Universitat Oberta de Catalunya

uoc.edu

Comparació a nivell d'estructura i de glicosilacions entre proteïnes "moonlighting" de microorganismes patògens i els seus homòlegs en altres microorganismes

Oriol Nualart Mundó
Màster Universitari en Bioinformàtica i Bioestadística
Àrea 2

Consultor: Luis Franco Serrano
Professor/a responsable: Marc Maceira Duch

08/06/2021

Oriol Nualart Mundó

FITXA DEL TREBALL FINAL

Títol del treball:	*Comparació a nivell d'estructura i de glicosilacions entre proteïnes "moonlighting" de microorganismes patògens i els seus homòlegs en altres microorganismes*
Nom de l'autor:	*Oriol Nualart Mundó*
Nom del consultor/a:	*Luis Franco Serrano*
Nom del PRA:	*Marc Maceira Duch*
Data de lliurament (mm/aaaa):	06/2021
Titulació:	*Màster Universitari en Bioinformàtica i Bioestadística*
Àrea del Treball Final:	*Àrea 2*
Idioma del treball:	*Català*
Nombre de crèdits:	*15*
Paraules clau	*Proteïnes moonlighting, virulència, homologia, glicosilacions, proteòmica*

Resum del Treball:

Les proteïnes "moonlighting" són aquelles que duen a terme més d'una funció bioquímica. En microorganismes patògens, s'ha vist que moltes d'aquestes proteïnes tenen funcions secundàries relacionades amb l'adhesió a l'hoste i la capacitat d'infecció. Conèixer quines regions de la proteïna tenen a veure amb la virulència pot ajudar a entendre millor els mecanismes que hi ha implicats.

Addicionalment, s'ha vist que moltes d'aquestes proteïnes relacionades amb la virulència tenen originalment funcions molt bàsiques per a la cèl·lula. Això ha suggerit la hipòtesi que aquestes proteïnes, estant evolutivament molt conservades en els hostes, podrien ser utilitzades pels patògens com una forma de camuflatge, exposant-les a la superfície cel·lular perquè l'hoste les identifiqui com a pròpies.

En aquest treball s'han buscat, en un conjunt de deu proteïnes moonlighting de patògens, elements estructurals, glicosilacions i motius de seqüència possiblement relacionats amb la virulència. Per fer-ho, s'han dut a terme, per a cada una de les proteïnes estudiades, comparacions amb diverses proteïnes homòlogues, tant d'altres patògens com de no patògens. La selecció d'elements sospitosos de tenir relació amb la virulència s'ha fet escollint aquells que es troben conservats exclusivament en els patògens. De forma semblant, en el cas de les glicosilacions, s'han seleccionat també aquelles conservades exclusivament en no patògens, com a sospitoses de tenir relació amb l'absència de virulència.

S'ha obtingut un total de 7 regions estructurals, 40 glicosilacions predites i 31 motius de seqüència. Aquests elements poden ajudar a dirigir futurs esforços de recerca, començant per la seva validació experimental.

Abstract:

"Moonlighting" proteins are those that carry out more than one biochemical function. In pathogen microorganisms, it has been observed that many of such proteins have secondary functions related to host adhesion and to the ability to infect. Knowing which regions of the protein are related to virulence can help better understand the mechanisms involved.

In addition, it has been observed that many of such proteins related to virulence originally have very essential cellular functions. This fact has suggested the hypothesis that those proteins, being evolutionarily very preserved in hosts, may be used by pathogens as a form of camouflage, exposing them to the cell surface so the host identifies them as its own.

With this project it has been sought, in a set of ten moonlighting proteins of pathogens, structural elements, glycosylations and sequence motifs possibly related to virulence. To do so, it has been carried out, for each of the studied proteins, a series of comparisons with several homologue proteins, both from other pathogens and from non-pathogens. The selection of elements suspicious of being related to virulence has been done choosing those preserved exclusively in pathogens. Similarly, in the case of glycosylations, have also been selected those preserved exclusively in non-pathogens, as suspicious of being related to the absence of virulence.

A total of 7 structural regions, 40 predicted glycosylations and 31 sequence motifs have been obtained. Those elements may help direct future research efforts, starting by its experimental validation.

Índex

Llista de figures

Llista de taules

1. Resum

Antecedents

Les proteïnes “moonlighting” són les que desenvolupen més d’una funció bioquímica a la cèl·lula. Moltes de les proteïnes moonlighting conegudes es troben en microorganismes patògens, i duen a terme funcions secundàries relacionades amb la capacitat d’infecció. A més, moltes d’elles són proteïnes la funció canònica de les quals és molt bàsica per al funcionament de la cèl·lula, com per exemple enzims del metabolisme primari.

Mètode

S’han seleccionat deu proteïnes moonlighting de patògens i s’han dut a terme comparacions amb els seus homòlegs d’altres espècies de patògens, de patògens oportunistes, de comensals/simbionts, i d’espècies que no es relacionen amb l’hoste.

Les comparacions s’han fet a tres nivells: a nivell de l’estructura tridimensional de la proteïna, a nivell de glicosilacions predites i a nivell de motius de seqüència. Per identificar els elements possiblement relacionats amb la virulència, s’han seleccionat aquells que estaven presents només en patògens. En el cas de les glicosilacions, també s’han seleccionat les que estaven presents només en no patògens, com a possiblement relacionades amb l’absència de virulència.

Per a les comparacions d’estructura tridimensional, s’ha alineat l’estructura de la proteïna estudiada amb la dels seus diferents homòlegs, utilitzant el servidor FATCAT. Per l’anàlisi de glicosilacions, s’ha utilitzat l’eina Map Sequon, que es pot trobar al servidor ProGlycProt. Per estudiar es motius de seqüència, s’ha utilitzat el servidor The MEME Suite.

Resultats

Entre les diferents proteïnes, s’han seleccionat un total de 7 regions estructurals, 40 glicosilacions predites i 31 motius de seqüència. Degut a les limitacions dels mètodes utilitzats, alguns d’aquests elements resulten més prometedors que altres, però és raonable pensar que part d’ells estan realment relacionats amb la virulència i que podran ser validats experimentalment.

No s’han trobat coincidències destacables en la posició dels elements seleccionats amb els diferents mètodes.

Tampoc no s’han trobat motius de seqüència comuns en més d’una de les proteïnes.

Conclusions

S’ha comprovat que els diferents mètodes utilitzats són útils per obtenir una sèrie d’elements sospitosos de tenir relació amb la virulència, si bé aquesta relació no és segur que sigui real.

Els mecanismes implicats en la virulència per part dels patògens són múltiples i complexes, i probablement no hi ha uns elements detectables que per sí sols determinin capacitat d’infecció. Tot i així, la possibilitat d’identificar elements associats a la virulència pot ajudar a entendre aquests mecanismes.

Els elements seleccionats en les anàlisis, tot i no estar garantida la seva relació amb la virulència, poden ser validats experimentalment.

Aportació

Aquest treball representa un possible punt de partida per a l'estudi de les proteïnes moonlighting utilitzant la seva relació d'homologia amb proteïnes d'altres espècies. Tant l'anàlisi de l'estructura tridimensional com la de les glicosilacions i la dels motius són mètodes prometedors. Els resultats obtinguts poden ser útils per dirigir futurs esforços de recerca.

La distinció utilitzada entre organismes patògens, patògens oportunistes, comensals/simbionts i no patògens pot ajudar en la interpretació de la possible funcionalitat dels elements seleccionats amb aquests mètodes o altres de semblants.

2. Introducció

2.1. Context i justificació del Treball

Les proteïnes moonlighting (multifuncionals) són proteïnes que desenvolupen més d'una funció bioquímica depenent de diversos factors. Hi ha funcions que es troben sempre actives i d'altres depenen del context de la cèl·lula o del compartiment cel·lular on es troba la proteïna.

Aquestes funcions poden ser molt variades i sovint no tenen res a veure amb les originalment conegudes. Per exemple, com a funcions secundàries d'enzims del metabolisme, en procariotes s'han identificat sobretot funcions d'adhesió a l'hoste (a l'epiteli, a matrius extracel·lulars o a plasminogen) i de regulació de la resposta immune de l'hoste (Wang et al., 2013).

En microorganismes patògens, s'ha observat que algunes de les proteïnes expressades a la superfície cel·lular i possiblement associades a la virulència són proteïnes moonlighting, l'altra funció de les quals és sovint part d'una tasca essencial en la cèl·lula, com ara el metabolisme primari (per exemple, veure Henderson B. et al., 2011). Proteïnes com la gliceraldehid-3-fosfat deshidrogenasa (GAPDH) o l'enolasa s'ha vist que en diversos organismes tenen altres funcions a més de les purament enzimàtiques per les que són principalment conegudes, i en patògens sovint estan relacionades amb la virulència. Després d'algunes observacions inicials, sovint casuals, com la troballa de GAPDH a la superfície cel·lular d'estreptococs (V Pancholi et al., 1992), en les últimes dècades s'han anat identificant nombrosos casos de multifuncionalitat lligada a la virulència en aquestes i en d'altres proteïnes.

Partint d'aquesta observació s'ha plantejat la següent hipòtesi (Franco-Serrano et al., 2018). Aquestes proteïnes, pel fet de dur a terme funcions biològicament essencials, estan molt conservades evolutivament entre espècies filogenèticament molt distants. És possible que en microorganismes patògens hagin adquirit la funció addicional de ser expressades en superfície, per tal que l'hoste, en reconèixer-hi patrons, les identifiqui com a pròpies i eviti generar una resposta immune, estalviant-se així possibles problemes d'autoimmunitat.

En l'estudi de la conservació de les proteïnes al llarg de l'evolució és essencial, lògicament, la comparació entre proteïnes homòlogues. Aquesta comparació es pot fer a diferents nivells, essent la seqüència d'aminoàcids el nivell més obvi però no l'únic. També són molt rellevants la conservació de l'estructura tridimensional de la proteïna o de determinats motius, que no té perquè quedar ben reflectida en la seqüència primària, a més de la conservació de les modificacions posttraduccionals.

Entre aquestes últimes, les glicosilacions sovint tenen un paper essencial en la funcionalitat de la proteïna, participant, entre d'altres funcions, en la senyalització cel·lular, la comunicació entre cèl·lules o el reconeixement entre patogen i hoste. L'estudi de les glicosilacions de proteïnes és un camp complex i relativament poc desenvolupat, i convé tenir en compte que hi ha una gran varietat de glicans diferents i que existeixen diversos mecanismes de glicosilació. A més, també és destacable el fet que hi ha diferències importants entre procariotes i eucariotes, tant en els mecanismes de glicosilació com en la funcionalitat de les glicosilacions.

En aquest treball s'han buscat, en proteïnes moonlighting de microorganismes patògens, patrons d'estructura i glicosilacions que potencialment estiguin relacionats amb la virulència. Durant el desenvolupament del treball s'ha afegit la cerca de motius de seqüència. Totes aquestes cerques s'han fet mitjançant comparacions amb proteïnes homòlogues d'altres microorganismes evolutivament propers, distingint entre aquells que son patògens i aquells que no ho son. Els elements identificats, si es consideren d'interès, podran ser estudiats amb més detall per mirar de confirmar la seva rellevància, servint així de punt de partida per a una millor comprensió dels mecanismes d'adhesió i infecció de l'hoste.

Un aspecte que cal tenir en compte a l'hora d'estudiar la possible relació amb virulència de les proteïnes és la distinció entre microorganismes patògens i no patògens. Al contrari del que es pot pensar d'entrada, aquesta distinció no sempre és del tot clara, ja que existeixen molts microorganismes que en general tenen una relació simbiòtica o comensal amb l'hoste, però que ocasionalment poden actuar com a patògens oportunistes. Igualment, la distinció entre organismes simbiòtics i comensals també té una certa ambigüitat, ja que els comensals solen competir amb patògens per l'adhesió a l'hoste, de manera que malgrat no aportar cap benefici específic a l'hoste, indirectament l'estaria beneficiant en evitar l'adhesió d'altres espècies.

El tipus de relació d'un microorganisme amb el seu hoste és rellevant, ja que espècies amb diferent grau (o amb absència) de virulència poden compartir mecanismes d'adhesió a l'hoste. La conservació de les proteïnes implicades en l'adhesió i en la virulència pot, doncs, veure's afectada per aquesta relació.

2.2. Objectius del Treball

2.2.1 Objectiu general

Identificar elements estructurals i patrons de glicosilació indicadors de virulència en proteïnes moonlighting de microorganismes patògens.

2.2.2 Objectius específics

a) Identificar diferències a nivell d'estructura entre proteïnes moonlighting associades a virulència i els seus equivalents en microorganismes no patògens.

b) Partint de les diferències detectades en l'estructura, identificar similituds associades a virulència entre proteïnes homòlogues de microorganismes patògens.

c) Identificar diferències a nivell de llocs predits de glicosilació entre proteïnes moonlighting associades a virulència i els seus equivalents en microorganismes no patògens.

d) Partint de les diferències detectades en els llocs predits de glicosilació, identificar similituds potencialment associades a virulència entre proteïnes homòlogues de microorganismes patògens.

e) Identificar diferències a nivell de l'estat de glicosilació entre proteïnes moonlighting associades a virulència i els seus equivalents en microorganismes no patògens.

f) Partint de les diferències detectades en l'estat de glicosilació, identificar similituds potencialment associades a virulència entre proteïnes homòlogues de microorganismes patògens.

g) Identificar, si n'hi ha, patrons generals indicadors de virulència en comú entre proteïnes no homòlogues a partir de les similituds i diferències detectades en els punts anteriors.

Els objectius *a*, *c* i *e* són passos necessaris per assolir els objectius *b*, *d* i *f*. Els objectius *b* i *d* s'han considerat prioritaris ja que són els que s'espera que permetin complir l'objectiu general. Els objectius *f* i *g*, tot i ser interessants, es parteix de la idea que possiblement no es puguin arribar a complir, per falta d'informació de base en el cas de *f*, i per possible falta de resultats en el cas de *g*.

2.3. Enfocament i mètode seguit

S'ha optat per buscar els indicadors de virulència a nivell estructural i a nivell de glicosilacions perquè a nivell de la seqüència d'aminoàcids ja s'han fet comparacions, i perquè aquesta per sí sola no té per què explicar les diferències en la funcionalitat de les proteïnes. Tot i així, durant el desenvolupament del treball s'ha afegit la cerca de motius de seqüència conservats, per complementar les altres cerques i per poder-los comparar entre proteïnes diferents.

La identificació dels indicadors de virulència s'ha fet mitjançant comparacions entre proteïnes homòlogues. La idea inicial era comparar primer proteïnes associades a virulència amb els seus homòlegs d'espècies no patògenes, i amb les diferències trobades buscar similituds amb els homòlegs d'altres patògens. Per això s'ha distingit en els objectius entre la comparació de patògens i la comparació de patogen amb no patogen. En la pràctica, l'ordre de les comparacions és irrellevant mentre es puguin detectar els elements conservats entre patògens i no conservats en no patògens. En qualsevol cas, s'ha considerat necessari fer els dos tipus de comparacions, per descartar almenys una part de les diferències detectades entre patògens i no patògens que no tenen a veure amb la virulència.

Durant la primera fase del desenvolupament del treball s'ha fet la selecció de les eines a utilitzar i s'han dut a terme les primeres comparacions.

Existeixen diversos programes i servidors d'ús lliure per treballar amb l'estructura tridimensional de les proteïnes. S'ha optat per utilitzar FATCAT (https://fatcat.godziklab.org/; Ye et al., 2004; Li et al., 2020), que detalla amb un codi de colors la distància entre els residus de les proteïnes superposades, cosa que resulta pràctica a l'hora de detectar el grau de semblança de les proteïnes en cada punt de la cadena.

Per estudiar les glicosilacions, hi ha principalment dos enfocaments possibles: utilitzar l'estat real de glicosilació en casos en què aquest sigui conegut, i utilitzar la predicció de llocs de glicosilació a partir de la seqüència d'aminoàcids. La idea inicial del treball era utilitzar els dos mètodes per poder-hi trobar coincidències o divergències. Com es detalla més endavant, però, no ha sigut possible dur a terme les comparacions de l'estat real de glicosilació per la poca disponibilitat de dades disponibles. Per a la comparació

de llocs predits de glicosilació s'ha utilitzat Map Sequon (http://www.proglycprot.org/map_sequon.php), que detecta llocs de glicosilació a partir de la seqüència d'aminoàcids de proteïnes de procariotes.

També s'ha considerat d'interès estudiar els motius de seqüència conservats, per la qual cosa s'ha afegit la seva anàlisi. Per a la seva detecció, s'ha buscat entre les possibles alternatives i s'ha vist que la opció més adequada era The MEME Suite (https://meme-suite.org; Bailey et al., 2015). S'ha utilitzat l'eina principal d'aquest servidor, MEME, per detectar motius comuns entre les seqüències proteiques de diverses proteïnes moonlighting. Aquests motius detectats s'han buscat en les diferents proteïnes moonlighting d'interès i en els seus homòlegs per estudiar-ne la conservació. També s'han comparat els motius detectats en les diferents proteïnes moonlighting d'interès per veure si hi havia coincidències.

La intenció a l'hora de fer els diferents tipus de comparacions és obtenir per a cada proteïna estudiada una sèrie d'elements sospitosos de tenir relació amb virulència. La possible coincidència en la posició d'alguns d'aquests elements, obtinguda per mètodes diferents, podria reforçar els resultats i ajudar en la seva interpretació.

L'esquema general dels procediments realitzats es pot veure a la Figura 1.

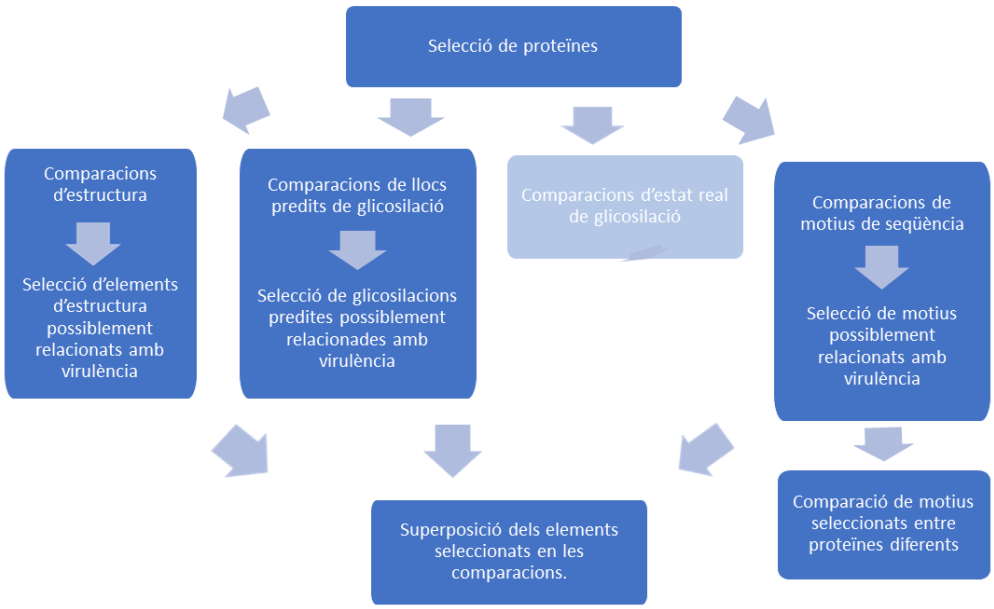

Figura 1. Esquema general dels procediments del treball.

2.4. Planificació del Treball

2.4.1 Tasques

El desenvolupament del treball s'ha organitzat en una sèrie de tasques per tal de compartimentalitzar la feina a fer i d'orientar-la a la consecució dels objectius marcats.

1. Selecció inicial de proteïnes a estudiar

2. Comparació a nivell d'estructura entre una proteïna associada a virulència i el seu equivalent en un o diversos microorganismes no patògens (objectiu específic *a*).

3. Comparació a nivell d'estructura entre la proteïna estudiada en el punt anterior i el seu equivalent en un o diversos microorganismes patògens (objectiu específic *b*).

4. Comparació a nivell de llocs predits de glicosilació entre una proteïna associada a virulència i el seu equivalent en un o diversos microorganismes no patògens (objectiu específic *c*).

5. Comparació a nivell de llocs predits de glicosilació entre la proteïna estudiada en el punt anterior i el seu equivalent en un o diversos microorganismes patògens (objectiu específic *d*).

6. Comparació a nivell de l'estat de glicosilació entre una proteïna associada a virulència i el seu equivalent en un o diversos microorganismes no patògens (objectiu específic *e*).

7. En cas que es conegui, comparació a nivell de l'estat de glicosilació entre una proteïna associada a virulència i el seu equivalent en un o diversos microorganismes no patògens (objectiu específic *f*).

8. Repetició de les tasques 2-7 amb proteïnes diferents. El nombre de proteïnes analitzades variarà en funció del temps real que requereixi fer els anàlisis (objectius específics *a*-*f*).

9. Comparació dels patrons indicadors de virulència identificats en les tasques anteriors (objectiu específic *g*).

2.4.2. Calendari

Per tenir una mica de marge a l'hora d'aprendre a fer els diferents tipus de comparacions, s'ha optat per donar una certa flexibilitat a la consecució de les tasques, superposant-les lleugerament.

Calendari inicial (Figura 2):

- Proposta de TFM: del 17 de febrer a l'1 de març.
- Pla de treball: del 2 al 16 de març.
- Tasca 1. Selecció de proteïnes: del 17 al 24 de març.
- Tasques 2-3. Primera comparació a nivell d'estructura: del 25 de març al 7 d'abril
- Tasques 4-5. Primera comparació a nivell de llocs predits de glicosilació: de l'1 al 14 d'abril.
- Tasques 6-7. Primera comparació a nivell d'estat real de glicosilació: del 8 al 21 d'abril.
- Tasca 8. Comparacions addicionals: del 15 d'abril al 12 de maig.
- Tasca 9. Comparació d'indicadors entre proteïnes diferents: del 6 al 17 de maig.
- Redacció de la memòria: del 18 de maig al 8 de juny.
- Elaboració de la presentació: del 9 al 13 de juny.
- Defensa pública: entre el 16 i el 23 de juny.

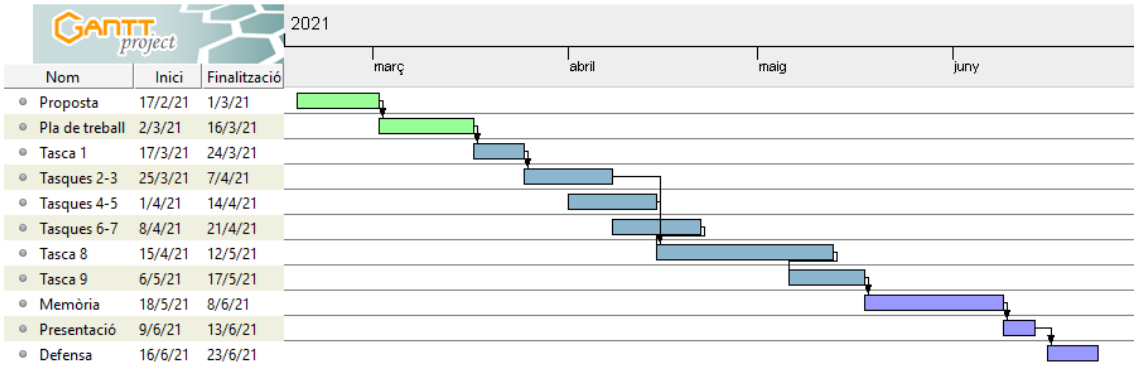

Figura 2. Diagrama de Gantt del calendari inicial.

Calendari modificat (Figura 3):

- ~~Tasques 6-7. Primera comparació a nivell d'estat real de glicosilació: del 8 al 21 d'abril.~~

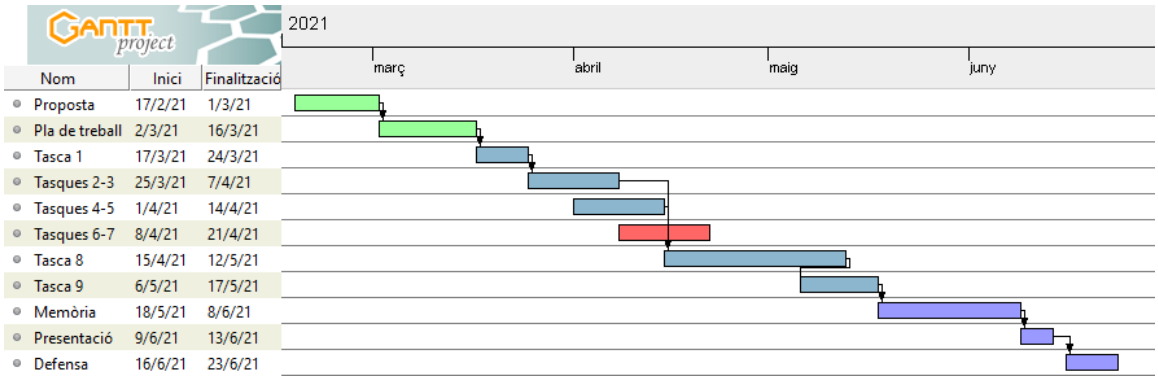

Figura 3. Diagrama de Gantt del calendari modificat.

2.5. Breu sumari de contribucions i productes obtinguts

Llistat de proteïnes moonlighting, cada una d'elles amb la posició d'alguns elements possiblement indicadors de virulència: els elements d'estructura, els llocs predits de glicosilació i els motius de seqüència.

Llistat de motius de seqüència possiblement indicadors de virulència comuns entre proteïnes no homòlogues.

2.6. Breu descripció dels altres capítols de la memòria

Estat de l'art

Aquest capítol pretén situar una mica el context de l'estudi de les proteïnes moonlighting i de la seva relació amb la virulència.

Metodologia

S'explica amb força detall tots els procediments duts a terme durant el treball, i com es pretenen utilitzar per obtenir uns resultats.

Resultats

En aquest apartat s'exposen els resultats obtinguts.

Discussió

En aquest capítol es valoren els resultats i s'intenta fer una lectura de les seves implicacions.

Conclusions

Es valora a partir dels objectius inicials quines conclusions s'han pogut extreure del treball.

3. Estat de l'art

L'estudi de les proteïnes moonlighting ha anat guanyant rellevància progressivament en les últimes dècades, a mesura que s'ha vist que les proteïnes multifuncionals són més comunes del que es creia. La idea que una proteïna té per regla general un única funció ha anat quedant desdibuixada.

Entre les proteïnes moonlighting conegudes, destaca la presència d'un gran nombre de proteïnes de patògens que apareixen expressades a la superfície cel·lular, que es consideren possiblement relacionades amb la virulència.

Per mirar d'entendre l'associació entre proteïnes moonlighting i la seva expressió en superfície (en molts casos indicadora de virulència), s'han buscat punts en comú entre elles. Per exemple, Vaishak Amblee i Constance J. Jeffery (2015) van estudiar les propietats biofísiques d'un total de 98 proteïnes moonlighting principalment intracel·lulars però expressades en superfície, per intentar trobar-hi propietats comunes. En la majoria de casos els paràmetres estudiats no diferien significativament de les proteïnes unifuncionals típiques, excepte pel fet que en la majoria de casos es tractava de proteïnes amb una llargada per sobre de la mitjana. Es van identificar 30 tipus de proteïnes, i en general aquestes mostraven les mateixes característiques físiques a l'interior de la cèl·lula i en superfície.

La hipòtesi plantejada a Franco-Serrano et al. (2018) donaria una explicació a l'expressió en superfície de determinades proteïnes amb funcions primàries intracel·lulars.

El present treball pretén buscar elements sospitosos d'estar relacionats amb la virulència en diverses proteïnes moonlighting. Si es consideren d'interès, aquests elements podrien ser estudiats posteriorment amb més detall per mirar de confirmar aquesta relació, ajudant a una millor comprensió del fenomen de la multifuncionalitat.

4. Metodologia

4.1. Consideracions generals

Per tal d'identificar els elements possiblement relacionats amb la virulència de les proteïnes d'interès, s'ha optat per comparar-les amb els seus homòlegs d'altres microorganismes patògens i amb els de microorganismes no patògens. La idea central del treball és que seleccionant aquells elements que estan conservats en patògens però no en no patògens, els que s'obtenen són sospitosos de tenir relació amb la virulència.

Al llarg del desenvolupament del treball s'ha constatat que la classificació dels microorganismes entre patògens i no patògens no sempre és del tot clara, havent-hi espècies en principi no infeccioses que es comporten ocasionalment com a patògens oportunistes. A més, s'ha vist que podia ser d'interès revisar aquesta classificació per poder distingir aquelles espècies que es comporten com a comensals o simbionts d'aquelles que no tenen relació coneguda amb cap hoste. Això hauria de permetre aprofundir una mica a l'hora d'interpretar els resultats de les comparacions, distingint aquells elements possiblement relacionats amb l'adhesió a l'hoste d'aquells que possiblement tinguin a veure amb la infecció.

Per aquesta raó s'ha decidit classificar les proteïnes en quatre grups, tal com es descriuen a continuació:

- Proteïnes de patògens (P). Aquelles presents en microorganismes que es relacionen amb l'hoste freqüentment en processos infecciosos.

- Proteïnes de patògens oportunistes (O). Aquelles presents en microorganismes que sovint es relacionen amb l'hoste com a comensals o simbionts, però que amb major o menor freqüència poden actuar com a patògens de forma oportunista.

- Proteïnes de comensals i simbionts (C). Aquelles presents en microorganismes que estableixen relacions de comensalitat o de simbiosi amb els seus hostes i no se'n coneix la capacitat d'infectar.

- Proteïnes de no patògens (N). Aquelles presents en microorganismes en els quals no s'ha descrit relació amb l'hoste.

Les proteïnes moonlighting seleccionades per analitzar han sigut totes del grup P, i s'ha intentat trobar-ne homòlegs de tots els grups.

A l'hora de fer les comparacions, tant d'estructura com de glicosilacions com de motius de seqüència, per seleccionar els elements considerats sospitosos de tenir relació amb la virulència, s'ha tingut en compte aquesta classificació. S'ha considerat com si aquesta fos una escala ordinal del grau de virulència o d'"agressivitat" en la seva relació amb l'hoste, essent P les proteïnes corresponents als organismes més virulents i N les corresponents als menys virulents.

Això ha permès destriar els elements no relacionats amb la virulència (aquells que estan presents en els homòlegs de tots els tipus, o només en alguns d'ells sense que aquesta presència o absència es correlacioni amb l'escala) dels possiblement relacionats (aquells presents només en els extrems de l'escala) (Figura 4).

Per exemple, una glicosilació present en homòlegs P i O però no en C i N s'ha considerat sospitosa i ha sigut seleccionada. En canvi, una de present en P i C però no en O i N no s'ha considerat sospitosa.

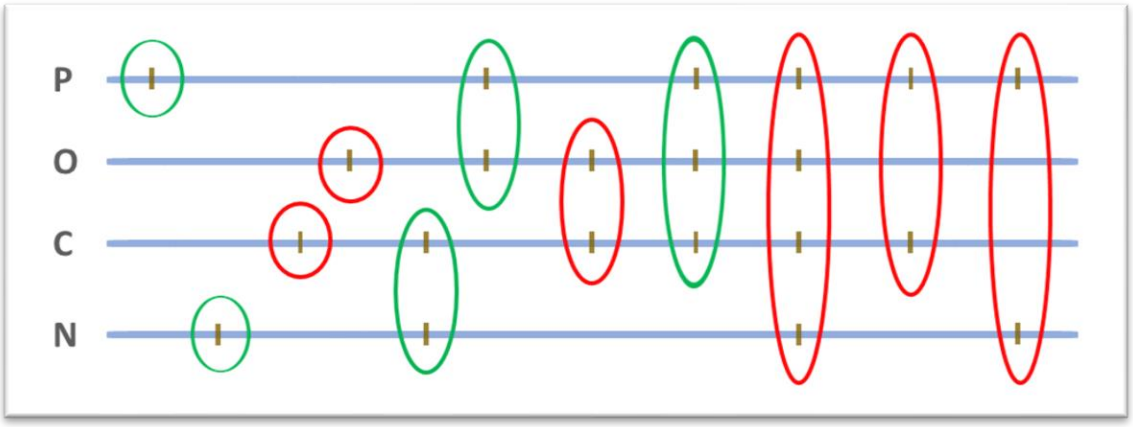

Figura 4. Representació de la selecció dels elements detectats segons la seva presència en els diferents tipus de proteïnes. Aquestes estan ordenades de major a menor "grau de virulència" envers l'hoste. Els cercles verds senyalen els elements seleccionats i els vermells els no seleccionats.

Un últim detall a tenir en compte en relació a les comparacions entre proteïnes és que els elements comparats (regions mal alineades estructuralment, llocs predits de glicosilació i motius de seqüència) han sigut registrats segons la seva posició en la cadena d'aminoàcids. Caldria destacar que la posició exacta no és rellevant per a l'objecte del treball, i que aquesta pot variar entre els homòlegs ja que l'alineament de la numeració dels aminoàcids no sol ser exacte. El que sí que és important és que els elements comparats siguin els equivalents entre els homòlegs, independentment de les petites variacions en la seva posició que hi pugui haver. Per exemple, una O-glicosilació en la mateixa posició que una N-glicosilació no serien equivalents, però dues N-glicosilacions en posicions separades per quatre aminoàcids probablement ho siguin. En alguns casos ha calgut revisar com s'alineen les proteïnes homòlogues per assegurar-se que es tractava del mateix element.

4.2. Selecció de proteïnes

Inicialment s'ha fet una selecció d'una sèrie de proteïnes moonlighting de la base de dades MultitaskProtDB-II (Franco-Serrano et. al., 2017) (Annex 1), atenent els següents criteris:

- Que corresponguessin a microorganismes patògens (grup P).
- Que entre les funcions "moonlighting" de la proteïna (les funcions no canòniques) hi hagués alguna relació amb l'adhesió a l'hoste o amb la virulència.
- Que no fossin homòlogues d'una altra proteïna seleccionada.
- Que pogués trobar proteïnes homòlogues d'almenys un altre patogen i d'almenys un no patogen propers filogenèticament (però no del mateix gènere) per fer les comparacions.

- A ser possible, que apareguessin com a “reviewed” a la base de dades d’UniProt (https://www.uniprot.org), és a dir que la informació disponible hagués sigut revisada individualment i no simplement anotada de forma automàtica.

Tanmateix, aquests criteris aviat s’han demostrat insuficients, especialment a l’hora d’estudiar l’estructura tridimensional de les proteïnes, ja que no totes les proteïnes tenen disponible la seva estructura, i les prediccions d’estructura que s’haurien pogut fer a partir de la seqüència són inexactes i per tant inadequades per als objectius del treball. Cal recordar que el que busquem són similituds i diferències petites, en proteïnes evolutivament properes.

A més, tal com s’ha explicat a l’apartat anterior, durant el desenvolupament del treball també s’ha vist que podia ser d’interès revisar la separació establerta inicialment dels organismes dels quals agafar els homòlegs a comparar, més enllà de la distinció entre patògens i no patògens, per incloure-hi patògens oportunistes i comensals/simbionts.

Per aquests motius, un cop establerts definitivament els mètodes per al desenvolupament del treball, s’han seleccionat noves proteïnes afegint els següents criteris:

- Disponibilitat de l’estructura tridimensional de la proteïna i dels homòlegs a comparar, preferiblement obtingudes experimentalment i no per homologia.

- Respecte a les proteïnes homòlogues a comparar, també s’ha intentat que com a mínim una d’elles correspongués a un microorganisme que es pogués classificar en cada un dels quatre grups descrits a l’apartat 4.1.

Les restriccions que suposen aquests criteris i el limitat nombre de proteïnes que els compleixen han dificultat força la selecció de les candidates a estudiar. Per facilitar aquesta tasca i per evitar passar per alt proteïnes vàlides, s’ha fet una revisió de la base de dades MultitaskProtDB-II. S’ha fet una llista extensa de proteïnes de possible interès per al treball, s’ha pres nota de la disponibilitat o no de la seva estructura tridimensional obtinguda experimentalment, i s’han classificat els organismes corresponents a les proteïnes segons el tipus de relació que tenen amb l’hoste, tal com s’explica a l’apartat anterior (Annex 2). A partir d’aquesta llista s’ha pogut anar identificant amb més facilitat les proteïnes que compleixen els criteris.

El llistat de proteïnes finalment estudiades és aquest:

- 60 KD chaperonin 2 de *Mycobacterium tuberculosis*
- ATP-dependent Clp protease proteolytic subunit de *Streptococcus pneumoniae*
- Chaperone protein DnaK d’*Escherichia coli O127:H6*
- Elongation factor Tu de *Pseudomonas aeruginosa*
- Enolase de *Streptococcus pneumoniae*
- Glyceraldehyde-3-phosphate dehidrogenase de *Streptococcus pneumoniae*
- Glutamine synthetase de *Mycobacterium tuberculosis*
- Malate Synthase G de *Mycobacterium tuberculosis*
- Phosphoglycerate kinase de *Streptococcus pneumoniae*
- Triosephosphate isomerase de *Paracoccidioides lutzii*

4.3. Comparacions d'estructura

Per fer les comparacions de l'estructura tridimensional de les proteïnes s'han valorat diverses eines, entre elles Phyre2 (Kelley, L.a. et. al., 2015) i Swiss-PdbViewer (Guex, N. et. al., 1997). La necessitat de poder localitzar amb facilitat les regions en què l'estructura proteica divergeix ha sigut decisiva per acabar optant per FATCAT (https://fatcat.godziklab.org/; Ye et al., 2004; Li et al., 2020). El servidor FATCAT permet visualitzar les dues cadenes proteiques superposades, acolorides amb una escala de colors que reflecteix la proximitat dels residus alineats (Figura 5). També mostra l'alineament de les cadenes amb aquesta mateixa escala de colors, permetent identificar de forma ràpida i intuïtiva les regions més mal alineades (Figura 6).

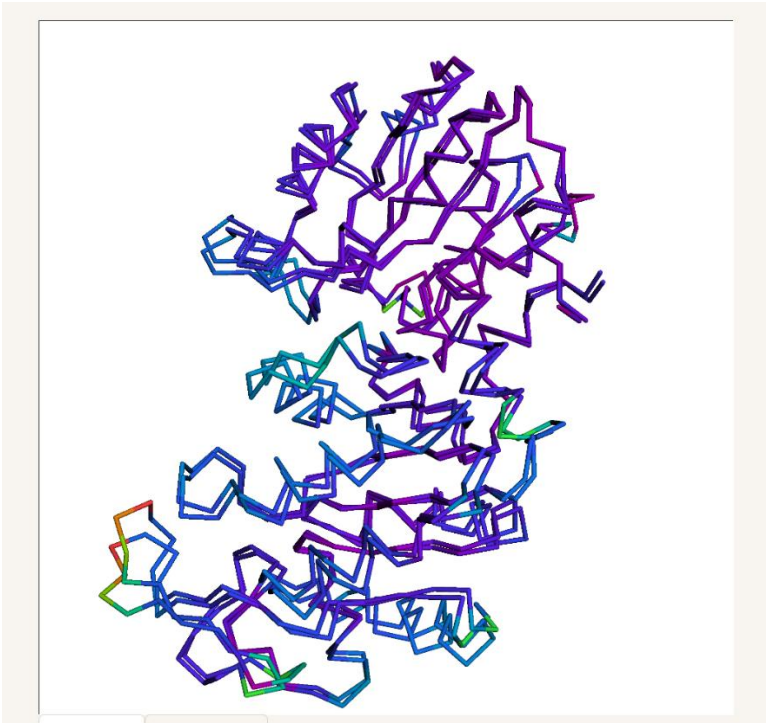

Figura 5. Alineament entre l'estructura tridimensional de la fosfoglicerat quinasa de *Streptococcus pneumoniae* i la de *Lactobacillus acidophillus*, utilitzant l'escala de colors que reflecteix la proximitat entre residus.

Figura 6. Alineament de seqüència de la fosfoglicerat quinasa de *Streptococcus pneumoniae* i la de *Lactobacillus acidophillus*, utilitzant l'escala de colors que reflecteix la proximitat estructural entre residus.

FATCAT utilitza un algoritme d'alineament flexible que permet introduir girs entre fragments alineats, obtenint resultats generalment millors que els algoritmes d'alineament rígids. Aquest algoritme intenta minimitzar l'arrel de l'error quadràtic mig (root mean square deviation, RMSD) i simultàniament minimitzar el nombre de girs.

Els resultats de l'alineament de FATCAT inclouen un *P*-valor com a indicador de la semblança entre les proteïnes comparades, essent aquest la probabilitat d'obtenir el mateix grau de semblança entre dues estructures generades a l'atzar. També inclouen el càlcul global de l'RMSD entre les dues estructures utilitzant el reordenament (els girs) necessari per a l'alineament (opt-RMSD), i sense utilitzar-lo (chain-RMSD). Per al treball, com a mesura del grau d'alineament estructural s'ha registrat l'opt-RMSD, ja que s'ha considerat que és el que millor reflecteix aquest alineament.

Per seleccionar les regions possiblement relacionades amb la virulència de la proteïna moonlighting estudiada, s'han fet comparacions individualment entre la proteïna i cada un dels homòlegs escollits, prenent nota de les regions mal alineades. S'han comparat les regions mal alineades de la proteïna respecte els diferents homòlegs, tenint en compte el grup en què s'ha classificat cada homòleg (P, O, C o N, d'acord amb el que s'explica a l'apartat 4.1.).

Han sigut considerades sospitoses aquelles regions mal alineades entre la proteïna d'interès (una proteïna moonlighting de patògen) i els seus homòlegs "de baixa virulència" (és a dir aquells classificats com a N, o aquells classificats com a N i com a C, o aquells classificats com a N, com a C i com a O), però ben alineades amb la resta d'homòlegs.

4.4. Comparacions de glicosilacions

Per estudiar les glicosilacions s'han plantejat dos possibles enfocaments: mirar l'estat real de glicosilació i mirar la predicció dels llocs de glicosilació a partir de la seqüència primària d'aminoàcids.

Per poder comparar l'estat real de glicosilació s'han buscat bases de dades que poguessin contenir aquest tipus d'informació. Per desgràcia, però, no se n'ha trobat cap que la recollís de forma consistent per a les proteïnes d'interès per al treball. ProGlycProt (http://proglycprot.org/; Choudhary P. et. al., 2019) recull informació de glicoproteïnes validades experimentalment, però són molt poques les proteïnes moonlighting que s'hi han pogut localitzar. A Li X. et al. (2020) es fa un repàs actualitzat i força exhaustiu de les principals bases de dades i eines bioinformàtiques relacionades amb la glicobiologia i la glicoproteòmica. Allí s'hi poden trobar diverses bases de dades de glicoproteïnes, però la majoria d'elles es centren en humans i en uns pocs organismes model. Entre elles, a GlyConnect (https://glyconnect.expasy.org/) i a GlyCosmos (https://glycosmos.org/) s'hi poden trobar glicoproteïnes de microorganismes, però són poques i en qualsevol cas insuficients per a les necessitats del treball. Per aquest motiu s'ha optat per deixar de banda aquest enfocament a l'hora de comparar les glicosilacions.

Les comparacions dels llocs predits de glicosilació sí que s'han pogut fer.

Cal senyalar que la predicció de glicosilacions és un camp força complex. Existeixen quatre tipus de glicosilacions: les O-glicosilacions, les N-glicosilacions, les C-manosilacions i l'ancoratge GPI, essent els dos primers els més habituals. És possible, fins a cert punt, predir possibles llocs de glicosilació a partir de la seqüència d'aminoàcids, però els sequons que senyalen els llocs de glicosilació són diferents segons el tipus, i en general no es coneixen bé les normes que estableixen quins residus seran glicosilats i quins no. A més, aquestes varien entre espècies. Els algoritmes de predicció de llocs de glicosilació solen basar-se en la intel·ligència artificial, com en el cas de Hamby S.E., Hirst JD. (2008).

S'ha optat per utilitzar l'eina Map Sequon (http://proglycprot.org/map_sequon.php; Aadil H. Bhat et. al., 2012), inclosa al servidor ProGlycProt. Aquesta busca a la seqüència proporcionada per l'usuari diversos sequons coneguts en procariotes. La simplicitat del mètode de cerca pot ser un inconvenient, en passar per alt part de la complexitat inherent a la predicció de glicosilacions, però s'ha considerat un factor positiu el fet que l'eina estigués centrada en procariotes, cosa que hauria de reduïr aquesta complexitat. La limitació a procariotes és un inconvenient poc rellevant perquè la gran majoria de proteïnes d'interès per al treball són de procariotes.

Map Sequon requereix introduir la seqüència de la proteïna en format FASTA i retorna un llistat de glicosilacions predites amb la seva posició, l'aminoàcid glicosilat i el sequon detectat.

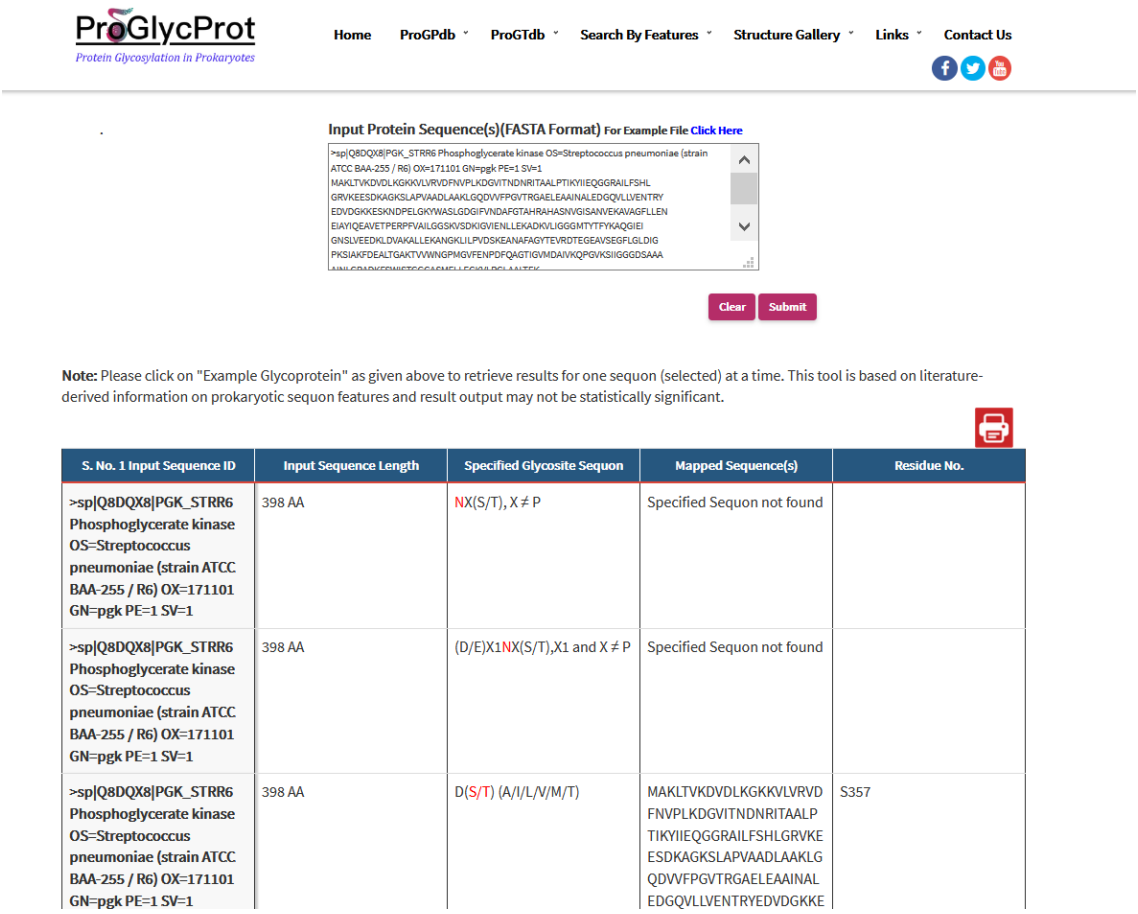

Figura 7. Pàgina amb els resultats de Map Sequon per a la fosfoglicerat quinasa de *Streptococcus pneumoniae*.

Aplicant la predicció a la proteïna moonlighting d'interès i als seus homòlegs, s'han comparat els llocs de glicosilació i s'han seleccionat aquells que estan conservats en funció del grau de virulència dels organismes a què corresponen, tal com s'explica a l'apartat 4.1. S'han seleccionat tant les glicosilacions conservades només en els microorganismes més virulents com les conservades només en els menys virulents, ja que s'ha considerat que la glicosilació tan aviat pot tenir funcions rellevants per la capacitat d'infectar (per exemple senyals d'exportació a la superfície cel·lular o funcions d'adhesió a l'hoste), com per l'absència de capacitat d'infectar (per exemple senyals per a la no exportació a la superfície cel·lular).

4.5. Comparacions de motius de seqüència

Per a l'anàlisi de motius de seqüència existeixen diversos servidors que permeten buscar o analitzar motius en seqüències proporcionades per l'usuari. Per exemple, InterPro (https://www.ebi.ac.uk/interpro/; Blum M. et al., 2020) permet identificar-hi dominis funcionals rellevants a partir de proteïnes de la mateixa família, i ho fa integrant informació de diverses bases de dades.

En el cas que ens ocupa, s'ha optat per utilitzar The MEME Suite (https://meme-suite.org; Bailey T.L. et al., 2015). Aquest servidor inclou l'eina MEME (Bailey T.L. et al., 2006), que busca seqüències repetides en diferents cadenes d'ADN, ARN o proteïnes per descobrir motius *de novo*. També inclou l'eina FIMO (Grant C.E. et al., 2011), que permet buscar ocurrències de determinats motius en cadenes d'ADN, ARN o proteïnes.

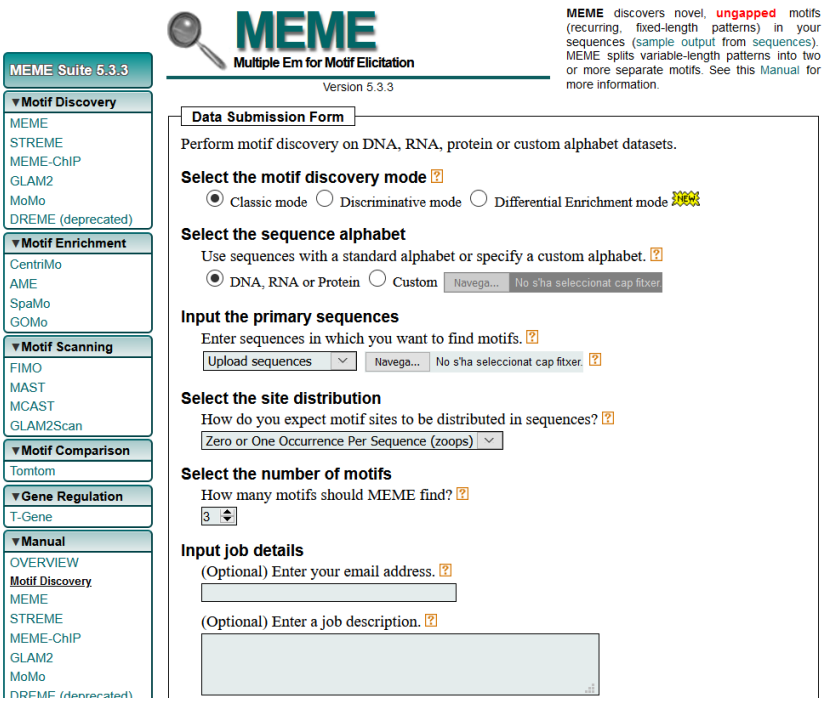

Figura 8. Pàgina principal del cercador de motius MEME.

L'anàlisi que es vol dur a terme consisteix en agafar un conjunt de motius de seqüència presents en diferents proteïnes moonlighting de patògens, i buscar aquests motius en les proteïnes moonlighting d'interès i en els seus homòlegs, per veure si hi estan presents i en quins casos es conserven.

Per generar els motius s'han seleccionat 38 proteïnes diferents de patògens de la base de dades MultitaskProtDB-II, incloent-hi les utilitzades en les comparacions (Annex 3), i s'ha utilitzat MEME per detectar motius que es repetissin entre elles. Entre els paràmetres de cerca, s'ha especificat que els motius havien de tenir entre 2 i 20 aminoàcids i que havien d'estar presents en almenys dues de les proteïnes. Per tenir una quantitat raonable de motius amb què treballar, s'ha indicat que se'n generessin 200.

Un cop generat el conjunt de motius, per cada proteïna moonlighting d'interès i els seus homòlegs s'ha fet una cerca utilitzant FIMO per mirar de detectar-los. Novament, a l'hora de fer la selecció s'ha tingut en compte la classificació segons el grau de virulència que s'explica a l'apartat 4.1. En aquest cas, els motius seleccionats han sigut els que estaven conservats només en els microorganismes virulents. S'ha considerat que la presència d'alguns dels motius només en microorganismes no patògens no era indicadora de relació amb l'absència de virulència, perquè els motius havien estat generats a partir de proteïnes de patògens (Figura 9).

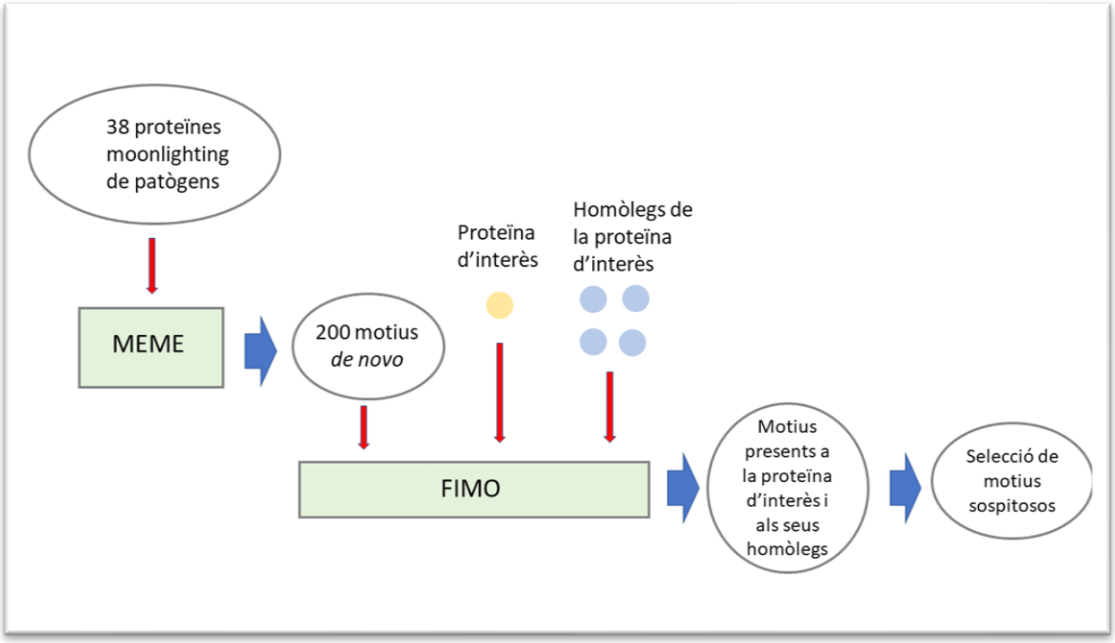

Figura 9. Procés de generació i selecció dels motius de seqüència.

4.6. Comparació de motius entre proteïnes moolighting no homòlogues

Per intentar trobar elements sospitosos de tenir relació amb la virulència comuns entre proteïnes diferents, s'han comparat els motius de seqüència detectats en l'anterior anàlisi. Senzillament, s'ha mirat quins dels motius s'havien seleccionat en les diferents proteïnes i si apareixien en més d'una d'elles. La presència de motius concrets en més d'una proteïna de patògens podria reforçar la sospita de la seva relació amb virulència.

5. Resultats

5.1. Comparacions d'estructura

Les comparacions d'estructura han de permetre complir els objectius específics *a* i *b* d'estudiar les similituds i diferències d'estructura entre proteïnes moonlighting de patògens i els seus homòlegs.

S'han fet comparacions individualment de les deu proteïnes moonlighting de patògens estudiades amb cada un dels seus homòlegs amb estructura tridimensional que s'ha trobat disponible. S'han seleccionat aquelles regions de la la proteïna estudiada que estan mal alineades amb seus homòlegs amb baix grau de virulència, però ben alineades amb els d'alt grau de virulència.

D'aquesta manera s'han identificat un total de 7 regions sospitoses de tenir relació amb virulència, entre 0 i 1 per proteïna en la majoria de casos (Annex 4). Només en el cas de la triosafosfat isomerasa de *Paracoccicoides lutzii* s'han identificat dues regions (Figura 10).

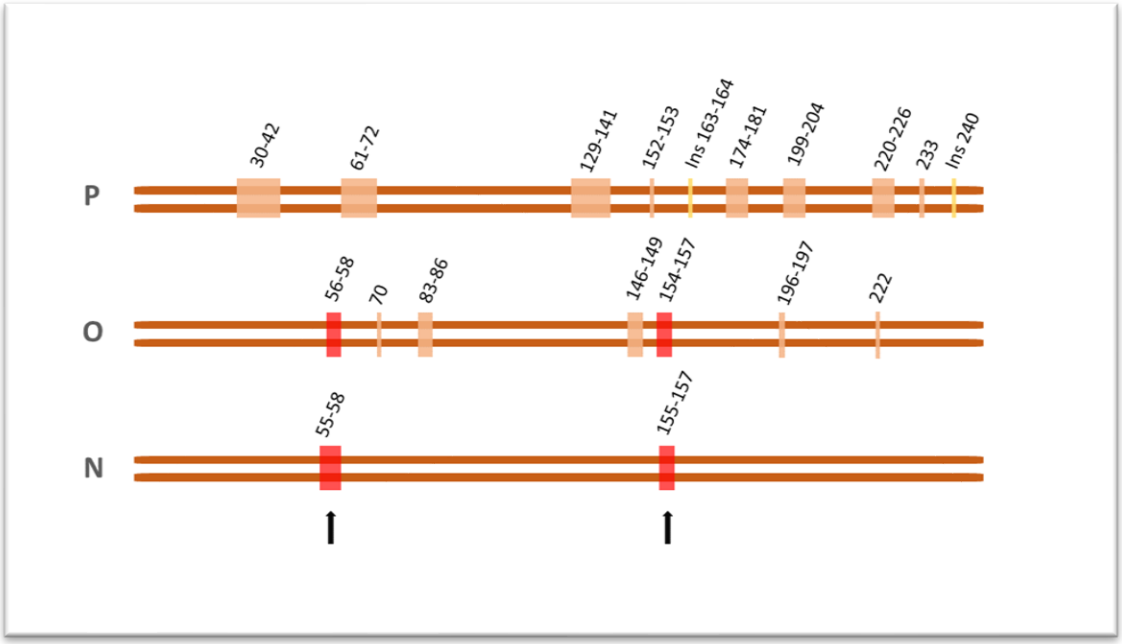

Figura 10. Posició de les regions mal alineades en les comparacions de la triosafosfat isomerasa de *Paracoccicoides lutzii* amb els seus homòlegs de *Coccidioides immitis* (P), *Saccharomyces cerevisiae* (O) i *Neurospora crassa* (N).

Com es pot observar, les dues regions seleccionades en el cas de la triosafosfat isomerasa de *Paracoccicoides lutzii* estan mal alineades respecte les proteïnes d'un patògen oportunista (*Saccharomyces cerevisiae*) i d'un no patògen (*Neurospora crassa*), però s'alineen correctament amb la d'un altre patogen (*Coccidioides immitis*).

Aquestes són les regions seleccionades per a cada proteïna:

Proteïna	**Posició dels fragments seleccionats**	
60 KD chaperonin 2	-	-
ATP-dependent Clp protease proteolytic subunit	28	-
Chaperone protein DnaK	44-46	-
Elongation factor Tu	-	-
Enolase	279	-
Glyceraldehyde-3-phosphate dehidrogenase	118-132	-
Glutamine synthetase	-	-
Malate Synthase G	207-213	-
Phosphoglycerate kinase	-	-
Triosephosphate isomerase	55-58	154-157

Taula 1. Posició de les regions estructurals seleccionades de cada cadena proteica.

A les figures 11 i 12 es poden observar les representacions tridimensionals d'alguns dels alineaments estructurals reaitzats.

Figura 11. Fragments de la representació tridimensional dels alineaments d'estructura:

A) Entre la ATP-dependent Clp protease proteolytic subunit de *S. pneumoniae* (P) i la de *C. hydrogenoformans* (N). La regió entre els fragments de color taronja correspon al fragment seleccionat.

B) Entre la Chaperone protein DnaK de *E. coli O127:H6* (P) i la de *S. oneidensis* (N). El fragment més acolorit al centre de la imatge correspon als aminoàcids 44-46, mal alineats en clar contrast amb l'alineament perfecte de la resta de la cadena.

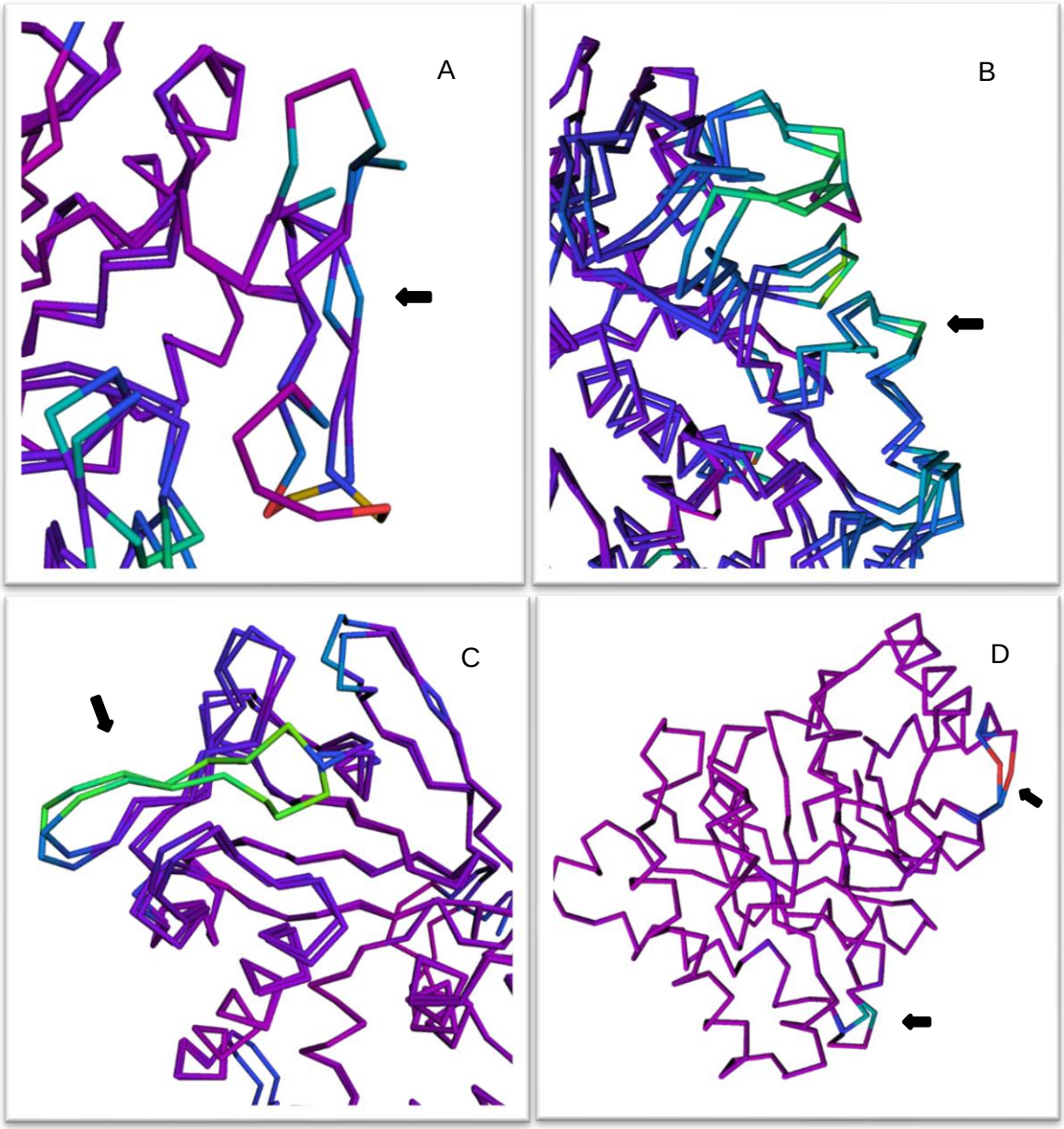

Figura 12. Fragments de la representació tridimensional dels alineaments d'estructura:

A) Entre la enolasa de *S. pneumoniae* (P) i la de *Listeria welshimeri* (N). Tot i que es poden veure diverses regions en què les cadenes no s'alineen correctament, la que s'ha seleccionat per no estar present en altres comparacions és l'obertura que s'observa en color blau a la dreta de la imatge, corresponent a l'aminoàcid 279.

B) Entre la gliceraldehid-3-fosfat deshidrogenasa de *S. pneumoniae* (P) i la de *G. Stearothermophilus* (N), que inclou la regió seleccionada entre els aminoàcids 118 i 132.

C) Entre la malat sintasa G de *M. tuberculosis* i la de *C. efficiens*. La regió verda correspon al fragment 207-213.

D) Entre la triosafosfat isomerasa de *P. lutzii* (P) i la de *N. crassa* (N). Els dos fragments més acolorits són els seleccionats, corresponents als aminoàcids 55-58 i 154-157 respectivament.

5.2. Comparacions de glicosilacions

Els objectius *c* i *d* consisteixen en identificar diferències en els llocs predits de glicosilació, i seleccionar aquells llocs en què es prediuen diferències entre les proteïnes de patògens i les de no patògens.

En aquest cas, per a algunes de les proteïnes s'han utilitzat més homòlegs que en el cas de les comparacions d'estructura, ja que no existia la limitació d'haver de tenir l'estructura tridimensional.

S'han obtingut un total de 40 llocs sospitosos de relació amb virulència o amb la seva absència (Annex 5). Aquests no estan repartits uniformement entre les proteïnes estudiades, ja que en alguns casos no se n'ha trobat cap i en d'altres se n'han trobat múltiples, fins a tretze en el cas de la glutamine syntethase de *Mycobacterium tuberculosis*. Convé esmentar també que en el cas de la triosephosphate isomerase de *Paracoccidioides lutzii* no s'han dut a terme aquestes comparacions ja que es tracta d'un organisme eucariota i l'eina utilitzada per predir els llocs de glicosilació, Map Sequon, està pensada per proteïnes de procariotes.

D'aquests 40 llocs sospitosos, 22 corresponen a glicosilacions en microorganismes patògens absents en no patògens, i 18 a glicosilacions en no patògens absents en patògens.

Aquests són els llocs predits de glicosilació seleccionats en la comparació de cada proteïna amb els seus homòlegs.

Proteïna	**Aminoàcid d'unió i posició de les glicosilacions seleccionades**												
60 KD chaperonin 2	(N468)	Y482	-	-	-	-	-	-	-	-	-	-	-
ATP-dependent Clp protease...	N192	-	-	-	-	-	-	-	-	-	-	-	-
Chaperone protein DnaK	(N62)	N187	(S208)	S234	(N276)	N314	(N497)	(N620)	-	-	-	-	-
Elongation factor Tu	-	-	-	-	-	-	-	-	-	-	-	-	-
Enolase	Y256	(N320)	(N422)	-	-	-	-	-	-	-	-	-	-
Glyceraldehyde-3-phosphate...	T49	(S49)	-	-	-	-	-	-	-	-	-	-	-
Glutamine synthetase	T127	S141	(N145)	S146	N149	(N262)	N271	T301	(N312)	N321	Y342	N346	S372
Malate Synthase G	S32	(S130)	S179	(N217)	S275	(S298)	N319	(S594)	(N608)	S689	(N705)	-	-
Phosphoglycerate kinase	-	-	-	-	-	-	-	-	-	-	-	-	-

Taula 2. Aminoàcid d'enllaç i posició de les glicosilacions predites seleccionades de cada proteïna. Les glicosilacions que apareixen entre parèntesi corresponen a les que estan absents en la proteïna en qüestió però presents en els seus homòlegs de baixa virulència.

5.3. Comparacions de motius de seqüència

La comparació de motius de seqüència s'ha afegit durant el desenvolupament del treball i no es correspon directament amb cap dels objectius específics inicials.

La selecció de motius a buscar s'ha fet a partir de 38 proteïnes moonlighting de patògens, incloent les deu proteïnes estudiades, tal com s'explica a l'apartat 4.5. S'ha obtingut un total de 200 motius presents en almenys dues de les proteïnes.

És important destacar que la gran majoria d'aquests motius apareixen amb un baix grau de significació estadística. En generar-se els motius s'han agafat els 200 amb un e-valor més petit. L'e-valor és una estimació del nombre de motius amb igual o major raó de versemblança logarítmica (log-likelihood ratio) del que s'esperaria trobar si les seqüències estiguessin generades aleatòriament. En el cas que ens ocupa, només dos dels motius apareixen amb un e-valor per sota de 0.05, cosa que vol dir que la majoria probablement no es corresponen amb una funcionalitat biològica real. Tot i així, s'ha considerat que podien ser indicadors de semblances i diferències entre les proteïnes. Es poden consultar els motius generats i la seva significació estadística a l'Annex 6, que és l'arxiu de sortida en format web de la cerca al servidor MEME.

S'ha fet una cerca dels motius generats en les 10 proteïnes estudiades i en els seus homòlegs, per tal de veure quins dels motius es mostraven conservats. S'han seleccionat un total de 31 motius presents en els homòlegs de més virulència que no es conserven en els de menys. El repartiment d'aquests motius no és uniforme entre les proteïnes estudiades, anant de 0 en el cas de la gliceraldehid-3-fosfat deshidrogenasa de *Streptococcus pneumoniae* a 7 en el cas de la Malat sintasa G de *Mycobacterium tuberculosis* (Taula 3; Annex 7).

Proteïna	**Posició dels motius seleccionats**						
60 KD chaperonin 2	281-299	-	-	-	-	-	-
ATP-dependent Clp protease...	79-88	110-115	121-124	163-166	166-171	-	-
Chaperone protein DnaK	422-426	447-461	533-539	540-544	605-608	-	-
Elongation factor Tu	36-40	271-285	-	-	-	-	-
Enolase	174-182	317-322	-	-	-	-	-
Glyceraldehyde-3-phosphate...	-	-	-	-	-	-	-
Glutamine synthetase	194-204	214-217	217-227	332-342	408-420	-	-
Malate Synthase G	67-70	68-73	101-116	235-241	320-337	331-336	510-515
Phosphoglycerate kinase	168-176	334-341	-	-	-	-	-
Triosephosphate isomerase	190-192	216-218	-	-	-	-	-

Taula 3. Posició dels motius de seqüència seleccionats en cada proteïna.

5.4. Comparació de motius entre proteïnes moonlighting no homòlogues

L'anàlisi de motius ha permès identificar fragments de seqüència potencialment rellevants per a la virulència dels microorganismes corresponents. La comparació d'aquests motius entre les diferents proteïnes estudiades és una manera de veure si hi ha elements en comú entre elles.

S'ha mirat quins dels 200 motius buscats apareixien seleccionats per a cada proteïna i no se n'ha trobat cap que es repetís en més d'una.

6. Discussió

Les diferents comparacions ens han proporcionat una sèrie d'elements que hem considerat sospitosos de tenir relació amb la virulència. Aquesta sospita no implica que la relació sigui real, ja que per a cada proteïna hem treballat amb pocs homòlegs, i per tant part de les coincidències i divergències que hem trobat entre els homòlegs virulents i no virulents podrien haver sorgit per atzar. En cas d'haver fet unes altres comparacions potser hauríem trobat uns altres elements sospitosos, i si haguéssim treballat amb més homòlegs per cada proteïna probablement hauríem hagut de descartar part dels que hem seleccionat.

En aquest sentit, tal com està plantejat, l'objectiu principal del treball d'identificar elements indicadors de virulència possiblement resulta una mica massa ambiciós, si es pretén afirmar amb una certa seguretat la relació dels elements trobats amb la virulència i necessitaria de un treball molt més llarg en el temps.

No obstant, els elements identificats donen una idea del tipus de resultats que es poden obtenir amb aquests mètodes. Segurament alguns d'aquests elements val la pena que siguin analitzats amb més detall per mirar de confirmar-ne o descartar-ne la relació amb la virulència.

6.1. Comparacions d'estructura

La principal limitació que hi ha hagut a l'hora de fer les comparacions d'estructura ha sigut la disponibilitat de l'estructura tridimensional, tant de proteïnes moonlighting d'interès per al treball com, un cop seleccionades les proteïnes, dels seus homòlegs. S'ha procurat treballar amb l'estructura obtinguda experimentalment, però són encara menys les proteïnes de les quals aquesta es pot trobar. En força casos s'ha hagut d'utilitzar l'estructura obtinguda per homologia, cosa que limita la fiabilitat dels alineaments realitzats.

De les deu proteïnes analitzades, en quatre d'elles no s'han trobat regions estructurals que estiguessin mal alineades amb homòlegs de baixa virulència i ben alineades amb homòlegs virulents. En total, només s'han trobat set d'aquestes regions entre les altres sis proteïnes. Això suggereix que, en cas de tenir un papar clau en la virulència de la proteïna, els patrons estructurals que hi estan implicats possiblement son pocs, o bé que els mitjans utilitzats, incloent en diversos casos estructures obtingudes per homologia, són poc adequats per aquest tipus d'anàlisi.

En qualsevol cas, les comparacions s'han pogut fer i s'ha obtingut una sèrie de regions estructurals associades a virulència, de manera que els objectius específics *a* i *b* es pot considerar que s'han acomplert.

- **ATP-dependent Clp protease proteolytic subunit** de *Streptococcus pneumoniae* (Figura 13)

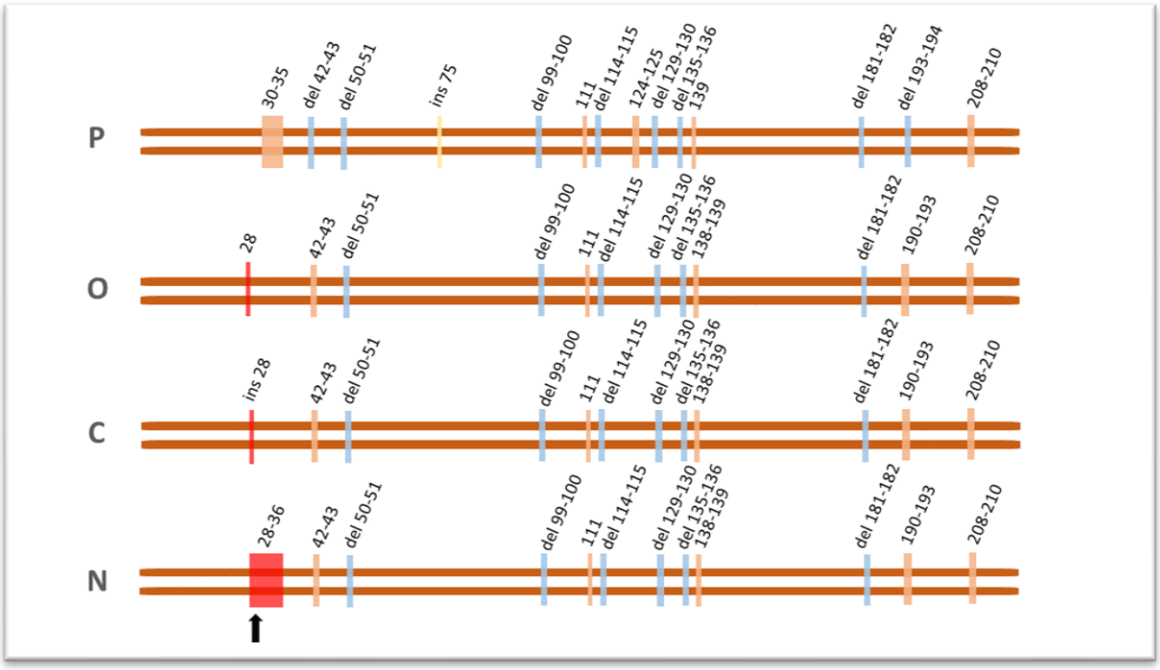

Figura 13. Posició de les regions mal alineades en les comparacions de la ATP-dependent Clp protease proteolytic subunit de *S. pneumoniae* amb els seus homòlegs de *L. monocytogenes* (P), *S. aureus* (O), *B. subtilis* (C) i *C. hydrogenoformans* (N).

Aquesta proteïna s'ha comparat amb un homòleg de cada tipus (P, O, C i N), corresponents, respectivament, a *Listeria monocytogenes*, *Staphylococcus aureus*, *Bacillus subtilis* i *Carboxydothermus hydrogenoformans*. En tots els casos excepte l'últim s'ha utilitzat l'estructura obtinguda experimentalment, de manera que en principi els alineaments haurien de ser força fiables.

La proteïna menys alineada, amb una opt-RMSD de 1,84Å, ha sigut la del patògen *Listeria monocytogenes* (les altres s'alineen amb una opt-RMSD d'entre 0,94 i 1,04Å). Això queda reflectit en les regions mal alineades, que són les mateixes en el cas dels altres tres homòlegs i algunes més en el cas de la proteïna de *L. monocytogenes*.

Com a excepció a aquest pitjor alineament estructural hi ha la regió que agafa l'aminoàcid 28, que es conserva només entre *S. pneumoniae* i *L. monocytogenes*.

El fet que aquesta regió es conservi entre els dos microorganismes patògens però no estigui conservada en un oportunista ni en un comensal, apunta la possibilitat que tingui relació amb la capacitat d'infecció i amb el grau de virulència dels patògens, i no tant amb l'adhesió a l'hoste, ja que en aquest cas probablement la regió estaria conservada en els tres tipus d'homòlegs d'espècies que s'adhereixen a l'hoste.

- **Chaperone protein DnaK** d'*Escherichia coli O127:H6* (Figura 14)

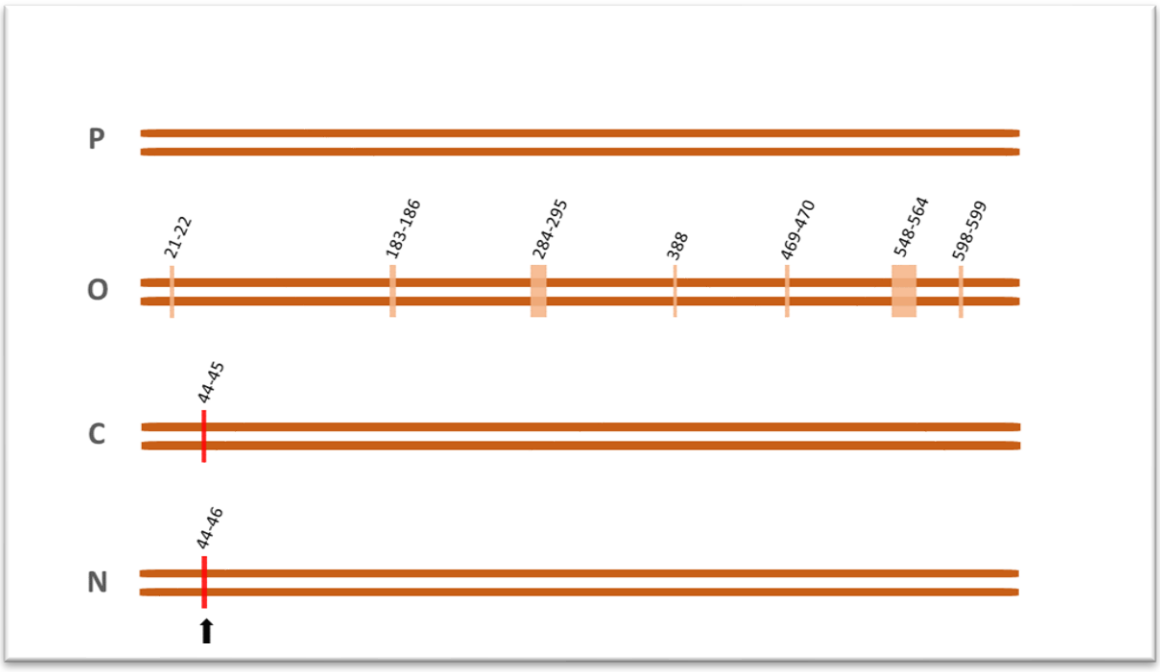

Figura 14. Posició de les regions mal alineades en les comparacions de la chaperone protein DnaK de *E. coli O127:H6* amb els seus homòlegs de *S. sonnei* (P), *P. mendocina* (O), *A. fischeri* (C) i *S. oneidensis* (N).

Només s'ha trobat una regió que estigués mal alineada exclusivament en no patògens, concretament en *Aliivibrio fischeri* (C) i en *Shewanella oneidensis* (N), entre els aminoàcids 44 i 46. Que la regió es conservi en el cas de P i O però no en el de C i N suggereix que la seva funció tingui a veure amb la capacitat d'infectar i no amb la capacitat d'adherir-se a l'hoste. Curiosament, l'homòleg que, amb diferència, menys s'alinea en aquest cas és el de *Pseudomonas mendocina* (O), que sí que alinea correctament aquest fragment.

- **Enolasa** de *Streptococcus pneumoniae* (Figura 15)

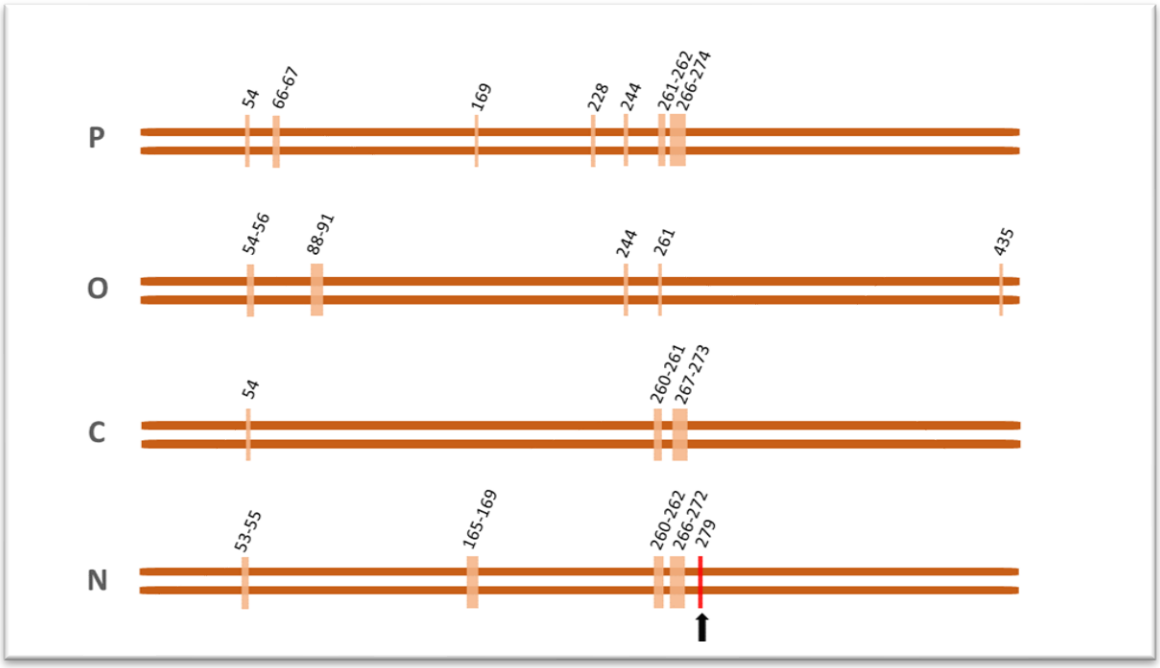

Figura 15. Posició de les regions mal alineades en les comparacions de l'enolasa de *S. pneumoniae* amb els seus homòlegs de *E. hyrae* (P), *S. aureus* (O), *B. subtilis* (C) i *L. welshimeri* (N).

S'ha localitzat un únic fragment d'estructura sospitós de relació amb virulència, corresponent a l'aminoàcid 279 de la cadena. Es troba mal alineat només en la comparació amb *Listeria welshimeri* (N), i per tant en cas de confirmar-se la relació amb virulència, podria ser per tenir una funció d'adhesió a l'hoste.

- **Gliceraldehid-3-fosfat deshidrogenasa** de *Streptococcus pneumoniae* (figura 16)

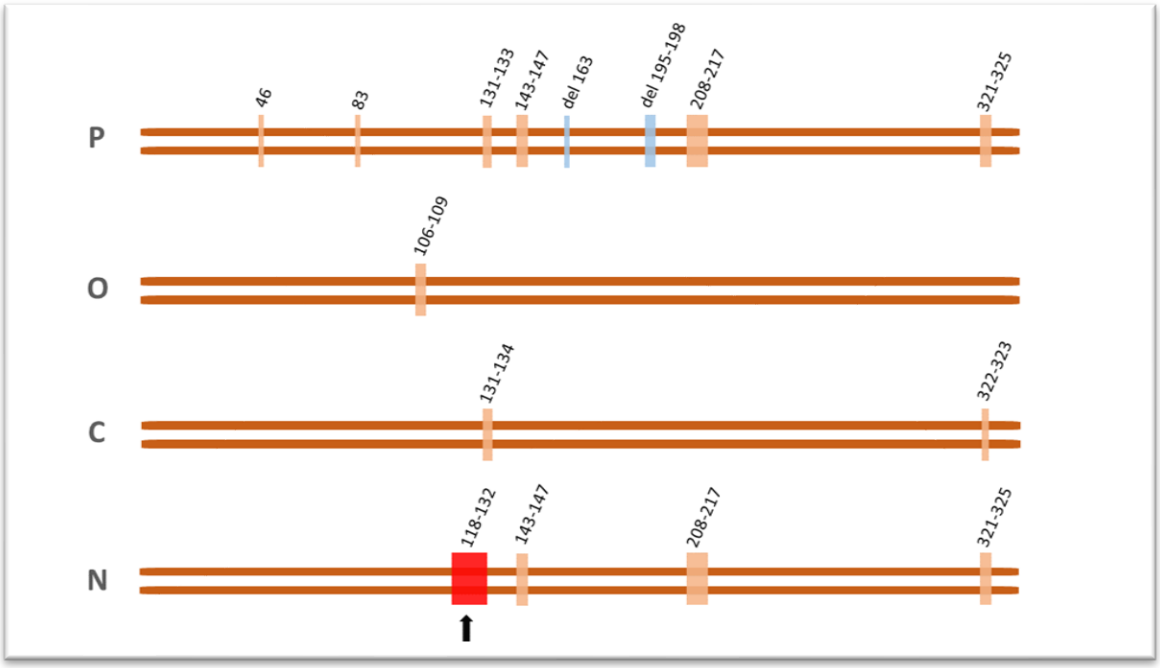

Figura 16. Posició de les regions mal alineades en les comparacions de la gliceraldehid-3-fosfat deshidrogenasa de *S. pneumoniae* amb els seus homòlegs de *B. anthracis* (P), *S. aureus* (O), L. *lactis* (C) i *G. stearotermophilus* (N).

En aquest cas també s'ha pogut fer l'alineament amb els quatre tipus d'homòlegs, tot i que en el cas de P i de C s'ha hagut d'utilitzar l'estructura obtinguda per homologia.

La única diferència seleccionada, entre les posicions 118 i 132 de la cadena, es troba només en l'alineament amb l'homòleg de *Geobacillus stearothermophilus*, microorganisme que no té cap relació amb l'hoste. Els altres alineaments mostren conservació de l'estructura en la mateixa zona (excepte en els aminoàcids 131 i 132, que sí que estan mal alineats en alguns dels altres homòlegs). Tenint en compte que un dels organismes que conserven l'estructura és un comensal sense capacitat d'infectar, això fa pensar que, en cas de confirmar-se la relació d'aquesta regió amb la virulència, probablement la seva funció tingui a veure amb l'adhesió a l'hoste i no amb la capacitat d'infecció.

- **Malate Synthase G** de *Mycobacterium tuberculosis* (Figura 17)

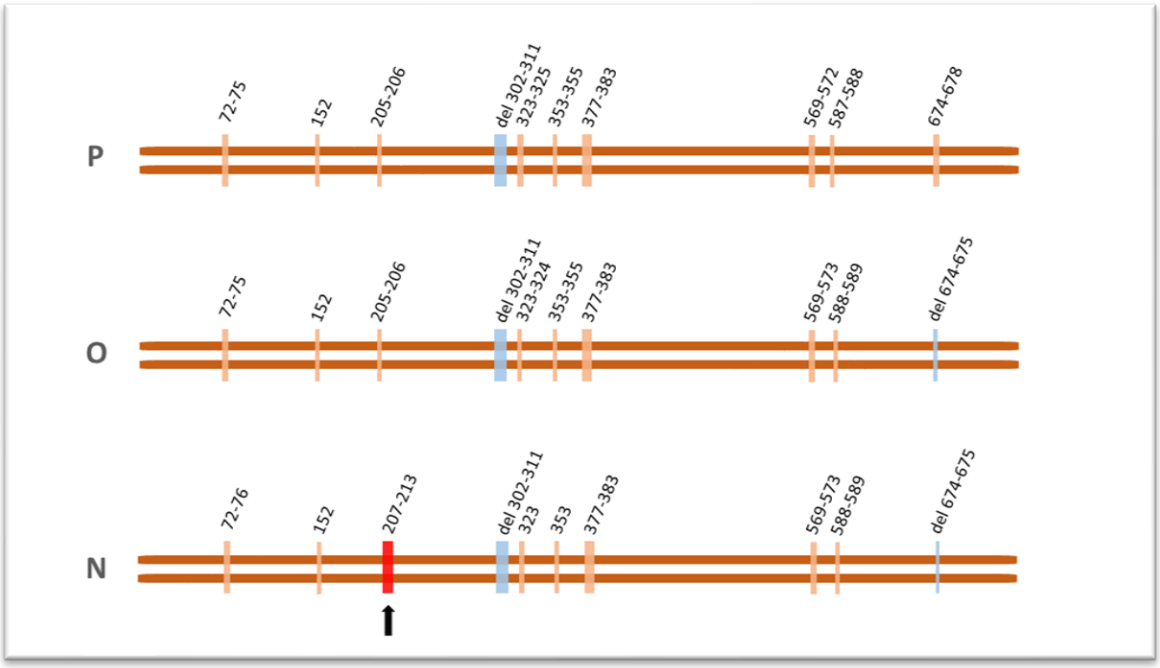

Figura 17. Posició de les regions mal alineades en les comparacions de la malat sintasa G de *M. tuberculosis* amb els seus homòlegs de *R. fascians* (P), *N. farcinica* (O) i *C. efficiens* (N).

De nou s'ha seleccionat una única regió, ara el fragment entre els aminoàcids 207 i 213. En trobar-se mal alineat només en la comparació amb *Corynebacterium efficiens* (N), i ben alineat en la resta de comparacions, aquest fragment seria sospitós de tenir relació amb funcions d'adhesió a l'hoste.

- **Triosephosphate isomerase** de *Paracoccidioides lutzii* (Figura 10)

Aquesta comparació és probablement la menys fiable, ja que no s'ha pogut trobar l'estructura tridimensional obtinguda experimentalment per a la proteïna moonlighting estudiada, i se n'ha utilitzat una obtinguda per homologia. Tampoc s'ha trobat un homòleg d'un organisme comensal amb l'estructura disponible per poder fer la comparació.

S'han trobat dues regions sospitoses de relació amb virulència: entre els aminoàcids 55 i 58 i entre els aminoàcids 154 i 157. En els dos casos, el mal alineament es dona tant en la comparació amb *Saccharomyces cerevisiae* (O) com amb *Neurospora crassa* (N), cosa que suggereix que la possible relació amb virulència tingui a veure amb l'agressivitat de la infecció i no amb la capacitat d'infectar.

6.2. Comparacions de llocs predits de glicosilació

En aquest cas, és important no oblidar que estem treballant amb prediccions, i les limitacions que això suposa. No podem afirmar amb seguretat que els llocs de glicosilació amb què treballem apareguin glicosilats en la proteïna real, i de fet el més probable és que una part d'ells no ho estigui, però la presencia de més llocs de glicosilació en microorganismes virulents respecte els no virulents o l'aparició de llocs diferents en els primers, pot suggerir que es tracta d'una regió important per a la virulència.

Les comparacions de llocs predits de glicosilació han donat un nombre de resultats més gran que les d'estructura: un total de 40 possibles glicosilacions, comptant les que es prediuen només en els organismes més virulents (22), juntament amb les que es prediuen només en els organismes menys virulents (18). El fet de recollir les glicosilacions presents en els dos extrems és perquè s'ha considerat que, de la mateixa manera que algunes glicosilacions podrien tenir una funció potenciadora de la virulència, també és possible que d'altres actuïn com a senyals inhibidores, per exemple evitant l'exportació a la superfície cel·lular de la proteïna en qüestió.

La gran diferència en el nombre de glicosilacions seleccionades per algunes de les proteïnes (entre 0 i 13 per proteïna) podria tenir a veure amb dos possibles factors.

D'una banda, cal destacar que per aquestes comparacions, en no dependre de la disponibilitat de l'estructura tridimensional sinó de la seqüència d'aminoàcids, en general s'ha disposat de més homòlegs per comparar. Tot i així, per algunes de les proteïnes el nombre d'homòlegs d'interès seguia sent molt reduït, possiblement per tractar-se de proteïnes en general menys estudiades. Per aquest motiu, per algunes de les proteïnes s'han fet més comparacions que per altres (fins a 10 en el cas de la fosfoglicerat quinasa de *Streptococcus pneumoniae*). El nombre de comparacions en general ha reduït el nombre de glicosilacions seleccionades, perquè sovint ha permès veure que algunes d'elles no es conservaven de forma consistent entre grups. Això explicaria en part que les proteïnes amb més glicosilacions seleccionades són també les que han sigut comparades amb menys homòlegs.

L'altre factor que pot haver influït en la selecció de les glicosilacions és l'espècie a què corresponen les proteïnes. Concretament, les dues proteïnes amb més glicosilacions seleccionades (13 i 11 respectivament) són de *Mycobacterium tuberculosis*, cosa que fa sospitar que potser hi ha organismes, o branques evolutives, en què les glicosilacions són més freqüents que altres, o senzillament que els sequons de glicosilació buscats pel servidor són més habituals en aquestes espècies. En tot cas, *Mycobacterium*

tuberculosis és el microorganisme que més temps porta infectant l'ésser humà i el més adaptat a saltar-se el sistema immune, de manera que la sospita que les seves proteïnes estan especialment modificades per poder infectar té força sentit.

Aquests són alguns dels resultats.

- **Chaperone protein DnaK** d'*Escherichia coli O127:H6* (Figura 18)

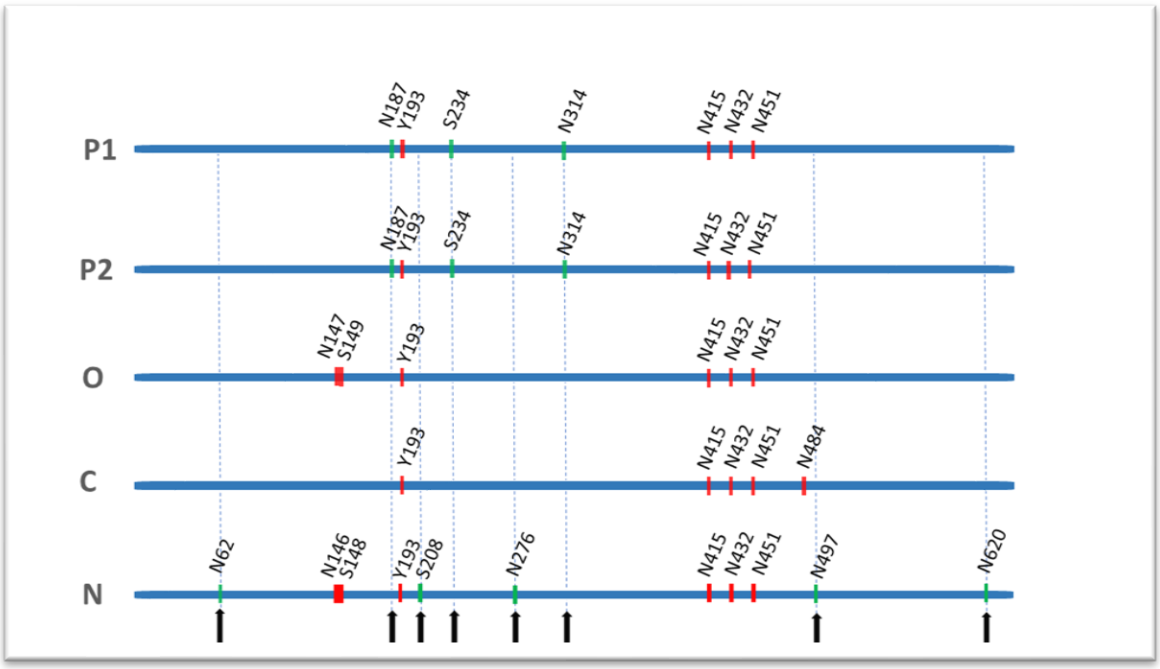

Figura 18. Posició de les glicosilacions predites a la chaperone protein DnaK de *E. coli O127:H6* (P1) i dels seus homòlegs de *S. sonnei* (P2), *P. mendocina* (O), *A. fischeri* (C) i *S. onedensis* (N).

S'han seleccionat fins a 8 glicosilacions. D'aquestes, 5 estan predites en l'homoleg de *Shewanella oneidensis*, que no té relació amb l'hoste, però no es prediuen en els altres homòlegs. Podria ser, doncs, que aquestes glicosilacions tinguessin a veure amb l'absència de capacitat d'adhesió a l'hoste. Les altres tres glicosilacions seleccionades estan presents en la proteïna d'*E. Coli* i en la d'un altre patogen, *Shigella sonnei*, cosa que les fa sospitoses de tenir relació amb la capacitat d'infectar d'aquests organismes.

- **Gliceraldehid-3-fosfat deshidrogenasa** de *Streptococcus pneumoniae* (Figura 19)

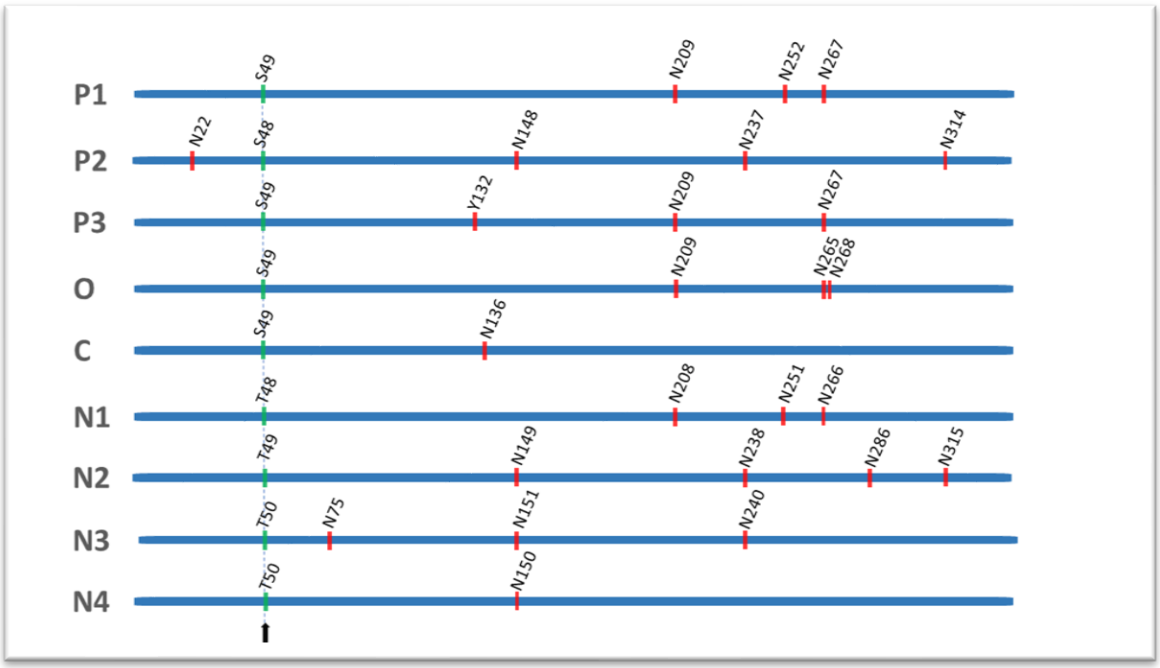

Figura 19. Posició de les glicosilacions predites a la gliceraldehid-3-fosfat deshidrogenasa de *S. pneumoniae* (P1) i dels seus homòlegs de *B. anthracis* (P2), *M. canis* (P3), *S. aureus* (O), *L. lactis* (C), *C. pasteurianum* (N1), *G. Stearotermophilus* (N2), *H. Hydrogeniformans* (N3) i *O. marismortui* (N4).

Resulta destacable la glicosilació predita a l'aminoàcid 49. Aquesta glicosilació apareix a les nou proteïnes homòlogues que s'han analitzat, però en les dels grups P, O i C estaria unida a una Serina (S) i en les del grup N a una Treonina (T). Aquesta diferència podria donar-se per atzar i no tenir rellevància funcional, però el fet que estigui ben delimitada entre els grups reforça la idea que pugui ser indicadora d'alguna variació en la funcionalitat de la proteïna.

- **Glutamine syntetase** de *Mycobacterium tuberculosis* (Figura 20)

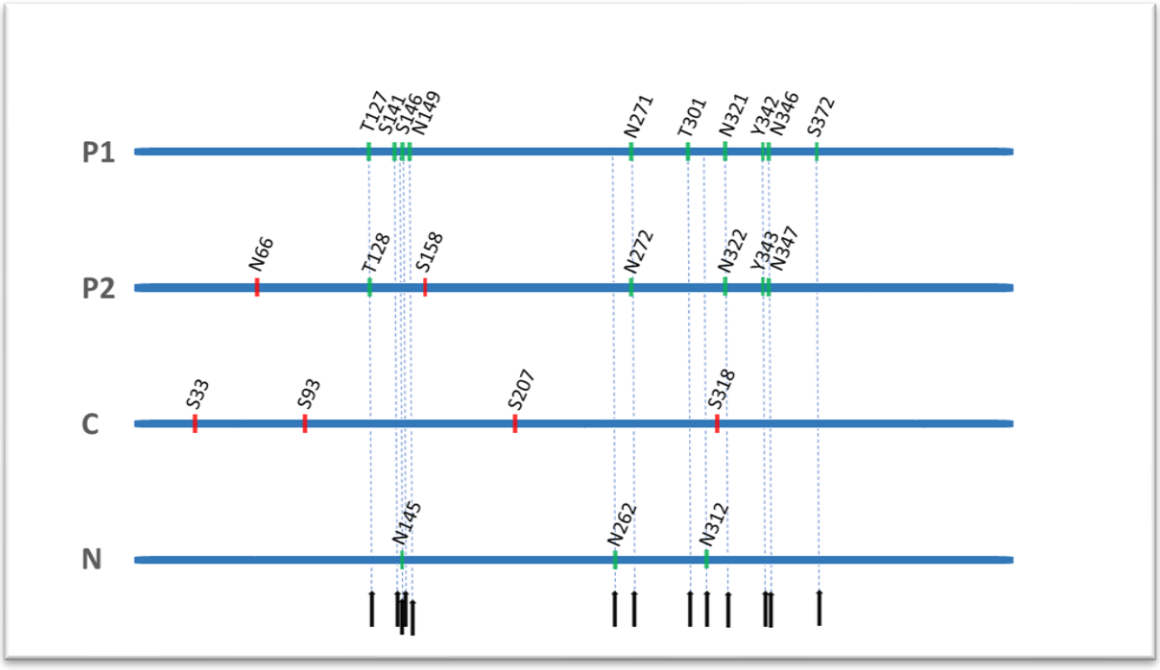

Figura 20. Posició de les glicosilacions predites a la glutamin sintetasa de *M. tuberculosis* (P1) i dels seus homòlegs de *B. adolescentis* (P2), *M. lylae* (C) i *S. coelicolor* (N).

Aquesta és la proteïna de la que més glicosilacions s'han seleccionat, 13, de les quals 10 estan presents només en patògens (grup P) i 3 només en no patògens (grup N).

Com s'ha dit, és possible que l'alt nombre es degui a que s'ha disposat de pocs homòlegs per comparar, descartant-se menys glicosilacions que en altres proteïnes, però també podria tenir a veure amb el fet que aquesta espècie tingui un nombre especialment alt d'adaptacions per infectar.

6.3. Comparacions de motius de seqüència

La gran majoria dels motius buscats no s'havien generat amb uns valors estadísticament significatius, cosa que vol dir que probablement molts d'ells no tinguin associada a una funció biològica i s'hagin trobat per atzar. Tot i així s'ha considerat d'interès veure quins d'ells es conserven entre homòlegs, perquè aquells que ho fan és més raonable pensar que realment obeeixen a una pressió selectiva, i que per tant sí que podrien tenir una funció.

No s'han seleccionat els motius que estiguessin diferencialment conservats en no patògens (amb la idea de trobar-ne de relacionats amb l'absència de virulència), perquè

els motius havien sigut generats a partir de proteïnes de patògens, i per tant les troballes que s'haguessin pogut fer haurien respost a altres funcions.

S'han seleccionat 31 motius de seqüència, entre 0 i 7 per proteïna. En la majoria de casos aquests apareixen només en la proteïna moonlighting d'interès i no en els homòlegs, cosa que no recolza la idea que els motius tinguin una funció real. De totes maneres, aquests són els resultats més destacats:

- **Chaperone protein DnaK** d'*Escherichia coli O127:H6* (Figura 21)

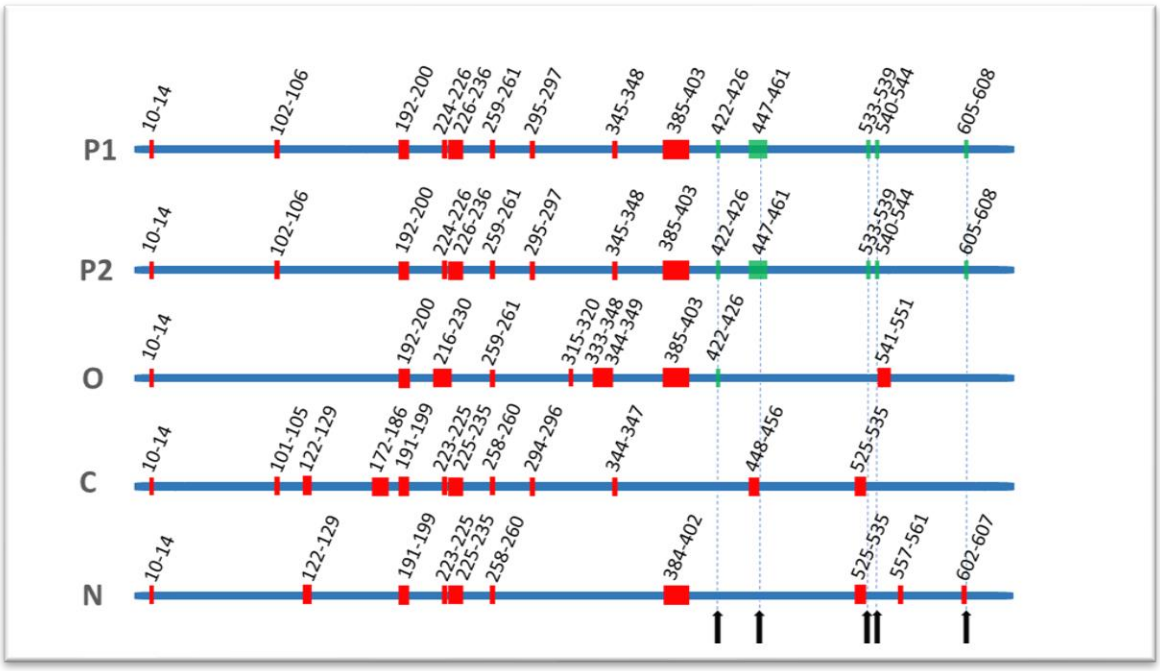

Figura 21. Posició dels motius de seqüència a la chaperone protein DnaK de *E. coli O127:H6* (P1) i dels seus homòlegs de *S. sonnei* (P2), *P. Mendocina* (O), *A. fischeri* (C) i *S. oneidensis* (N).

En aquest cas sí que s'han trobat diversos motius conservats en un altre patogen. Concretament, quatre dels motius es conserven en l'homòleg de *Shigella sonnei* (P), i no en la resta d'homòlegs, cosa que fa sospitar que tinguin relació amb la capacitat d'infectar. Un dels motius es conserva també en *Pseudomonas mendocina*, patogen oportunista.

- **Malate Synthase G** de *Mycobacterium tuberculosis* (Figura 22)

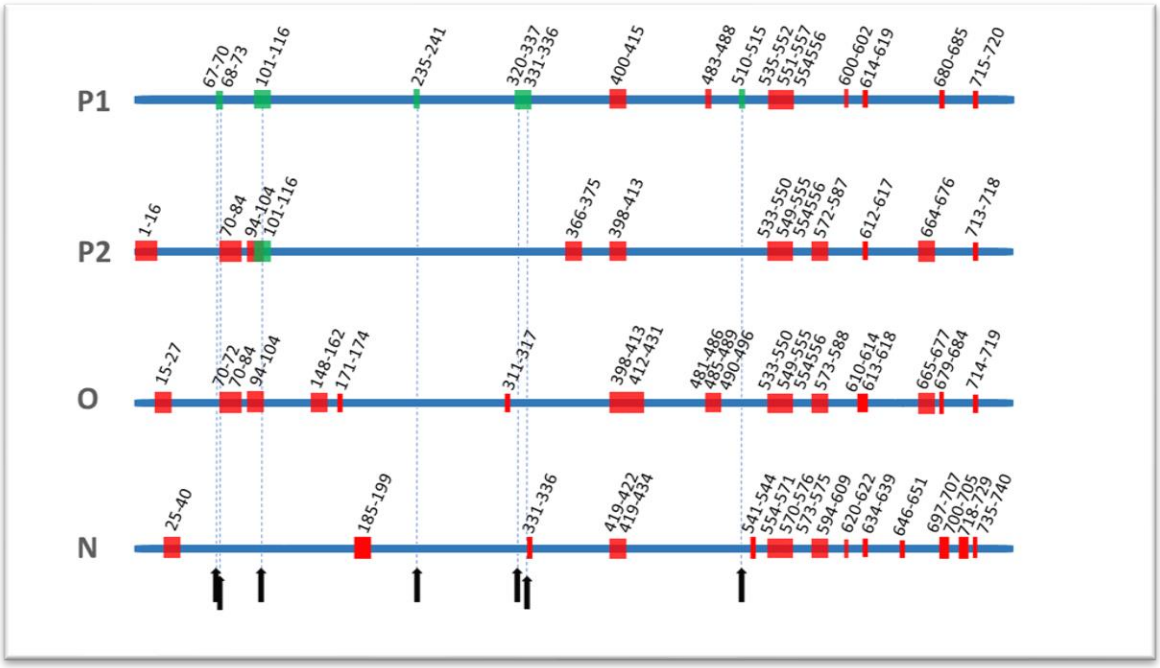

Figura 22. Posició dels motius de seqüència a la malat sintasa G de *M. tuberculosis* (P1) i dels seus homòlegs de *R. fascians* (P2), *N. farcinica* (O) i *C. efficiens* (N).

S'han trobat fins a set motius sospitosos, sis d'ells presents només en la proteïna de *Mycobacterium tuberculosis* i un també present en l'homòleg de *Rhodococcus fascians* (classificat com a P). El fet que la majoria només es trobin en un dels patògens resta verosimilitud a la suposició que els motius es corresponguin amb regions funcionals relacionades amb la capacitat d'infectar. De totes maneres, el nombre comparativament elevat de motius seleccionats (respecte les altres proteïnes) encaixaria amb la suposició, ja mencionada, que les proteïnes de *Mycobacterium tuberculosis* podrien estar especialment adaptades per infectar.

6.4. Superposició dels elements seleccionats en les comparacions

S'ha considerat interessant comprovar si hi ha coincidències en la posició dels elements que s'han seleccionat en les diferents comparacions, ja que aquestes podrien reforçar els resultats i ajudar a interpretar-los.

Dels múltiples elements sospitosos de relació amb virulència seleccionats en les deu proteïnes, només en un cas hi ha superposició entre ells. Es tracta de la glutamin

sintetasa de *Mycobacterium tuberculosis*, en què la glicosilació predita a l'aminoàcid 342 coincideix en posició amb el motiu predit entre els aminoàcids 332 i 342. Si ens fixem en el sequon utilitzat per predir la glicosilació i en el motiu concret, però, observem que, tot i que no es contradiuen, tenen molt poc en comú. És probable, doncs, que la superposició sigui una coincidència, i que o bé la glicosilació no sigui real, o bé el motiu no sigui funcional.

També hi ha el cas de l'enolasa de *Streptococcus pneumoniae*, en què s'ha seleccionat una glicosilació a l'aminoàcid 320 que coincidiria amb el motiu detectat entre les posicions 317 i 322. En aquest cas, però, la superposició no és real ja que la glicosilació està predita en un homòleg no virulent i no a l'enolasa de *S. pneumoniae* (s'ha seleccionat per la seva possible relació amb l'absència de virulència).

En definitiva, la superposició dels elements no sembla aportar nova informació. Si de cas, el poc grau de coincidència reforça la idea que no tots els elements seleccionats tenen realment funcions relacionades amb la virulència, i que per tant aquests resultats necessitarien confirmació experimental.

6.5. Comparació de motius entre proteïnes moonlighting no homòlogues

En relació amb l'objectiu *g* d'identificar patrons comuns entre proteïnes no homòlogues, s'han utilitzat els motius de seqüència seleccionats per mirar de trobar-ne. En fer la comparació, no s'ha detectat cap motiu que es repetís en diferents proteïnes.

Això pot ser degut a que realment no hi hagi cap motiu en comú entre les proteïnes seleccionades, o a que els mètodes utilitzats no hagin sigut capaços de trobar-los.

7. Conclusions

7.1. Conclusions

A grans trets, aquest treball mostra que és possible utilitzar l'estructura tridimensional de les proteïnes i la predicció de glicosilacions per, mitjançant comparacions amb proteïnes homòlogues, seleccionar elements possiblement relacionats amb la virulència. Això no garanteix que la relació sigui real però serveix per descartar molts altres elements.
L'absència d'uns resultats especialment clars indica que probablement no existeixen uns elements que determinin per sí sols la presència o absència de virulència, i que aquesta segurament ve influïda per molts factors. D'altra banda, també reflecteix les limitacions dels mètodes utilitzats: la poca disponibilitat de l'estructura tridimensional de les proteïnes d'interès, l'ús de prediccions com a mitjà per estudiar les glicosilacions, i l'ús de motius generats *de novo* a partir d'altres proteïnes moonlighting.

Aquest treball partia amb l'objectiu general de "Identificar elements estructurals i patrons de glicosilació indicadors de virulència en proteïnes moonlighting de microorganismes patògens". L'objectiu ha quedat assolit en la mesura en què s'ha obtingut un conjunt d'elements sospitosos de relació amb virulència. No hem d'oblidar que els resultats no garanteixen aquesta relació, però alhora convé posar en valor la seva utilitat de cara a dirigir futurs esforços de recerca, començant per la possible validació experimental dels resultats més prometedors.
Pel que fa als objectius específics, s'han pogut complir els objectius *a* i *b* que fan referència a les similituds i diferències d'estructura, obtenint-se un conjunt de fragments possiblement associats a virulència. El nombre de fragments és reduït i les limitacions en la disponibilitat de l'estructura tridimensional de les proteïnes han suposat una dificultat afegida, però els resultats obtinguts són prometedors de cara a la seva possible validació experimental.
Igualment, els objectius *b* i *c*, que fan referència a les similituds i diferències en els llocs predits de glicosilació, partien amb algunes limitacions, en aquest cas les pròpies de treballar amb prediccions. Si bé no és segur que tots els llocs de glicosilació seleccionats corresponguin a glicosilacions reals, sí que és raonable pensar que molts d'ells sí que ho siguin, i això fa pensar que les glicosilacions probablement sí que tenen un paper rellevant en la presència o absència de virulència.
Tal com ja s'ha explicat a l'apartat de metodologia, no ha sigut possible complir els objectius *e* i *f* que fan referència a l'estat real de glicosilació, per falta de dades disponibles per a les proteïnes d'interès. Seria interessant, quan en un futur es tingui accés a aquesta informació, poder-la utilitzar tant per fer comparacions entre proteïnes homòlogues com per comparar una mateixa proteïna en localitzacions cel·lulars diferents. Això permetria avaluar si la glicosilació realment té un paper en atorgar una nova funcionalitat a la proteïna.
L'objectiu *g* s'ha intentat dur a terme utilitzant els motius de seqüència com a possibles elements comuns entre proteïnes diferents. No s'han trobat motius comuns, cosa que ja es sospitava que podia passar, però no queda clar fins a quin punt això es deu a l'absència real d'elements comuns o a les limitacions dels mètodes utilitzats per obtenir i seleccionar els motius.

7.2. Línies de futur

D'una banda, es podria refer part del treball buscant alternatives a algunes de les eines utilitzades. En el cas de les glicosilacions, per exemple, el servidor Map Sequon possiblement no és l'eina més indicada, i algun altre sistema que predigui les glicosilacions amb algoritmes més sofisticats podria donar resultats més fiables.

Seria interessant també ampliar la cerca d'elements sospitosos de relació amb virulència utilitzant noves proteïnes moonlighting, especialment a mesura que se'n vagin coneixent més.

Pel que fa als resultats obtinguts, pot ser d'interès mirar de validar els elements seleccionats més prometedors. Si es pogués confirmar experimentalment la relació amb virulència d'alguns d'aquests elements, podria ser un pas important per entendre els mecanismes que intervenen en el procés d'infecció.

7.3. Seguiment de la planificació

La planificació s'ha seguit en general correctament, completant les tasques en els plaços previstos. L'única excepció significativa ha sigut la part d'anàlisi de l'estat real de glicosilació. En un principi, en no localitzar les dades necessàries per dur a terme aquesta tasca, s'ha allargat el temps destinat a buscar-les. Tot i així, no s'han pogut trobar dades de l'estat real de glicosilació per a les proteïnes d'interès, i la tasca no s'ha completat. En tot cas, el plantejament inicial contemplava la possibilitat que això passés, i per tant no representa una desviació realment destacable.

La metodologia escollida per dur a terme les anàlisis, amb les limitacions que s'han explicat, ha sigut en general adequada per a l'assoliment dels objectius. El principal canvi ha sigut l'addició de l'anàlisi de motius de seqüència, que ha servit per complementar les altres.

8. Glossari

Comensalisme. Relació entre dues espècies en què una d'elles, l'espècie comensal, obté un benefici, i l'altra, l'hoste, no obté cap benefici ni perjudici.

Funció moonlighting. En una proteïna multifuncional o moonlighting, és una funció no canònica, adquirida secundàriament per la proteïna.

Glicoproteïna. Proteïna sobre la qual s'ha produït una o més glicosilacions.

Glicosilació. Addició d'un carbohidrat a una altra molècula. En el cas de les proteïnes, es tracta d'una modificació cotranslacional o posttranslacional.

Homologia. Biològicament, és la relació que hi ha entre dos elements orgànics diferents que tenen un mateix origen evolutiu. A nivell de proteòmica, és la relació que existeix entre dues proteïnes diferents amb el mateix orígen evolutiu.

Motiu de seqüència. Patró de seqüència de nucleòtids o d'aminoàcids que és força comú i que sovint es té una funció biològica.

Patògen. Organisme que es relaciona amb el seu hoste infectant-lo. La relació entre els dos organismes implica un perjudici per a l'hoste.

Proteïna moonlighting. Proteïna que du a terme més d'una funció bioquímica sense veure alterada la seva seqüència d'aminoàcids.

Proteòmica. És l'estudi de les proteïnes a nivell del proteoma, és a dir l'estudi del conjunt de proteïnes d'una espècie determinada.

Sequon. Seqüència d'aminoàcids que marca la presència d'una glicosilació.

Simbiosi. Relació entre dues espècies en què ambdues obtenen un benefici.

Virulència. Capacitat d'un organisme patogen per infectar un hoste.

9. Bibliografia

Aadil H. Bhat, Homchoru Mondal, Jagat S. Chauhan, Gajendra P. S. Raghava, Amrish Methi, Alka Rao (2012). "ProGlycProt: a repository of experimentally characterized prokaryotic glycoproteins", *Nucleic Acids Research*, Volume 40, Issue D1, 1 January 2012, Pages D388–D393, https://doi.org/10.1093/nar/gkr911

Amblee V., Jeffery C.J. (2015). "Physical Features of Intracellular Proteins that Moonlight on the Cell Surface" *PLoS One*. 2015; 10(6): e0130575. https://dx.doi.org/10.1371%2Fjournal.pone.0130575

Bailey TL., Johnson J., Grant CE., Noble WS. "The MEME Suite". *Nucleic Acids Res.* 2015 Jul 1;43(W1):W39-49. https://doi.org/10.1093/nar/gkv416. Epub 2015 May 7.

Bailey TL, Williams N, Misleh C, Li WW. (2006). "MEME: discovering and analyzing DNA and protein sequence motifs". *Nucleic Acids Res.* 2006 Jul 1;34(Web Server issue):W369-73. https://doi.org/10.1093/nar/gkl198.

Blum M., Chang H., Chuguransky S., Grego T., Kandasaamy S., Mitchell A., Nuka G., Paysan-Lafosse T., Qureshi M., Raj S., Richardson L., Salazar G.A., Williams L., Bork P., Bridge A., Gough J., Haft D.H., Letunic I., Marchler-Bauer A., Mi H., Natale D.A., Necci M., Orengo C.A., Pandurangan A.P., Rivoire C., Sigrist C.J.A., Sillitoe I., Thanki N., Thomas P.D., Tosatto S.C.E., Wu C.H., Bateman A. and Finn R.D. (2020). "The InterPro protein families and domains database: 20 years on". *Nucleic Acids Research,* Nov 2020. https://doi.org/10.1093/nar/gkaa977

Choudhary P., Nagar R., Singh V., Bhat AH., Sharma Y., Rao A. "ProGlycProt V2.0, a repository of experimentally validated glycoproteins and protein glycosyltransferases of prokaryotes" *Glycobiology.* 2019 Jun 1;29(6):461-468. https://doi.org/10.1093/glycob/cwz013.

Franco-Serrano L., Cedano J., Perez-Pons J.A., Mozo-Villarias A., Piñol J., Amela I., Querol E. (2018). "A hypothesis explaining why so many pathogen virulence proteins are moonlighting proteins". *Pathogens and Disease*, (vol. 76, Issue 5, July 2018). https://doi.org/10.1093/femspd/fty046

Franco-Serrano L., Hernández S., Calvo A., Severi M.A., Ferragut G., Pérez-Pons J.A., Piñol J., Pich O., Mozo-Villarias A., Amela I., Querol E., Cedano J. (2017) "MultitaskProtDB-II: an update of a database of multitasking/moonlighting proteins". *Nucleic AcidsResearch*. 46: D1D645-D648. https://doi.org/10.1093/nar/gkx1066

Grant C.E., Bailey T.L., Noble W.S. (2011). "FIMO: scanning for occurrences of a given motif". Bioinformatics. 2011 Apr 1;27(7):1017-8. https://doi.org/10.1093/bioinformatics/btr064. Epub 2011 Feb 16.

Guex, N., Peitsch, M.C. (1997) "SWISS-MODEL and the Swiss-PdbViewer: An environment for comparative protein modeling". *Electrophoresis* 18, 2714-2723. https://doi.org/10.1002/elps.1150181505

Hamby S.E., Hirst JD. (2008) "Prediction of glycosylation sites using random forests". *BMC Bioinformatics*. 2008 Nov 27;9:500. doi: https://10.1186/1471-2105-9-500.

Henderson B., Martin A. (2011) "Bacterial Virulence in the Moonlight: Multitasking Bacterial Moonlighting Proteins Are Virulence Determinants in Infectious Disease". *Infection and Immunity* Sep; 79(9): 3476–3491 https://dx.doi.org/10.1128%2FIAI.00179-11

Kelley L., Mezulis S., Yates C., Wass M., Sternberg M. (2015) "The Phyre2 web portal for protein modeling, prediction and analysis". *Nat Protoc* 10, 845–858. https://doi.org/10.1038/nprot.2015.053

Li X., Xu Z., Hong X., Zhang Y., Zou X. "Databases and Bioinformatic Tools for Glycobiology and Glycoproteomics" *Int J Mol Sci.* 2020 Sep 14;21(18):6727. doi: https://doi.org/10.3390/ijms21186727.

Li Z., Jaroszewski L., Iyer M., Sedova M., Godzik A. "FATCAT 2.0: towards a better understanding of the structural diversity of proteins" *Nucleic Acids Res.* 2020 Jul 2;48(W1):W60-W64. doi: https://doi.org/10.1093/nar/gkaa443.

Pancholi,V., Fischetti V.A. (1992) "A major surface protein on group A streptococci is a glyceraldehyde-3-phosphate-dehydrogenase with multiple binding activity". *J Exp Med* 176 (2): 415–426. https://doi.org/10.1084/jem.176.2.415

Wang G., Xia Y., Cui J., Gu Z., Song Y., Chen Y.Q., Chen H., Zhang H., Chen W. (2013) "The Roles of Moonlighting Proteins in Bacteria". *Curr Issues Mol Biol.* 2014;16:15-22. https://www.caister.com/cimb/v/v16/15.pdf

Ye Y., Godzik A. (2004). "FATCAT: a web server for flexible structure comparison and structure similarity searching" *Nucleic Acids Res.* 2004 Jul 1;32(Web Server issue): W582-5. doi: https://doi.org/10.1093/nar/gkh430.

Annexos

Annex 1. Llista inicial de proteïnes moonlighting a estudiar

Annex 2. Llista de proteïnes moonlighting seleccionades a partir de la base de dades MultitaskProtDB-II

Annex 3. Llista de proteïnes moonlighting utilitzades en la descoberta de motius de seqüència

Annex 4. Resultats de les comparacions d'estructura

Annex 5. Resultats de les comparacions dels llocs predits de glicosilació

Annex 6. Motius de seqüència de proteïnes moonlighting de patògens

Annex 7. Resultats de les comparacions de motius de seqüència