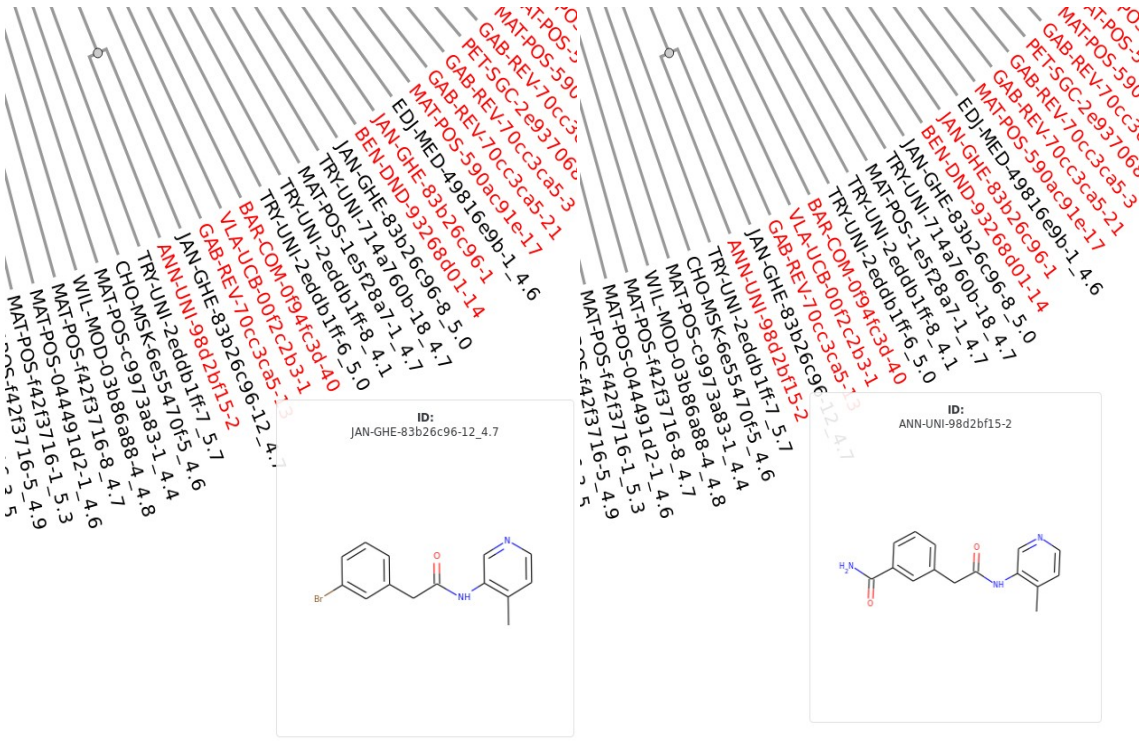
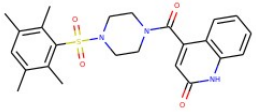
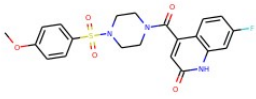
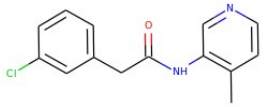
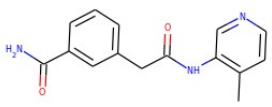
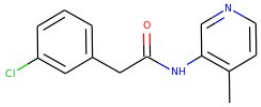
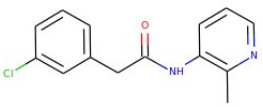
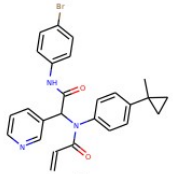
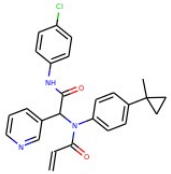
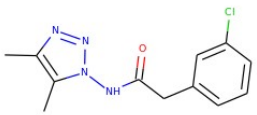
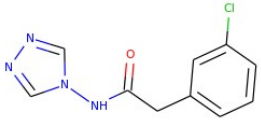
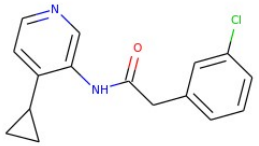
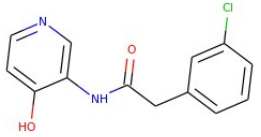
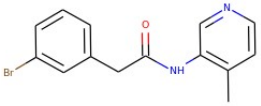
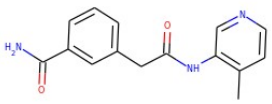
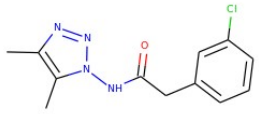
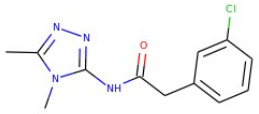
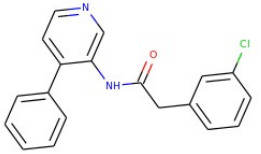
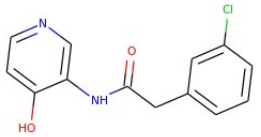


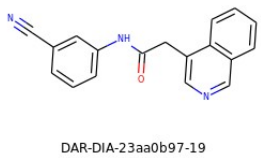
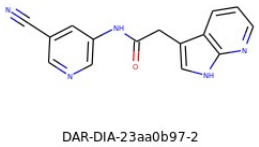
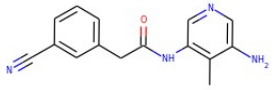
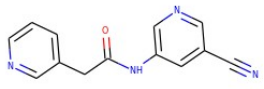
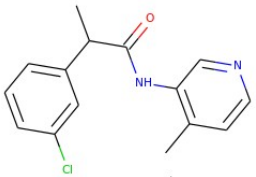
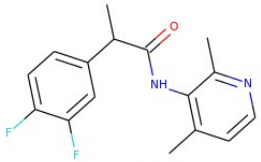
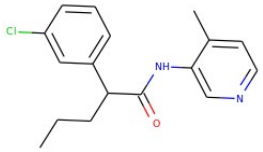
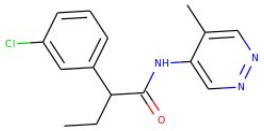
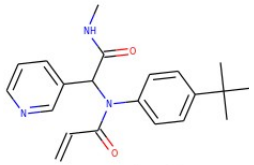
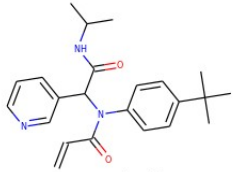
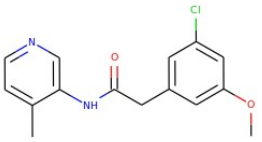
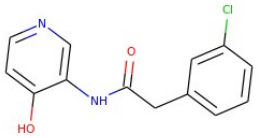
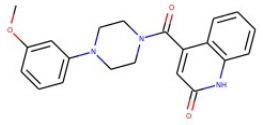
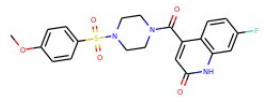
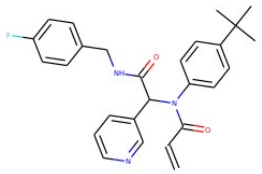
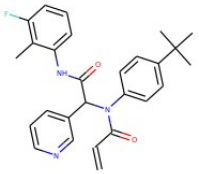
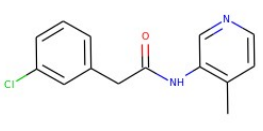
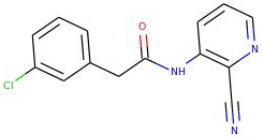
Figura 15. Troballa d'un activity cliff. *La possibilitat de moure's per el gràfic fent zoom i la observació de l'estructura de les molècules fa que la tasca sigui més fàcil d'efectuar.*

4. Resultats i discussió

Mitjançant una cerca per els dos arbres filogenètics generats es van trobar 18 *activity cliffs* entre les molècules dels dos arxius. Aquests es troben a la Taula 2, on també es mostra l'activitat en pIC_{50} de la molècula activa i els indexs de Tanimoto de les parelles obtinguts utilitzant el fingerprint de Morgan amb el radi per defecte (radi=2) i el fingerprint MACCS166. Els índexs de Tanimoto s'han buscat en les matrius de distàncies generades per l'aplicació.

L'activitat es mesura en pIC_{50}, que s'obté calculant el -log de la concentració a la que un fàrmac pot inhibir un procés biològic en un 50% (IC_{50}), obtenint un número que representa la activitat inhibitòria de manera natural, ja que contra més alt sigui el número més activitat inhibitòria tindrà una molècula.

Activa	Inactiva	Activa	Inactiva	Activa	Inactiva
 MAT-POS-916a2c5a-3	 MAT-POS-590ac91e-65	 TRY-UNI-714a760b-6	 ANN-UNI-98d2bf15-2	 TRY-UNI-714a760b-6	 EDG-MED-0da5ad92-14
pIC_{50}: 5,3	Inactiva	pIC_{50}: 4,6	Inactiva	pIC_{50}: 4,7	Inactiva
Índex de Tanimoto MORGAN: 0,5075 Índex de Tanimoto MACCS166: 0,8400		Índex de Tanimoto MORGAN: 0,7083 Índex de Tanimoto MACCS166: 0,7838		Índex de Tanimoto MORGAN: 0,6458 Índex de Tanimoto MACCS166: 1,0000	
 LON-WEI-d1c9908a-11	 LON-WEI-d1c9908a-2	 JAN-GHE-5a013bed-1	 JAN-GHE-5a013bed-5	 ALP-POS-95b75b4d-5	 ALP-POS-95b75b4d-6
pIC_{50}: 4,4	Inactiva	pIC_{50}: 4,5	Inactiva	pIC_{50}: 5,1	Inactiva
Índex de Tanimoto MORGAN: 0,8413 Índex de Tanimoto MACCS166: 0,9545		Índex de Tanimoto MORGAN: 0,5870 Índex de Tanimoto MACCS166: 0,8750		Índex de Tanimoto MORGAN: 0,6939 Índex de Tanimoto MACCS166: 0,6809	
 JAN-GHE-83b26c96-12	 ANN-UNI-98d2bf15-2	 JAN-GHE-5a013bed-1	 JAG-UCB-a3ef7265-15	 ALP-POS-0c2c77e1-1	 ALP-POS-95b75b4d-6
pIC_{50}: 4,7	Inactiva	pIC_{50}: 4,5	Inactiva	pIC_{50}: 4,5	Inactiva
Índex de Tanimoto MORGAN: 0,7083 Índex de Tanimoto MACCS166: 0,7838		Índex de Tanimoto MORGAN: 0,5386 Índex de Tanimoto MACCS166: 0,8302		Índex de Tanimoto MORGAN: 0,7083 Índex de Tanimoto MACCS166: 0,7272	

 DAR-DIA-23aa0b97-19  DAR-DIA-23aa0b97-2	 TRY-UNI-714a760b-19  DAR-DIA-23aa0b97-6	 TRY-UNI-714a760b-18  BAR-COM-0f94fc3d-46
pIC_{50}: 4,7 Inactiva	pIC_{50}: 4,6 Inactiva	pIC_{50}: 4,7 Inactiva
Índex de Tanimoto MORGAN: 0,5082 Índex de Tanimoto MACCS166: 0,7368	Índex de Tanimoto MORGAN: 0,4655 Índex de Tanimoto MACCS166: 0,8000	Índex de Tanimoto MORGAN: 0,4286 Índex de Tanimoto MACCS166: 0,9063
 JAN-GHE-83b26c96-8  JAN-GHE-83b26c96-1	 LON-WEI-adc59df6-3  LON-WEI-adc59df6-21	 MAT-POS-c9973a83-1  ALP-POS-95b75b4d-6
pIC_{50}: 5,0 Inactiva	pIC_{50}: 4,6 Inactiva	pIC_{50}: 4,4 Inactiva
Índex de Tanimoto MORGAN: 0,4333 Índex de Tanimoto MACCS166: 0,7073	Índex de Tanimoto MORGAN: 0,8148 Índex de Tanimoto MACCS166: 0,9737	Índex de Tanimoto MORGAN: 0,5091 Índex de Tanimoto MACCS166: 0,8085
 BEN-DND-7e92b6ca-12  MAT-POS-590ac91e-65	 LON-WEI-adc59df6-23  LON-WEI-d1c9908a-9	 TRY-UNI-714a760b-6  AGN-NEW-c7b24fe3-4
pIC_{50}: 4,9 Inactiva	pIC_{50}: 5,0 Inactiva	pIC_{50}: 4,7 Inactiva
Índex de Tanimoto MORGAN: 0,4925 Índex de Tanimoto MACCS166: 0,6267	Índex de Tanimoto MORGAN: 0,6479 Índex de Tanimoto MACCS166: 0,8723	Índex de Tanimoto MORGAN: 0,5577 Índex de Tanimoto MACCS166: 0,9143

Taula 2. Activity cliffs trobats utilitzant els fingerprints MACCS166 i Morgan d'RDKit. Per cada parella es mostra l'activitat dels inhibidors en pIC_{50} i els índex de Tanimoto obtinguts amb els dos fingeprtints. Les imatges de les molècules s'han obtingut utilitzant RDKit.

El primer que crida l'atenció dels resultats són les diferencies que hi han entre els índexs de Tanimoto obtinguts amb el fingerprint Morgan i els obtinguts amb el fingerprint MACCS166. Els índexs de Tanimoto obtinguts amb el fingerprint Morgan mostren valors més baixos en pràcticament totes les parelles, cosa que podia esperar-se tenint en compte com es calculen les cadenes de bits en cadascun d'aquests fingerprints i el nombre de bits que té cada fingerprint, ja que el fingerprint Morgan (circular) té un vector de 1024 bits i mapeja bits a partir de les descripcions del voltant de cada àtom de la molècula, cosa que fa que augmenti la sensibilitat i capacitat descriptiva d'aquest fingerprint, mentre que el fingerprint MACCS166 té un vector de 166 bits i només busca si una molècula té les subestructures que el fingerprint pot identificar o no. Però, en alguns casos la diferència és molt elevada, com ara en la parella TRY-UNI-714a760b-6 i EDG-MED-0da5ad92-14, on aquesta és de 0,3452 i l'únic canvi entre les molècules és la posició del nitrogen de la piridina. De fet l'índex de Tanimoto d'aquesta parella en el cas del fingeprint MACCS166 és d'1 degut a que només es produeix un canvi posicional que no canvia la subestructura que reconeix el fingerprint, pel que es genera el mateix vector de bits per a les dues molècules. Als gràfics aquestes diferencies es tradueixen en que en un dendrograma la parella de molècules es troba molt propera i en l'altre la parella està allunyada. Altres *activity cliffs* que presenten una diferència elevada entre els índexs de Tanimoto dels dos fingerprints són les parelles JAN-GHE-5a013bed-1 i JAG-UCB-a3ef7265-15 on la diferència és de 0,2916 i TRY-UNI-714a760b-18 i BAR-COM-0f94fc3d-46 on la diferència és de 0,4777. En general, si tenim en compte els índexs de Tanimoto obtinguts, el fingerprint que descriu millor la similitud entre molècules és el fingerprint MACCS166.

De totes les parelles trobades, hi han vuit que presenten un canvi en tota la molècula, i tres d'aquestes vuit presenten canvis en un sol àtom de la molècula activa, que són la parella LON-WEI-d1c9908a-11 i LON-WEI-d1c9908a-2, on un bromobenzè canvia per un clorobenzè, la parella JAN-GHE-6a013bed-1 i JAG-UCB-a3ef7265-15, on el 1, 2, 3 triazol canvia per un 1, 2, 4 triazol i la parella TRY-UNI-714a760b-6 i EDG-MED-0da5ad92-14 ja mencionada prèviament. Les altres cinc parelles presenten canvis en un grup funcional de la molècula activa per un altre grup funcional diferent.

En els *activity cliffs* restants les parelles presenten múltiples petits canvis. Es podria discutir si alguna d'aquestes parelles de molècules es podrien definir com *activity cliffs* o no, personalment penso que tot i que hi hagin tres o quatre canvis aquests són petits i

segueix sent interessant preguntar-se com aquests afecten a la bio-activitat de la molècula.

Respecte a l'aplicació, personalment crec que aquesta té diferents punts que la fan una eina interessant:

- Aquesta ofereix molts paràmetres per a dur a terme el procés de creació de l'arbre. Implementa pràcticament tots els fingerprints que la llibreria ChemFP de codi obert suporta, i a més a més implementa diferents mètodes de càlcul de distàncies per al clústering jeràrquic. Els únics fingerprints de ChemFP que l'aplicació no pot implementar són els de la llibreria OpenEye, ja que aquesta requereix una llicencia, però en cas de tindre-la només s'hauria d'afegir una línia de codi per a que l'aplicació els pugues implementar.
- Aquesta ofereix una visualització interactiva. La possibilitat de moure l'arbre, ampliar-ne l'amplada o alçada, fer-hi zoom i canviar la forma d'un dendrograma a un arbre filogenètic i viceversa fa que la navegació pel gràfic estigui a mans de l'usuari i sigui el més flexible i còmoda possible. Hi han altres opcions de personalització com ara la coloració de les branques de l'arbre per clústers i els canvis en la mida de la font dels nodes i de la font mateixa d'aquests.
- Aquesta té l'opció de veure l'estructura de les molècules mentre es navega l'arbre. Aquest és un dels punts que crec que fan que l'aplicació tingui un salt de qualitat. Tot i que sigui una cosa que sembla secundaria, aquesta funció ha estat essencial per a trobar *activity cliffs*, i trobo que també podria ser útil en altres escenaris. A l'hora d'implementar aquesta funció no li vaig donar la importància que actualment crec que té, i vaig estar a punt de deixar-la estar ja que em va costar donar amb la forma d'implementació correcta.

L'aplicació ha permès trobar 9 *activity cliffs* entre molècules de propostes de diferents entitats de COVID moonshot, com ara les parelles JAN-GHE-6a013bed-1 i JAG-UCB-a3ef7265-15, TRY-UNI-714a760b-6 i EDG-MED-0da5ad92-14 i TRY-UNI-714a760b-18 i BAR-COM-0f94fc3d-46. Aquest es un altre punt fort de l'aplicació, ja que trobar *activity cliffs* entre molècules d'una mateixa proposta de molècules de COVID moonshot era una cosa que es podia esperar, però les troballes de parelles entre molècules de propostes diferents són les interessants, per que s'aconsegueixen relacionar molècules de propostes de diferents grups que d'altre manera no s'haguessin pogut relacionar tan fàcilment.

En definitiva, crec que l'aplicació ha estat realment útil, i tot i que hi han altres eines disponibles que permeten fer algunes de les funcions que aquesta ofereix, per exemple hi ha moltes eines que permeten la creació de dendrogrames, el fet de tindre les funcions de creació d'un arbre a partir d'un o més arxius i la visualització interactiva de l'arbre creat en una sola aplicació fa que aquesta sigui una eina innovadora.

Tot i això, l'aplicació no està acabada al 100%, ja que encara falta dur a terme un deployment d'aquesta al núvol per a que sigui accessible públicament. Per a fer el deployment encara s'ha d'implementar un sistema d'usuaris i una base de dades per a que es puguin guardar els arxius que els usuaris pugen i els arbres que creen.

6. Conclusions

S'ha creat una aplicació que permet la creació de dendrogrames que representin la similitud de les molècules de dos arxius. Alhora, aquesta aplicació permet visualitzar quines molècules pertanyen a cada arxiu i l'estructura 2D de cada molècula, cosa que ha permès que es trobin 18 *activity cliffs* entre un arxiu de compostos que contenia molècules que inhibien la M^{pro} del SARS-CoV-2 i un arxiu que contenia molècules que no l'inhibien. Els activity cliffs que s'han trobats podrien permetre entendre millor la relació entre l'estructura i l'activitat dels inhibidors de la M^{pro} i ser d'utilitat en el disseny de nous inhibidors més actius. Tot i això, l'aplicació pot tindre altres utilitats a banda de trobar *activity cliffs*, ja que aquesta es pot utilitzar per exemple en un procés de cribratge virtual.

El pròxim pas a dur a terme és aconseguir que l'aplicació sigui accessible públicament per a que el grup de Quimioinformàtica i nutrició i altres grups la puguin fer servir.

7. Bibliografia

1. Cui J, Li F, Shi ZL. Origin and evolution of pathogenic coronaviruses. Nat Rev Microbiol [Internet]. 2019;17(3):181–92. Available from: http://dx.doi.org/10.1038/s41579-018-0118-9
2. Su S, Wong G, Shi W, Liu J, Lai ACK, Zhou J, et al. Epidemiology, Genetic Recombination, and Pathogenesis of Coronaviruses. Trends Microbiol. 2016;24(6):490–502.
3. Frequently Asked Questions About SARS @ www.cdc.gov [Internet]. [cited 2020 Oct 27]. Available from: http://www.cdc.gov/nchhstp/socialdeterminants/faq.html
4. middle-east-respiratory-syndrome-coronavirus-(mers-cov) @ www.who.int [Internet]. [cited 2020 Oct 27]. Available from: https://www.who.int/es/news-room/fact-sheets/detail/middle-east-respiratory-syndrome-coronavirus-(mers-cov)
5. COVID-19 Map - Johns Hopkins Coronavirus Resource Center [Internet]. [cited 2020 Oct 24]. Available from: https://coronavirus.jhu.edu/map.html
6. WHO. Coronavirus @ Www.Who.Int [Internet]. Who. 2020 [cited 2020 Oct 27]. Available from: https://www.who.int/health-topics/coronavirus
7. Wu F, Zhao S, Yu B, Chen YM, Wang W, Song ZG, et al. A new coronavirus associated with human respiratory disease in China. Nature. 2020;579(7798):265–9.
8. Wu A, Peng Y, Huang B, Ding X, Wang X, Niu P, et al. Genome Composition and Divergence of the Novel Coronavirus (2019-nCoV) Originating in China. Cell Host Microbe [Internet]. 2020;27(3):325–8. Available from: https://doi.org/10.1016/j.chom.2020.02.001
9. Huang Y, Yang C, Xu X feng, Xu W, Liu S wen. Structural and functional properties of SARS-CoV-2 spike protein: potential antivirus drug development for COVID-19. Acta Pharmacol Sin [Internet]. 2020;41(9):1141–9. Available from: http://dx.doi.org/10.1038/s41401-020-0485-4
10. Astuti I, Ysrafil. Severe Acute Respiratory Syndrome Coronavirus 2 (SARS-CoV-2): An overview of viral structure and host response. Diabetes Metab Syndr Clin Res Rev. 2020;14(4):407–12.
11. V'kovski P, Kratzel A, Steiner S, Stalder H, Thiel V. Coronavirus biology and replication: implications for SARS-CoV-2. Nat Rev Microbiol [Internet]. 2020; Available from: http://www.ncbi.nlm.nih.gov/pubmed/33116300%0Ahttp://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC7592455

12. Bzówka M, Mitusińska K, Raczyńska A, Samol A, Tuszyński JA, Góra A. Structural and evolutionary analysis indicate that the sars-COV-2 mpro is a challenging target for small-molecule inhibitor design. Int J Mol Sci. 2020;21(9).

13. Jin Z, Du X, Xu Y, Deng Y, Liu M, Zhao Y, et al. Structure of Mpro from SARS-CoV-2 and discovery of its inhibitors. Nature [Internet]. 2020;582(7811):289–93. Available from: http://dx.doi.org/10.1038/s41586-020-2223-y

14. Romano M, Ruggiero A, Squeglia F, Maga G, Berisio R. A Structural View of SARS-CoV-2 RNA Replication Machinery: RNA Synthesis, Proofreading and Final Capping. Cells. 2020;9(5).

15. Mirza MU, Froeyen M. Structural elucidation of SARS-CoV-2 vital proteins: Computational methods reveal potential drug candidates against main protease, Nsp12 polymerase and Nsp13 helicase. J Pharm Anal. 2020;10(4):320–8.

16. Song J. 2019-nCoV 3C-Like Protease carries an activity-enhancing T285 /A variation which may contribute to its high infectivity. 2020;

17. Gimeno A, Mestres-Truyol J, Ojeda-Montes MJ, Macip G, Saldivar-Espinoza B, Cereto-Massagué A, et al. Prediction of novel inhibitors of the main protease (M-pro) of SARS-CoV-2 through consensus docking and drug reposition. Int J Mol Sci. 2020;21(11).

18. Tang B, He F, Liu D, Fang M, Wu Z, Xu D. AI-aided design of novel targeted covalent inhibitors against SARS-CoV-2. bioRxiv Prepr Serv Biol. 2020;

19. Kneller DW, Phillips G, O'Neill HM, Jedrzejczak R, Stols L, Langan P, et al. Structural plasticity of SARS-CoV-2 3CL Mpro active site cavity revealed by room temperature X-ray crystallography. Nat Commun [Internet]. 2020;11(1):7–12. Available from: http://dx.doi.org/10.1038/s41467-020-16954-7

20. Stumpfe D, Hu H, Bajorath J. Evolving Concept of Activity Cliffs. ACS Omega. 2019;

21. Ojeda-Montes MJ, Gimeno A, Tomas-Hernández S, Cereto-Massagué A, Beltrán-Debón R, Valls C, et al. Activity and selectivity cliffs for DPP-IV inhibitors: Lessons we can learn from SAR studies and their application to virtual screening. Med Res Rev. 2018;38(6):1874–915.

22. Cereto-Massagué A, Ojeda MJ, Valls C, Mulero M, Garcia-Vallvé S, Pujadas G. Molecular fingerprint similarity search in virtual screening. Methods. 2015;71(C):58–63.

23. Rogers D, Hahn M. Extended-connectivity fingerprints. J Chem Inf Model. 2010;50(5):742–54.

24. Lance GN. A general theory of classificatory sorting strategies: II. Clustering systems. Comput J. 1967;10(3):271–7.

25. Mauthe A, Thomas P. File Formats. Prof Content Manag Syst. 2005;(June):121–32.

26. Weininger D. SMILES, a Chemical Language and Information System: 1: Introduction to Methodology and Encoding Rules. J Chem Inf Comput Sci. 1988;28(1):31–6.

27. HTML @ developer.mozilla.org [Internet]. [cited 2020 Nov 5]. Available from: https://developer.mozilla.org/en-US/docs/Glossary/HTML

28. JavaScript @ developer.mozilla.org [Internet]. [cited 2020 Nov 5]. Available from: https://developer.mozilla.org/es/docs/Web/JavaScript

29. CSS @ developer.mozilla.org [Internet]. [cited 2020 Nov 5]. Available from: https://developer.mozilla.org/es/docs/Web/CSS

30. JQuery. Index @ Jquery.Com [Internet]. 2018 [cited 2020 Nov 5]. Available from: https://jquery.com/

31. chemfp 1.5 documentation @ chemfp.readthedocs.io [Internet]. [cited 2020 Nov 5]. Available from: https://chemfp.readthedocs.io/en/chemfp-1.5/

32. The RDKit Documentation @ rdkit.readthedocs.io [Internet]. [cited 2020 Nov 5]. Available from: https://rdkit.readthedocs.io/en/latest/

33. Main_Page @ openbabel.org [Internet]. [cited 2020 Nov 5]. Available from: http://openbabel.org/wiki/Main_Page

34. cluster @ docs.scipy.org [Internet]. [cited 2020 Nov 5]. Available from: https://docs.scipy.org/doc/scipy/reference/cluster.hierarchy.html

35. skbio @ scikit-bio.org [Internet]. [cited 2020 Nov 5]. Available from: http://scikit-bio.org/docs/0.5.6/generated/skbio.tree.TreeNode.from_linkage_matrix.html#skbio.tree.TreeNode.from_linkage_matrix

36. reference_tree @ etetoolkit.org [Internet]. [cited 2020 Nov 5]. Available from: http://etetoolkit.org/docs/latest/reference/reference_tree.html?highlight=from_skbio#master-tree-class

37. Phylotree.js documentation @ phylotree.hyphy.org [Internet]. [cited 2020 Nov 5]. Available from: http://phylotree.hyphy.org/documentation/

38. COVID Moonshot @ covid.postera.ai [Internet]. [cited 2020 Nov 6]. Available from: https://covid.postera.ai/covid

8. Autoavaluació

En aquest treball he hagut d'utilitzar coneixements apresos tant al grau de Biotecnologia com en el grau d'enginyeria informàtica. L'aplicació ha estat la meva primera experiència en el desenvolupament web, cosa que m'ha descobert un mon de l'enginyeria informàtica que es molt vast i que, sobretot, es pot utilitzar per a crear eines biotecnològiques, com és el cas de l'aplicació. En el desenvolupament d'aquesta també he agut d'adquirir coneixements de quimioinformàtica per a saber què havia de fer en cada moment.

Ha estat molt satisfactori poder utilitzar l'aplicació amb el fi específic de trobar *activity cliffs* pel fet de que he pogut veure de primera mà que les funcions d'aquesta han estat útils, ja que m'ha donat la sensació de que tot el temps i esforç que he invertit en desenvolupar-la han valgut la pena.

En resum crec que el projecte ha estat una bona unió entre els graus de Biotecnologia i Enginyeria informàtica i estic segur de que els coneixements que he adquirit duent-lo a terme em seran molt útils en un futur.

Annexe 1. Arbre fingerprint MACCS166

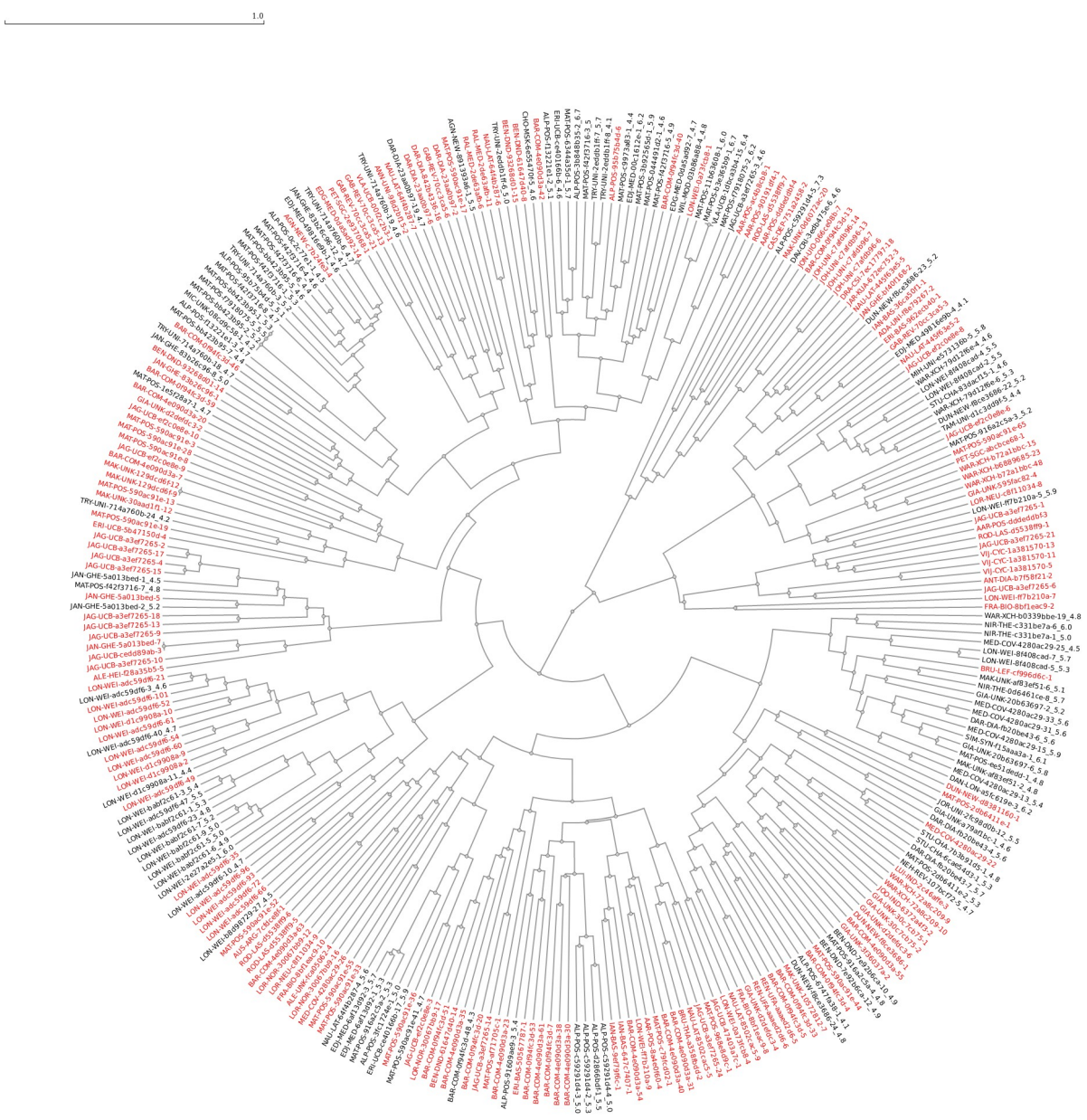


Figura 16. Figura de l'arbre filogenètic obtingut utilitzant el fingerprint MACC166 i el mètode de càlcul de distàncies del clústering aglomeratiu average

Annexe 2. Pantalles i funcionalitats de l'aplicació

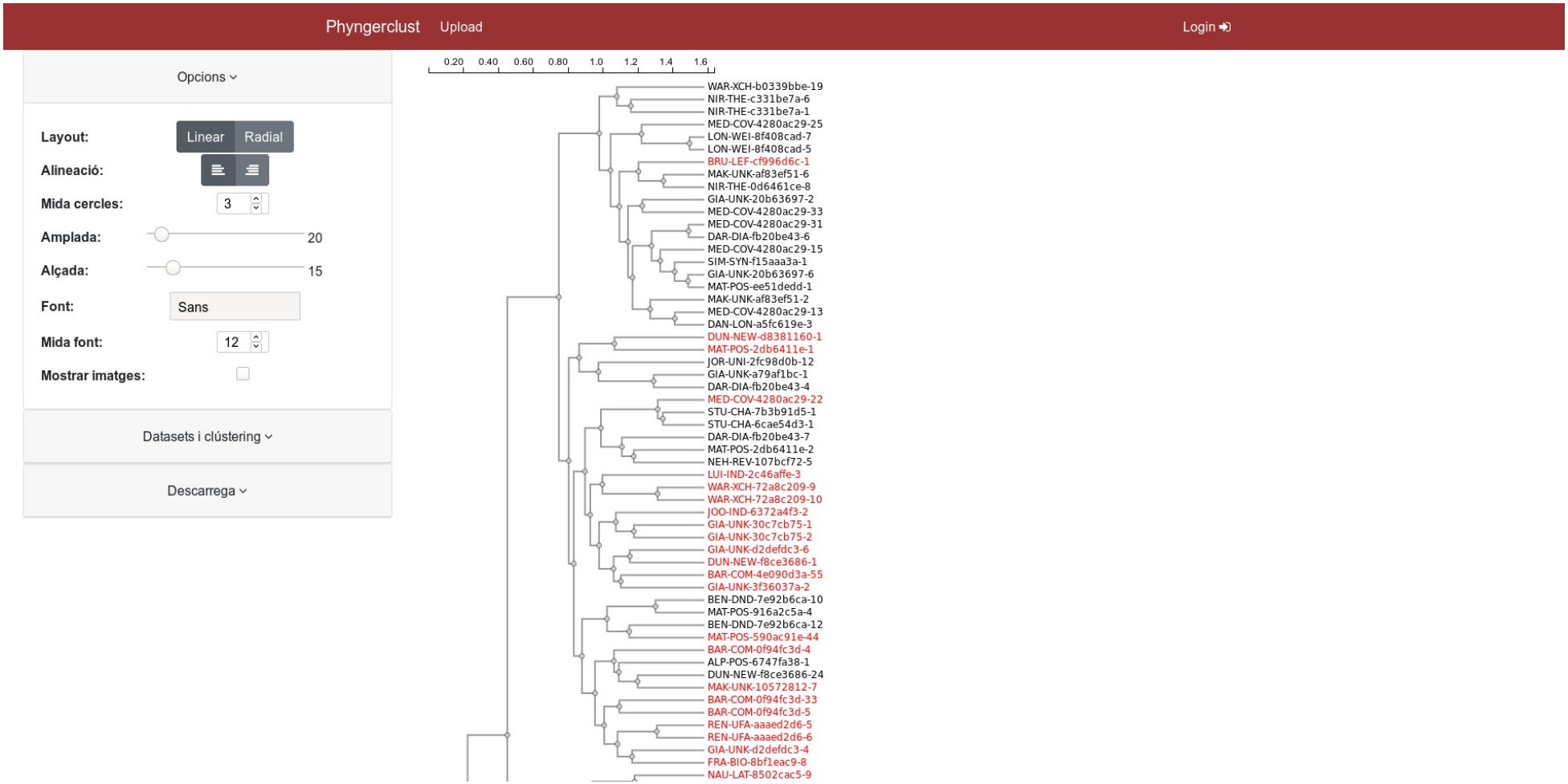


Figura 17. Visualització de la pàgina un cop s'ha creat l'arbre. A l'esquerra es mostren els menús que s'han ensenyat a l'apartat de metodologia. El tipus de visualització per defecte és un dendrograma.

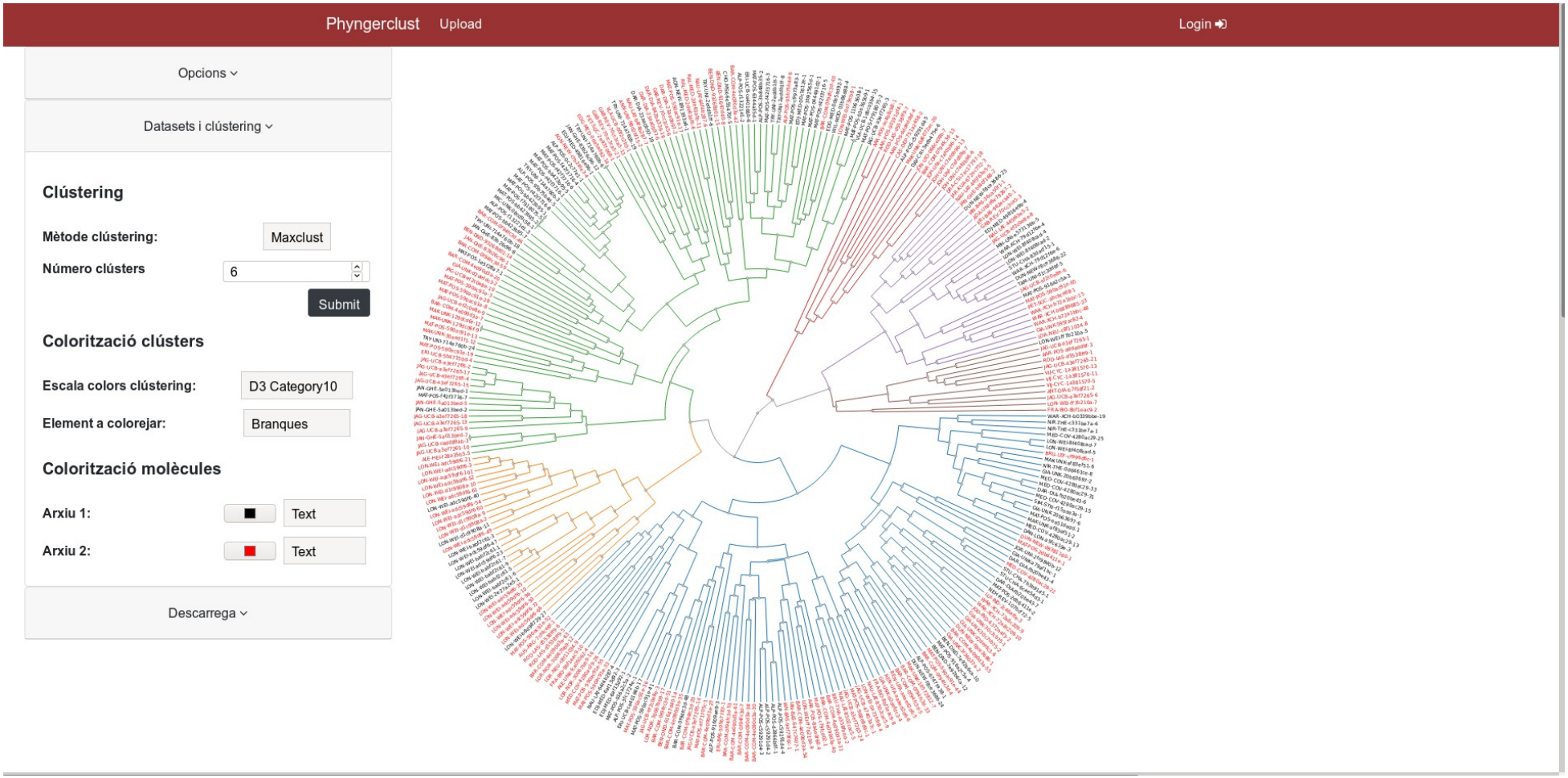


Figura 18. En aquesta figura es mostra com es poden generar clústers i pintar-los. En la figura s'han creat 6 clústers i s'ha utilitzat la escala de colors Category10 de la llibreria D3.