



This is the **published version** of the bachelor thesis:

García Monreal, Lluís; Erill, Ivan, dir. Inferència de circuits de control genètic en virus. 2024. (Grau en Enginyeria Informàtica)

This version is available at <https://ddd.uab.cat/record/298944>

under the terms of the  license

Inferència de circuits de control genètic en virus

Lluís García Monreal

2 de juliol de 2024

Resum– Els bacteriòfags, o "fags", són virus que infecten bacteris i tenen el potencial de controlar les poblacions bacterianes, una característica crucial davant l'augment de les resistències bacterianes als antibiòtics. Malgrat això, els fags també poden codificar toxines perjudicials per als humans. Aquest projecte té com a objectiu desenvolupar un programa informàtic per identificar nous circuits de lisis-lisogènia en fags mitjançant el descobriment de motius. Utilitzant l'algorisme MEME per a la detecció de patrons en seqüències de nucleòtids.

Paraules clau– Bacteriòfags, lisis, lisogènia, motius, BLAST, MEME.

Abstract– Bacteriophages, or "phages", are viruses that infect bacteria and have the potential to control bacterial populations, a crucial feature in the face of increased bacterial resistance to antibiotics. However, phages can also encode toxins that are harmful to humans. This project aims to develop a computer program to identify new phage lysis lysogeny circuits by discovering motifs. Using the MEME algorithm for detecting patterns in nucleotide sequences.

Keywords– Bacteriophages, lysis, lysogeny, motifs, BLAST, MEME

1 INTRODUCCIÓ - CONTEXT DEL TREBALL

ELS bacteris són els responsables d'una àmplia gamma de malalties i actualment estan desenvolupant resistències als antibiòtics, limitant les maneres de combatre'ls. Per aquest motiu cada vegada s'estan duent a terme més treballs de recerca per a trobar alguna solució. En aquest punt apareixen els bacteriòfags o "fags", que són virus que ataquen només els bacteris. Existeixen més de 10E30[1] bacteriòfags a la terra i, entre d'altres, tenen la funció de controlar la població de bacteris al nostre món[2], [3], [4]. D'altra banda, no totes les seves funcions són beneficioses per al cos humà, ja que aquests en alguns casos, poden codificar gens per a toxines que poden ser perjudicials per als humans, com poden ser la toxina botulínica, diftèrica, del còlera o Shiga[5].

A continuació es pot observar el cicle de vida d'un fag anomenat lambda, desde la primera fase d'inserció al bacteri fins al cicle lític.

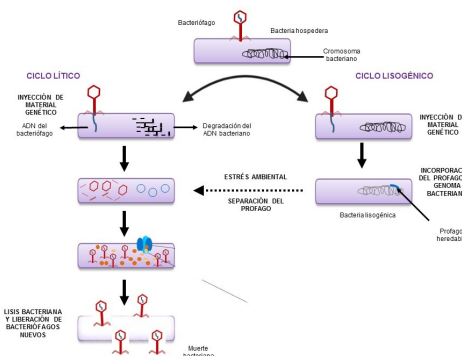


Fig. 1: Cicle de vida d'un fag

1.1 Cicle de vida

El cicle de vida d'un fag consta de tres fases principals:

Unió a l'hoste: El fag s'adhereix a la superfície del bacteri i introdueix el seu genoma viral a l'interior de la cèl·lula hoste.

Elecció del destí: Després de la infecció, el fag decideix entre dues vies:

- Fase lítica: El genoma viral es replica ràpidament, utilitzant els recursos de la cèl·lula hoste per produir noves partícules virals, que eventualment causen la lisi cel·lular.
- Fase lisogènica: El genoma viral s'integra en el geno-

• E-mail de contacte: 1532569@uab.cat
 • Menció realitzada: Tecnologies de la Informació
 • Treball tutoritzat per: Ivan Erill Sagales (Àrea de Ciències de la Computació i Intel·ligència Artificial)
 • Curs 2023/24

ma de la cèl·lula hoste i es replica juntament amb ell sense destruir la cèl·lula immediatament, romanent latent com a pròfag.[5]

Regulació del cicle de vida: Les proteïnes reguladores com C1 i Cro determinen l'entrada en la fase lítica o lisogènica, unint-se a regions específiques de l'ADN del bacteri. Aquestes regions sovint contenen seqüències conservades, conegudes com a motius, on s'uneixen aquestes proteïnes.

1.2 Motius

Els motius o "motifs" són patrons curts i freqüents en l'ADN que es presumeix que tenen una funció biològica.[6] Aquests motius es troben en la zona anomenada reguladora, que es troba entre les dues proteïnes d'unió C1 i Cro en el cas del fag Lambda.

A continuació es pot observar una imatge de les zones reguladores de diversos fags, com es pot veure apareixen rectangles que senyalen patrons o motius dins de la seqüència. Solen seguir unes pautes, en els quatre exemples aquesta imatge apareixen tres vegades durant la seqüència, són palíndroms i normalment l'espai entre els tres és molt semblant.

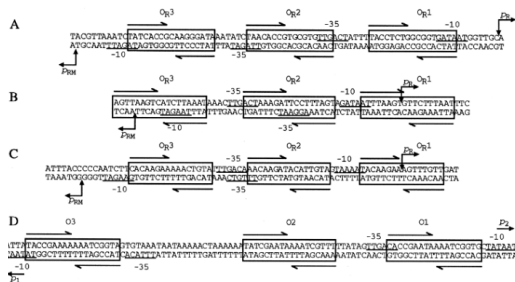


Fig. 2: Comparació d'elements reguladors d'homòlegs de cI de diversos fags[7]

1.3 Descobriment de motius

El descobriment de motius és un procés crucial en bioinformàtica per identificar seqüències de nucleòtids comunes que poden tenir funcions biològiques importants, com llocs d'unió per a proteïnes. Aquestes seqüències poden ser curtes i estar localitzades a milers de bases aigües amunt dels gens d'interès, i la seva identificació pot ser complexa a causa de la seva naturalesa sovint degenerada.

Hi ha diversos mètodes per al descobriment de motius, que es poden agrupar en tres enfocaments principals:

Enumeració: Examina exhaustivament totes les possibles seqüències dins d'un model de motiu concret.

Optimització determinista: Utilitza algorismes matemàtics per trobar el motiu que millor s'ajusta a un conjunt de dades.

Optimització probabilística: Utilitza mètodes estadístics per trobar el motiu més probable que hagi generat les dades.

Per aquest estudi, utilitzarem l'algoritme MEME (Multiple EM for Motif Elicitation), que pertany a l'enfocament d'optimització determinista. MEME utilitza un algorisme de Maximització d'Esperança (EM) per identificar motius

en seqüències de nucleòtids. Funciona de la següent manera:

Realitza una única iteració per cada n-mer en les seqüències objectives.

Selecciona el millor motiu d'aquest conjunt inicial.

Itera fins a la convergència, evitant quedar-se atrapat en òptims locals.

MEME és popular per la seva robustesa i eficiència en la identificació de motius en seqüències d'ADN, proporcionant resultats detallats que poden ajudar a descobrir patrons rellevants.

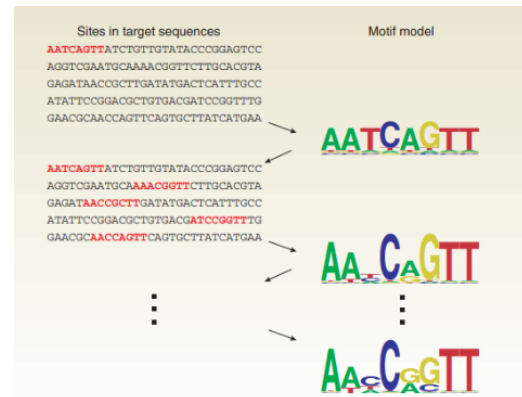


Fig. 3: Representació de MEME[8]

2 OBJECTIUS

L'objectiu principal d'aquest projecte és desenvolupar un programa informàtic que identifiqui nous circuits de lisis-lisogènia mitjançant la combinació del descobriment de motius amb la disposició de circuits de control.

Per poder arribar a complir l'objectiu final anteriorment esmentat, s'ha d'anar complint les fites o objectius secundaris reunits en el següent llistat.

- Documentació i familiarització amb el context del treball
- Documentació dels objectius del treball i metodologia que s'utilitza.
- Planificació del treball i tasques a realitzar durant el projecte.
- Familiaritzar-se amb les tecnologies de descobriment de motius i començar a identificar ortòlegs repressors. (BLAST)
- Filtratge dels resultats de BLAST, obtenció de les regions reguladores, i utilització de MEME.
- Filtratge dels resultats de MEME i obtenció de les conclusions.
- Elaboració del seguiment del projecte.

3 METODOLOGIA

Per a aquest projecte s'ha decidit utilitzar la metodologia ScrumBan, aquesta metodologia es caracteritza en la unió de Scrum i Kanban, per la part de Scrum aquest projecte està dividit en diferents esprints que tenen com a objectiu final lliurar un informe cada vegada més avançat sobre el

document final, durant aquests esprints hi ha diferents reunions planificades amb el tutor per supervisar el treball, per una altra part, Kanban servirà per a dur a terme el seguiment de la planificació, també té un petit apartat per a anar registrant les diferents reunions amb els temes que s'han comentat i la data d'aquesta. A part d'aquestes 2 metodologies, s'usaran diferents eines per a seguir el cicle de vida d'aquest treball. En primer lloc, s'usarà Trello com a eina de la metodologia Kanban que com s'ha comentat anteriorment serà l'encarregat de registrar tots els passos que es fan durant el projecte i registrar les reunions que es duuguin a terme. D'altra banda, s'usarà l'eina Zotero, aquesta és un gestor de referències bibliogràfiques que ens ajuda a tenir el control de totes les bibliografies consultades durant tot el projecte i en concret en el primer esprint. Finalment, s'usarà l'eina GitHub com a controlador de les versions del codi que en els següents esprints s'anirà implementant, aquest serà la columna vertebral del treball i l'objectiu final.

4 PLANIFICACIÓ

El projecte es dividirà en esprints, cadascun amb un objectiu lliurable específic. Tot això monitorat amb l'eina Trello.

- 1r Esprint:

Data: 18/02/2024 fins 10/03/2024

Objectiu: Documentació i familiarització amb el context del treball.

Tasques:

- Recopilar informació exhaustiva i actualitzada sobre bacteris, fags, lisis/lisogènia, circuits de control, motius, representació i descobriment de motius per a la creació de l'estat de l'art.

- Definició dels objectius, metodologia i planificació del projecte.

Lliurament: Informe inicial que documenta la informació anterior.

- 2n Esprint:

Data: 11/03/2024 fins 21/04/2024

Objectiu: Familiaritzar-se amb les tecnologies de descobriment de motius i començar a identificar ortòlegs repressors.

Tasques:

- Reunió de seguiment amb el tutor per discutir l'informe inicial.

- Aprenentatge sobre l'eina BLAST.

- Identificació ortòlegs repressors mitjançant BLAST.

- Agrupació dels ortòlegs per identitat de seqüència.

- Corregir l'informe inicial

Lliurament: Informe de progrés que resumeix les tasques fetes i la correcció de l'informe inicial.

- 3r Esprint:

Data: 22/04/2024 fins 26/05/2024

Objectiu: Investigar seqüències d'ADN, identificar regions reguladores, funció de MEME

Tasques:

- Reunió de seguiment amb el tutor per discutir l'informe de progrés i comentar els avanços

- Utilitzar els ortòlegs com a llavors per buscar seqüències properes amb BLAST.

- Identificar regions reguladores dins de les seqüències identificades.

- Filtrar seqüències molt similars per excloure motius insignificants.

- Identificar motius amb MEME.

Lliurament: Segon informe de progrés que resumeixi els resultats obtinguts i l'informe de progrés amb les seves correccions.

- 4t Esprint:

Data: 27/05/2024 fins 16/06/2024

Objectiu: Realitzar un filtratge final per identificar els motius més significatius, analitzar la informació i extreure conclusions.

Tasques:

- Aplicar criteris de filtratge addicionals per identificar els motius més significatius.

- Analitzar la informació obtinguda sobre els motius i els circuits de control.

- Extreure conclusions sobre els nous circuits de lisis-lisogènia identificats.

Lliurament: Proposta de l'informe final amb la informació de l'últim informe de progrés i la nova informació extreta en aquest esprint.

- 5t Esprint:

Data: 16/06/2024 fins 10/07/2024

Objectiu: Finalitzar el Treball de Fi de Grau (TFG) amb la defensa oral i lliurament de tots els materials requerits.

Tasques: - Revisió i correcció de la proposta d'informe final. - Elaboració de la proposta de presentació. - Confecció del dossier. - Disseny del pòster. - Preparació de la defensa del TFG

Lliurament: Dossier, presentació i pòster.

5 SEGUIMENT

En aquest apartat es parla de l'evolució del projecte, l'explicació del treball desenvolupat durant l'esprint i el seguiment de la planificació.

Durant la primera fase del projecte i com es va planificar en primer moment, es va dedicar el temps en la comprensió i aprenentatge del tema a tractar com era els bacteriòfags, cicle de vida d'aquests i dels motius, i sintetitzar tota aquella informació en l'estat d'art del projecte. Per una altra banda, es va seleccionar la metodologia que més s'adaptava a aquest projecte i es va definir la planificació a seguir.

Durant aquest temps es van fer dues reunions de seguiment per a l'aprenentatge sobre el tema en context i per ajudar a elaborar una planificació viable.

Seguidament, després del tancament del primer esprint es va fer una reunió per discutir com va anar el primer esprint i veure si la planificació s'anava seguint adequadament, a part es va corregir l'informe lliurat en l'anterior esprint. Per una altra banda, es va començar a parlar sobre la primera eina que és s'anava a utilitzar que era BLAST, per poder començar a treballar amb aquesta tecnologia primer el tutor en va compartir uns fitxers d'entrenament per donar les primeres passes amb BLAST. A partir d'aquí es va començar a implementar el programa final, el primer pas va ser crear el JSON amb els paràmetres d'inicialització (e-value, BD, coverage, proteïnes, taxonomia, etc.), en segon pas es va crear la funció de main que és l'encarregada de llegir el fitxer JSON anterior i inicialitzar les variables i trucar a la funció searchBlast. La següent tasca va ser crear les funcions searchBlast, i obtenirSecuenciaProteinaPorId, com es pot observar en els seus noms la primera és l'encarregada de fer la cerca a BLAST, la segona d'obtenir la seqüència d'RNA a partir de la id de la proteïna necessària per fer la cerca a BLAST. Durant la implementació es va fer una reunió per resoldre uns dubtes, sobre la implementació. A part de la programació en aquest esprint s'ha corregit l'informe inicial, s'ha afegit el descobriment de motius i la secció de seguiment.

El tercer esprint va començar amb una reunió per corregir l'informe de l'anterior esprint i parlar sobre els següents objectius per aquest esprint i analitzar si la planificació estava seguint adequadament. Per part del treball de documentació les tasques executades han sigut, la correcció de l'informe anterior, creació del seguiment d'aquest esprint i elaboració de l'informe en format article, per començar a familiaritzar-se amb el format que s'ha de lliurar al final del projecte.

Per part de la programació, com s'explica a la planificació s'ha realitzat l'obtenció dels resultats de la cerca a BLAST, l'obtenció de les seqüències dels "hits" de BLAST, l'alineament de les seqüències i el filtratge de les seqüències similars. Per una altra part, s'ha realitzat les variables i el fitxer d'entrada per l'algoritme MEME. Durant la realització d'aquestes funcions han aparegut dubtes que s'han solucionat amb l'ajuda del tutor, entre elles l'obtenció de la seqüència reguladora de cada "hit", i la trucada a MEME des del programa.

A continuació es pot observar el diagrama de flux del programa que s'està realitzant.

El quart esprint, com tots els esprints anteriors començava amb una reunió amb el tutor per parlar sobre l'esprint anterior, l'informe entregat. Per aquest últim esprint es va decidir fer una reunió setmanal per portar un millor seguiment per a la recta final. Abans de començar a treballar es va dur a terme una altra reunió per parlar dels següents passos a fer, entre ells es va parlar de com utilitzar els serveis de MEME, els paràmetres a utilitzar per a la crida i els paràmetres que s'utilitzaran per ranquejar els motius. Per la part de programació s'han corregit errors tant en l'alineament de les seqüències, ja que estava utilitzant una biblioteca obsoleta, també s'han corregit errors en el filtratge de les seqüències. Les noves funcions realitzades han sigut la crida a MEME, el parseig del resultat de MEME i

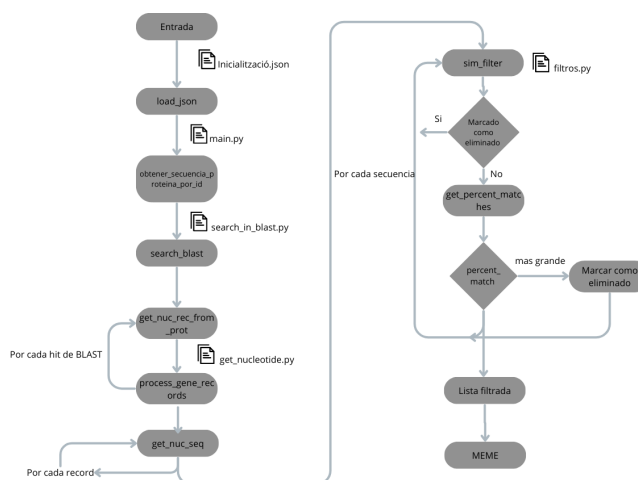


Fig. 4: Representació del programa

la classificació dels motius. Per poder fer la crida a MEME s'ha utilitzat una imatge de MEME amb Docker per poder utilitzar els seus serveis.

Per part de l'informe s'ha elaborat l'apartat de resultats, per poder elaborar-lo durant l'última reunió de seguiment es va parlar dels bacteriòfags que anàvem a utilitzar per poder provar el programa, vam buscar les id de les proteïnes repressores dels bacteriòfags seleccionats, com pot ser L-Lactococcus phage BK5-T": [ÇAC80176.1", ÇAC80178.1"], aquestes són uns dels paràmetres inicials per poder corre el programa, també es van pautar tota la resta de paràmetres inicials.

A continuació es va executar el programa per a cada bacteriòfag, tota aquesta informació s'ha recollit en les taules que es poden observar en l'apartat de resultats.

Finalment, es va elaborar la conclusió del projecte, tenint en compte la part teòrica i bibliogràfica i els resultats obtinguts. Per a poder realitzar aquesta conclusió es va realitzar una reunió amb el tutor per a comentar els resultats aconseguits i monitoritzar les conclusions obtingudes.

6 RESULTATS

En aquesta secció es recull tota la informació rellevant extreta dels diferents resultats del programa elaborat per aquest treball.

Abans d'entrar als resultats anem a explicar els diferents paràmetres utilitzats durant l'execució del programa. Aquest paràmetres estan dividits per quatre grups:

- Bacteriòfag: Com indica el nom, recull els identificadors de les proteïnes repressores del bacteriòfag.
- BLAST: Aquest grup el componen tots els paràmetres per a la crida i recerca de BLAST, com poden ser, la id taxonomica per exemple: 2 (Bacteries) o 10239(Virus) que han sigut les dues utilitzades, la base de dades on es buscarà, el e-value, el coverage i el màxim de hits a buscar.
- Filtratge: en aquest grup tenim els paràmetres que tenen a veure amb el filtratge de seqüències com pot ser les màxim de similitud o la longitud mínima de la seqüència.
- MEME: per últim tenim els paràmetres que tenen a veure amb la crida a MEME, com poden ser la longitud

mínima i màxima dels motius, si han de ser palíndroms o no, número de motius, etc.

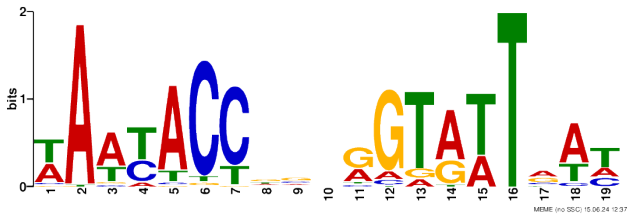
En aquest cas els paràmetres que hem utilitzat són la taxonomia del bacteri, és a dir estem buscant a BLAST proteïnes similars a la seleccionada dins d'una base de dades de bacteris, ja que si es buscava a la base de dades de Virus els "hits" trobat eren pocs en molt casos, per això s'ha preferit utilitzar la taxonomia del bacteri.

TAULA 1: RESULTATS

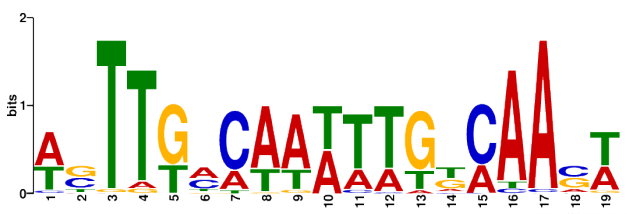
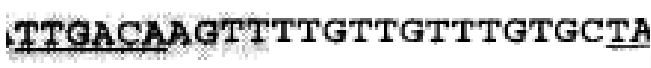
Bacteriòfag	Proteïnes	NºSeqüències	NºMotius
Escherichia phage Lambda	["ALA45761.1", "ABD60142.1"]	54	10
Salmonella phage P22	["AAF75024.1", "AAM81424.1"]	28	10
Escherichia phage 434	["ÜMO75315.1", "ÜMO75316.1"]	72	10
Lactobacillus phage A2	["NP680505.1", "NP680506.1"]	182	10
Lactobacillus phage phig1e	["NP695143.1", "NP695142.1"]	106	10
Lactococcus phage BK5-T	["ÇAC80176.1", "ÇAC80178.1"]	57	10
Lactococcus phage phiLC3	["NP996677.1", "NP996678.1"]	49	10
Lactococcus phage r1t	["NP695032.1", "NP695031.1"]	48	10

Finalment, es poden observar les taules amb els resultats del motiu més significatiu de cada bacteriòfag. En les dues primeres files de la taula es pot observar el motiu, la mitjana d'instàncies per seqüència del motiu i el nombre de seqüències on apareix el motiu entre el nombre total de seqüències. En la següent fila es pot observar una imatge del motiu extreta d'un informe del bacteriòfag, amb aquesta imatge i la imatge anterior del motiu extret pel programa es pot comparar si coincideix l'extret amb el programa i la teoria. Per una altra banda, en les següents files es pot observar per a cada cas de quantitat d'instàncies, per exemple en el cas que la seqüència té 3 instàncies, la mitja de les dues distàncies que existeixen entre les instàncies del motiu, la desviació estàndard de la distància i per últim en nombre de casos de seqüències en aquest exemple 3. Per finalitzar, l'última fila és un altre cas d'estudi, en aquest s'ha decidit calcular les distàncies i les desviacions estàndard tenint en compte que totes les seqüències amb el motiu tenen el màxim d'instàncies del motiu. Per poder fer això si una seqüència no tenia el nombre màxim d'instàncies es creaven de manera aleatòria entre l'última instància i el final de la seqüència i tenint en compte la dimensió del motiu, noves instàncies fins a arribar al màxim. Per exemple com es pot observar en la primera taula, la seqüència del bacteriòfag lambda amb més instàncies és 3, aquest és el màxim, aleshores en totes les altres seqüències on existeix el motiu s'han creat noves instàncies fins a arribar a les 3. Això és far per poder agrupar totes les seqüències on està el motiu i poder fer un cas estadístic de totes com iguals.

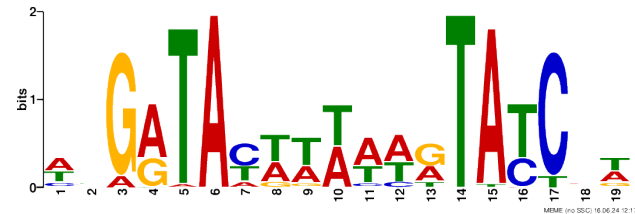
TAULA 2: ESCHERICHIA PHAGE LAMBDA

Motiu	Mitjana d'instàncies	Nºinstàncies/ NºSeqüències	-
	2.76	37/54	-
 [7]	-	-	-
Nºinstàncies	Distàncies	Desviació estàndard	Nº de casos
3 instàncies	(5.45, 6.32)	(0.50, 0.78)	31/37
2 instàncies	5.00	0.00	3/37
1 instància	-	-	3/37
Cas 2	Distàncies	Desviació estàndard	Nº de casos
-	(5.65, 5.30)	(0.94, 8.72)	-

TAULA 3: LACTOCOCCUS PHAGE R1T

Motiu	Mitjana d'instàncies	Nºinstàncies/ NºSeqüències	-
	2.11	27/48	-
 [9]	-	-	-
Nºinstàncies	Distàncies	Desviació estàndard	Nº de casos
3 instàncies	(45.12, 25.50)	(68.82, 33.70)	8/27
2 instàncies	10.00	8.96	14/27
1 instància	-	-	5/27
Cas 2	Distàncies	Desviació estàndard	Nº de casos
-	(31.70, 22.37)	(51.30, 24.17)	-

TAULA 4: LACTOBACILLUS PHAGE PHIG1E

Motiu	Mitjana d'instàncies	Nºinstàncies/ NºSeqüències	
	3.33	44/106	
5'-GATATGATACGTTTGGTATCCTTAT-3' [10]	-	-	
Nºinstàncies	Distàncies	Desviació estàndard	Nº
5 instàncies	(40.86, 11.71, 10.86, 10.00)	(75.77, 2.31, 0.83, 5.29)	
4 instàncies	(29.12, 40.06, 10.88)	(33.46, 35.33, 13.04)	
3 instàncies	(58.29, 35.43)	(47.28, 35.43)	
2 instàncies	51.83	51.34	
1 instància	-	-	
Cas 2	Distàncies	Desviació estàndard	Nº
-	(37.23, 28.86, 8.05, 3.16)	(47.41, 33.26, 15.51, 12.84)	

TAULA 5: LACTOCOCCUS PHAGE BK5-T

Motiu	Mitjana d'instàncies	Nºinstàncies/ NºSeqüències	-
	2.39	28/57	-
TATACCGAAAAAATCGGTAGT TATGGCTTTTATTAGCCATC [7]	-	-	-
Nºinstàncies	Distàncies	Desviació estàndard	Nº de casos
3 instàncies	(8.87, 20.73)	(3.30, 0.85)	15/28
2 instàncies	27.56	13.62	9/28
1 instància	-	-	4/28
Cas 2	Distàncies	Desviació estàndard	Nº de casos
-	(20.75, 19.71)	(18.36, 22.23)	-

Ara que ja hem vist els resultats obtinguts, podem fer unes petites comparacions i extraure informació de tot el conjunt. Primer, es pot parlar de la longitud del motiu. Com es pot observar en la imatge del motiu en qualsevol de les taules, la longitud d'aquest és molt semblant en tots els casos, ja que la seva longitud mitjana és de 20 nucleòtids. Aquesta longitud uniforme pot suggerir que aquests elements tenen una funció conservada a través de les diferents seqüències analitzades, entre altres teories com que necessiten una longitud mínima o màxima per poder realitzar la seva funció biològica. En alguns casos les desviacions estàndard de la distància entre les instàncies dels motius són bastant petites, això vol dir que en la gran part dels casos la distància és molt semblant en totes les seqüències on podem trobar aquest motiu. En gran part dels casos les mitjanes de les distàncies i les desviacions augmenten quan estem parlant dels casos on no tenim el màxim d'instàncies en una seqüència, moltes vegades això pot passar perquè s'ha perdut algun motiu que estava entre els motius trobats. Un exemple pot ser en el cas del bacteriòfag "Escherichia phage 434", quan calculem les estadístiques en el cas de dues instàncies la distància i la desviació augmenta considerablement, en aquest cas el més segur és que entre les dues instàncies trobades realment abans existia un altre que feia més constant com en el cas dels casos amb tres instàncies la distància i la desviació estàndard.

Una altra observació que es pot veure reflectit en els resultats del segon cas estadístic és l'alteració de les desviacions estàndard. Com es pot observar en el fag Lambda, la desviació estàndard de la segona distància s'ha vist greument alterada, ja que ha passat de 0,78 a 8,72, això passa, perquè com estem instanciant 6 noves distàncies de manera aleatòria, aquesta estadística es veu greument afectada, en el cas de la distància mitjana no es veu tan clar, pel fet que aquestes noves distàncies és només una petita part de la mostra.

Una altra dada que podem extreure d'aquests resultats és que la mitjana total de nombre d'instàncies per seqüència del motiu més significatiu és de 2,58, en la majoria dels casos el nombre de seqüències en el nombre d'instàncies que més hi havia és de 3, com es pot veure reflectit en la mitja, i com es podien veure en els distints estudis comentats durant la introducció dels motius i com es pot observar en la Fig 2 amb l'exemple de 4 bacteriòfags diferents.

Finalment, un dels resultats més clars que es poden observar en totes les taules excepte una és la coincidència entre el motiu extret pel programa i els motius descoberts teòricament. Com es pot observar en les taules, hi ha coincidències entre les dues fonts d'informació. Si ens fixem per exemple en *Lactobacillus phage phig1e* es pot observar la seqüència GATA i 7 posicions més endavant TATC, si seguidament observem la imatge de sota la seqüència que està en negreta també es pot observar les mateixes seqüències i a la mateixa distància. Això també es pot observar a la taula del *Lactococcus phage BK5-T* en aquest cas CG seguidament de 7 A i TCG, això es pot observar tant en l'extret com en la imatge. Tot això s'observa en 7 dels 8 casos que hem estudiat, per això tinc la certesa que en el cas del bacteriòfag *Lactococcus phage r1t*, durant l'elaboració de l'informe científic hi va haver alguna equivocació alhora de la identificació del motiu d'aquest bacteriòfag.

7 CONCLUSIONS

Com s'ha esmentat a la introducció, els motius són seqüències específiques que es troben en la regió reguladora de l'ADN d'un virus o bacteri. En aquest estudi, s'han cercat motius en bacteris. Aquests motius solen aparèixer repetidament dins de la seqüència i mantenen una distància similar entre ells.

Les imatges dels resultats de tres bacteriòfags diferents mostren que els motius es repeteixen al llarg de les seqüències amb una distància constant. Tot i que no sempre apareixen tres vegades i en algunes seqüències no es poden observar, en més del 50% i en molts casos més del 70% de les seqüències aquest motiu sol aparèixer. A més del motiu més significatiu, es poden identificar altres motius que no tenen gran quantitat d'instàncies, però que apareixen en distintes seqüències diferents. Això suggereix que els bacteriòfags comparteixen patrons de seqüències similars.

També s'observa que els motius més significatius trobats en els diferents bacteriòfags comparteixen característiques similars: solen aparèixer tres vegades en una seqüència, tenen una longitud comparable i presenten una desviació estàndard de distància similar.

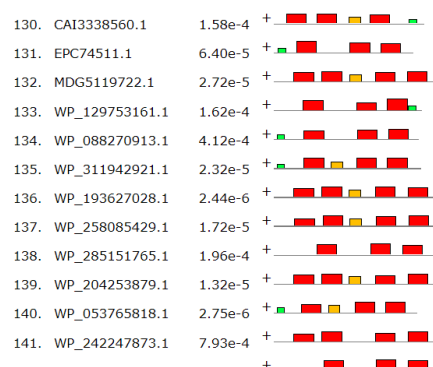


Fig. 5: Representació del motiu més significatiu del *Lactobacillus phage A2*

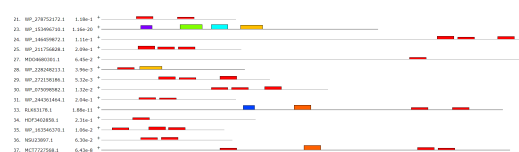


Fig. 6: Representació del motiu més significatiu del *Lactococcus phage r1t*

Els resultats obtinguts són un nou descobriment i validen teories sobre els bacteriòfags com poden ser:

La regulació gènica comuna: La presència de motius similars en la regió reguladora dels bacteriòfags pot indicar mecanismes de regulació gènica conservats entre diferents bacteriòfags. Això podria ajudar a comprendre com els bacteriòfags controlen l'expressió gènica en les seves cèl·lules hoste.

L'evolució convergent: Els patrons similars podrien suggerir una evolució convergent, on diferents bacteriòfags han desenvolupat mecanismes similars per adaptar-se a entorns similars o hostes bacterians. I ens poden ajudar per a futures teràpies, ja que identificar motius comuns pot ser útil

per al desenvolupament de teràpies basades en bacteriòfags. Si es coneixen els patrons reguladors, es poden dissenyar fàrmacs o teràpies genètiques que interfereixin amb aquests mecanismes, millorant l'eficàcia del tractament.

En conclusió, els resultats obtinguts suggereixen que els bacteriòfags tenen la característica de tenir, dins de la seva regió reguladora, patrons de tres fragments idèntics amb una distància constant entre ells. Això no només reforça la idea que els bacteriòfags comparteixen patrons reguladors similars, sinó que també obre la porta a diverses aplicacions pràctiques i àrees d'investigació futura.

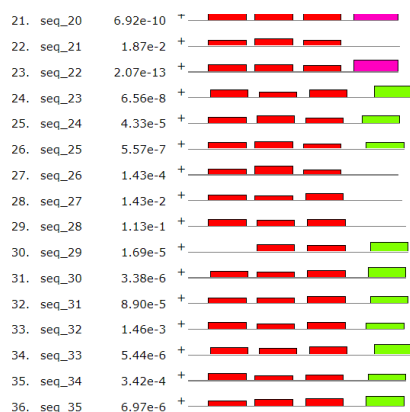


Fig. 7: Representació del motiu més significatiu del Escherichia phage Lambda

REFERÈNCIES

- [1] S. Minot, A. Bryson, C. Chehoud, G. D. Wu, J. D. Lewis, y F. D. Bushman, «Rapid evolution of the human gut virome», *Proc. Natl. Acad. Sci. U. S. A.*, vol. 110, n.o 30, pp. 12450-12455, jul. 2013, doi: 10.1073/pnas.1300833110.
- [2] A. Martínez, «Bacteriòfags, ¿una alternativa a los antibióticos?», AINIA. Accedido: 3 de marzo de 2024. [En línea]. Disponible en: <https://www.ainia.es/ainia-news/bacteriofagos/>
- [3] R. O. A. Jariah y M. S. Hakim, «Interaction of phages, bacteria, and the human immune system: Evolutionary changes in phage therapy», *Rev. Med. Virol.*, vol. 29, n.o 5, p. e2055, sep. 2019, doi: 10.1002/rmv.2055.
- [4] F. Safari et al., «The interaction of phages and bacteria: the co-evolutionary arms race», *Crit. Rev. Biotechnol.*, vol. 40, n.o 2, pp. 119-137, mar. 2020, doi: 10.1080/07388551.2019.1674774.
- [5] J. Roughgarden, «Lytic/Lysogenic Transition as a Life-History Switch», *Virus Evol.*, vol. 10, n.o 1, p. veae028, 2024, doi: 10.1093/ve/veae028.
- [6] P. D'haeseleer, «What are DNA sequence motifs?», *Nat. Biotechnol.*, vol. 24, n.o 4, pp. 423-425, abr. 2006, doi: 10.1038/nbt0406-423.
- [7] J. D. Boyce, B. E. Davidson, y A. J. Hillier, «Identification of prophage genes expressed in lysogens of the *Lactococcus lactis* bacteriophage BK5-T.», *Appl. Environ. Microbiol.*, vol. 61, n.o 11, pp. 4099-4104, nov. 1995.
- [8] P. D'haeseleer, «How does DNA sequence motif discovery work?», *Nat. Biotechnol.*, vol. 24, n.o 8, pp. 959-961, ago. 2006, doi: 10.1038/nbt0806-959.
- [9] A. Nauta, O. van Sinderen, H. Karsens, E. Smit, G. Venema, y J. Kok, «Inducible gene expression mediated by a repressor-operator system isolated from *Lactococcus lactis* bacteriophage r1t», *Mol. Microbiol.*, vol. 19, n.o 6, pp. 1331-1341, 1996, doi: 10.1111/j.1365-2958.1996.tb02477.x.
- [10] M. Kakikawa, S. Ohkubo, M. Syama, A. Taketo, y K.-I. Kodaira, «The genetic switch for the regulatory pathway of *Lactobacillus plantarum* phage g1e: characterization of the promoter PL, the repressor gene cpg, and the cpg-encoded protein Cpg in *Escherichia coli*», *Gene*, vol. 242, n.o 1, pp. 155-166, ene. 2000, doi: 10.1016/S0378-1119(99)00531-4.
- [11] J. M. Blatny, P. A. Risøen, D. Lillehaug, M. Lunde, y I. F. Nes, «Analysis of a regulator involved in the genetic switch between lysis and lysogeny of the temperate *Lactococcus lactis* phage phi LC3», *Mol. Genet. Genomics MGG*, vol. 265, n.o 1, pp. 189-197, mar. 2001, doi: 10.1007/s004380000407.
- [12] V. Ladero, P. García, J. C. Alonso, y J. E. Suárez, «A2 Cro, the Lysogenic Cycle Repressor, Specifically Binds to the Genetic Switch Region of *Lactobacillus casei* Bacteriophage A2», *Virology*, vol. 262, n.o 1, pp. 220-229, sep. 1999, doi: 10.1006/viro.1999.9903.

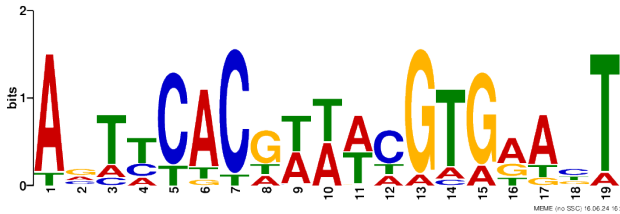
8 AGRAÏMENTS

Vull agrair tot l'esforç, educació i suport que m'han donat els meus pares durant tots aquests anys i especialment aquest últim semestre, també vull agrair la paciència de la meua parella i el suport durant cada tarda fent aquest treball. Finalment, donar les gràcies al meu tutor Ivan Erill per tenir la paciència d'explicar-me més d'una vegada tot el context d'aquest treball. També gràcies pel temps dedicat durant aquests últims mesos, i per tenir sempre disponibilitat per poder fer reunions i ajudar-me durant tot el procés. Per acabar, vull tornar a agrair a tota la meua família pel suport que m'han brindat durant tots aquests anys del grau. Moltes gràcies.

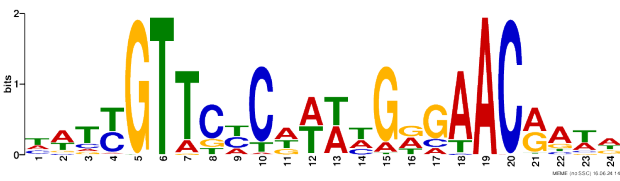
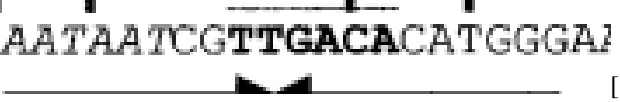
9 APÈNDIX

En aquest apartat es poden observar les taules restants dels resultats obtinguts i esmentats en la secció Resultats.

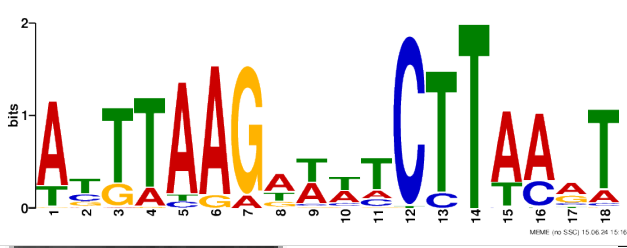
TAULA 6: LACTOCOCCUS PHAGE PHILC3

Motiu	Mitjana d'instàncies	Nºinstàncies/ NºSeqüències	-
	1.96	26/49	-
	-	-	-
Nºinstàncies	Distàncies	Desviació estàndard	Nº de casos
2 instàncies	29.20	8.45	25/26
1 instància	-	-	1/26
Cas 2	Distàncies	Desviació estàndard	Nº de casos
-	28.58	8.85	-

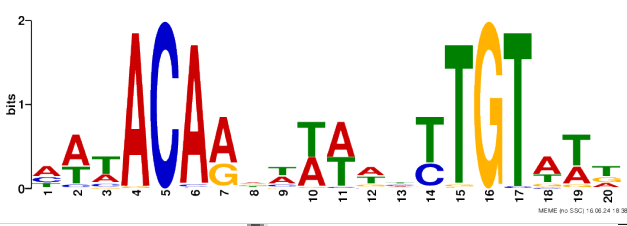
TAULA 7: LACTOBACILLUS PHAGE A2

Motiu	Mitjana d'instàncies	Nºinstàncies/ NºSeqüències	-
	2.78	76/182	-
	-	-	-
Nºinstàncies	Distàncies	Desviació estàndard	Nº de casos
4 instàncies	(9.71, 39.57, 14.86)	(0.70, 0.73, 0.35)	7/76
3 instàncies	(33.61, 18.02)	(13.09, 9.83)	54/76
2 instàncies	47.89	61.57	9/76
1 instància	-	-	6/76
Cas 2	Distàncies	Desviació estàndard	Nº de casos
-	(31.30, 18.66, 9.08)	(26.28, 11.69, 11.69)	-

TAULA 8: SALMONELLA PHAGE P22

Motiu	Mitjana d'instàncies	Nºinstàncies/ NºSeqüències	-
	2.84	20/28	-
 [7]	-	-	-
Nºinstàncies	Distàncies	Desviació estàndard	Nº de casos
4 instàncies	(1.00, 10.00, 12.00)	(0.0.0)	1/20
3 instàncies	(13.29, 10.50)	(6.34, 2.23)	14/20
2 instàncies	34.50	0.50	4/20
1 instància	-	-	1/20
Cas 2	Distàncies	Desviació estàndard	Nº de casos
-	(17.11, 9.00, 0.89)	(10.85, 3.45, 7.93)	-

TAULA 9: ESCHERICHIA PHAGE 434

Motiu	Mitjana d'instàncies	Nºinstàncies/ NºSeqüències	-
	2.75	48/72	-
 [7]	-	-	-
Nºinstàncies	Distàncies	Desviació estàndard	Nº de casos
3 instàncies	(3.72, 3.42)	(0.51, 0.49)	36/48
2 instàncies	19.92	9.20	12/48
1 instància	-	-	0/48
Cas 2	Distàncies	Desviació estàndard	Nº de casos
-	(7.77, 5.38)	(8.40, 3.59)	-