SEQÜÈNCIES AMB FI

ANNA SANJUAN VILAPLANA
MÀSTER EN BIOINFORMÀTICA I BIOESTADÍSTICA
BIOINFORMÀTICA FARMACÈUTICA

MARIA JESÚS MARCO GALINDO
MELCHOR SANCHEZ MARTINEZ

05/06/2018

FICHA DEL TRABAJO FINAL

Título del trabajo:	*Seqüències amb fi*
Nombre del autor:	*Anna Sanjuan Vilaplana*
Nombre del consultor/a:	*Melchor Sanchez Martinez*
Nombre del PRA:	*Maria Jesús Marco Galindo*
Fecha de entrega (mm/aaaa):	06/2018
Titulación:	*Màster en Bioinformàtica i Bioestadística*
Área del Trabajo Final:	*Bioinformàtica Farmacèutica*
Idioma del trabajo:	*Català*
Palabras clave:	*Machine Learning (ML); Base de dades (BD); Seqüència polipeptídica (SP)*

Resumen del Trabajo (máximo 250 palabras):

Els pèptids es presenten com a unes de les molècules terapèutiques prometedores donats els avantatges que presenten front altres molècules: penetració cel·lular, toxicitat, vida mitjana, solubilitat i immunogenicitat...

La predicció *in silico* de la toxicitat, interacció pèptid-proteïna i la funció biològica té molt de pes a l'estadi inicial del procés d'obtenció de pèptids terapèutics.

No existeix un mètode estandarditzat que ens indiqui quin és l'algoritme òptim per realitzar aquestes prediccions això com quina metodologia d'anàlisi s'hauria de seguir. Tanmateix, l'algoritme *Support Vector Machine* (SVM) és el més utilitzat a l'hora de predir les anteriors característiques peptídiques juntament amb l'anàlisi de les seqüències atenent a la seva composició aminoacídica i dipeptídica.

Els resultats d'aplicar els mètodes de *Machine Learning* (*ML*) dependran en gran mesura, d'una banda, dels paràmetres que es fixen a l'hora de executar aquests, i d'altra, de les característiques i estructura que presenta el nostre conjunt de dades a analitzar. La creació d'una base de dades comú no redundant i baix un mateix format milloraria l'avanç d'aquest camp en l'estudi de la predicció *in silico* de pèptids terapèutics.

Abstract (in English, 250 words or less):

Peptides are presented as one of the promising therapeutic molecules given the advantages over other molecules: cell penetration, toxicity, mèdium-life, solubility and immunogenicity ...

In silico prediction of peptides' toxicity, peptide-protein interaction and biological function has a lot of weight in the initial stage of the process of obtaining therapeutic peptides.

There is not a standardized method with the optimal algorithm performing these predictions; neither the method of analysis that should be followed. However, *Support Vector Machine* algorithm (*SVM*) has been the mostly used and performed algorithm. In turn, to predicting the previous peptide characteristics researchers usually performthe analysis of the sequences based on their amino acid and dipeptide composition.

The results of applying *Machine Learning* (*ML*) methods will depend, on the parameters that are set to execute these ones, and, on the characteristics and structure that our dataset has. Creating a common, non-redundant database under the same format would improve the progress of this field in the study of *in silico* sequence-based prediction of therapeutic peptides.

Índice

Lista de figuras

1. Introducción

1.1 Context i justificació del TFM

En els darrers anys, el *Machine Learning* (*ML*) s'ha aplicat a un considerablement nombre de diferents àrees de coneixement biològic per millorar l'assistència sanitària[1].

Han passat més de vint anys des que es va introduir el primer pèptid terapèutic dissenyat com a producte farmacèutic[2]. Així doncs, moltes línies d'investigació científica s'ha centrat en desenvolupar teràpies basades en la interacció pèptid-proteïna pel tractament de malalties[3].
Els pèptids es presenten com a molècules terapèutiques prometedores donats els avantatges que presenten front altres molècules de baix pes molecular. Doncs, es caracteritzen per ser fàcils de produir, presentar alta penetració i, alta especificitat front proteïnes diana crítiques a moltes malalties[2,3].

Un dels handicaps en el tractament amb pèptids és la seva toxicitat. Així doncs, la predicció de la toxicitat i*n silico* de les possibles interaccions terapèutiques pèptids-proteïnes abans de la seva síntesi es presenta com un pas crític i molt important per estalviar temps i diners durant el desenvolupament de fàrmacs basats en aquests[1–3]. Per tant, la predicció de les interaccions proteïna-pèptid basada només en la informació de la seqüència peptídica és un altre dels importants reptes de la biologia computacional[4–6].

I, tot i que, al principi sols uns quants grups d'investigació participaven en el desenvolupament d'algoritmes d'aprenentatge per a aquest mètode predictiu, cada vegada més grups de recerca i centres implementen mètodes de *ML* i *Deep learning* per analitzar, predir i classificar grans produccions de informació (*big data*). Per aquestes raons i per altres, és un enfocament molt universal tant computacional com experimentalment[3].

Com s'ha indicat anteriorment, d'una banda, les bases de dades així com les eines dels servidors de predicció tenen un rol clau al disseny de nous pèptids amb un paper terapèutic[1].

D'altra banda, en el present i en el futur el nostre coneixement de noves proteïnes és i serà majoritàriament fruit de la predicció[7,8]. No ens sorprèn que la predicció de funcions biològiques com d'altre tipus d'anotacions[9] de proteïnes i pèptids sigui un altre dels reptes important en el camp de la bioinformàtica. Però, aquesta és molt més complexa que altres

classificacions, ja que la similitud de la seqüència aminoàcidica entre pèptids amb la mateixa funció és molt pobra[8,10].
Així doncs, aquest projecte es presenta com un repte combinat on, gràcies a les tècniques, anàlisi i algoritmes de ML intentarem dur a terme una predicció acurada de la toxicitat i activitat de pèptids problema en base a les seves característiques de seqüència i interacció proteïna-pèptid.

Els mètodes de predicció basats en seqüència a nivell aminoacídic s'ha tornat molt popular, doncs, sols requereix conèixer la seqüència d'aminoàcids dels pèptids o proteïnes a analitzar[11]. Així doncs, aquests nous mètodes d'estudi i investigació són i seran una eina molt útil per suplementar estudis presents i futurs en l'àrea de la proteòmica[11].

1.2 Objectius del treball

L' objectiu del TFM primerament era, donada una nova seqüencia polipeptídica curta (pèptid), predir la toxicitat, interaccions i possible funció biològica que aquesta pot tenir i exercir.

Per poder dur a terme totes aquestes prediccions diferents d'una banda extrauríem informació de distintes bases de dades (BD) per generar-ne d'una pròpia, la qual, ens permetrà classificar les seqüències problema atenent a la seva: Toxicitat, Interaccions amb altres amb cadenes polipeptídiques i funció biològica.

Per a aconseguir-ho, caldria crear diversos algoritmes de *Machine Learning* (*ML*) que pugui fer la feina de predicció que necessitem per a cada cas atenent a la informació i classificació que em recolliríem prèviament a una base de dades pròpia.

Finalment, s'establirà la predicció i categorització de la seqüència d'aminoàcids atenent a les característiques de la seva cadena polipeptídica.

1.4 Planificació del treball

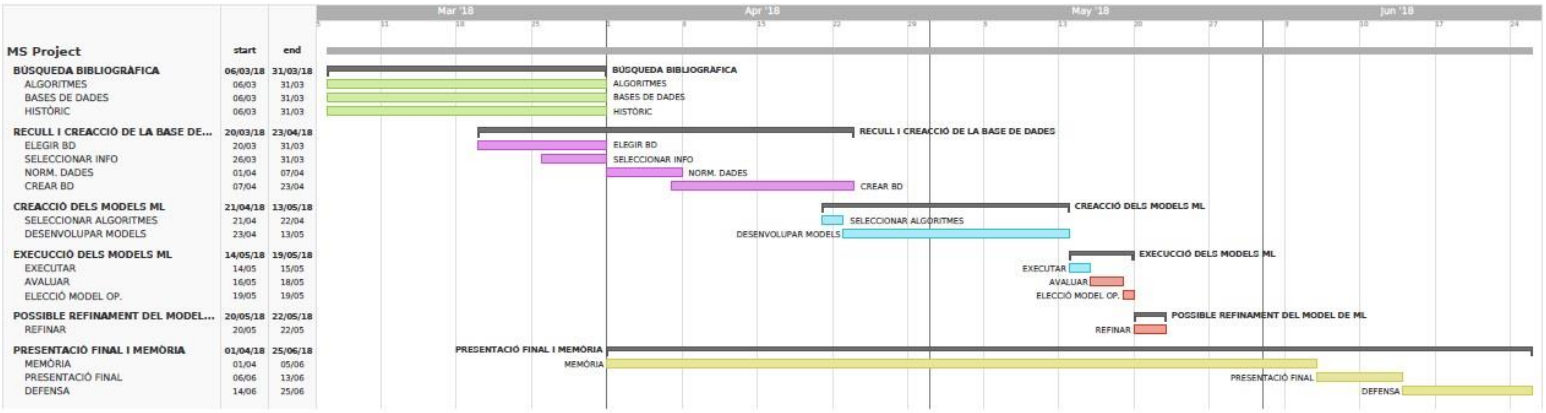
Per a dur a terme aquestes estratègies de treball i assolir els objectius marcats duríem a terme les següents tasques enumerades a l'esquema temporal que les segueix:

1. Cerca bibliogràfica
    a. Buscar informació sobre toxicitat, interacció pèptid-proteïna i activitat peptídica
    b. Buscar informació dels possibles algoritmes de *ML*
    c. Trobar bases de dades de toxicitat, interacció pèptid-proteïna, activitat
2. Recull i creació de la base de dades
    a. Elegir les bases de dades d'extracció

b. Seleccionar-ne la informació útil
c. Normalitzar les dades
d. Desenvolupar la base de dades

3. General el model
   a. Seleccionar els models de *ML* a desenvolupar
   b. Desenvolupar els models de *ML*
4. Execució dels models de *ML*
   a. Executar els models de *ML*
   b. Avaluar els models de *ML*
5. Possible refinament del model de *ML*
   a. Elegir el model òptim
   b. Refinar el model de *ML*

Imatge 1.



1.5 Breu sumari dels productes obtinguts

Els productes que s'adjunten amb aquest treball són:

- *BDatos.txt*: Base De Dades Inicial, recull de la informació col·lapsada a la base de dades *SwissProt*.
- *BD.csv*: Base De Dades Amb La Metodologia D'anàlisi De Seqüència aplicada
- *TFM_AnnaSanjuanVilaplana.Rmd*: Arxiu de *Markdown* amb els *scripts* de programació exemple.

1.6 Breu descripció dels altres capítols de la memòria

2.1. Bases de dades existents

Descripció de les bases de dades que s'han fet servir per construir les distintes bases de dades que hem consultat per a cada tipus d'anàlisis: Toxicitat, Interacció pèptid-proteïna i funció biològica.

2.2 Metodologia d'anàlisi de seqüència

Descripció dels distints anàlisis que s'han dut a terme per tal d'extreure la màxima quantitat d'informació útil de les seqüències polipeptídiques.

2.3 Llenguatges de programació

Breu descripció dels llenguatges de programació més utilitzats per realitzar els anàlisis y prediccions en base a algoritmes de *ML*.

2.4. Algoritmes de predicció

Descripció breu dels avantatges i inconvenients de cada algoritme que s'ha fet servir a l'anàlisi i predicció de seqüències polipeptídiques atenent a la característica que es vulgui estudiar.

2. Resta de capítols

2.1 Bases de dades existents

Toxicitat

La toxicitat dels fàrmacs és fruit, principalment d'efectes secundaris i productes indesitjables que es formen a conseqüència del seu metabolisme. Ara bé, als fàrmacs peptídics, aquesta pot estar generada per dues raons, per Immunogenicitat o per agregació[2]. Així doncs, el disseny d'un pèptid terapèutic passa per analitzar *in silico* que aquest no sigui tòxic[3].

Es coneixen actualment diverses bases de dades i/o predictors d'informació envers la toxicitat de proteïnes i pèptids, alguns són: *Swiss-Prot*[3,12,13], una base de dades en línia que conté informació bioquímica i farmacològica sobre medicaments (*DrugBank*)[14], Base de dades de toxines animals provinents de *UniProtKB/Swiss-Prot* (*Uni-database platform for animal toxins, ATDB*)[3,15], Servidor web de toxines bacterianes (*BTXpred*)[3,13], Servidor web de neurotoxines (*NTXpred*)[3,16], etc.

La majoria de les anotacions en quant a activitat toxicològica es refereix, és troben col·lapsades a la base de dades del servidor web d'*SwissProt*[3,13,15,16], tot i així, com hem observat existeix un conjunt de bases de dades i servidors on també es pot consultar aquesta informació.

Protein-protein Interaction o Interaccions proteïna-proteïna (PPI)

Es parla de *PPI* quan dos o més proteïnes s'uneixen entre elles, normalment per permetre o inhabilitar, a totes o a una, dur a terme la seva funció biològica[5] o procés[17].

Normalment, les bases de dades per *PPI* precisen d'una curació de les dades fruit de col·lapsar diverses fons en una mateixa[5] que ens servirà per realitzar el model de predicció d'aquestes *PPI*.

Actualment, existeixen moltes bases de dades pel que fa a informació de PPI, atenent: formen part del sistema immune *InnateDB* (*PPIs in the Immune System*)[5,18,19]; Formen part de les interaccions a la matriu extracel·lular (*MatrixDB, Extracellular Matrix Interactions Database*)[5,20,21]; Interaccions de proteïnes (*DIP Database of Interacting Proteins*)[5,22,23]; Interaccions moleculars (*MINT, Molecular Interactions Database*)[5,24,25]; Repositori biològic general d'interaccions (*BioGRID, Biological General Repository for Interaction Datasets*)[5,26,27]; Servidor d'accés obert de dades d'interacció molecular (*IntAct*)[5,28,29]; Interaccions entre dominis i pèptids (*A Database of Domain-Peptide Interactions, DOMINO*)[30]; Plataforma d'11 bases de dades d'interaccions moleculars (*MIntAct*)[31,32];

Col·lecció d'interaccions proteïna-proteïna i les seves homòlogues en una o més espècies (*Homologous INTeractions database, HINTdb*)[33]; etc.

D'altra banda, la majoria de les dades es poden exportar en format tabular *PSI-MI*[28,34], cosa que facilita l'obtenció de dades dels distints servidors, normalment baix una estructura *XML*.

Funció biològica

Les funcions d'un producte genètic s'entenen com les habilitats que aquests poden desenvolupar. Mentre que, un procés biològic implica més d'una activitat derivada de productes gènics. En aquest sentit, els productes gènics tenen diferents habilitats o funcions i, treballen conjuntament, per aconseguir diferents processos[35]. Per tant, els processos biològics es defineixen com totes aquelles transformacions - processos físics o reaccions químiques entre d'altres - que es produeixen en sistemes biològics fruit de l'alteració, el consum, la producció i/o la formació d'entitats[36].

Aquestes funcions són molt més complexes de predir que altres característiques de proteïnes i pèptids, doncs, la similitud de la seqüència no és un atribut significatiu entre les proteïnes que tenen la mateixa funció[8,10].

En aquest sentit la base de dades *Gene Ontology Annotation Database* (*GOA*) es presenta com un servidor d'anotacions d'elevada qualitat de l'ontologia gènica (*Gene Ontology, GO*)[35]. El qual, ens proporciona vocabularis controlat per a les anotacions de les característiques moleculars de diferents models d'organismes classificades en els següents grups: *Molecular Function*, *Biological Process* i *Cellular Component*[33].

2.2 Metodologia d'anàlisi de seqüència

Executada anteriorment: Bibliogràfica

Conjoint triad feature[4,11] (*CTF*). Aquest mètode analitza el % de la composició aminoacídica de les seqüències gèniques atenent a les característiques de les cadenes laterals dels aminoàcids que les conformen. Per tant, els aminoàcids es classificaran, atenent a la polaritat i volum de les seves cadenes laterals en els següents grups: "A,G,V"; "I,L,F,P"; "Y,M,T,S"; "H,N,Q,W"; "R,K"; "D,E"; i, finalment, "C".

Amino Acid Composition-based model[3,8,13]. Anàlisi de la composició dels aminoàcids que componen la cadena peptídica atenent a la seva llargada. És a dir, per a cada un del 20 aminoàcids existents quina fracció compon la seqüència.
Dipeptide Composition-based model[3,8,13]. S'aparellen els aminoàcids en els 400 (20x20) possibles dipèptids que poden formar-se'n i es calcula la

fracció que compon a cada un dels possibles dipèptids consecutius que se'n deriven de cada seqüència peptídica.

Ab inition paterns[8]. Es calcula la freqüència dels 16000 possibles tetrapèptids (20x20x20x20) que pot estar present a cada seqüència polipeptídica.

Residu preference[3]. S'analitzen els extrems C-terminal i N-terminal en busca de posicions de preferència d'algun aminoàcid.

Altres estudis han fet servir anàlisis basats en multialineaments de seqüència amb eines com *CLUSTAL-W* o *PSI-BLAST*[13,16].

A més a més, també es fa servir la informació de les propietats fisicoquímiques d'aquestes seqüències com el punt isoelèctric o l'aromaticitat d'aquestes envers als aminoàcids que la componen[1]. Aquest són sols algunes de les moltes metodologies seguides per tal d'analitzar tota la informació útil continguda a les seqüencies polipeptídiques que, normalment, se solen combinar entre elles per obtenir millors resultats de predicció[3,8,13,16].

2.3 Llenguatges de programació

Hi ha una gran varietat de llenguatges de programació que s'utilitzen hui per hui al món científic. Però, els darrer any hi ha dos que són més capdavanters. D'una banda *R*[37], àmpliament conegut pels estadistes i economistes. D'altra banda, *Python*[38], el qual es presenta com a possible substitut del llenguatge *Perl*[39] pel que respecta en el camp de la bioinformàtica.

Python

Hi te descrits infinitats de mòduls envers l'anàlisi de seqüències polipeptídiques i les seves característiques. La majoria de les eines i ferramentes d'aquest llenguatge de programació al camp de la bioinformàtica es troben recollides al conjunt de programaris d'accés obert conegut com *BioPython*[40].

Com a exemple de les aplicacions dels mòduls presents a aquest programari, trobem la ferramenta *ProteinAnalysis* del paquet *SeqUtils* present al mòdul *Bio*. Aquesta ferramenta ens permet obtenir, entre altres, els valors dels punt isoelèctric de les seqüencies polipeptídiques.

R

D'altra banda, *R* ens presenta el programari de desenvolupament obert *Bioconductor*[41], el qual ens dona infinitats de programes i eines per realitzar anàlisi de seqüència.

Anàlogament al cas anterior, el paquet seqinr[42] d'*R* ens permet obtindré el punt isoelèctric d'una seqüència polipeptídica, dada que ens podria ser útil per incorporar-la a l'algoritme de predicció com han fet alguns estudis[4].

Cal dir que la interfície d'*R*, anomenada *RStudio*[43] dota a *R* d'una eina de programació literària, gràcies als paquets *Knitr*[44] i *RMarkdown*[45,46], on es poden combinar scripts dels dos llenguatges de programació anteriors encara que la majora de programadors recomanen fer ús sols d'un llenguatge de programació per minimitzar els errors que se'n poden ocasionar.

2.4 Algoritmes de predicció

Tipus de mètodes de predicció amb *ML*

Taula1.

	AVANTATGES	INCONVENIENTS
SVM (SUPPORT VECTOR MACHINE)	S'adapta a la classificació i/o predicció de problemes numèrics	El millor model requereix provar diverses combinacions de diversos *Kernels* i paràmetres
	No és desajusta per dades amb molt de soroll de fons i no sol sobre ajustar les dades	D'aprenentatge lent a major complexitat del conjunt de dades
	Molt popular, ràpid i amb alts beneficis a concursos de *data mining*	Model de caixa negra, complex i difícil d'interpretar
RT (RANDOM FOREST)	D'ús múltiple, funciona bé per a la majoria de problemes	Precisa d'un ajust del model al data set
	Pot fer servir dades amb soroll de fons, *NA*'s, variables categòriques i/o contínues	Model no intuïtiu
	Selecciona sols les variables més importats	
	Pot fer servir dades amb moltes variables o casos (registres/exemples)	
	Classificador universal, funciona bé en la majoria de problemes	
	Procés d'aprenentatge altament automàtic	
DT (DECISION TREE)	Pot fer servir *NA*'s, característiques numèriques i/o nominals	Esbiaixat a les divisions amb característiques de nivells
	Exclou les característiques que no són importants	Modelatge dependent de les divisions de l'eix paral·lel
	Pot fer-se servir en bases de dades petites i grans	Fàcil de superposar i desajustar
	Es pot interpretar sense coneixement matemàtic	Petits canvis a les dades d'entrenament poden produir grans canvis lògics de decisió
	És més eficient que altres models més complexos	Els arbres grans són difícils d'interpretar i les seves decisions poden semblar contra-intuïtives
NB (NAÏVE BAYES)	Simple, ràpid i molt efectiu	Es basa en la assumpció que totes les característiques són igualment importants i independents
	Funciona bé amb soroll de fons i presència de *NA*'s	No es l'ideal per un conjunt de dades amb moltes variables numèriques
	Requereix poques mostres per al subconjunt d'entrenament i també treballa bé amb conjunts amb moltes mostres	Les probabilitats estimades són menys fiables que les classes predites
	És fàcil obtindre'n una estimació de probabilitat per a una predicció	
NN (NEURAL NETWORK)	És capaç de modelar patrons molt complexos	És un model de caixa negra complex i difícil d'interpretar
	No necessita moltes restriccions pel que fa a les relacions subjacents de les dades	Propens a sobre ajustar les dades d'entrenament

Toxicitat

En investigacions anteriors, s'han realitzat estudis comparatius amb diferents mètodes de *ML* per investigar la toxicitat de proteïnes i pèptids. Entre altres s'han fet servir: *SVM*[1,3,13,16], *NN*[1,16] i *RF*[1].

Però, és l'algoritme *SVM* al que majoritàriament recorren els investigadors a l'hora de realitzar aquests mètodes d'anàlisi[3,13] de toxicitat en pèptids, tot i que se solen realitzar models híbrids amb *SVM* per realitzar-los[1].

PPI

Si recorrem a la bibliografia existent podem observar que s'han realitzat estudis comparatius amb diferents mètodes de *ML* per investigar aquestes *PPI*[17]. Entre altres s'han fet servir: *SVM*[4,5,11,17,33], *NB*[5,17], *ANN*[17], *DT*[5,17] i *RF*[5,17].

Igualment que amb l'elecció del mètode de *ML* per l'anàlisi de la toxicitat, és el primer de tots els anteriorment citat és al que majoritàriament recorren els investigadors a l'hora de realitzar aquests mètodes d'anàlisi[4,11,33] i amb el que millor es realitza la predicció d'aquestes interaccions proteiques, en termes majoritàriament d'exactitud i precisió[5,17], seguit pel mètode del *RF*[17] o el *NB*[5].

Funció biològica

Pel que fa a la predicció de l'activitat de proteïnes i pèptids, així com d'altres anotacions, mitjançant mètodes de *ML*, hem vist que s'han fet servir majoritàriament mètodes de *SVM*[8,9].

2.5 Enfocament i metodologia

Bases de dades i recursos

Per obtenir les dades de treball s'ha recorregut al servidor d'*SwissProt*, d'on s'ha obtingut el conjunt de dades recollit a l'arxiu *BDatos.txt*. Aquest arxiu conté les seqüències polipeptídiques tòxiques (TOXIC) i no tòxiques (NOTOXIC).

La cerca d'aquestes seqüències s'ha basat en seleccionar totes aquelles entrades revisades d'*Uniprot* (*SwissProt*) que contenien la paraula clau KW-0800 per les seqüències "TOXIC" i NOT KW-0800 per les "NOTOXIC", com s'ha fet a altres estudis[3].

Elecció de la metodologia d'anàlisi a seguir

Anàlisi basat en les característiques de la cadena lateral dels aminoàcids. S'analitza el % de la composició aminoacídica de les

seqüències atenent a les característiques de les cadenes laterals dels aminoàcids que les conformen. L'estudi de les característiques de les cadenes laterals dels aminoàcids que componen una cadena ens dona molta informació sobre com aquesta interaccionarà en un futur amb altres molècules i amb ella mateixa (conformació espacial). És a dir ens pot ajudar a predir les interaccions que pot arribar a establir i la funció que pot dur a terme.

Amino Acid Composition-based model. És un dels mètodes més emprats.

Aquestos anàlisis es troben recollits al conjunt de dades anomenat *BD.csv*. Caracteritzada per contenir els pèptids del conjunt de dades BDatos.txt amb una longitud menor a 35 aminoàcids. A més a més, aquesta base de dades conté els vectors fruits de l'anàlisi basat en les característiques de la cadena lateral dels aminoàcids que componen les seqüències no redundants, i la seva composició aminoacídica (1024 "TOXIC" i 4577 "NOTOXIC").

Llenguatge de programació

Per realitzar aquest treball s'ha volgut fer servir el llenguatge de programació *R*, concretament la interfície *RStudio*.

Algoritme de predicció

Atenent a la informació recollida de la bibliografia existent, en aquest treball s'ha considerat que, atenent a les característiques del nostre conjunt de dades i a l'objectiu d'aquest els models de *ML* a utilitzar seran: *SVM, DC* i *RF*.

2.6 Relació de les desviacions en la temporalització i assoliment d'objectius

Malauradament, en el transcurs d'aquest treball no s'han aconseguit executar correctament els scripts programats per estudiar i analitzar la base de dades creada.

Tanmateix, hem generat un arxiu en format *Markdown* per tal de compartir els scripts generats per tal de predir, a mode d'exemple, la toxicitat del nostre conjunt de dades, atenent a la informació prèvia obtinguda de *SwissProt*.

3. Conclusions

Tot i que en el transcurs de la realització d'aquest treball no hem arribat a assolir els objectius d'aquests hem aprés que la predicció *in silico* de la toxicitat, interacció pèptid-proteïna i la funció biològica té molt de pes a l'estadi inicial del procés d'obtenció de pèptids terapèutics. Doncs, amb millores de les nostres capacitats de comprensió i algoritmes computacionals, el disseny *in silico* de pèptids es presenta des de fa anys com una de les teràpies biològiques més prometedores. Aquests però, hauran de ser provats *in vivo* en estadis posteriors, és a dir, per millorar les seves habilitats i característiques com: penetració cel·lular, toxicitat, vida mitjana, solubilitat i immunogenicitat...etc.

D'una banda posar en manifest que no hi ha un mètode estandarditzat així com no hi ha un algoritme òptim per realitzar aquesta predicció. Tanmateix, l'algoritme *SVM* és el més utilitzat juntament amb l'anàlisi de les seqüències atenent a la seva composició aminoacídica i dipeptídica.

D'altra banda, no hi ha que oblidar que els resultats d'aplicar els mètodes de *ML* dependran en gran mesura, d'una banda, dels paràmetres que es fixen a l'hora de executar aquests, i d'altra, de les característiques i estructura que presenta el nostre conjunt de dades a analitzar.

Finalment, incidir en la importància de crear una base de dades global que col·lapsi la major quantitat d'informació biològica que es pugui en quant a pèptids i proteïnes, doncs, si reduïm l'heterogeneïtat de les bases de dades d'inici pel que fa als anàlisis de predicció, s'unificarà l'avanç en l'estudi d'aquestes teràpies biològiques.

4. Bibliografia

1. Shah, Y., Sehgal, D. & Valadi, J. K. Recent trends in antimicrobial peptide prediction using machine learning techniques. **13,** 415–416 (2017).
2. Roy, A., Nair, S., Sen, N., Soni, N. & Madhusudhan, M. S. In silico methods for design of biological therapeutics. *Methods* **131,** 33–65 (2017).
3. Gupta, S. *et al.* In Silico Approach for Predicting Toxicity of Peptides and Proteins. *PLoS One* **8,** (2013).
4. Shen, J. *et al.* Predicting protein-protein interactions based only on sequences information. *Proc. Natl. Acad. Sci. U. S. A.* **104,** 4337–41 (2007).
5. Saha, I. *et al.* Ensemble learning prediction of protein–protein interactions using proteins functional annotations. *Mol. Biosyst.* **10,** 820–830 (2014).
6. Huang, Y.-A., You, Z.-H., Chen, X., Chan, K. & Luo, X. Sequence-based prediction of protein-protein interactions using weighted sparse representation model combined with global encoding. *BMC Bioinformatics* **17,** 184 (2016).
7. Devos, D. & Valencia, A. Practical limits of function prediction. *Proteins Struct. Funct. Genet.* **41,** 98–107 (2000).
8. Saha, S. & Raghava, G. P. S. VICMpred: An SVM-based method for the prediction of functional proteins of gram-negative bacteria using amino acid patterns and composition. *Genomics, Proteomics Bioinforma.* **4,** 42–47 (2006).
9. Hua, S. & Sun, Z. Support vector machine approach for protein subcellular localization prediction. *Bioinformatics* **17,** 721–728 (2001).
10. Hannenhalli, S. S. & Russell, R. B. Analysis and prediction of functional sub-types from protein sequence alignments. *J. Mol. Biol.* **303,** 61–76 (2000).
11. Wang, Y., Wang, J., Yang, Z. & Deng, N. Sequence-based protein-protein interaction prediction via support vector machine. *J. Syst. Sci. Complex.* **23,** 1012–1023 (2010).
12. Boutet, E., Lieberherr, D., Tognolli, M., Schneider, M. & Bairoch, A. UniProtKB/Swiss-Prot. *Methods Mol. Biol.* **406,** 89–112 (2007).
13. Saha, S. & Raghava, G. P. S. BTXpred: prediction of bacterial toxins. *In Silico Biol.* **7,** 405–412 (2007).
14. Law, V. *et al.* DrugBank 4.0: Shedding new light on drug metabolism. *Nucleic Acids Res.* **42,** D1091–D1097 (2014).
15. He, Q. Y. *et al.* ATDB: A uni-database platform for animal toxins. *Nucleic Acids Res.* **36,** 293–297 (2008).
16. Saha, S. & Raghava, G. P. S. Prediction of neurotoxins based on their function and source. *In Silico Biol.* **7,** 369–387 (2007).
17. Saha, I. *et al.* Evaluation of Machine Learning Algorithms on Protein-Protein Interactions. in *Man-Machine Interactions 3* (ed. Gruca, D. A., Czachórski, T. & Kozielski, S.) 211–218 (Springer International Publishing, 2014). doi:10.1007/978-3-319-02309-0
18. Korb, M. *et al.* The innate immune database (IIDB). *BMC Immunol.* **5,** 9–7 (2008).
19. Breuer, K. *et al.* InnateDB: Systems biology of innate immunity and beyond - Recent updates and continuing curation. *Nucleic Acids Res.* **41,**

D1228–D1233 (2013).
20. Chautard, E., Fatoux-Ardore, M., Ballut, L., Thierry-Mieg, N. & Ricard-Blum, S. MatrixDB, the extracellular matrix interaction database. *Nucleic Acids Res.* **39,** (2011).
21. Chautard, E., Ballut, L., Thierry-Mieg, N. & Ricard-Blum, S. MatrixDB, a database focused on extracellular protein-protein and protein-carbohydrate interactions. *Bioinformatics* **25,** 690–691 (2009).
22. Xenarios, I. DIP: the Database of Interacting Proteins. *Nucleic Acids Res.* **28,** 289–291 (2000).
23. Salwinski, L. The Database of Interacting Proteins: 2004 update. *Nucleic Acids Res.* **32,** 449D–451D (2004).
24. Chatr-aryamontri, a & Chatr-aryamontri, A. MINT: the Molecular INTeraction database. *Nucleic Acids Res.* **35,** D572–D574 (2007).
25. Licata, L. *et al.* MINT, the molecular interaction database: 2012 Update. *Nucleic Acids Res.* **40,** D857–D861 (2012).
26. Stark, C. BioGRID: a general repository for interaction datasets. *Nucleic Acids Res.* **34,** D535–D539 (2006).
27. Chatr-Aryamontri, A. *et al.* The BioGRID interaction database: 2017 update. *Nucleic Acids Res.* **45,** D369–D379 (2017).
28. Kerrien, S. *et al.* IntAct–open source resource for molecular interaction data. *Nucleic Acids Res.* **35,** D561–D565 (2007).
29. Kerrien, S. *et al.* The IntAct molecular interaction database in 2012. *Nucleic Acids Res.* **40,** D841–D846 (2012).
30. Ceol, A. *et al.* DOMINO: A database of domain-peptide interactions. *Nucleic Acids Res.* **35,** D557–D560 (2007).
31. Orchard, S. *et al.* The MIntAct project - IntAct as a common curation platform for 11 molecular interaction databases. *Nucleic Acids Res.* **42,** D358–D363 (2014).
32. Licata, L. & Orchard, S. The MintAct project and molecular interaction databases. *Methods Mol. Biol.* **1415,** 55–69 (2016).
33. Urquiza, J. M. *et al.* Using machine learning techniques and genomic/proteomic information from known databases for defining relevant features for PPI classification. *Comput. Biol. Med.* **42,** 639–650 (2012).
34. Kerrien, S. *et al.* Broadening the horizon - Level 2.5 of the HUPO-PSI format for molecular interactions. *BMC Biol.* **5,** 44 (2007).
35. The Gene Ontology Consortium. Gene ontology consortium: Going forward. *Nucleic Acids Res.* **43,** D1049–D1056 (2015).
36. Mossio, M., Montévil, M. & Longo, G. Theoretical principles for biology: Organization. *Prog. Biophys. Mol. Biol.* **122,** 24–35 (2016).
37. R Development Core Team. *R Language Definition V. 3.1.1*. *R Development Core Team* (2015).
38. Van Rossum, G. & Et Al. The Python programming language. *Python Softw. Found.* (2010).
39. Perl. The Perl Programming Language - www.perl.org. *Dr Dobbs Journal* (2011).
40. Chang, J. *et al.* Biopython Tutorial and Cookbook. *Update* 15–19 (2010). doi:10.1145/360262.360268
41. Gentleman, R. *et al.* Bioconductor: open software development for computational biology and bioinformatics. *Genome Biol.* **5,** R80 (2004).

42. Charif, D. *et al.* Package seqinr. *R Packag.* 218 (2015). doi:10.1093/nar/12.1Part1.121>.License
43. RStudio. RStudio: Integrated development environment for R. *J. Wildl. Manage.* (2012). doi:10.1002/jwmg.232
44. Package, T. Package ‘ knitr ’. (2018).
45. Stander, J. & Dalla Valle, L. On enthusing students about big data and social media visualization and analysis using R, RStudio, and RMarkdown. *J. Stat. Educ.* **25,** 60–67 (2017).
46. JJ Allaire, Yihui Xie, Jonathan McPherson, Javier Luraschi, Kevin Ushey, Aron Atkins, Hadley Wickham, Joe Cheng, Winston Chang, Jeff Allen, Roy Storey, Rob Hyndman, Ruben Arslan, RStudio, Inc., jQuery F. Package ‘rmarkdown’. (2018). doi:10.1080/00031305.1980.10483031>.License