FACULTAT
DE CIÈNCIES
I TECNOLOGIA
UVIC | UVIC·UCC

Research Group in
Bioinformatics and
Medical Statistics
UVIC

Treball de Final de Grau

IMPLEMENTACIÓ D'UN PROGRAMA DE CLUSTERITZACIÓ DE PROTEÏNES BASAT EN LA SIMILUT DE SEQÜÈNCIA I APLICAT EN LA CARACTERITZACIÓ DELS RECEPTORS CD300

Alba Vilalta Carrera

Grau en Biotecnologia

Tutor/a: Luis Agulló

Vic, Juny del 2019

Vull agrair a en Jordi Villà, a en Luis Agulló i a en Martin Floor, líder i membres del Computational Biochemistry and Biophysics Lab (CBBL) , el fet d'haver-me permès realitzar el Treball de Final de Grau amb ells. Agrair-los-hi també la paciència que han mostrat i la quantitat de coneixements que m'han ensenyat al llarg d'aquesta estada. Gràcies Martin per haver-me ensenyat tan sobre programació i per les hores que has dedicat en fer-ho. Gràcies Luis per tot el suport que m'has aportat com a tutor de la universitat i gràcies Jordi per fer que, a quart de carrera, em plantegés encarar el meu futur cap al món de la bioinformàtica.

Vull agrair també a la Malu, coordinadora del Research Group in Bioinformatics and Medical Statistics (BEM), l'estada en un dels laboratoris que conformen el seu grup.

ÍNDEX DE CONTINGUT

ÍNDEX DE FIGURES

ÍNDEX DE TAULES

RESUM

Títol: Implementació d'un programa de clusterització de proteïnes basat en la similitud de seqüència i aplicat en la caracterització dels receptors CD300

Paraules clau: clusterització, seqüència, CD300, xarxes de similitud

Autora: Alba Vilalta Carrera

Tutors: Dr. Luis Agulló Rueda (UVic – UCC) i Dra. M. Luz Calle Rosingana (BEM)

Data: Juny 2019

L'augment exponencial de dades obtingudes en biologia molecular degut a la implementació de tècniques d'alt rendiment ha impulsat la necessitat de desenvolupar mètodes que permetin el tractament i organització d'aquestes. Les biomolècules més abundants en l'organisme són les proteïnes i, degut a l'interès que desperten, s'obté d'elles una gran quantitat d'informació que ha esdevingut en un desenvolupament d'eines que en permeten l'anàlisi i el tractament. Un tipus de tractament consisteix en l'agrupament per similitud de seqüència, el qual ens permet associar potencials característiques a una proteïna de la qual només se'n coneix la seqüència, mitjançant la similitud amb altres biomolècules d'aquest grup ja conegudes. Aquesta associació es basa en el fet que seqüències similars semblen esdevenir en estructures semblants, les quals determinen la funció de la proteïna.

L'objectiu d'aquest treball és implementar una eina informàtica que permeti agrupar les proteïnes en base a aquesta similitud per tal de generar Xarxes de Similitud de Seqüència. Per crear-la, s'ha utilitzat el llenguatge de programació Python per generar un script que pren com a input inicial un conjunt de seqüències i genera un fitxer que permet visualitzar-se, usant el software Cytoscape, com una xarxa basada en un llindar de similitud que adjudica l'usuari. S'ha fet una prova inicial per estudiar el funcionament d'aquesta eina amb la família de receptors CD300, un conjunt molècules interessants pel Laboratori de Bioquímica i Biofísica Computacional (CBBL), on s'ha dut a terme aquest treball.

Com ha resultat s'ha obtingut una eina informàtica que permet treballar amb un nombre elevat de seqüències suposant un baix cost computacional gràcies a que treballa amb alineaments per parelles i que, aplicada a la família CD300 dona resultats molt similars als obtinguts mitjançat arbres filogènics. L'avantatge que presenta és que, al tractar-se d'un script generat *de novo*, permet futures implementacions de *metadata,* gràcies a les qual també es podrien classificar les molècules segons la seva funció, entre d'altres.

ABSTRACT

Title: Implementation of a protein clustering program based on the sequence similarity and applied in the characterization of the CD300 receptors

Key words: clustering, sequence, CD300, sequence similarity networks

Author: Alba Vilalta Carrera

Tutors: Dr. Luis Agulló Rueda (UVic – UCC) and Dr. M. Luz Calle Rosingana (BEM)

Date: June 2019

The exponential increase of data obtained in molecular biology due to the implementation of high throughput techniques has led to the need to develop methods that allow the treatment and organization of these. The most abundant biomolecules in the organism are proteins and, due to the interest they arouse, we have obtained a large amount of information that has caused in the development of tools that allow the analysis and the treatment. One type of treatment consists in grouping by sequence similarity, which allows us to associate characteristic potentials with a protein whose no much information is known, by means of the similarity with other known biomolecules of this group. This association is based on the fact that similar sequences appear to become in similar structures, which determine the function of the protein.

The objective of this work is to implement a computer tool that allows grouping the proteins based on this similarity to generate sequence similarity networks. To create it, the Python programming language has been used to generate a script that takes as input a set of sequences and generates a file that allows to view, using the Cytoscape software, a network based on a threshold of similarity adjudged by the user. An initial test has been done to study the operation of this tool with the family of CD300 receptors, a set of interesting molecules for the Laboratory of Biochemistry and Computational Biophysics (CBBL), where this work has been carried out.

As a result, it has been obtained a computer tool that allows to work with a high number of sequences supposing a low computational cost thanks to that it works with alignments by pairs and, applied to the CD300 family gives results very similar to those obtained through phylogenetic trees. The advantage that it presents is that, since it is a new generated script, it allows for future metadata implementations, thanks to which, molecules could also be classified according to their function, among others.

1. INTRODUCCIÓ

L'any 2013, l'Institut Europeu de Bioinformàtica (EBI) emmagatzemava 20 petabytes (1 petabyte equival a 1015 bytes) de dades i còpies de seguretat sobre gens, proteïnes i petites molècules i es preveia que aquest número es duplicaria cada any. Aquesta enorme quantitat de dades, paradigma de *Big data* i de la complexitat que poden presentar, ha comportat la necessitat de desenvolupar mètodes que permetin el seu emmagatzematge, manipulació i processament. L'arribada de tècniques d'alt rendiment, anomenades tècniques *high throughput*, ha permès als científics obtenir una gran quantitat de dades però alhora ha proporcionat un conjunt d'informació massiu que cal estructurar i organitzar (Howe and Rhee, 2008; Marx, 2013). Un exemple en són les proteïnes, un conjunt de biomolècules que es troben en constant descobriment i que aporten un gran conjunt de dades.

De manera senzilla, les proteïnes es poden definir com cadenes d'aminoàcids que es pleguen adquirint una estructura tridimensional funcional. El material genètic contingut en els gens de les cèl·lules, és a dir l'ADN, codifica la informació de les proteïnes, que són sintetitzades pels ribosomes, uns orgànuls presents en totes les cèl·lules, a excepció dels espermatozoides, en un procés anomenat traducció (Figura 1). Aquestes molècules es troben en tots els sistemes vius que van des dels bacteris i virus fins als vertebrats i mamífers superiors com els humans. Les proteïnes constitueixen més del 50% del pes sec de les cèl·lules i es troben presents en major quantitat que qualsevol altra biomolècula (Whitford, 2005).

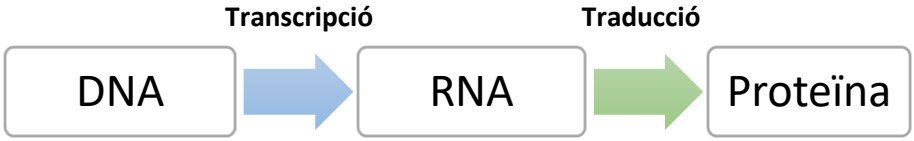


Figura 1. Dogma central de la biologia molecular. *La informació genètica es transfereix de l'ADN a l'ARN, mitjançant un procés anomenat transcripció i les proteïnes es sintetitzen a partir de l'ARN a través de la traducció.*

Cada proteïna està definida per una seqüència única de residus d'aminoàcids que passa per un seguit de nivells d'organització fins a arribar a ser una molècula funcional (Figura 2). Aquesta seqüència linear de residus d'aminoàcids al llarg de la cadena polipeptídica s'anomena estructura primària i sorgeix de l'enllaç covalent d'aminoàcids individuals a través d'enllaços peptídics. Aquesta seqüència condueix a l'estructura secundària, una conformació adoptada segons la relació espacial dels residus d'aminoàcids. En aquest nivell, les dues estructures que es poden adoptar són l'alfa hèlix o la fulla beta. El plegament tridimensional d'aquestes dóna

lloc a l'estructura terciària, una conformació globular o fibril·lar que es manté estable gràcies a diverses forces que formen enllaços.

Moltes proteïnes contenen més d'una cadena polipeptídica i, la interacció entre elles genera l'últim nivell d'organització anomenat estructura quaternària. Les interaccions que presenten són exactament les mateixes que s'observen en l'estructura terciària, amb la diferència que, en aquest cas, ocorren entre una o més cadenes polipeptídiques (Whitford, 2005).

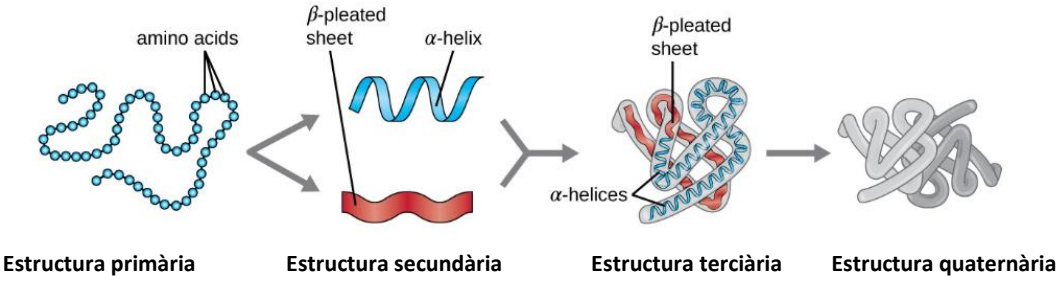


Figura 4. Nivells d'estructuració de les proteïnes*. Estructura primària, seqüència de la cadena d'aminoàcids; estructura secundària, plegament local de la cadena polipeptídica en forma d'hèlix alfa o fulla beta; estructura terciària, plegament tridimensional d'una proteïna mitjançant interaccions en la cadena; estructura quaternària, proteïna formada per més d'una cadena d'aminoàcids. Font: Lumen Learning Microbiology*

Degut a que l'estructura funcional de les proteïnes parteix d'una cadena linear d'aminoàcids, s'ha establert que, en certs casos, proteïnes que presenten una seqüència d'aminoàcids similar poden desenvolupar funcions semblants. Aquesta particularitat permet estudiar molècules poc conegudes a partir de molècules de les quals se'n sap més informació.

Al llarg de la seva evolució, les proteïnes es conserven o s'eliminen en funció de la seva utilitat. Una proteïna pot perdre o guanyar un o més residus d'aminoàcids degut als diferents mètodes de recombinació genètica, ja que duplicacions en els gens o segments d'aquests i variacions puntuals poden produir canvis en el producte sintetitzat pel gen (Munro, 1969). Aquestes biomolècules poden presentar dos tipus d'evolució. Per un banda poden experimentar una evolució divergent, la qual descriu un procés biològic on un ancestre comú va experimentant canvis al llarg del temps. Aquest procés permet explicar casos on les proteïnes presenten dominis estructurals similars, duen a terme funcions relacionades però tenen poca similitud en la seqüència. L'altre tipus d'evolució s'anomena convergent i permet descriure l'evolució independent d'un seguit de proteïnes que acaben presentant similituds funcionals i estructurals (Bork, Sander and Valencia, 1993; Graumann and Marahiel, 1996).

El gran nombre de proteïnes conegudes ha augmentat la necessitat de mètodes especialitzats per agrupar-les i així poder-les estudiar de manera més senzilla. Dos dels mètodes més usats són els arbres filogenètics i els clústers, els quals es detallen a continuació.

1.1 Mètodes d'agrupació de proteïnes

1.1.1 Arbres filogenètics

La relació d'un conjunt de seqüències de proteïnes relacionades pot expressar-se mitjançant un arbre filogenètic i la precisió d'aquest depèn de l'alineament de les seqüències. Els arbres filogenètics són gràfics que mostren relacions evolutives i es generen a partir d'una matriu que conté les distàncies genètiques entre les seqüències (Figura 3). Com més properes estiguin dues molècules entre elles, més curta serà la branca de l'arbre que les separa (Feng and Doolittle, 1990; Mahapatro *et al.*, 2012).

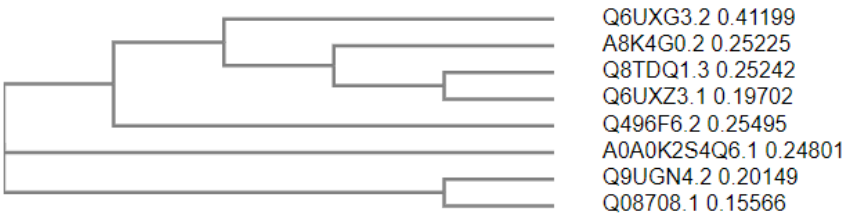


***Figura 5. Arbre filogenètic**. Exemple d'arbre filogenètic generat a través de l'aplicació en línia de l'Institut Europeu de Bioinformàtica (EMBL-EBI).*

1.1.2 Clústers

El procés de clusterització és l'organització d'un conjunt de dades, per exemple, proteïnes, en subgrups anomenats clústers. Aquesta agrupació es fa en base a característiques determinades que presenten i, per tant, els patrons dins d'un clúster vàlid seran més similars entre sí que en comparació amb un patró que pertany a un clúster diferent (Jain, Murty and Flynn, 2000; Mahapatro *et al.*, 2012).

La Figura 4 es mostra un exemple de clusterització on s'observa un conjunt de molècules que seguidament es troben organitzades tenint en compte alguna característica concreta.

Figura 6. Clusterització. *Exemple del procés de clusterització on primer s'observa com un conjunt de molècules s'acaben agrupant en clústers. Les proteïnes A, C i D formen part d'un sol clúster mentre que E i B formen clústers per separat.*

En aquest senzill exemple es pot observar com, després de produir-se la clusterització, les proteïnes A, C i D s'han agrupat en un clúster, mentre que E i B formen un clúster cada una per separat.

Aquest dos mètodes d'agrupació aporten un seguit d'informació molt útil pels científics però, i si es pogués unir tota aquesta informació en una sola representació? Aquesta hipòtesis és possible gràcies a la Xarxes de Similitud de Seqüència, les quals permeten representar molècules agrupades en base a la seva similitud de seqüència, és a dir, com de prop es troben evolutivament i alhora distingir-les segons unes característiques que presenten. Més endavant, es parlarà en detall d'aquest tipus d'agrupació de proteïnes.

1.2 Xarxes

Una xarxa és, en la seva forma més senzilla, una col·lecció de punts units per línies. En la nomenclatura correcta, aquests punts s'anomenen nodes i les línies vores (Figura 5). Molts sistemes d'interès en les ciències físiques, biològiques i socials poden considerar-se xarxes i, un bon exemple, n'és la internet, una xarxa de dades en la qual els nodes són computadores que s'uneixen mitjançant cables de fibra òptica que transporten connexions.

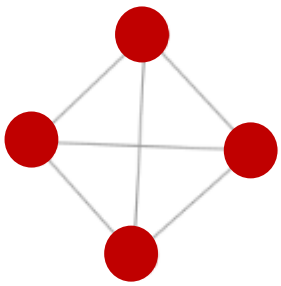


Figura 7. Xarxa. *Exemple d'una xarxa visualitzada amb Cytoscape.*

En la branca de la biologia, les xarxes apareixen com una forma de representació de patrons d'interacció entre elements biològics. En la biologia molecular, per exemple, s'utilitzen per representar patrons de reaccions químiques, mentre que en la neurociència recreen patrons de connexions entre les cèl·lules cerebrals. Un altre exemple de xarxes biologies són les esmentades Xarxes de Similitud de Seqüència, que es descriuen amb detall a continuació (Newman, 2018).

1.2.1 Xarxes de Similitud de Seqüència

El ràpid creixement de les bases de dades de informació de proteïnes proporciona noves oportunitats per l'anàlisi i l'agrupament per similitud. A mesura que aquests conjunts esdevenen més grans i els seus membres més divergents, la seva ràpida exploració es torna menys factible utilitzant enfocs tradicionals com alineaments i arbres filogènics.

Les Xarxes de Similitud de Seqüències permeten l'anàlisi i visualització de les relacions entre estructura i funció en grans conjunts de dades de proteïnes, ja que agrupen un seguit de proteïnes individuals per un anàlisis més complex, al mateix temps que també resumeixen les relacions de connectivitat entre aquests (Barber and Babbitt, 2012). Aquest mètode ofereix un seguit de característiques que fan que eclipsi els mètodes tradicionals:

- Proporciona un marc de treball ràpid i fàcil de calcular per poder observar les relacions entre conjunts molt grans de proteïnes relacionades evolutivament.
- Permet la percepció de tendències en la informació ortogonal, és a dir, en la informació de la funció, gràcies a la visualització de la xarxa.
- Satisfà la necessitat d'accedir a aquestes relacions de manera intuïtiva i fàcil de manipular.
- Aporta avantatges no coberts pels arbres filogenètics, ja que proporciona la visualització de conjunts extremadament grans de seqüències relacionades.
- Permet observar totes les relacions que tenen una puntuació superior al límit de similitud definit per l'usuari, en lloc de només un petit nombre de connexions de puntuació òptima.
- El fet de visualitzar una xarxa enlloc d'analitzar-la numèricament, permet sobreposar diversos tipus d'informació sobre aquesta.
- Per la mateixa quantitat de computació, permet analitzar un conjunt molt més gran de seqüències en comparació amb un arbre.

Tot aquest conjunt de característiques permet, gràcies a l'ús d'un software interactiu per visualitzar la xarxa, enllaçar altres tipus de informació, per exemple el procés biològic al qual

pertanyen, el que proporciona una agrupació de molècules tan per similitud de seqüència com per procés al qual participen. Aquest fet permet l'avaluació d'individus i conjunts que proporciona un grau més d'informació a l'investigador (Atkinson *et al.*, 2009).

1.2.2 Mètode actual per generar Xarxes de Similitud de Seqüència

El mètode més utilitzat per generar Xarxes de Similitud de Seqüència s'anomena Pythoscape, un marc computacional implementat a Python per generar i processar grans xarxes de proteïnes.

Pythoscape ofereix diverses opcions per calcular similituds per parelles de seqüències o estructures, aplica filtres i defineix conjunts d'eixos similars per la compressió de la informació de la xarxa. A més, genera arxius de sortida formatejats per permetre la seva visualització (Barber and Babbitt, 2012).

Aquest mètode però, presenta aspectes a millorar a l'hora de generar la xarxa. Principalment, el que fa que no s'hagi usat Pythoscape en aquest treball és que, per fer els alineaments, utilitza BLAST, un algoritme informàtic per generar alineaments disponible en línia pel Centre Nacional de Informació Biotecnològica (NCBI). Aquesta eina utilitza segments curts de cada seqüències per crear agrupacions d'alineaments, és a dir, no compara tota la seqüència global, sinó certes regions d'aquesta. Degut a aquest mètode, per un mateix grup de seqüències, els resultats obtinguts de l'alineament poden variar (Lobo, 2008).

En aquest treball es relacionaran les Xarxes de Similitud de Seqüència amb una família de molècules anomenades receptors de membrana CD300, que formen part del sistema immunitari i que generen un interès dins el Laboratori de Bioquímica i Biofísica Computacional (CBBL), laboratori en el qual s'ha dut a terme aquest treball i que es troba integrat dins el grup de Recerca en Bioinformàtica i Estadística Mèdica (BEM) de la Universitat de Vic – Universitat Central de Catalunya.

1.3 Receptors CD300

Les molècules CD300 són un conjunt de receptors de membrana que es troben tan en els llinatges mieloides com limfoides i presenten la capacitat de modular la resposta immunitària mitjançant les seves capacitats estimuladores i inhibidores (Vitallé *et al.*, 2018).

1.3.1 Components

En humans, la família CD300 està formada per 8 molècules, els gens codificants de les quals es troben localitzats al llarg del cromosoma 17. Aquestes molècules s'han anomenat

alfabèticament segons el seu ordre dins del cromosoma (CD300a, CD300b, CD300c, CD300d, CD300e, CD300f, CD300g i CD300h) i presenten ortòlegs en els ratolins, on es troben codificades per 9 gens localitzats al cromosoma 11 (Figura 4) (Clark *et al.*, 2009; Borrego, 2013). La nomenclatura usada per anomenar aquest grup de molècules és molt variada; a l'annex I s'hi poden trobar dues taules amb les diferents nomenclatures que presenten els receptors CD300 en humans i en ratolins.

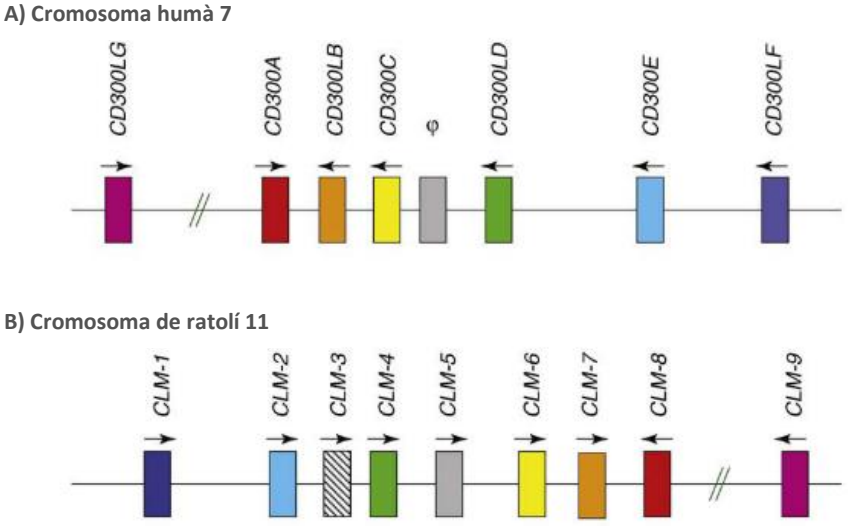


***Figura 10. Diagrama de l'organització dels gens de CD300**. Diagrama esquemàtic de l'organització dels complexes del gen CD300 (a) humà i (b) de ratolí. Els ortòlegs genètics es troben subratllats de la mateixa manera. Les fletxes indiquen la direcció de la transcripció Font: "The CD300 family of molècules are evolutionarily significant regulators of leukocyte functions", Clark et al., no date. .*

1.3.2 Estructura

Les molècules CD300 són unes glicoproteïnes de transmembrana de tipus I que presenten un domini d'immunoglobulina (Ig) extracel·lular de tipus variable (V), un domini de transmembrana i una cua citoplasmàtica, que varia de mida segons la seva funcionalitat (Vitallé *et al.*, 2018). Aquest domini extracel·lular es troba dins de la superfamília de les immunoglobulines i és el que els aporta especificitat en la reacció amb els antígens. Està format per una estructura anomenada plegament d'Ig, constituïda per cadenes β antiparal·leles disposades en dues fulles unides per un pont disulfur (Figura 5) (Barclay, 2002).

La cua citoplasmàtica que posseeixen aquestes molècules pot presentar dues formes diferents, fet que divideix els components d'aquesta família en dos grups. Un conjunt de molècules (CD300a i CD300f) presenta un motiu inhibidor basat en tirosines immunoreceptores (ITIM) i la resta (CD300b, CD300c, CD300d, CD300e i CD300h) posseeixen un residu de transmembrana

bàsic que permet l'associació amb proteïnes adaptadores que contenen motius activadors basats ens tirosines immunoreceptores (ITAM). Les molècules del primer grup presenten cues citoplasmàtiques llargues i el motiu ITIM que contenen es requereix per la senyalització inhibidora, mentre que el residu d'aminoàcid, com que ocupa menor espai, es troba en cues més curtes i permet, degut a l'associació amb dominis ITAM, que es generin senyals d'activació (Niizuma *et al.*, 2015; Zenarruzabeitia *et al.*, 2016; Vitallé *et al.*, 2018).

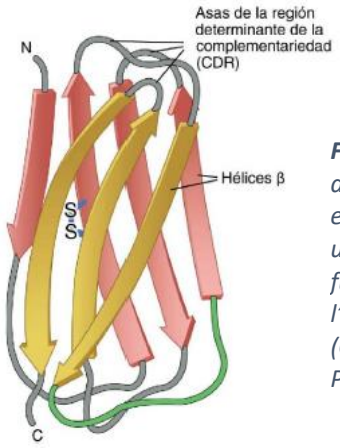


***Figura 13. Estructura d'un domini d'Ig**. Cada domini està compost de dues fulles beta antiparal·leles (groc i vermell) unides entre si per un enllaç disulfur formant una estructura en forma de Sandwich. S'observa un domini constant (C) que conté tres i quatre cadenes β en les dues fulles. Tres nanses en cada domini variable contribueixen a la unió de l'antigen i s'anomenen regions determinants de la complementarietat (CDR). Font: "Inmunología célular y molecular", Abbas, Lichtman and Pillai, 2015.*

Dos casos excepcionals són els de les molècules CD300f i CD300g. La primera, a part de presentar un motiu ITIM, també és capaç d'entregar senyals d'activació a través de motius que s'uneixen a subunitats reguladores de certes proteïnes, és a dir , la molècula CD300f presenta doble funció. Per altra banda, el receptor CD300g, presenta un domini extracel·lular similar al de les mucines (proteïnes altament glicosilades produïdes per les cèl·lules epitelials) i manca de motius estructurals indicatius de potencial estimulant o inhibidor a la cua intracel·lular (Vitallé *et al.*, 2018).

1.3.3 Expressió

En humans, a nivell cel·lular, la família CD300 presenta quatre patrons d'expressió; les molècules CD300a i CD300c s'expressen àmpliament en la majoria dels llinatges dels leucòcits, mentre que CD300b, CD300d, CD3000f i CD300h es restringeixen als llinatges mieloides i a les cèl·lules dendrítiques. CD300e s'expressa principalment a la superfície de cèl·lules mieloides madures i CD300g únicament es troba a la superfície de cèl·lules epitelials i endotelials, essent l'única molècula d'aquesta família que no s'expressa ni en llinatges mieloides ni en cèl·lules dendrítiques (Clark *et al.*, 2009; Niizuma *et al.*, 2015).

1.3.4 Lligands

Tot i que els lligands específics de cada membre de de la família CD300 encara es desconeixen, s'ha demostrar que són capaços d'unir-se als lípids i que, segons el receptor amb el qual s'uneixin, la senyal variarà.

Per exemple, els receptors CD300a, CD300c i CD300h humans reconeixen la fosfatidilserina (PS) i la fosfatidiletanolamina (PE). Tan en cèl·lules vives com en cèl·lules en repòs, PS i PE es localitzen a la regió interna de la membrana plasmàtica, mentre que ambdós lípids es desplacen a la zona externa quan les cèl·lules es sotmeten a apoptosis o s'infecten. CD300a s'uneix a PE amb major afinitat que amb PS, igual que el seu ortòleg en ratolins, mentre que CD300c ho fa amb la mateixa afinitat. Els receptors CD300b i CD300f de ratolí també tenen la capacitat d'unir-se a PS i, a més, s'ha descrit que altres lligands no lipídics es poden unir a CD300b, com per exemple les immunoglobulines de cèl·lules T receptores TIM-1 i TIM-4, en un procés dependent de PS (Borrego, 2013; Niizuma *et al.*, 2015; Vitallé *et al.*, 2018).

També s'ha demostrat que tan CD300f humà com CD300e de ratolí s'uneixen a esfingomielina, un esfingolípid que es troba a la membrana de les cèl·lules animals, mentre que només el primer s'uneix també a la ceramides, un altre tipus de lípids. CD300b en ambdues espècies es pot unir a LPS però només el de ratolí també interacciona amb 3-O-sulfo-β-D-galactosilceramida C24:1. Finalment, només els ortòlegs de CD300f i CD300d en ratolins són capaços d'unir-se amb Norovirus, en una interacció que es detallà més endavant (Borrego, 2013; Vitallé *et al.*, 2018).

1.3.5 Funció

Les diferents molècules que formen part de la família CD300 duen a terme diverses funcions. Les més importants s'expliquen a continuació.

Infecció viral

La família de receptors CD300 es troba involucrada en la patogènesis de moltes malalties. Concretament, aquestes molècules participen en els mecanismes utilitzats pels virus durant la infecció de cèl·lules hostes. El receptor CD300a actua com a factor d'unió per partícules virals associades a PS i PE i, tot i que encara es requereix investigació per aclarir la seva implicació, es pot afirmar que té un paper important en la unió de virus que utilitzen el mimetisme apoptòtic com a mecanisme per ingressar en cèl·lules hostes.

A part d'unir-se a les partícules virals associades a PS i PE, també s'ha demostrat que les molècules CD300 poden facilitar l'entrada de virus de manera independent a aquests lligands ja

que s'ha observat que certes proteïnes d'adenovirus s'uneixen directament als receptors CD300a i CD300c, que presenten homologia en els seus dominis extracel·lulars.

Una altra molècula involucrada en aquest procés és el receptor CD300f de ratolins, anomenat CD300lf, que s'ha classificat com a component crucial en el mecanisme d'infecció del Norovirus murí (MNV), una tipus de norovirus que afecta a aquesta espècie. S'ha demostrat que el bloqueig d'aquest receptor en ratolins mitjançant anticossos policlonals va causar resistència a la infecció per MNV, el que demostra que CD300lf és necessari en el procés d'infecció d'aquest norovirus concret (Figura 5) (Vitallé *et al.*, 2018).

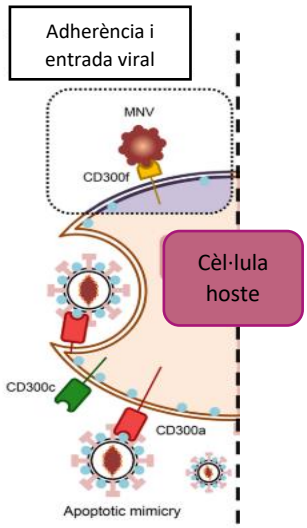


Figura 16. Molècules CD300 en l'entrada viral. *A través de la mímica apoptòtica, els virus tanquen les seves càpsides en una bicapa lipídica obtinguda de la membrana plasmàtica de les cèl·lules hoste, el que porta a la incorporació de PS i PE a l'embolcall víric. Degut a això, els receptors superficials, com el CD300a, expressats en cèl·lules hoste s'uneixen a partícules virals que contenen aquets dos lípids i poden promoure la captació viral. Altres molècules que s'uneixen a PS i PE, com ara CD300c, tenen el potencial d'unir-se als virus que expressen aquests fosfolípids al seu embolcall. Per últim, el receptor CD300f de ratolí (CD300lf) s'uneix a les partícules de norovirus murí i promou la infecció de manera independent de PS i PE. Font: "CD300 receptor family in viral infections", Vitallé et al., 2018.*

Apoptosis

L'apoptosi o mort cel·lular és una forma de regeneració de cèl·lules d'un organisme que es duu a terme al llarg de la vida d'aquest. No es tracta d'un fenomen aleatori, sinó d'un procés actiu, ben definit genèticament, on les cèl·lules estan destinades a morir en un temps fixat (Jordán, 2003). Aquestes cèl·lules apoptòtiques exposen tan fosfatidilserina com fosfatidiletanolamina, dos lligands dels receptors CD300, a la zona externa de la membrana plasmàtica, que actuen com a senyals de "menja'm" que avisen als fagòcits (Borrego, 2013).

Les molècules CD300b i CD300f actuen com receptors de la PS per eliminar les cèl·lules apoptòtiques. La senyal d'activació intervinguda per CD300b es transmet a través de l'associació d'aquesta amb una molècula adaptadora DAP12 i requereix un motiu funcional DAP12 ITAM. Un cop el receptor CD300b reconeix la fosfatidilserina, s'activa la ruta PI3K / Akt, que és la principal via de senyalització utilitzada pels macròfags (Borrego, 2013; Murakami *et al.*, 2014). Pel que fa a CD300f, el seu paper en l'apoptosi és més complex, ja que és l'únic receptor d'aquesta família que presenta tan activitat estimuladora com inhibidora.

La cua citoplasmàtica de CD300f conté cinc regions de fosforilació de tirosines que resideixen en motius de senyalització; un d'ells recluta proteïnes capaces d'iniciar senyals que promouen la fagocitosis, mentre que la resta de motius bastats en tirosina (4 motius) s'uneixen a proteïnes que inicien senyals inhibidores. Un cop s'ha dut a terme el reconeixement cèl·lules apoptòtiques a través de PS, es produeix la fosforilació d'una de les tirosines de CD300f, el que comporta el reclutament de p85α, que, com a subunitat reguladora de PI3K, l'activa. La fosforilació de les altres quatre tirosines restants resulta en la inhibició de l'activació de PI3K, el que dona lloc a la inhibició de la fagocitosis (Tian *et al.*, 2014).

Una altra molècula d'aquesta família, la CD300a, inhibeix la captació de cèl·lules apoptòtiques a través de la unió a PS i PE, el que genera la transmissió de senyals inhibidores que impedeixen l'engoliment de les cèl·lules mortes pels macròfags (Park and Kim, 2017; Vitallé *et al.*, 2018). El receptor CD300c també és capaç de reconèixer PS i PE en cèl·lules en fase d'apoptosis, però no ho fa amb tanta força com CD300a (Takahashi *et al.*, 2013).

Altres funcions possibles

S'ha observat de forma in vivo que la molècula CD300d és capaç d'interaccionar amb la resta de receptors CD300 coneguts, amb excepció del CD300c. És per això, que es creu que el receptor CD300d podria tenir un paper important com a regulador en la formació de complexes CD300 i en l'expressió d'aquestes molècules a la membrana extracel·lular (Casellas, no date; Comas-Casellas *et al.*, 2012). En ratolins però, el seu ortòleg (CD300ld), pot inhibir la proliferació cel·lular en cèl·lules B induïda per receptors BCR i TRL-9, mitjançant un mecanisme que involucre el reclutament de la fosfatasa SHP-1 (Casellas, no date).

En general, la senyalització de CD300 efectua la mobilització de calci i activa certes proteïnes quinases (Clark *et al.*, no date). No obstant, s'ha demostrat que CD300e pot desencadenar la mobilització de calci intracel·lular i la secreció de ROS (espècies reactives de l'oxigen) en monòcits, el que recolza la teoria que aquest receptor pot regular l'activitat microbicida dels monòcits (Brckalo *et al.*, 2010).

Tot i presentar una alta identitat de seqüència amb els dominis d'immunoglobulina, només s'ha demostrar que CD300g és capaç d'unir-se a Ig, concretament s'uneix a les IgA i IgM humanes (Clark *et al.*, no date).

1.3.6 Malalties relacionades

La capacitat que presenten les molècules de la família CD300 per regular la funció dels leucòcits i les respostes immunitàries obre un ventall d'opcions per explotar aquests receptors com a

dinares terapèutiques en malalties inflamatòries cròniques, al·lèrgies i altres malalties. A més, la selecció de gens ha demostrat les funcions generals de les molècules adaptadores (com DAP12) i de les fosfatases (SHP1/SHP2) en el desenvolupament de certes malalties autoimmunitàries i òssies i, donat que els receptors CD300 interactuen àmpliament amb aquestes molècules, és probable que algun d'ells contribueixi a la protecció o agreujament d'aquestes malalties patologies.

Psoriasis

S'han publicat evidències d'alteracions en l'expressió i funció d'algunes molècules de la família CD300 en pacients que pateixen psoriasis. Això és degut a que el complex de gens que codifica per aquests receptors està vinculat a PSORS2, un locus per la malaltia de la psoriasis. Aquesta regió també s'ha relacionat de forma variable amb la dermatitis atòpica, l'artritis reumatoide i altres trastorns de cutanis (Clark *et al.*, no date).

Respostes al·lèrgiques

Es creu que la presència de CD300a en eosinòfils i mastòcits pot destacar el seu important potencial per les reaccions al·lèrgiques o d'hipersensibilitat. Aquest receptor mostra una major expressió en els pulmons, el que suggereix un paper en la inflamació pulmonar (Clark *et al.*, 2009).

Tumors o infeccions víriques

Les cèl·lules Natural Killer (NK) són limfòcits derivats de la medul·la òssia que poden destruir un ampli ventall de patògens i tumors. L'activació d'aquestes cèl·lules resulta d'un balanç de senyals oposades que, per una banda, activen les cèl·lules NK permetent que erradiquin cèl·lules tumorals o infectades per virus i, per una altra banda, les inhibeixen per evitar la mort d'un mateix.

La molècula CD300a es troba expressada en la superfície de les cèl·lules NK desenvolupant una funció inhibitòria però no tots els clons de NK poden emetre senyals inhibidores, degut probablement a la presència de CD300c. Com que CD300a i CD300c són difícils de distingir quan es troben a la superfície cel·lular, encara queden moltes preguntes per respondre sobre la seva activitat cap a les cèl·lules NK (Lankry *et al.*, 2010).

2. OBJECTIUS

L'objectiu d'aquest Treball de Final de Grau és desenvolupar una eina de creació de Xarxes de Similitud de Seqüència que permeti analitzar múltiples molècules a la vegada i suposi un baix cost computacional.

Donat l'interès del CBBL per la família de receptors CD300, un cop desenvolupada l'eina, es pretén estudiar-ne l'eficàcia mitjançant la seva aplicació en la caracterització d'aquestes molècules.

3. MATERIALS I MÈTODES

La metodologia per realitzar aquest treball s'ha dividit en tres fases. Primer de tot s'ha generat un programa informàtic, seguidament s'ha establert un conjunt de molècules per ser analitzades mitjançant xarxes i finalment, aquestes xarxes s'han generat amb el programa i posteriorment s'han visualitzat amb el software Cytoscape.

3.1 Generació del programa

S'ha generat un programa informàtic que permet crear Xarxes de Similitud de Seqüència a partir d'un conjunt de seqüències guardades en un fitxer en format *fasta*. S'ha utilitzat el llenguatge de programació Python per escriure i generar el programa. Es tracta d'una gran plataforma per la computació científica que usa un llenguatge entenedor amb una sintaxis fàcil d'aprendre. A més, consta d'una gran gama de llibreries que aporten funcions addicionals a aquest llenguatge (Lindstorm, 2005 ; Cock *et al.*, 2009). En concret, s'ha fet ús de les següents llibreries i utilitats:

- **Biopython:** És un projecte que es va iniciar l'any 1999 com una col·laboració per recopilar i produir eines de bioinformàtica de codi obert escrites en Python. Gran part del seu desenvolupament s'ha centrat en escriure codis que puguin recuperar dades de bases biològiques i analitzar-les en una estructura de dades de Python. Així doncs, Biopython s'ha convertit en una gran col·lecció de mòduls destinats a la programació de biologia computacional i bioinformàtica (Chapman and Chang, 2000; Cock *et al.*, 2009).

Dins de Biopython s'han usat dos mòduls concrets:

- **Pairwise2**: mòdul que alinea parells de seqüències mitjançant un algoritme de programació dinàmica. Proporciona funcions per obtenir alineament globals i locals. En aquest programa però, només s'ha usat l'alineament global, el qual permet trobar la major concordança entre tots els aminoàcids de dues seqüències (*Bio.pairwise2*, no date).

- **SeqIO**: mòdul que proporciona una interfície senzilla per llegir i escriure arxius de seqüències biològiques en diversos formats (Cock *et al.*, 2009).

- **Jupyter:** Projecte de codi obert que pot treballar amb diferents llenguatges de programació. L'aplicació principal oferta per aquest projecte és *Jupyter Notebook*, una plataforma de computació interactiva en web que permet als usuaris publicar codis, resultats i explicacions en un format llegible i executable. Per realitzar l'script (document que conté instruccions escrites

en codis de programació) de Python s'ha utilitzat aquesta plataforma interactiva a través de la qual s'accedeix mitjançant un navegador web (Perez *et al.*, no date; Kluyver *et al.*, 2016).

- **Numpy:** *Numerical Python* és el paquet fonamental per la computació científica en Python. Les matrius *NumPy* són la representació estàndard per dades numèriques i permeten la implementació eficient de càlculs numèrics. Proporciona una eina per calcular, llegir i escriure conjunts de dades basats en matrius (McKinney, no date; Van Der Walt, Colbert and Varoquaux, 2011). Aquest paquet s'ha utilitzat per generar i manipular les diferents matrius que proporciona el programa.

3.2 Cerca de seqüències

Un cop s'ha generat el programa, s'han cercat les seqüències de les proteïnes que contenen el domini d'immunoglobulina de tipus variable (IgV) a través de la base de dades UniProt, una base de dades proteica totalment gratuïta que permet navegar per una gran quantitat de seqüències i informació funcional. Se n'han obtingut 514, 20 de les quals formen part de la família CD300 (Annex I) i s'han descarregat en format *fasta* directament de la base de dades.

UniProt KnowledgeBase (UniProtKB) és el recurs central d'aquesta base de dades i combina dues seccions. La secció de la base de dades que conté les entrades revisades i curades manualment es coneix com UniProtKB / Swiss-Prot i actualment conté més de mig milió de seqüències. Per aquestes entrades, la informació experimental s'ha extret de la literatura i s'ha organitzat i resumit i va creixent a mesura que noves proteïnes es caracteritzen experimentalment. Les seqüències restants es recopilen en la secció sense revisar coneguda com UniprotKB / TrEMBL, que actualment disposa de més de 150 milions de seqüències. Les seqüències que conté s'han derivat, en gran part, de la seqüenciació de l'ADN d'alt rendiment i, tot i que aquestes entrades no es curen manualment, es complementen amb una anotació generada automàticament (The UniProt Consortium, 2017).

Per tal d'evitar un excés de redundància en el conjunt de seqüències obtingudes d'UniProt, s'ha utilitzat un programa de clusterització anomenat CD-HIT (Figura 9). Aquest programa utilitza un algoritme que pren com a primera seqüència representativa la seqüència més llarga del conjunt i la compara amb la resta de seqüències del fitxer. Si les seqüències presenten una identitat superior a la fixada per l'usuari, s'agrupen en el clúster de la seqüència representativa. Un cop s'han comparat totes les seqüències contra la més llarga, CD-HIT pren la segona seqüència més llarga com a representativa d'un nou clúster i torna a comprar les seqüències restants amb aquest (Fu *et al.*, 2012). El procés es repeteix fins a definir totes les seqüències en clústers i

s'obtenen dos fitxers *fasta;* un que conté els clústers que s'han generat i un altre amb només les seqüències representatives de de cada un d'ells. Tal i com es pot veure a la següent figura, en aquest treball s'ha executat CD-HIT a un llindar d'identitat de 99, amb la qual cosa, de totes les molècules que presentaven una identitat superior al 0.99% només se n'ha guardat una com a representativa.

```
os.system('cd-hit -i '+file_name+' -o cd_'+file_name+'.fasta -c 0.99 -n 5 -d 0')
```

***Figura 17. Línia d'execució de CD-HIT**. Línia usada per executar CD-HIT on l'input (-i) és 'file_name', variable que conté l'arxiu fasta inicial i l'output (-o) és el mateix però afegint '_cd' a l'inici del nom i especificant que es vol en format fasta. La variable -n defineix el llindar d'identitat en percentatge i en aquest cas té un valor de 0.99. La variable -n fa referència al nombre de caràcters que agafa el programa per fer la comparació entre seqüències i venen definits segons el llindar utilitzat. En aquest cas s'utilitza un valor de 5, ja que es treballa amb un llindar entre 0.7 i 1.0. La variable -d especifica la longitud de la descripció del fitxer que conté els clústers que es generen (.clst). Si el seu valor és zero, com en aquest cas, utilitza el mateix nom que el fitxer festa importat i s'atura al primer espai.*

3.3 Visualització i anàlisis

S'han generat xarxes a diferents nivells de similitud de seqüència, concretament els valors utilitzats són els compresos entre 25 i 70. Aquestes xarxes s'han visualitzat gràcies a l'eina Cytoscape, una aplicació d'escriptori Java utilitzada per la visualització de xarxes biològiques i per la integració de dades.

Cytoscape permet visualitza xarxes que es troben representades per nodes i eixos que els uneixen. Les dades s'integren en aquesta xarxa mitjançant atributs, que assignen característiques específiques als nodes, com els nivells d'expressió gènica o les funcions de les proteïnes. Aquests valors d'atributs es poden usar per controlar aspectes visuals, per exemple la forma o el color, així com per realitzar recerques complexes a la xarxa, operacions de filtratge o altres anàlisis (Smoot *et al.*, 2011).

Un cop obtinguts els resultats de les xarxes, s'ha utilitzat la base de dades de proteïnes Protein Data Bank (PDB) per cercar les estructures tridimensionals que presten tan les molècules CD300 com les molècules TREML4, que han destacat en els resultats. Aquesta base de dades utilitzada es tracta d'una base d'estructures de macromolècules biològiques que va representar una de les primeres col·leccions de dades de biologia molecular impulsades per la comunitat quan es va establir l'any 1971 (Berman *et al.*, 2000). Les estructures trobades s'han tractat mitjançant l'UCSF Chimera, un sistema extensible de visualització de models 3D que permet analitzar les estructures gràcies a un seguit d'eines implantades que presenta (Pettersen *et al.*, 2004).

Finalment, per identificar si els resultats obtinguts mitjançant l'eina desenvolupada són coherents, s'ha generat un arbre filogènic, utilitzant ClustalW, amb un subgrup de proteïnes CD300 i proteïnes d'altres clústers escollides a l'atzar sota un llindar de 29. Aquesta eina usada proporciona un mètode d'alineament múltiple de seqüències de proteïnes o nucleòtids que s'aconsegueix mitjançant tres passos: l'alineament per parelles, la generació d'un arbre filogènic i l'alineament progressiu (Li, 2003). Primer de tot, l'algoritme realitza alineaments per parelles de seqüències i es genera una matriu amb els valors de similitud de cada parell. Seguidament, aquests valors són convertits a valors de distàncies, els quals són usats per l'algoritme per generar un arbre filogènic mitjançant el mètode Neighbour-Joining. Finalment, l'últim pas és construir un alineament múltiple de seqüència mitjançant un alineament progressiu de les seqüències més properes d'acord amb els resultats de l'arbre generat prèviament (Daugelaite, O' Driscoll and Sleator, 2013). També s'ha usat ClustalW per generar un alineament múltiple entre CD300 i TREML4.

4. RESULTALTS

4.1 Programa

El primer resultat obtingut ha estat l'script generat mitjançant Python que permet, a partir d'un seguit de seqüències d'interès, generar una Xarxa de Similitud de Seqüència, la qual anomenarem xarxa d'IgV. A la Taula 1 s'esmenten les diferents funcions que s'executen al programa, els arguments que necessiten, els outputs que generen i una breu descripció de la seva funció. A més, tan l'script del programa com el de les diferents funcions que s'hi importen es poden trobar als annexes II i III.

El programa rep com a input un fitxer de seqüències en format *fasta*, el llegeix i en fa un alineament global per parelles. Els valors més alts d'identitat que s'obtenen mitjançant l'alineament s'organitzen en una matriu que té tantes files i columnes com seqüències conté el fitxer inicial. Seguidament, en funció d'un llindar d'identitat (*threshold)* definit per l'usuari, la matriu adopta la seva forma binària i finalment, aquesta és convertida al format que requereix Cytoscape per tal de ser visualitzada, on s'especifiquen les seqüències de sortida d'eixos i les d'entrada.

Taula 1. *Funcions utilitzades pel programa*

Funció	Arguments	Output	Descripció
readFasta	'fasta_file'	'seqs'	Llegeix l'arxiu *fasta* i n'extreu el llista de seqüències
seqID	'seq1' 'seq2'	'max(values)'	Genera l'alineament global de seqüències mitjançant el mòdul *pairwise2* de Biopython
seqIDmatrix	'sequences'	'idmatrix'	Genera una matriu d'identitat
matrixPId	'idmatrix' 'threshold'	'bmatrix'	Converteix la matriu d'identitat en la seva forma binària en funció del valor del llindar
binary2Cytoscape	'fasta_file' 'bmatrix' 'sequences' 'output_file'	fitxer en format .txt	Converteix la matriu binària en un fitxer amb el format adequat per poder ser visualitzada la xarxa

Llistat de les funcions que s'han desenvolupat per generar el programa de creació de Xarxes de Similitud de Seqüència. De cada una se'n defineixen els arguments i l'output (objecte que genera) i se'n descriu breument la seva funció.

4.2 Clusterització de les molècules CD300

Tal i com mostra la Taula 3, les primeres molècules en agrupar-se han estat els ortòlegs de rata i ratolí. Seguidament s'han agrupat les proteïnes de l'espècie *Mus musculus* i finalment s'han organitzat conjuntament els ortòlegs d'aquesta i d'*Homo sapiens.* No ha estat fins a un llindar de 41 quan s'han començat a generar clústers més grans i l'agrupació de tots els receptors CD300 en un sol clúster ha estat a un llindar de 29. Cal observar que, des d'un llindar de 37 (Figura 10A) fins a arribar a 29 (Figura 10B), les úniques molècules de la família CD300 que no s'han unit amb la resta han estat els ortòlegs d'humà, ratolí i boví CLM9.

Taula 2. *Agrupacions de CD300 en funció del llindar de similitud*

Valor del llindar de similitud	Molècules que s'han agrupat
64	CLM-1 *(Mm)* i CLM-1 *(Rn)*
61	CLM-8 *(Mm)* i CLM-8 *(Rn)*
59	CLM-6 *(Mm)* i CLM-4 *(Mm)*
57	CLM-5 *(Mm)* i CLM-3 *(Mm)*
47	CLM-2 *(Hs)* i CLM-2 *(Mm)*
44	CLM-6 *(Hs)* i CLM-8 *(Hs)*
42	CLM-9 *(Hs)* i CLM-9 *(Mm)* CLM-7 *(Hs)* i CLM-7 *(Mm)* CLM-8 *(Mm),* CLM-8 *(Rn),* CLM-6 *(Mm)* i CLM-4 *(Mm)*
41	CLM-8 *(Mm),* CLM-8 *(Rn),* CLM-6 *(Mm),* CLM-4 *(Mm)*, CLM-6 *(Hs)* i CLM-8 *(Hs)*
39	CLM-8 *(Mm),* CLM-8 *(Rn),* CLM-6 *(Mm),* CLM-4 *(Mm)*, CLM-6 *(Hs)* i CLM-8 *(Hs)* i CD300H CLM-1 *(Mm)* i CLM-1 *(Rn),* CLM-5 *(Mm),* CLM-3 *(Mm)*, CLM-5 (*Hs)* i CLM-1 (*Hs)* CLM-9 *(Hs),* CLM-9 *(Mm)* i CLM-9 (*Bt)*
38	CLM-8 *(Mm),* CLM-8 *(Rn),* CLM-6 *(Mm),* CLM-4 *(Mm)*, CLM-6 *(Hs)* i CLM-8 *(Hs),* CD300H, CLM-2 *(Hs)* i CLM-2 *(Mm)*
37	CLM-8 *(Mm),* CLM-8 *(Rn),* CLM-6 *(Mm),* CLM-4 *(Mm)*, CLM-6 *(Hs)* i CLM-8 *(Hs),* CD300H, CLM-2 *(Hs)* i CLM-2 *(Mm),* CLM-1 *(Mm),* CLM-1 *(Rn),* CLM-5 *(Mm)* i CLM-3 *(Mm)*, CLM-5 (*Hs),* CLM-1 (*Hs),* CLM-7 *(Hs)* i CLM-7 *(Mm)*
29	CLM-8 *(Mm),* CLM-8 *(Rn),* CLM-6 *(Mm),* CLM-4 *(Mm)*, CLM-6 *(Hs)* i CLM-8 *(Hs),* CD300H, CLM-2 *(Hs)* i CLM-2 *(Mm),* CLM-1 *(Mm),* CLM-1 *(Rn),* CLM-5 *(Mm)* i CLM-3 *(Mm)*, CLM-5 (*Hs),* CLM-1 (*Hs),* CLM-7 *(Hs)* i CLM-7 *(Mm)*, CLM-9 *(Hs),* CLM-9 *(Mm)* i CLM-9 (*Bt)*

Ms: Mus musculus (ratolí), *Rn: Rattus norvegicus* (rata), *Hs: Homo sapiens* (humà), *Bt: Bos taurus* (boví)

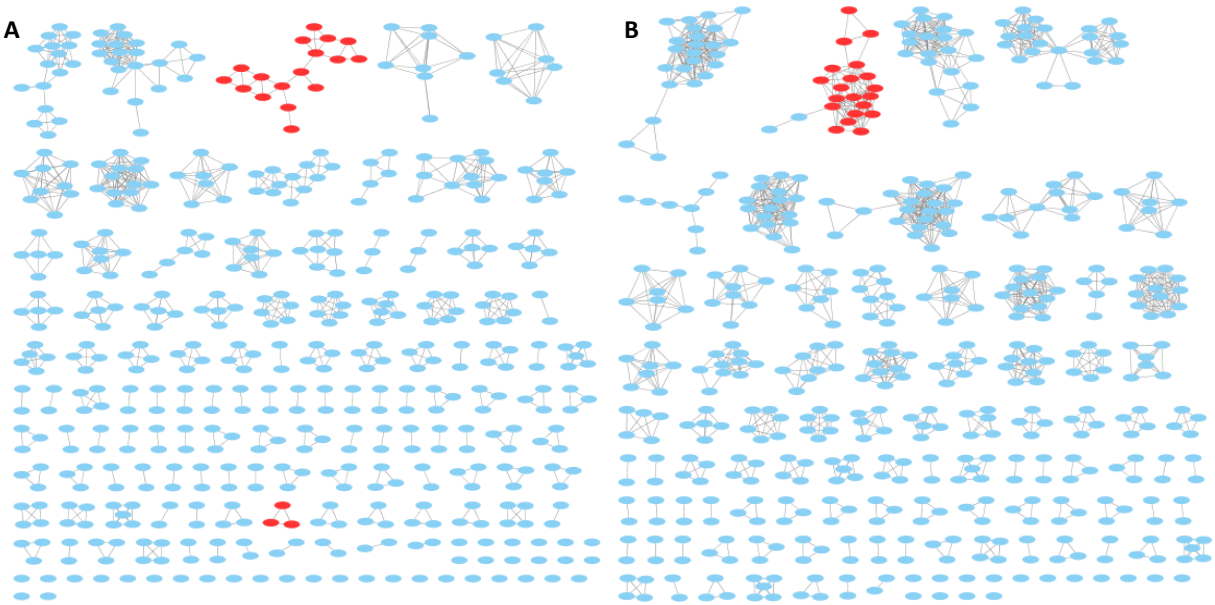


Figura 20. Xarxa d'IgV a diferents llindars de similitud. *En ambdues xarxes es mostres les molècules de la família CD300 remarcades de color vermell. A) Xarxa d'IgV a un llindar de similitud de 37. B) Xarxa d'IgV a un llindar de similitud de 29.*

S'han buscat al PDB les estructures tridimensionals que presenten les molècules CD300, els identificadors de les quals es mostren a la Taula 4 juntament amb el grau de similitud que presenten amb les seqüències. S'observa que totes les proteïnes presenten un dels cinc models 2NMS, 2Q87, 1ZOX, 6C74 i &E47 amb diferents graus de similitud.

Taula 3. *Estructures tridimensionals de les molècules CD300*

Molècula CD300	Espècie	Identificador PDB de l'estructura 3D	% identitat
CLM1	*Homo sapiens*	2NMS	98,36
CLM2	*Homo sapiens*	2Q87	51,85
CLM5	*Homo sapiens*	2NMS	73,15
CLM6	*Homo sapiens*	2Q87	91,74
CLM7	*Homo sapiens*	2NMS	62,60
CLM8	*Homo sapiens*	2Q87	98,18
CLM9	*Homo sapiens*	2Q87	41,67
CD300H	*Homo sapiens*	2Q87	50,00
CLM1	*Mus musculus*	1ZOX	99,09
CLM2	*Mus musculus*	2Q87	53,77
CLM3	*Mus musculus*	6C74	84,40

CLM4	*Mus musculus*	2Q87	49,53
CLM5	*Mus musculus*	1ZOX	87,39
CLM6	*Mus musculus*	2Q87	53,77
CLM7	*Mus musculus*	2NMS	56,48
CLM8	*Mus musculus*	2Q87	49,53
CLM9	*Mus musculus*	6E47	37,96
CLM1	*Rattus norvegicus*	1ZOX	75,89
CLM8	*Rattus norvegicus*	2Q87	53,77
CLM9	*Bos taurus*	2Q87	38,53

Les estructures tridimensionals trobades s'han alineat entre elles per entendre la similitud estructural que presenten. A la Figura 11A s'observen els cinc models 3D obtinguts. 1ZOX, 6C74 i 2NMS presenten una estructura molt similar, mentre que 6E47 i 2Q87 difereixen de la resta. L'alineament es mosta a la Figura 11B i, observant la Figura 11C es pot veure la regió compartida alineada.

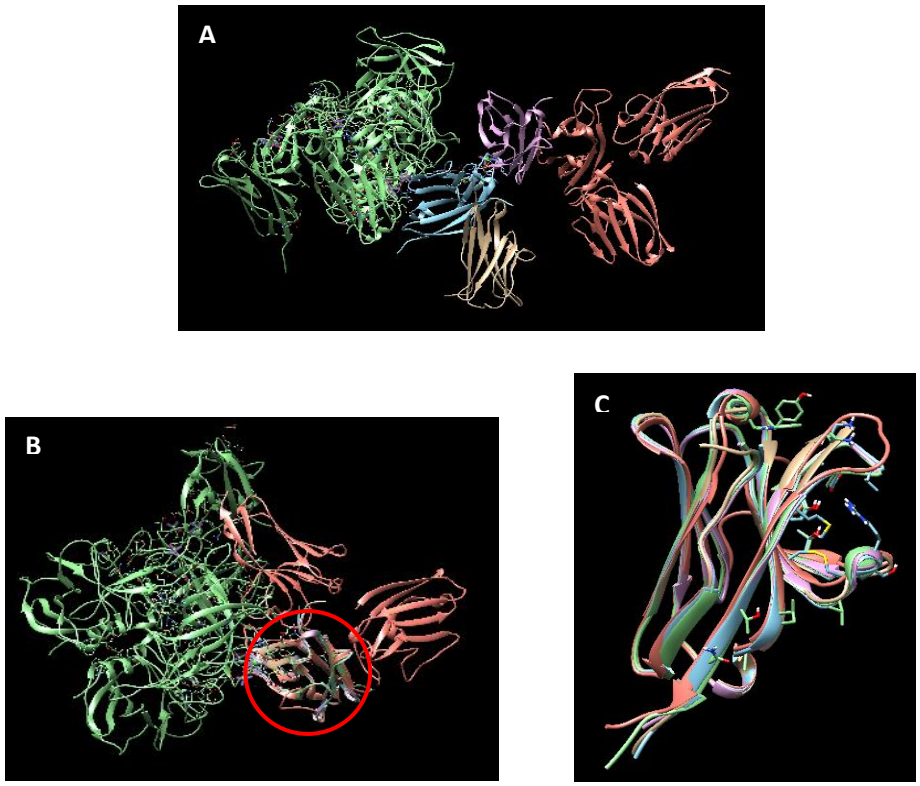


Figura 11. Models tridimensionals de les molècules CD300*. Estructures tridimensionals dels receptors CD300 generats mitjançant el software Chimera. A) Estructures 3D que presenten les diferents molècules de la família CD300: 2NMS (marró), 2Q87 (vermell), 6C74 (blau), 1ZOX (violeta) i 6E47 (verd). B) Estructures alineades. En vermell s'encercla la regió d'interès. C) Regió de les estructures que*

4.3 Clusterització de CD300 i altres molècules no pertanyents a aquesta família

Tal i com es mostra a la Figura 12, en el mateix llindar on les molècules CD300 s'han unit en un sol clúster, també ho han fet dues molècules que no pertanyen a aquesta família, anomenades TREML4 de ratolí i TREML4 d'humà.

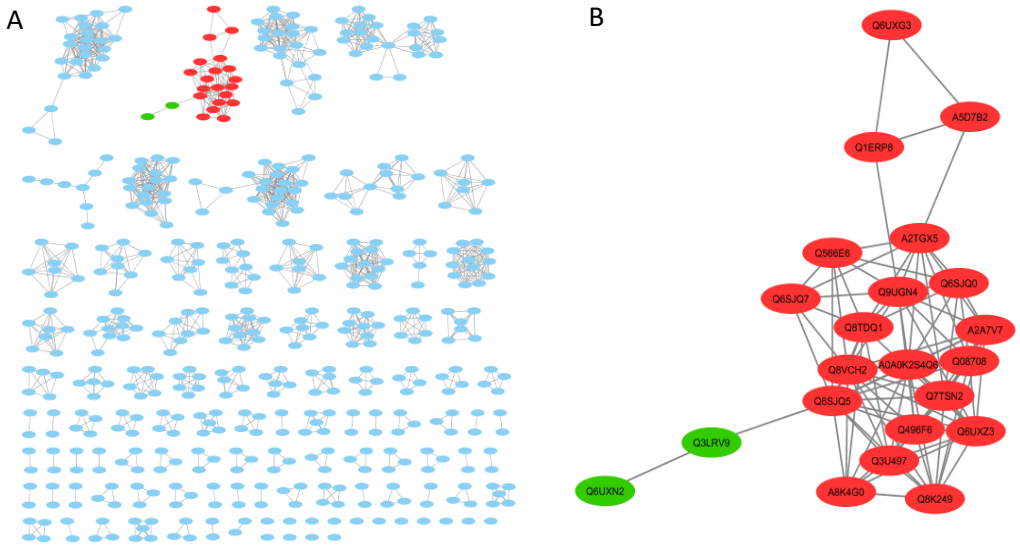


Figura 12. Imatges de la xarxa de molècules amb domini IgV a un llindar d'identitat de 25. *A) Xarxa de Similitud de Seqüència de les molècules que contenen el domini variable d'immunoglobulina. En vermell s'observen les molècules de la família CD300 i en verd TREML4 humà i de ratolí. B) Clúster de receptors CD300 ampliat on es poden observar les dues molècules TREML4 que també s'hi uneixen.*

Per comparar les estructures de CD300 i TREML4, s'han dut a terme el mateix procés esmentat anteriorment; s'han buscat al PDB les estructures tridimensionals que presenten aquestes molècules i s'han alineat entre elles per obtenir un model consens. Per les molècules TREML4 s'ha obtingut una estructura tridimensional que presenta l'identificador PDB 1HKF i els valors d'identitat per TREML4 humà i de ratolí han estat 48,54 i 46,23%, respectivament. A la Figura 13 s'observa que 1HKF té una estructura similar a 1ZOX, 6C74 i 2NMS i que, un cop fet l'alineament, presenten una regió compartida.

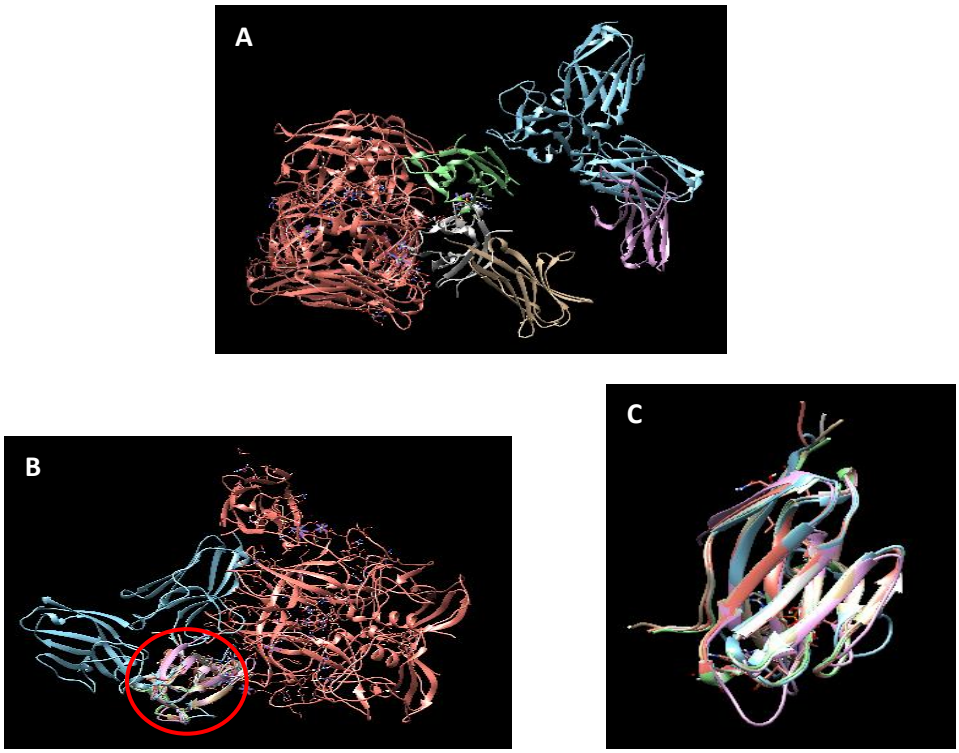


Figura 13. Models tridimensionals de les molècules CD300 i TREML4*. Estructures tridimensionals dels receptors CD300 i TREML4 generats mitjançant el software Chimera. A) Estructures 3D que presenten les diferents molècules de la família CD300: 2NMS (marró), 2Q87 (blau), 6C74 (gris), 1ZOX (verda) i 6E47 (vermell) i la molècula TREML4: 1HKF (violeta). B) Estructures alineades. En vermell s'encercla la regió d'interès. C) Regió de les estructures que s'alinea.*

4.4 Eina implementada vs. Arbre filogènic

El resultat obtingut de la construcció de arbre filogènic es mostra a la Figura 14 i les molècules escollides a l'atzar utilitzades per construir-lo es mostren a la remarcades a la Figura 15A, sota un llindar de 29, i a la Figura 15B sota un llindar de 37.

Observant l'arbre filogènic es poden diferenciar tres blocs que parteixen d'un ancestre comú. Un primer grup és el format per cinc molècules CD300 (vermell), entre les quals es troben CD300G (blau cel), i les dues molècules TREML4 (marró). El segon grup està constituït per dues molècules PSG (violeta) i el tercer per les molècules restants.

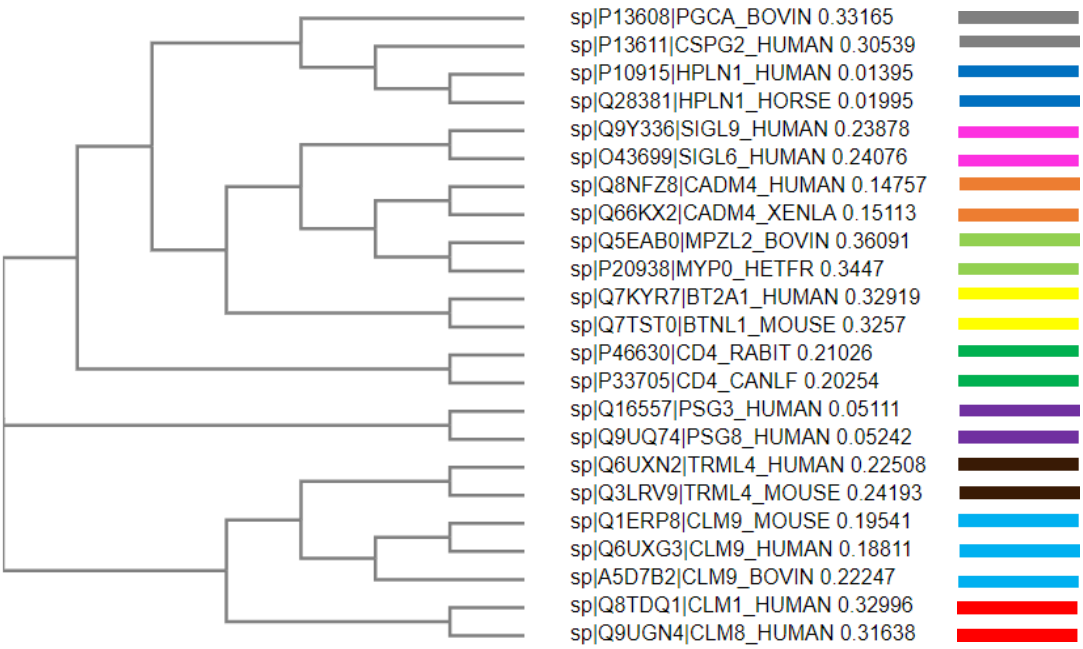
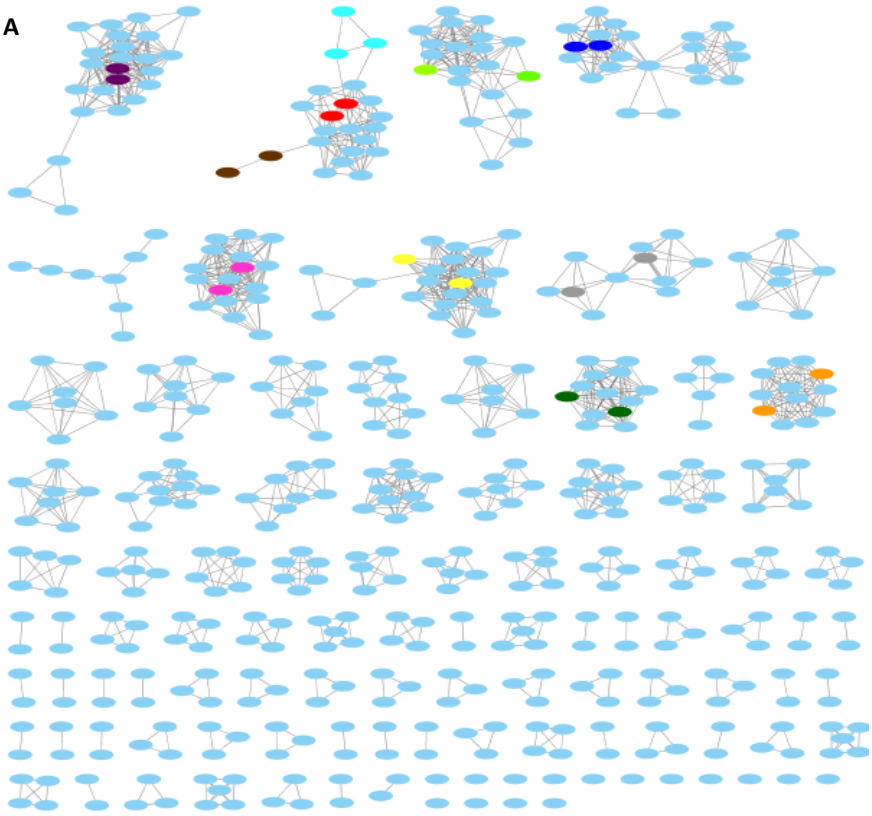


Figura 14. Arbre filogènic. *Arbre filogènic de vint molècules escollides a l'atzar de la xarxa de similitud creada anteriorment creat mitjançant l'aplicació web Clustalw. A la dreta de cada identificador es mostren els colors que els identifiquen en la següent figura (Figura 15).*



B

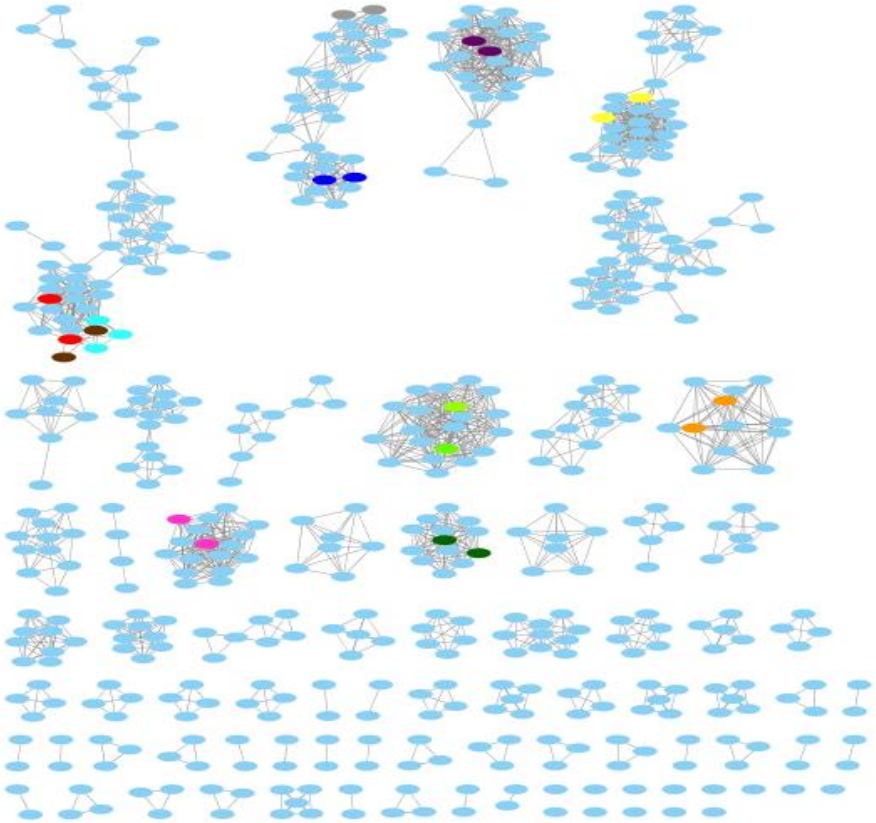


Figura 23. Xarxa d'IgV. *Es mostren ressaltats amb diferents colors les molècules que s'han usat per generar l'arbre filogènic A) sota un llindar de similitud de 29 B) i 37.*

5. DISCUSSIÓ

Programa

El programa informàtic desenvolupat en aquest treball permet treballar amb un gran nombre de seqüències de manera més còmode en comparació amb altres eines de generació de Xarxes de Similitud de Seqüències, com per exemple el software Pythoscape. Això es deu a que el programa desenvolupat utilitza un mòdul d'alineament per parelles (*pairwise2)*, fet que redueix en gran mesura tan el nombre d'alineaments com la quantitat d'informació obtinguda si és compara amb Pythoscape, que realitza alineaments múltiples amb l'eina BLAST.

Aquest programa permet, gràcies al fet de relacionar visualment un conjunt de molècules en relació a la seva similitud de seqüència, estudiar la funció de seqüències desconegudes mitjançant la seva relació amb molècules de funció definida. També, l'observació de la xarxa a diferents llindars de similitud, ens permet intuir quin tipus d'evolució han patit les diferents molècules.

Clusterització de les molècules CD300

Tot i formar part de la mateixa família s'ha observat mitjançant les xarxes generades que les molècules CD300 no s'agrupen en un sol clúster fins a arribar a un llindar d'identitat considerat límit, com és el valor de 29. Això pot ser degut principalment a que per generar aquesta xarxa s'ha utilitzat la seqüència completa de cada molècula enlloc d'únicament la regió del domini IgV. Com que les molècules CD300 presenten dominis diferents entre elles, és d'esperar que presentin una baixa similitud de seqüència si s'analitza tota aquesta. Així doncs, per obtenir un resultat més acurat, una opció possible de continuació del treball seria realitzar de nou la xarxa amb les mateixes molècules però utilitzant únicament les regions on es trobin els dominis IgV. D'aquesta manera es podrà observar si aquests dominis presenten variabilitat entre els receptors de la família CD300 o, pel contrari, romanen estables.

La visualització de la clusterització dels receptors CD300 a diferents llindars de similitud permet identificar com a divergent el tipus d'evolució que han sofert. A llindars superiors les proteïnes es troben separades individualment o en petits clústers de CD300 i a mesura que el llindar va disminuint, les molècules es van agrupant fins a formar un sol clúster. Aquest esquema és comparable al recorregut de l'evolució divergent, que pren un ancestre comú que va patint canvis al llarg del temps fins a obtenir diverses molècules amb funció i estructura similar però

amb variacions en la seqüència. Aquest fet també podria explicar el baix grau de similitud que presenten aquestes molècules entre elles.

Un altre fet destacable de l'agrupació de CD300 ha estat que en llindars superiors, els receptors es troben agrupats en diversos clústers que s'acaben unint en un de sol a un llindar de 37, a excepció dels ortòlegs CLM-9 humà, de ratolí i boví, que es mantenen separats en un clúster de tres nodes fins al llindar 29, tal com mostra la Figura 11. Probablement, aquest fet estigui degut a que aquestes molècules són les úniques de la seva família que presenten un domini extracel·lular similar al de les mucines, anomenat 'Mucin-like domain'. Per tant, el fet que s'agrupin per separat fins a baixos llindars esdevé un resultat coherent.

Clusterització de CD300 i altres molècules no pertanyents a aquesta família

Tal i com mosta la Figura 12, quan els receptors CD300 s'unifiquen en un sol clúster a un llindar de 29, també s'hi uneixen dues molècules que no formen part d'aquesta família, els ortòlegs TREML4 humà i de ratolí que pertanyen a la família TREM (receptors desencadenants expressats en cèl·lules mieloides). Aquestes proteïnes presenten un domini extracel·lular de tipus IgV, un domini de transmembrana i una cua citoplasmàtica curta que no presenta cap motiu de senyalització conegut (Ramirez-Ortiz *et al.*, 2015). El fet que s'agrupin en un mateix clúster pot ser degut a que ambdues famílies de receptors comparteixen la característica d'unió a lípids i seqüències similars (annex V), per tant, poden presentar un domini d'unió similar (Cannon, O'driscoll and Litman, 2012). Aquesta hipòtesi es reforça a l'observar-se la Figura 13, on s'aprecia una similitud entre estructures de CD300 i TREML4.

Eina implementada vs. Arbre filogènic

Comparant la xarxa i l'arbre, es pot establir que el primer mètode ha funcionat correctament, ja que presenta uns resultats molt similars als de l'arbre, pel que fa a l'organització de les molècules. Per una banda, es pot observar a la Figura 15 com, sota un llindar de similitud de 29, les molècules que en l'arbre es troben unides al primer bloc, es troben agrupades al mateix clúster. Per altra banda, tal i com mostra la Figura 16, quan el llindar disminueix fins a un valor de 27, un subgrup de proteïnes del tercer bloc (gris i blau) s'agrupen en un mateix clúster, les quals també apareixen unides per un ancestre més proper a l'arbre filogènic.

Aquesta comparativa permet verificar els resultats obtinguts mitjançant la Xarxa de Similitud de Seqüència, ja que s'obtenen resultats molt semblants si es compara amb un arbre filogènic generat mitjançant ClustalW.

6. CONCLUSIONS

Aplicant l'eina desenvolupada a un a unes molècules d'exemple com és la família de receptors CD300 i comparant els resultats amb els obtinguts mitjançant la generació d'un arbre filogènic, es pot concloure que l'eina implementada permet una visualització de la clusterització de proteïnes en funció a la seva similitud de seqüència. A més, aquesta eina permet treballar amb un nombre elevat de seqüències sense suposar un alt cost computacional.

Per tal de fer més precisos els resultats obtinguts, una possible modificació seria la de realitzar de nou la xarxa utilitzant com a seqüències únicament les regions que contenen el domini comú, és a dir el domini IgV, per tal d'observar els possibles canvis que pot haver sofert al llarg de la seva evolució.

Finalment, es proposa com a millora de l'eina desenvolupada la implementació de *metadata* a la xarxa, per exemple mitjançant els *GO terms*, que permeti, a més de la classificació per similitud de seqüència, una diferenciació i classificació funcional dels diferents clústers i molècules de la xarxa.

7. REFERÈNCIES

Atkinson, H. J. *et al.* (2009) 'Using sequence similarity networks for visualization of relationships across diverse protein superfamilies', *PLoS ONE*, 4(2). doi: 10.1371/journal.pone.0004345.

Barber, A. E. and Babbitt, P. C. (2012) 'Pythoscape: A framework for generation of large protein similarity networks', *Bioinformatics*, 28(21), pp. 2845–2846. doi: 10.1093/bioinformatics/bts532.

Barclay, A. N. (2002) 'Ig-like domains: Evolution from simple interaction molecules to sophisticated antigen recognition', *Proceedings of the National Academy of Sciences*, 96(26), pp. 14672–14674. doi: 10.1073/pnas.96.26.14672.

Berman, H. M. *et al.* (2000) 'The Protein Data Bank www.rcsb.org', *Nucleic acids research*, 28(1), pp. 235–242. doi: 10.1093/nar/28.1.235.

Bio.pairwise2 (no date). Available at: https://biopython.org/DIST/docs/api/Bio.pairwise2-module.html (Accessed: 15 May 2019).

Bork, P., Sander, C. and Valencia, A. (1993) *Convergent evolution of similar enzymatic function on different protein folds: The hexokinase, ribokinase, !&!I and galactokinase families of sugar kinases*, *Protein Science*. Cambridge University Press. Available at: https://onlinelibrary.wiley.com/doi/pdf/10.1002/pro.5560020104 (Accessed: 4 June 2019).

Borrego, F. (2013) 'The CD300 molecules: An emerging family of regulators of the immune system', *Blood*, 121(11), pp. 1951–1960. doi: 10.1182/blood-2012-09-435057.

Brckalo, T. *et al.* (2010) 'Functional analysis of the CD300e receptor in human monocytes and myeloid dendritic cells', pp. 722–732. doi: 10.1002/eji.200939468.

Cannon, J. P., O'driscoll, M. and Litman, G. W. (2012) 'Specific lipid recognition is a general feature of CD300 and TREM molecules', *Immunogenetics*. doi: 10.1007/s00251-011-0562-4.

Casellas, E. C. (no date) 'Molecular and functional characterization of the immunoreceptors CD300d and CD300f Caracterització molecular i funcional dels immunoreceptors MOLECULAR AND FUNCTIONAL CHARACTERIZATION OF THE IMMUNORECEPTORS CD300d and CD300f'.

Chapman, B. and Chang, J. (2000) 'Biopython : Python tools for computation biology Parsers for Biological Data', *ACM SIGBIO Newsletter*, (August), pp. 1–8.

Clark, G. J. *et al.* (2009) 'The CD300 molecules regulate monocyte and dendritic cell functions', *Immunobiology*, 214(9–10), pp. 730–736. doi: 10.1016/j.imbio.2009.06.004.

Clark, G. J. *et al.* (no date) 'The CD300 family of molecules are evolutionarily significant regulators of leukocyte functions'. doi: 10.1016/j.it.2009.02.003.

Cock, P. J. A. *et al.* (2009) 'Biopython: freely available Python tools for computational molecular biology and bioinformatics', 25(11), pp. 1422–1423. doi: 10.1093/bioinformatics/btp163.

Comas-Casellas, E. *et al.* (2012) 'Cloning and characterization of CD300d, a novel member of the human CD300 family of immune receptors', *Journal of Biological Chemistry*, 287(13), pp. 9682–9693. doi: 10.1074/jbc.M111.279224.

Daugelaite, J., O' Driscoll, A. and Sleator, R. D. (2013) 'An Overview of Multiple Sequence Alignments and Cloud Computing in Bioinformatics', *ISRN Biomathematics*, 2013, pp. 1–14. doi: 10.1155/2013/615630.

Feng, D.-F. and Doolittle, R. F. (1990) '[23] Progressive alignment and phylogenetic tree construction of protein sequences', *Methods in Enzymology*. Academic Press, 183, pp. 375–387. doi: 10.1016/0076-6879(90)83025-5.

Fu, L. *et al.* (2012) 'CD-HIT: Accelerated for clustering the next-generation sequencing data', *Bioinformatics*, 28(23), pp. 3150–3152. doi: 10.1093/bioinformatics/bts565.

Graumann, P. and Marahiel, M. A. (1996) 'A case of convergent evolution of nucleic acid binding modules', *BioEssays*. John Wiley & Sons, Ltd, 18(4), pp. 309–315. doi: 10.1002/bies.950180409.

Howe, D. and Rhee, S. Y. (2008) 'The Future of biocuration', *Nature*, 455(October), p. 5.

Jain, A. K., Murty, M. N. and Flynn, P. J. (2000) *Data Clustering: A Review*. Available at: http://eprints.iisc.ernet.in/273/1/p264-jain.pdf (Accessed: 21 May 2019).

Jordán, J. (2003) *Apoptosis: muerte celular programada ÁMBITO FARMACÉUTICO*, *OFFARM*. Available at: https://previa.uclm.es/profesorado/jjordan/pdf/review/10.pdf (Accessed: 4 April 2019).

Kluyver, T. *et al.* (2016) 'Jupyter Notebooks-a publishing format for reproducible computational workflows'. doi: 10.3233/978-1-61499-649-1-87.

Lankry, D. *et al.* (2010) 'Expression and Function of CD300 in NK Cells', *The Journal of Immunology*, 185(5), pp. 2877–2886. doi: 10.4049/jimmunol.0903347.

Li, K.-B. (2003) 'ClustalW-MPI: ClustalW analysis using distributed and parallel computing', *BIOINFORMATICS APPLICATIONS NOTE*, 19(12), pp. 1585–1586. doi: 10.1093/bioinformatics/btg192.

Lobo, I. (2008) 'Basic Local Alignment Search Tool (BLAST)', *Nature Education*, p. 9. Available at: http://csc.columbusstate.edu/carroll/7840/private/papers/BasicLocalAlignmentSearchTool-BLAST.pdf (Accessed: 20 May 2019).

Mahapatro, G. *et al.* (2012) 'Phylogenetic Tree Construction for DNA Sequences using Clustering Methods', *Procedia Engineering*, 38, pp. 1362–1366. doi: 10.1016/j.proeng.2012.06.169.

Marx, V. (2013) 'Biology: The big challenges of big data', *Nature*. Available at: http://pic.b.qs1401.com/42548/pdf/bigbioldata_nature13.pdf (Accessed: 3 June 2019).

McKinney, W. (no date) *Python for data analysis*. Available at: https://books.google.es/books?hl=es&lr=&id=v3n4_AK8vu0C&oi=fnd&pg=PR3&dq=Python+for+data+analysis:+Data+wrangling+with+Pandas,+NumPy,+and+IPython&ots=rgIM3ouyqv&sig=UnI4bnsAEPccel6g7sRSVAYNl8I#v=onepage&q=Python for data analysis%3A Data wrangling with Pandas%2C NumPy%2C and IPython&f=false (Accessed: 17 May 2019).

Munro, H. N. (Hamish N. (1969) *Mammalian protein metabolism. Volume III*. Available at: https://books.google.es/books?hl=es&lr=&id=FDHLBAAAQBAJ&oi=fnd&pg=PA21&dq=protein+evolution&ots=blgsXKX4gA&sig=fhd-SNtjM6Shmb5j9KgBT4vKsm0#v=snippet&q=convergent&f=false (Accessed: 4 June 2019).

Murakami, Y. *et al.* (2014) 'CD300b regulates the phagocytosis of apoptotic cells via phosphatidylserine recognition', *Cell Death and Differentiation*. Nature Publishing Group, 21(11), pp. 1746–1757. doi: 10.1038/cdd.2014.86.

Newman, M. (2018) *Networks*. Oxford University Press. doi: 10.1093/oso/9780198805090.001.0001.

Niizuma, K. *et al.* (2015) 'Identification and Characterization of CD300H , a New Member of the Human CD300 Immunoreceptor Family *', 290(36), pp. 22298–22308. doi: 10.1074/jbc.M115.643361.

Park, S.-Y. and Kim, I.-S. (2017) 'Engulfment signals and the phagocytic machinery for apoptotic cell clearance.', *Experimental & molecular medicine*. Korean Society for Biochemistry and Molecular Biology, 49(5), p. e331. doi: 10.1038/emm.2017.52.

Perez, F. *et al.* (no date) *Project Jupyter: Computational Narratives as the Engine of Collaborative Data Science*. Available at: http://www-01.ibm.com/software/analytics/spss (Accessed: 17 May 2019).

Pettersen, E. F. *et al.* (2004) 'UCSF Chimera-A Visualization System for Exploratory Research and Analysis', *J Comput Chem*, 25, pp. 1605–1612. doi: 10.1002/jcc.20084.

Ramirez-Ortiz, Z. G. *et al.* (2015) 'TREML4 amplifies TLR7-mediated signaling during antiviral responses and autoimmunity HHS Public Access', *Nat Immunol*, 16(5), pp. 495–504. doi: 10.1038/ni.3143.

Smoot, M. E. *et al.* (2011) 'Cytoscape 2.8: New features for data integration and network visualization', *Bioinformatics*, 27(3), pp. 431–432. doi: 10.1093/bioinformatics/btq675.

Takahashi, M. *et al.* (2013) 'Human CD300C delivers an Fc receptor-γ-dependent activating signal in mast cells and monocytes and differs from CD300A in ligand recognition', *Journal of Biological Chemistry*, 288(11), pp. 7662–7675. doi: 10.1074/jbc.M112.434746.

The UniProt Consortium (2017) 'UniProt: the universal protein knowledgebase.', *Nucleic acids research*. Oxford University Press, 45(D1), pp. D158–D169. doi: 10.1093/nar/gkw1099.

Tian, L. *et al.* (2014) 'P85α recruitment by the CD300f phosphatidylserine receptor mediates apoptotic cell clearance required for autoimmunity suppression', *Nature Communications*. Nature Publishing Group, 5, pp. 1–15. doi: 10.1038/ncomms4146.

Vitallé, J. *et al.* (2018) 'CD300 receptor family in viral infections', *European Journal of Immunology*, 2, p. eji.201847951. doi: 10.1002/eji.201847951.

Van Der Walt, S., Colbert, S. C. and Varoquaux, G. (2011) 'The NumPy array: A structure for efficient numerical computation', *Computing in Science and Engineering*, 13(2), pp. 22–30. doi: 10.1109/MCSE.2011.37.

Whitford, D. (2005) *Proteins : structure and function*. J. Wiley & Sons.

Zenarruzabeitia, O. *et al.* (2016) 'The expression and function of human CD300 receptors on blood circulating mononuclear cells are distinct in neonates and adults', *Nature Publishing Group*. Nature Publishing Group, (September), pp. 1–12. doi: 10.1038/srep32693.

ANNEX

Annex I. Llistat de molècules CD300 usades per l'estudi

Taula 4. *Taula on es mostren les 20 molècules CD300 usades en l'estudi.*

UniProt ID	Nomenclatura molècula (UniProt)	Espècie
Q6SJQ7	CLM1 (CD300LF)	*Mus musculus*
Q566E6	CLM1 (CD300F)	*Rattus norvegicus*
Q8TDQ1	CLM1 (CD300F)	*Homo sapiens*
Q8K249	CLM2 (CD300LE)	*Mus musculus*
Q496F6	CLM2 (CD300E)	*Homo sapiens*
Q6SJQ5	CLM3 (CD300LD3)	*Mus musculus*
Q7TSN2	CLM4 (CD300C2)	*Mus musculus*
Q8VCH2	CLM5 (CD300LD)	*Mus musculus*
Q6UXZ3	CLM5 (CD300D)	*Homo sapiens*
A2A7V7	CLM6 (CD300C)	*Mus musculus*
Q08708	CLM6 (CD300C)	*Homo sapiens*
Q3U497	CLM7 (CD300LB)	*Mus musculus*
A8K4G0	CLM7 (CD300B)	*Homo sapiens*
Q6SJQ0	CLM8 (CD300LA)	*Mus musculus*
A2TGX5	CLM8 (CD300A)	*Rattus norvegicus*
Q9UGN4	CLM8 (CD300A)	*Homo sapiens*
Q1ERP8	CLM9 (CD300LG)	*Mus musculus*
Q6UXG3	CLM9 (CD300LG)	*Homo sapiens*
A5D7B2	CLM9 (CD300LG)	*Bos taurus*
A0A0K2S4Q6	CD300H	*Homo sapiens*

Se'n especifica l'identificador de la base de dades UniProt, la seva nomenclatura i l'espècie a la qual pertanyen.

Annex II: Nomenclatura de les molècules CD300

Taula 5. *Nomenclatures per anomenar les molècules CD300 humanes*

Nomenclatura CD	Noms alternatius	HGNC
CD300a	CLM8, CMRF-35H, IRp60, IRC1, IRC2, IGSF12, CMRF-35-H9	CD300A
CD300b	CD300lb, IREM-3, TREM5, CLM7	CD300LB
CD300c	CLM6, CMRF-35A, LIR, IGSF16	CD300C
CD300d	CLM5, CD300ld, CMRF35A4	CD300LD
CD300e	CLM2, IREM-2, CLM2	CD300E
CD300f	CLM1, CD300lf, IREM-1, IgSF13, CLM1, NKIR	CD300LF
CD300g	CLM9, CD300lg, nepmucin	CD300LG
CD300h		CD300H

HGNC: Human Gene Nomenclature Comitte

Taula 6. *Nomenclatures per anomenar les molècules CD300 de ratolí*

Nomenclatura NCBI i MGI	Noms alternatius
CD300a	CLM-8, LMIR-1, MAIR-I
CD300lb	CLM-7, LMIR-5, CD300b
CD300c	CLM-6
CD300ld	CLM-5, LMIR-4, MAIR-IV
AF251705	CLM-4, LMIR-2, MAIR-II, DIgR1, CD300d
CD300lh	CLM-3, LMIR-7
CD300e	CLM-2
CD300lf	CLM-1, DIgR2, LMIR-3, MAIR-V
CD300lg	CLM-9, nepmucin

NCBI: National Center for Biotechnology Information; MGI: Mouse Genome Informatics

Annex III: Script del programa generat

```
#!/usr/bin/env python
# coding: utf-8


# Import functions
import NPS_FUNCTIONS
from Bio.UniProt import GOA

fasta_file = "igv-domain.fasta"
out_file = 'igv-dom'

# Read fasta file
sequences = NPS_FUNCTIONS.readFasta(fasta_file)

# Identity matrix
matrix = NPS_FUNCTIONS.seqIDmatrix(sequences)

# Binary matrix
BM = NPS_FUNCTIONS.matrixPId(matrix, threshold=37)

# Binary matrix to Cytoscape
NPS_FUNCTIONS.binary2Cytoscape(fasta_file, BM, sequences, output_file=out_file+'_37.txt')
```

Annex IV: Script de les funcions utilitzades pel programa

```
#!/usr/bin/env python
# coding: utf-8

# Import functions
from Bio import SeqIO
import numpy as np
from Bio import pairwise2
import json


def readFasta(fasta_file):
    """

    Function to read a fasta file and save the sequences.

    Inputs:
        fasta_file (fasta): Fasta file that contains sequences.

    Output:
        seqs (list): List of sequences.


    """
    if not isinstance(fasta_file,str):
        raise ValueError('The input file is not a valid fasta file')

    seqs = []
    for seq_record in SeqIO.parse(fasta_file, "fasta"):
        seqs.append(seq_record)
    return seqs


def seqID(seq1, seq2):
    """

    Function to align the sequences using the module pariwise2 global alignment.
    Also, this function save the maximum identity value of each alignment.

    Inputs:
        seq1,seq2 (strings, Biopython sequence objects or lists): Aminoacid sequences from a molecule.

    Output:
        values (list): List of maximum identity values from each alignment.


    """
    alignments = pairwise2.align.globalxx(seq1.seq, seq2.seq)
    values = []
    for aln in alignments:
        min_len = float(min(len(aln[0]), len(aln[1])))
        identity = (aln[2]/min_len)*100
        values.append(identity)
    return max(values)
```

```python
def seqIDmatrix(sequences):
    """

    Function to create a numpy array with the values of values list (maximum identity values).

    Inputs:
        sequences (list): List of sequences from a fasta file.

    Output:
        idmatrix (np.array): Numpy array containing the identity values.


    """
    if not isinstance(sequences,list):
        raise ValueError('The input is not valid')

    idmatrix = np.zeros((len(sequences), len(sequences)))
    for i in range(len(sequences)):
        for j in range(i,len(sequences)):
            if i == j:
                idmatrix[i,j] = 100.0
            else:
                idmatrix[i,j] = seqID(sequences[i], sequences[j])
                idmatrix[j,i] = idmatrix[i,j]
    return idmatrix
```

```python
def matrixPId(idmatrix,threshold=None):
    """

    Function to convert a matrix into its binary form based on a threshold value.

    Inputs:
        idmatrix (np.array): Square matrix conataining identity values.
        threshold (float): Value to filter the identity values

    Output:
        bmatrix (np.array): Binary matrix


    """
    if not isinstance(idmatrix,np.ndarray):
        raise ValueError('The input matrix is not a valid numpy array')

    if threshold == None:
        raise ValueError('Please input a threshold value to use matrixPId')

    #Check matrix quadrature
    assert idmatrix.shape[0] == idmatrix.shape[1]

    #Create zero filled matrix
    bmatrix = np.zeros_like(idmatrix)
    for i in range(idmatrix.shape[0]):
        for j in range(i, idmatrix.shape[0]):
            if i == j:
                bmatrix[i,j] = 1
            else:
                if idmatrix[i,j] >= threshold:
                    bmatrix[i,j] = 1
                    bmatrix[j,i] = 1

    return bmatrix
```

```
def binary2Cytoscape(fasta_file, bmatrix, sequences, output_file=None):
    """

    Function to convert the binary matrix to the format required by Cytoscape to visualize it.

    Inputs:
        bmatrix (np.array): binary matrix.
        sequences (list): List of sequences from a fasta file.
        output_file (txt): Text file.

    Output:
        bmatrix (np.array): Binary matrix


    """

    if not isinstance(bmatrix,np.ndarray):
        raise ValueError('The input matrix is not a valid numpy array')

    if output_file == None:
        raise ValueError('Please enter an output file')

    idValues = []
    for seq_record in SeqIO.parse(fasta_file, "fasta"):
        idValues.append(seq_record.id.split('|')[1])

    written = []
    with open(output_file,'w') as of:

        of.write('source\ttarget\n')

        for i in range(len(sequences)):
            for j in range(i+1,len(sequences)):
                if bmatrix[i,j] == 1:
                    of.write(idValues[i]+'\t'+idValues[j]+'\n')
                    if idValues[i] not in written:
                        written.append(idValues[i])
                    if idValues[j] not in written:
                        written.append(idValues[j])
        for v in idValues:
            if v not in written:
                of.write(v+'\n')
```

Annex V: Alineament múltiple entre CD300 i TREML4