

LATVIJAS UNIVERSITĀTE



Elīna Černooka

**Bakteriofāga Enc34 neraksturotu konservatīvu
proteīnu strukturāli un funkcionāli pētījumi**

**Structural and functional studies of uncharacterized
conserved proteins from bacteriophage Enc34**

PROMOCIJAS DARBA KOPSAVLIKUMS

Zinātnes doktora grāda iegūšanai Bioloģijas nozarē
Molekulārās bioloģijas apakšnozarē

SUMMARY OF DOCTORAL THESIS

For the degree of Doctor of Science in Biology
Subfield of Molecular Biology

Rīga, 2022

Latvijas Universitāte
Bioloģijas fakultāte

Elīna Černooka

**BAKTERIOFĀGA ENC34 NERAKSTUROTU
KONSERVATĪVU PROTEĪNU STRUKTURĀLI
UN FUNKCIONĀLI PĒTĪJUMI**

Promocijas darba kopsavilkums

Zinātnes doktora grāda iegūšanai Bioloģijas nozarē
Molekulārās bioloģijas apakšnozarē

Rīga, 2022

Promocijas darbs tika izstrādāts Latvijas Biomedicīnas pētījumu un studiju centrā no 2016. līdz 2022. gadam.



Latvijas Biomedicīnas
pētījumu un studiju centrs
biomedicīnas pētījumi un izglītība no ģenēiem līdz cilvēkam

Šis darbs ir realizēts ar SIA “MikroTiks” (MikroTik) stipendijas, ko administrēja Latvijas Universitātes fonds, un Eiropas Sociālā fonda granta nr. 8.2.2.0/20/I/006 atbalstu.



NACIONĀLAIS
ATTĪSTĪBAS
PLĀNS 2020



EIROPAS SAVIENĪBA
Eiropas Sociālais
fonds

IEGULDĪJUMS TAVĀ NĀKOTNĒ

Darba forma: publikāciju kopa Molekulārās bioloģijas apakšnozarē

Darba zinātniskais vadītājs: Dr. biol. Andris Kazāks

Darba recenzenti:

- 1) Dr. biol. Kalvis Brangulis, Latvijas Biomedicīnas pētījumu un studiju centrs
- 2) Prof., Dr. chem. Kristaps Jaudzems, Latvijas Organiskās sintēzes institūts
- 3) Asst. prof. Emilio Parisini, PhD, Istituto Italiano di Tecnologia un Latvijas Organiskās sintēzes institūts

Promocijas darba aizstāvēšana notiks Latvijas Universitātes Bioloģijas nozares promocijas padomes atklātā sēdē, 2023. gada 12. maijā plkst. 13:00 Latvijas Biomedicīnas pētījumu un studiju centrā, Rātsupītes ielā 1 k-1.

Ar promocijas darbu un tā kopsavilkumu var iepazīties Latvijas Universitātes bibliotēkā Rīgā, Kalpaka bulvārī 4.

Latvijas Universitātes Bioloģijas zinātņu nozares promocijas padomes priekšsēdētājs _____ / Prof., Dr. biol. Nils Rostoks / sekretāre _____ / Dr. biol. Daina Eze /

© Latvijas Universitāte, 2022

© Elīna Černooka, 2022

ANOTĀCIJA

Vairumam fāgu būtiska to genomu daļa pieder t.s. “biosfēras tumšajai matērijai”, kas uzrāda zemu vai nenosakāmu līdzību ar raksturotām genomiskajām sekvencēm un tādēļ ietver daudzus saistošus mērķproteīnus eksperimentālai izpētei, kas galu galā varētu atklāt šo proteīnu bioloģiskos aspektus vai praktiskā pielietojuma potenciālu. *Hafnia* vīruss Enc34, ko atklāja mūsu zinātniskā grupa, kodē vairākus šādus proteīnus, tostarp atvērtā lasīšanas rāmja (*open reading frame*, ORF) 6 hipotētisko DNS replikācijas proteīnu un ORF39 iespējamo endolizīnu, kas abi pārstāv evolūcijas gaitā saglabājušās, bet līdz šim neraksturotas proteīnu saimes. Lai gūtu labāku izpratni par šiem proteīniem, tika producēti, attīrīti un kristalizēti to natīvie, mutantie un selenometionīna aizvietotie varianti, un tika eksperimentāli noteiktas to trīsdimensionālās struktūras. Tālākas strukturālās homoloģijas analīzes atklāja, ka Enc34 ORF6 proteīns ir fāga T7 tipa vienpavediena DNS saistošo proteīnu (*single-stranded DNA-binding proteins*, SSBs) saimes diverģents loceklis. Tā DNS saistīšanas īpašības tika pārbaudītas *in vitro*, un tika noteikta arī proteīna kristālstruktūra kompleksā ar DNS oligonukleotīdu. Šī kompleksa struktūra pirmoreiz atklāja DNS saistības molekulāros pamatus lielai SSB saimei, kas sastopama gan fāgos, gan eikariotu vīrusos. Enc34 ORF39 proteīna gadījumā, pamatojoties uz genoma kontekstu, tika izvirzīta hipotēze, ka tas ir jauna veida fāgu endolizīns – peptidoglikānu noārdošs enzīms, kas no iekšpuses sagrauj baktērijas šūnas sienīgu. Tā hipotēzētā enzīmātiskā aktivitāte tika apstiprināta *in vitro* testos, un proteīna enzīmātiski aktīvā domēna trīsdimensionālā struktūra atklāja izteikti diverģentu lizocīma folda variāciju. Struktūras analīze parādīja, ka Enc34 ORF39 proteīns ir tikai attāli radniecīgs vienam citam fāga endolizīnam un diviem baktēriju autolizīniem, kuriem, neraugoties uz būtiskām atšķirībām, ir kopīgs konservatīvs trīs aminoskābju motīvs substrāta šķelšanai. Šī darba rezultātā nezināmas funkcijas domēna (*domain of unknown function*, DUF) 2815 un fāgu proteīnu saimes PHA02564 anotācijas beigs pastāvēt sekvenču datos kā proteomu tumšā matērija, jo esam tos raksturojuši kā attiecīgi SSB un endolizīnus.

SATURS

SAĪSINĀJUMI	5
IEVADS	6
Promocijas darba mērķi.....	6
Darba uzdevumi.....	7
Tēzes.....	7
REZULTĀTU APROBĀCIJA	8
Zinātniskās publikācijas.....	8
Konferences un semināri.....	8
Stipendijas un apbalvojumi.....	10
1. LITERATŪRAS APSKATS	11
1.1. Fāgu bioloģija.....	11
1.2. <i>Hafnia</i> vīruss Enc34.....	13
2. MATERIĀLI UN METODES	16
2.1. Sekvences analīze.....	16
2.2. Plazmīdu veidošana.....	16
2.3. Proteīnu producēšana.....	17
2.4. Proteīnu attīrīšana.....	17
2.5. Kristalizēšana un datu iegūšana.....	18
2.6. Struktūras noteikšana.....	18
2.7. Struktūras analīze.....	19
2.8. Elektroforētiskās kustības nobīdes tests.....	19
2.9. Optiskā blīvuma samazināšanās tests.....	20
3. REZULTĀTI UN DISKUSIJA	21
3.1. Vienpavediena DNS saistošā proteīna struktūras raksturošana: bakteriofāga Enc34 ORF6 proteīna gadījuma analīze.....	21
3.2. <i>Enterobacter</i> fāga Enc34 vienpavediena DNS saistošā proteīna DNS atpazīšanas strukturālie pamati.....	23
3.3. Lizocīma folda daudzveidība: neparasta endolizīna, ko kodē fāgs Enc34, katalītiskā domēna struktūra.....	25
3.4. Enc34 ORF3, ORF4 un ORF5 proteīnu AlphaFold2 struktūras un turpmākā darba perspektīvas.....	28
SECINĀJUMI	31
PATEICĪBAS	32

SAĪSINĀJUMI

Å	angstrēms
AF2	AlphaFold2
ĀM	ārējā membrāna
BMC	Latvijas Biomedicīnas pētījumu un studiju centrs
CDC	ASV Slimību kontroles un profilakses centrs
CRISPR	grupēti, regulāri izkliedēti īsi palindromiskie atkātojumi
Dali	attālumu matricas salīdzināšana
DSB	divpavedienu pārrāvums
dpDNS	divpavedienu dezoksiribonukleīnskābe
DTT	ditiotreitols
DUF	nezināmas funkcijas domēns
EAD	enzimātiski aktīvais domēns
ECDC	Eiropas slimību profilakses un kontroles centrs
GH	glikozīdu hidrolāze
GlcNAc	N-acetilglikozamīns
IPTG	izopropil-β-D-1-tiogalaktopiranozīds
kbp	kilobāzu pāri
MurNAc	N-acetilmuramīnskābe
NCBI	Nacionālais biotehnoloģiju informācijas centrs
OB	oligonukleotīdu/oligosaharīdu saistīšana
OD600	optiskais blīvums pie 600 nanometriem
ORF	atvērtais lasīšanas rāmis
PBS	fosfātu buferšķīdums
PCR	polimerāzes ķēdes reakcija
PDB	Proteīnu datu banka
PHA	fāgu proteīns
pLDDT	prognozēto lokālo attālumu starpības tests
SeMet	selenometionīns
SSB	vienpavediena dezoksiribonukleīnskābes saistošais proteīns
TAE	Tris-acetāts-etilēndiamīntetraetiķskābe
TEV	tabakas kodinājuma vīruss
TN	Tris-nātrija hlorīds
vpDNS	vienpavediena dezoksiribonukleīnskābe
vpRNS	vienpavediena ribonukleīnskābe
WHO	Pasaules Veselības organizācija

IEVADS

Bakteriofāgi ir iecienītas modeļsistēmas dabaszinātnēs, un aizvien notiekošie pētījumi ar paradigmiskajiem kolifāgiem λ , T4, M13 un dažiem citiem ir devuši daudzus fundamentālus atklājumus bioloģijā un neskaitāmus rīkus zinātnei un rūpniecībai. Genomiskā un metagenomiskā sekvenēšana ir atklājusi, ka visdažādākie fāgi ir sastopami it visur dabā, kur tie pastāv kā nozīmīgs ekoloģisko un evolūcijas procesu virzītājspēks, tomēr daudzi fāgu bioloģijas aspekti, domājams, joprojām nav atklāti. Antimikrobiālās rezistences globālā izplatība ir vēl vairāk veicinājuši interesi par fāgiem kā līdzekļiem cīņai pret bakteriālām infekcijām un kontamināciju. Sagaidāms, ka daudzi inovatīvi antibakteriālie līdzekļi tiks radīti, iedvesmojoties no fāgu kodētajām šūnu līzes mašīnērijām, kas ir izrādījušās daudzveidīgākas, nekā iepriekš tika uzskatīts, un joprojām ir nepietiekami izpētītas.

Pieaugošajā bakteriofāgu sekvenču krājumā lielākajai daļai prognozēto proteīnu aizvien trūkst uzticamu funkcionālo anotāciju, jo to līdzība ar eksperimentāli raksturotiem gēnu produktiem ir neliela vai nenosakāma, un tādējādi tie veido, t.s., biosfēras tumšo matēriju. Tomēr daži no šiem noslēpumainajiem proteīniem ir ievērojami saglabājušies evolūcijas gaitā kopā ar citu vīrusu, prokariotu vai pat eikariotu homoloģiem, un tāpēc tie ir lieliski mērķi strukturāliem un funkcionāliem pētījumiem, kas var palīdzēt izziņāt veselas proteīnu saimes ar acīmredzami būtiskām bioloģiskām funkcijām. Mūsu modeļfāgs Enc34, kas tika izolēts BMC, tā genoma moduļos DNS replikācijai un saimniekšūnas līzei kodē vairākus šādus proteīnus, kas pēc to identifikācijas ir radījuši interesi tālākai eksperimentālai izpētei.

Promocijas darba mērķi

Mana promocijas darba mērķi bija iegūt jaunas zināšanas par *Hafnia* vīrusa Enc34 proteoma līdz šim neraksturoto proteīnu struktūrām un funkcijām, izziņāt nepietiekami izpētītas proteīnu saimes un sniegt pienesumu bioloģisko sekvenču funkciju anotēšanā.

Darba uzdevumi

Mērķu sasniegšanai tika izvirzīti sekojoši darba uzdevumi:

- Identificēt konservatīvus neraksturotus mērķproteīnus Enc34 proteomā, izmantojot *in silico* sekvenču analīzes rīkus;
- Izstrādāt producēšanas, attīrīšanas un kristalizēšanas shēmas izvēlēto proteīnu natīvajiem un selenometionīna aizvietotajiem variantiem;
- *De novo* noteikt un analizēt izvēlēto proteīnu trīsdimensionālās struktūras un izvirzīt hipotēzes par to funkcijām;
- Izpētīt mērķproteīnu hipotētizētās funkcijas, veicot papildus strukturālos pētījumus un *in vitro* testus.

Tēzes

Šī darba galvenās tēzes bija:

1. Bakteriofāga Enc34 ORF6 proteīns ir vīrusa DNS replikācijas mašīnērijas diverģenta sastāvdaļa.
2. Bakteriofāga Enc34 ORF39 kodē endolizīnu ar netipisku arhitektūru un izteikti diverģentu enzimatiski aktīvo domēnu.
3. Mērķtiecīgi strukturālie pētījumi, kopā ar *in vitro* testiem, var atklāt vīrusu “tumšās matērijas” proteīnu funkcijas un īpašības.
4. Konservatīvu neizpētītu proteīnu izzināšana var radīt priekšstatus par lielākām proteīnu saimēm un uzlabot radniecīgu sekvenču funkciju anotāciju.

REZULTĀTU APROBĀCIJA

Zinātniskās publikācijas

- I. **Černooka, E.**, Rūmnieks, J., & Kazāks, A. (2021). Structural characterization of a single-stranded DNA-binding protein: a case study of the ORF6 protein from bacteriophage Enc34. *Methods in molecular biology* (Clifton, N.J.), 2281, 343–373. https://doi.org/10.1007/978-1-0716-1290-3_23
- II. **Černooka, E.**, Rūmnieks, J., Tārs, K., & Kazāks, A. (2017). Structural basis for DNA recognition of a single-stranded DNA-binding protein from *Enterobacter* phage Enc34. *Scientific reports*, 7(1), 15529. <https://doi.org/10.1038/s41598-017-15774-y>
- III. **Černooka, E.**, Rūmnieks, J., Zrelavs, N., Tārs, K., & Kazāks, A. (2022). Diversity of the lysozyme fold: structure of the catalytic domain from an unusual endolysin encoded by phage Enc34. *Scientific reports*, 12(1), 5005. <https://doi.org/10.1038/s41598-022-08765-1>

Konferences un semināri

1. **Černooka, E.**, Rūmnieks, J., Zrelavs, N., Tārs, K., Kazāks, A. Structural analysis of a divergent endolysin from *Hafnia* bacteriophage Enc34. Instruct Biennial Structural Biology Conference. Stenda referāts. 2022, Utrehta, Nīderlande
2. **Černooka, E.**, Rūmnieks, J., Zrelavs, N., Kazāks, A. Proteīnu struktūru un funkciju pētījumi baktēriju vīrusu modeļsistēmā. Latvijas Universitātes 78. Starptautiskā zinātniskā konference, Latvijas Universitātes fonda mecenātu atbalstīto doktorantūras studiju programmas stipendiātu zinātniskā sesija. Mutiska prezentācija. 2020, Rīga, Latvija

3. Zrelovs, N., **Černooka, E.**, Dišlers, A., Kazāks, A. Genomic and proteomic studies of novel bacteriophages from Latvian Biomedical Research and Study Centre's phage collection. 9th International Conference 'Bioresources and Viruses'. Mutiska prezentācija. 2019, Kijiva, Ukraina
4. **Černooka, E.**, Rūmnieks, J., Zrelovs, N., Tārs, K., Kazāks, A. Structural studies of conserved uncharacterized proteins in the genomes of novel DNA phages. FEBS3+ Conference. Mutiska prezentācija. 2019, Rīga, Latvija
5. **Černooka, E.**, Rūmnieks, J., Zrelovs, N., Tārs, K., Kazāks, A. Jaunatklātu DNS fāgu raksturošana un to proteīnu struktūras un funkciju pētījumi. Latvijas Universitātes 77. Starptautiskā zinātniskā konference, Latvijas Universitātes fonda mecenātu atbalstīto doktorantūras studiju programmas stipendiātu zinātniskā sesija. Mutiska prezentācija. 2019, Rīga, Latvija
6. **Černooka, E.**, Kazāks, A. DNS bakteriofāgu nestrukturālo proteīnu pētījumi. Instruct-ERIC seminārs. Mutiska prezentācija. 2019, Rīga, Latvija
7. **Černooka, E.**, Kazāks, A. Jaunatklātu DNS bakteriofāgu genomu raksturošana un to kodēto proteīnu strukturāla un funkcionāla izpēte. Latvijas Universitātes 76. Starptautiskā zinātniskā konference, Molekulārās bioloģijas sekcijas sēde. Mutiska prezentācija. 2018, Rīga, Latvija
8. **Černooka, E.**, Tārs, K., Kazāks, A. *Enterobacter* fāga Enc34 hipotētiska replikācijas proteīna strukturāla un funkcionāla izpēte. Latvijas Universitātes 75. Starptautiskā zinātniskā konference, Molekulārās bioloģijas sekcijas sēde. Mutiska prezentācija. 2017, Rīga, Latvija

Stipendijas un apbalvojumi

- Zrelavs, N., Černooka, E., Rūmnieks, J., Kazāks, A., Dišlers, A. Latvijas Zinātņu akadēmijas atzinības raksts par “Latvijai jauna DNS bakteriofāgu izolēšanas un padziļinātas izpētes virziena attīstību”, 2021
- Černooka E. SIA “Mikrotīkls” stipendija doktorantūras studentiem eksakto un medicīnas zinātņu jomā, 2017-2020
- Černooka E. Mundheim ģimenes Izcilības stipendija, 2014-2015

1. LITERATŪRAS APSKATS

No pierādīšanas, ka DNS ir iedzimtības molekula, līdz pat grupētu, regulāri izkļiedētu īsu palindromisko atkārtojumu (CRISPR) atklāšanai – pētījumi ar bakteriofāgiem vienmēr ir bijuši molekulārās bioloģijas priekšplānā, sniedzot svarīgas atziņas un radot instrumentus, ko šobrīd plaši izmanto zinātnē un rūpniecībā. Daudzi no ietekmīgākajiem atklājumiem un pielietojumiem ir radušies, pateicoties zinātkārei par pašiem fāgiem, kā arī dažādu bioloģisko procesu izpausmēm bakteriofāgu modeļsisistēmās (Salmond, Fineran, 2015). Tomēr šie sasniegumi, iespējams, ir tikai “aisberga redzamā daļa”, jo izpētīta ir pavisam niecīga daļa no visiem dabā sastopamajiem fāgiem un tikai neliela vīrusu genofonda daļa ir labi definēta un izziņāta (Aevarsson et al., 2021). Kā dabiski baktēriju antagonisti fāgi var kalpot arī kā biokontroles aģenti, un to lītiskos mehānismus var izmantot, lai radītu jaunus farmaceitiskos līdzekļus baktēriju infekciju ārstēšanai. Šis pētniecības virziens nesēn īpaši aktualizējies, ņemot vērā pieaugošo antimikrobiālās rezistences izplatību bakteriālo patogēnu vidū un ievērojamu sabiedrības veselības aģentūru paziņojumus par nenovēršamo “post-antibiotiku laikmetu”, kā arī aicinājumus izstrādāt jaunus antibakteriālus līdzekļus (WHO, 2015; CDC, 2019; ECDC, WHO, 2022). Neapšaubāmi, vēl daudzi atklājumi un inovācijas tiks gūti gan fundamentālos, gan lietišķos fāgu pētījumos, kas, kā aizvien, turpinās sniegt ieguldījumu zinātnē, tehnoloģijās un medicīnā.

1.1. Fāgu bioloģija

Bakteriofāgi inficē visur esošās baktērijas, tāpēc tie ir vieni no mazākajām un vienkāršākajām, tomēr dabā visplašāk sastopamajām radībām, kam ir sava ģenētika, evolucionārā vēsture un zināma pašregulācijas spēja (Yolles, Frieden, 2022). Fāgi ir atklāti visās prokariotu dzīvotnēs – gan ūdenī, gan uz sauszemes, gan asociācijā ar augiem un dzīvniekiem, un to skaits pārsniedz visu biosfēras organismu kopskaitu (Cobián Güemes et al., 2016). Fāgi kā visplašāk pārstāvētās bioloģiskās būtnes uz Zemes spēlē nozīmīgas lomas bioģeoķīmiskajā aprītē (Suttle, 2007), baktēriju ekosistēmu veidošanā (Chevallereau et al., 2021) un mikrobu evolūcijas virzīšanā (Hampton et al., 2020).

Fāgi atšķiras ne tikvien pēc virionu struktūras, kas var būt ar vai bez astes, pavedienveida, vai iekļauta apvalkā, bet arī pēc nukleīnskābes tipa, kas veido vīrusa genomu, jo pašlaik ir zināmi gan vienpavediena (vp), gan divpavedienu (dp) DNS un RNS fāgi (Ofir, Sorek, 2018). Līdz ar sekvenēšanas tehnoloģiju attīstību genomika kļuva par aizvien lielāku un nozīmīgāku fāgu izpētes daļu (Brüssow, Hendrix, 2002), un tā joprojām ir jaunu fāgu atklāšanas un raksturošanas stūrakmens. Genomikas pētījumiem paplašinoties, tika atklāts, ka dažādu fāgu genoma izmēri un sekvenču ievērojami atšķiras un ka fāgu genomiem ir mozaīkveida struktūra, proti, tie sastāv no gēnu blokiem jeb moduļiem, kas aizgūti no citiem fāgiem, pateicoties plašai ģenētiskai apmaiņai populācijas ietvaros (Hatfull, 2008). Šobrīd GenBank ģenētisko sekvenču datubāzē ir vairāk nekā 16 000 dažādu fāgu sekvenču (organizējot pēc Cook et al., 2021, metodikas), šī kolekcija turpina divkāršoties katru divu līdz trīs gadu laikposmā, un tas, domājams, drīzumā neapstāsies vai nepalēnināsies. Tomēr tikai neliela daļa baktēriju un ar tām saistīto fāgu ir viegli kultivējami laboratorijas apstākļos, tāpēc vīrusu metagenomikas attīstība un nesenis uzplaukums ir ļāvis vēl precīzāk novērtēt fāgu izplatību dažādos biotopos (Paez-Espino et al., 2016; Gregory et al., 2019), un tas ir radījis vēl pilnīgāku priekšstatu par ģenētiskās daudzveidības plašumu vīrusu pasaulē (Dion et al., 2020). Vēl joprojām liela daļa fāgu genomikas un metagenomikas datu veido t.s. biosfēras tumšo matēriju, kas uzrāda zemu vai neuzrāda nekādu līdzību ar šobrīd zināmajām sekvencēm un tādējādi ietver veselus genomu moduļus un mīklainus atsevišķus gēnus, kas vēl gaida savu kārtu zinātniskajā izpētē (Hatfull, 2015; Santiago-Rodriguez, Hollister, 2022).

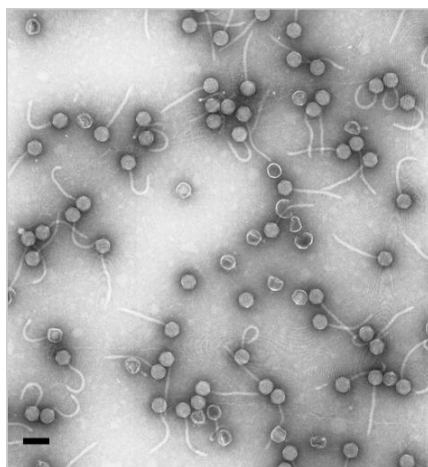
Fāgi ar astēm un DNS genomu aptuveni 25-100 kilobāzu pāru (kbp) intervālā ir visizplatītākie (Zrelovs et al., 2020a), un parasti tie kodē desmitiem proteīnu izteikti modulārā izkārtojumā, visticamāk pateicoties selektīvam spiedienam uz efektīva kompakta genoma saglabāšanu (Berg, Roux, 2021). Funkcionālajos moduļos vienkopus atrodas proteīni, kas nodrošina lizogēniju, DNS replikāciju, genoma pakošanu, fāga galvas un astes morfoģenēzi un saimniekšūnas līzi (Brüssow, Hendrix, 2002), kas lielākos genomos bieži vien satur iestarpinātus hiperplastiskus reģionus, kuri kodē jaunus proteīnus ar potenciāli adaptīvām funkcijām (Comeau et al., 2008).

1.2. *Hafnia* vīruss Enc34

Latvijas Biomedicīnas pētījumu un studiju centrā (BMC) jau izsenis tiek veikti fāgu pētījumi, kas vēsturiski un joprojām ir vērsti uz vpRNS bakteriofāgiem un to strukturālajiem proteīniem (Rūmnieks, Tārs, 2011; Rūmnieks, Tārs, 2014; Rūmnieks, Tārs, 2017; Rūmnieks et al., 2020). Meklējot jaunus RNS fāgus izpētei no dažādiem vides avotiem, mums faktiski sāka veidoties kolekcija ar “astainajiem” DNS vīrusiem, kas jau tad iekļāva vairāk nekā 40 izolātu. Kādu laiku šie fāgi netika tālāk pētīti, līdz DNS sekvenēšana kļuva mums daudz pieejamāka. Tas rosināja interesi par mūsu atklāto vīrusu raksturošanu, un galu galā tika aizsākts pavisam jauns DNS fāgu izpētes virziens BMC un Latvijā kopumā. Šobrīd BMC fāgu kolekcija ir tikusi papildināta ar vēl vairākiem desmitiem izolātu, kas iegūti no kukaiņu, notekūdeņu un pat Antarktīkas augsnes paraugiem u.c. avotiem. Kopumā GenBank datubāzē līdz šim ir deponētas 18 fāgu pilna genoma sekvences, un mēs esam atklājuši pirmos izolētos vīrusus baktērijām *Virgibacillus halotolerans* (Zrelavs et al., 2020b) un *Erwinia persicina* (Zrelavs et al., 2020c), diverģentu *Lactococcus* bakteriofāgu (Zrelavs et al., 2021), *Morganella* fāgu ar atšķirīgu genoma pakošanas veidu (Zrelavs et al., 2022a) un *Psychrobacillus* vīrusus, kas mitinās Antarktīkas augsnē (Zrelavs et al., 2022b); saprotams, ka katram no šiem izolātiem piemīt kas unikāls un interesants.

Tomēr pats pirmais pilnībā sekvenētais izolāts no kolekcijas bija bakteriofāgs Enc34 (Kazāks et al., 2012), par kuru tolaik tika uzskatīts, ka tas inficē *Enterobacter cancerogenus*, pamatojoties uz bioķīmiskajiem testiem, bet jaunākas molekulārās analīzes parādīja, ka tā saimniekbaktērija ir neparasts *Hafnia alvei* celms (skat. publikāciju III). *H. alvei* ir Gram-negatīva *Enterobacterales* kārtas *Hafniaceae* dzimtas baktērija (Adeolu et al., 2016), un, lai gan parasti to uzskata par dzīvnieku mikrobioma komensālu sastāvdaļu (Song et al., 2017), šī suga ir identificēta arī kā oportūnistisku infekciju izraisītāja cilvēkiem (Ionescu et al., 2022) un kā bišu, zivju un citu dzīvnieku patogēns (Padilla et al., 2014). *Hafnia* vīrusa Enc34 virionam ir galvas-astes struktūra ar garu nekotraktīlu asti (1. attēls, pa kreisi), un vīrusa genoms ir lineāra dpDNS molekula, kuras garums ir aptuveni 60 kbp ar 5' pārkarēm, ko sauc par kohezīvajiem galiem jeb cos un kuras norāda uz 5' cos genoma pakošanas

stratēģiju (Casjens, Gilcrease, 2009). Fāga oficiālais nosaukums ir *Enchivirus Enc34*, un pašlaik tas ir vienīgais vīruss savā ģintī, kura pieder *Casjensviridae* dzimtai (1. attēls, pa labi). Šajā dzimtā vēl ietilpst liela *Salmonella* vīrusu grupa, kas ir apvienoti *Chivirus* ģintī, kā arī daudzas mazas ģintis, katra ar vienu vai dažiem izolētiem fāgiem. Tie savukārt inficē virkni dažādu Gram-negatīvu baktēriju, tostarp *Burkholderia* (Lynch et al., 2012), *Providencia* (Onmus-Leone et al., 2013), *Xylella* (Ahern et al., 2014), *Proteus* (Nguyen et al., 2019), *Pectobacterium* (Zaczek-Moczydłowska et al., 2020) un citas, bet, neraugoties uz to, šiem vīrusiem ir līdzīgi genoma izmēri un sintētisks gēnu izkārtojums.



Vīrusa klasifikācija

Impērija:	<i>Duplodnaviria</i>
Valsts:	<i>Heunggongvirae</i>
Tips:	<i>Uroviricota</i>
Klase:	<i>Caudoviricetes</i>
Dzimta:	<i>Casjensviridae</i>
Ģints:	<i>Enchivirus</i>
Suga:	<i>Enchivirus Enc34</i>

Sinonīmi

Hafnia fāgs Enc34, agrāk saukts
Enterobacter fāgs Enc34

1. attēls. *Hafnia* vīruss Enc34. Virionu transmisijas elektronmikroskopija (pa kreisi) un fāga taksonomija un nomenklatūra (pa labi). Marķieris, 100 nm (nepublicēti dati).

Enc34 genoms satur 80 atvērto lasīšanas rāmjus (ORF), kas ir sakārtoti funkcionālos moduļos un atbild par attiecīgi vīrusa DNS replikāciju, genoma pakošanu, viriona morfoģenēzi un saimniekšūnas līzi, tomēr 45 ORF neuzrādīja nekādu līdzību ar zināmiem gēnu produktiem, tāpēc tie tika anotēti kā hipotētiski (Kazāks et al., 2012) (GenBank ID: JQ340774.2). Šajā darbā es vērsu uzmanību uz neraksturotajiem proteīniem fāga Enc34 DNS replikācijas modulī, kura vidū acīmredzami atradās intriģējošs gēnu bloks, kas secīgi kodēja četrus hipotētiskos proteīnus (ORF3 līdz 6), kā arī uz šūnu līzes moduli, kurā šķietami atradās jauna tipa baktēriju šūnu sienīņu noārdošs enzīms jeb endolizīns (ORF39). Šim proteīnam tika prognozēta neparasta arhitektūra

ar C-terminālu transmembrānas domēnu un diverģents enzimatiski aktīvais domēns. Izvēlētie proteīni ir evolucionāri konservatīvi, tomēr neviens no identificējamajiem sekvenču homoloģiem tajā laikā nebija eksperimentāli izpētīts. Šāda veida mērķproteīnu pētījumi tāpat ne tikai atklātu būtiskus paša proteīna struktūras un funkcijas aspektus, bet arī raksturotu veselas proteīnu saimes, izceļot tās no “vīrusu tumšās matērijas” (Galperin, Koonin, 2004). Mans promocijas darbs detalizēti skaidro un ilustrē *de novo* struktūras noteikšanu un funkcijas pārbaudi ORF6 hipotētiskajam DNS replikācijas proteīnam un ORF39 hipotētiskajam endolizīnam. Nobeigumā tiek prezentētas arī *in silico* atrisinātās ORF3, ORF4 un ORF5 proteīnu struktūras un tiek īsi izklāstītas turpmākā darba perspektīvas.

2. MATERIĀLI UN METODEDES

Lai pētītu mūsu modeļfāga Enc34 proteīnu struktūras un funkcijas, es izmantoju dažādu molekulārās bioloģijas, mikrobioloģijas un bioinformātikas metožu klāstu. Šajā nodaļā es sniedzu kopsavilkumu par galvenajām metodēm un procedūrām, kas tika izmantotas darbā ar Enc34 ORF6 un ORF39 proteīniem. Sīkāka informācija ir atrodamā attiecīgajos manuskriptos, jo īpaši publikācijā I, kas pati par sevi ir visaptverošs struktūrbioloģijas pētījuma ceļvedis.

2.1. Sekvences analīze

Enc34 proteīnu sekvences tika pārbaudītas ar BLASTP (Boratyn et al., 2013), izmantojot Nacionālā biotehnoloģiju informācijas centra (NCBI) ne-redundanto proteīnu sekvenču datubāzi, kā arī HHpred (Gabler et al., 2020) un PDB_mmCIF70 datubāzi. Transmembrānu spirāles tika identificētas ar TMHMM (Krogh et al., 2001) un DeepTMHMM (Hallgren et al., 2022), bet signālpeptīdi tika prognozēti ar SignalP (Teufel et al., 2022). Vairāku sekvenču salīdzinājumi turpmākai homoloģijas un filoloģētiskajai analīzei tika ģenerēti, izmantojot Clustal Omega (Sievers, Higgins, 2021).

2.2. Plazmīdu veidošana

Mērķproteīnu kodējošās sekvences tika pavairotas ar PCR no Enc34 genomiskās DNS (GenBank ID: JQ340774). ORF6 tika amplificēts, izmantojot oligonukleotīdus, kas saturēja tabakas kodinājuma vīrusa (TEV) proteāzes šķelšanas saita sekvenci un restrikcijas saitus klonēšanai un klonu atlasei. Amplificētā DNS tika klonēta pETDuet-1 vektorā (Novagen) aiz N-gala heksahistidīnu iezīmes (His-Tag). Iegūtais vektors, kas tātad saturēja afinitātes iezīmi, tam sekojošu TEV saitu, kā arī strupinātā gala restrikcijas saitu inserta izgriešanai, tika izmantots, lai veidotu visas pārējās plazmīdas, ieskaitot ORF39 kodējošo konstruktus. Lai sekmētu strupināto galu klonēšanu, pirms PCR tiešie oligonukleotīdi tika fosforilēti ar T4 polinukleotīdu kināzi (Thermo Scientific). Plazmīdas C-termināli saīsinātajiem proteīna variantiem tika izveidotas, izmantojot atbilstošus apgrieztos oligonukleotīdus un to pašu pETDuet-1 atvasināto vektora bāzi.

2.3. Proteīnu producēšana

Natīvie proteīni un to saīsinātie varianti tika producēti *E. coli* BL21(DE3) šūnās, izmantojot optimizētu protokolu, ko mēs izstrādājām, lai palielinātu mērķproteīnu šķīdību un iznākumu. Īsumā, ar atbilstošu plazmīdu transformētas šūnas tika audzētas 2xTY barotnē ar 50 µg/mL ampicilīnu 25 °C temperatūrā līdz OD600 0,6-0,8. Pēc tam kultivēšanas temperatūra tika samazināta līdz 22 °C un mērķa proteīna producēšana tika inducēta ar 0,01 mM izopropil-β-D-1-tiogalaktopiranozīda (IPTG) šķīdumu, turpinot kultivēšanu pa nakti. Ar selenometionīnu (SeMet) aizvietotie proteīni savukārt tika producēti *E. coli* B834, izmantojot reaģentu komplektu (Molecular Dimensions) un nedaudz modificētu protokolu, kas mūsu pētījumos uzlaboja mērķa proteīnu šķīdību. Vispirms šūnas tika audzētas 2xTY barotnē ar 50 µg/mL ampicilīnu 25 °C temperatūrā līdz OD600 0,8-1,0, pēc tam tās tika pārnestas uz speciālu komerciālu barotni ar barības vielām un inkubētas 25 °C temperatūrā 2 h. Tad tika pievienots SeMet šķīdums un produkcija tika inducēta ar 0,1-1,0 mM IPTG 25 °C temperatūrā, turpinot inkubāciju pa nakti. Šūnu biomasa tika ievākta ar centrifugēšanu un uzglabāta -20 °C temperatūrā līdz izmantošanai.

2.4. Proteīnu attīrīšana

Lai uzsāktu mērķproteīnu izolēšanu, šūnu masa tika atkausēta, resuspendēta TN buferšķīdumā (20 mM Tris-HCl (pH 8,0), 300 mM NaCl) un lizēta ar sonikāciju. Lizāta šķīstošā frakcija tika iegūta centrifugējot, un tālāk to izlaida cauri HisTrap afinitātes hromatogrāfijas kolonnai (GE Healthcare). Kolonna tika skalota ar TN buferšķīdumu, kas saturēja 20 mM imidazolu, un saistītais proteīns tika eluēts TN buferšķīdumā ar pievienotu 300 mM imidazolu. His-Tag iezīme tika atdalīta ar TEV proteāzi, kam sekoja vēl viens afinitātes hromatogrāfijas solis, lai izolētu iezīmi nesaturošo proteīnu. Lai iegūtu kristalizācijai pietiekami tīru preparātu, ORF6 proteīns tika papildus attīrīts ar augstas izšķirtspējas jonu apmaiņas hromatogrāfiju, izmantojot MonoQ kolonnu (GE Healthcare). SeMet proteīni tika izolēti un attīrīti pēc tāda paša protokola kā attiecīgie natīvie proteīni, izņemot to, ka visi buferšķīdumi tika papildināti ar ditiotreitolu (DTT), lai samazinātu proteīnu agregāciju.

2.5. Kristalizēšana un datu iegūšana

Attīrīto proteīnu preparātiem buferšķīdumi tika nomainīti uz 20 mM Tris-HCl (pH 8,0), 100 mM NaCl (ar 1 mM DTT SeMet proteīniem), un proteīni tika koncentrēti līdz ~10 mg/mL, izmantojot atbilstošas molekulmasas filtrus (Millipore). Pēc tam istabas temperatūrā tika veikti “sēdošā piliena” tvaika difūzijas kristalizēšanas eksperimenti, izmantojot komerciāli pieejamus kristalizācijas šķīdumu komplektus, piemēram, JCSG-plus, Structure screen 1 + 2 un PACT premier (Molecular Dimensions), un pipetēšanas robotu (Tecan). Kristālu audzēšanas apstākļi tika replicēti un manuāli optimizēti, lai iegūtu vairāk kristālu, kurus pēc tam pārnesa uz krioprotektanta šķīdumu, sasaldēja šķidrā slāpekļī un tur uzglabāja līdz rentgenstaru datu ievākšanai. Lai iegūtu proteīna-liganda kompleksus, es galvenokārt izmantoju ko-kristalizāciju, mērķa proteīnu sajaucot ar ligandu dažādās molārajās attiecībās tieši pirms skrīninga vai turpmākajiem kristalizācijas eksperimentiem. Enc34 endolizīna gadījumā tika izmēģināta arī proteīna kristālu mērcēšana ligandu šķīdumos, tomēr ne šī, ne arī ko-kristalizācijas metode iegūtajās struktūrās neatklāja saistītu ligandu. Difrakcijas dati no visiem kristālu paraugiem tika iegūti vairākos komandējumos MAX-lab un MAX IV (Lundas Universitātē) un BESSY II (Helmholca centrā, Berlīnē).

2.6. Struktūras noteikšana

Lai noteiktu ORF6 proteīna struktūru, SeMet kristāla difrakcijas attēli tika apstrādāti ar MOSFLM (Battye et al., 2011) un mērogoti, izmantojot SCALA (Evans, 2006), no CCP4 programmatūras (Potterton et al., 2018). Fāzēšana tika veikta, izmantojot SHELX C/D/E programmas (Usón, Sheldrick, 2018), kam sekoja blīvuma modificēšana ar DM (Cowtan, 2010). Sākotnējais modelis tika ģenerēts ar BUCCANEER (Cowtan, 2006), un pēc tam tas tika manuāli pārbūvēts COOT programmā (Casañal et al., 2020) un uzlabots ar REFMAC5 (Kovalevskiy et al., 2018). C-termināli saīsinātā ORF6 proteīna struktūra tika atrisināta ar molekulāro aizvietošanu, izmantojot MOLREP (Vagin, Teplyakov, 2010), un proteīna-DNS kompleksa struktūra vēlāk tika izrēķināta, izmantojot PHASER (McCoy et al., 2007). Savukārt ORF39 proteīna difrakcijas attēli tika apstrādāti ar XDS (Kabsch, 2010a), un pēc tam vairākas SeMet datu kopas tika mērogotas kopā, izmantojot

XSCALE (Kabsch, 2010b) caur XDSAPP (Sparta et al., 2016). Struktūra tika atrisināta ar AutoSol Wizard (Terwilliger et al., 2009) no Phenix programmatūras (Liebschner et al., 2019), un modelis tika izmantots, lai fāzētu nedaudz augstākas izšķirtspējas natīvo datu kopu. Vēlāki ko-kristalizācijas un kristālu mērcēšanas eksperimenti rezultējās ar vēl daudzu citu augstas izšķirtspējas struktūru noteikšanu, izmantojot PHASER, un no tām vispilnīgākais modelis tika tālāk uzlabots ar COOT un REFMAC5 un izmantots publicēšanai. Visas deponētās ORF6 un ORF39 struktūras ir validētas ar COOT un MolProbity (Williams et al., 2018).

2.7. Struktūras analīze

Modeļu pārbaude un detalizēta izpēte tika veikta COOT un PyMOL programmās (Schrödinger, DeLano, 2020). ORF6 un ORF39 strukturālie homologi tika identificēti, izmantojot attālumu matricas salīdzināšanas (Dali) serveri (Holm, 2022). Starpmolekulārās mijiedarbības kristālos tika pārbaudītas ar PDBePISA (Krissinel, Henrick, 2007). ORF6 proteīnam elektrostatiskā virsma tika aprēķināta ar PyMOL papildus programmu APBS Tools (Jurrus et al., 2018), un “omit” elektronblīvuma karte tika izveidota ar REFMAC un FFT (Murshudov et al., 1999) no CCP4. ORF39 proteīna aminoskābju konservativitātes analīzei un kartēšanai tika izmantots ConSurf (Ashkenazy et al., 2016). Modeļu vizualizācija un attēlu sagatavošana tika veikta ar PyMOL.

2.8. Elektroforētiskās kustības nobīdes tests

ORF6 proteīna DNS saistīšanas īpašības tika pētītas ar elektroforēzes eksperimentiem agarozes gēlā. Gan pilna garuma, gan saīsinātais proteīna variants tika testēts ar vpDNS un dpDNS fragmentiem, kas tika iegūti no Enc34 genoma. DNS alikvotas tika sajauktas ar ORF6 proteīnu dažādās tā koncentrācijās (molāro attiecību diapazonā no 1:156 līdz 1:2500), un šie maisījumi 20 minūtes tika inkubēti istabas temperatūrā. Tālāk tos ienesa 1% natīvajā agarozes gēlā, kas saturēja etīdija bromīdu, Tris-acetāta-EDTA (TAE) buferšķīdumā, un tika veikta elektroforēze 20 min pie 10 V/cm, kuras beigās DNS tika vizualizēta ultravioletajā gaismā.

2.9. Optiskā blīvuma samazināšanās tests

ORF39 proteīna peptidoglikāna šķelšanas aktivitāte tika noteikta spektrofotometriski, testējot proteīnu pret baktēriju šūnu substrātu ar ārēji eksponētu peptidoglikāna slāni. Lai iegūtu šādu substrātu, Gram-negatīvas *Escherichia coli* W3100, *Pseudomonas aeruginosa* PAO1 un Enc34 jutīgas *Hafnia alvei* šūnas (visas no laboratorijas kolekcijas) tika apstrādātas ar hloroforma piesātinātu 50 mM Tris-HCl (pH 7,7), pēc tam tās skalojot un resuspendējot fosfātu buferšķīdumā (PBS) optiskajā blīvumā OD600 0,6-1,0. Reakciju maisījumi tika sagatavoti 96 bedrīšu mikroplātē, membrānas permeabilizētajām šūnām pievienojot dažādus ORF39 proteīna daudzumus, un tālāk to absorbcija pie 600 nm tika mērīta ar mikroplašu lasītāju (BioTek) 3 h ar 3 min starplaikiem. Pēc tam enzimatiskā aktivitāte tika aprēķināta, izmantojot ActivityCalculator izklājlapu no <https://www.biw.kuleuven.be/logt/ActivityCalculator.htm> (Briers et al., 2007). ORF39 proteīns tika testēts arī uz ārējo membrānu saturošām Gram-negatīvām šūnām, kā arī Gram-pozitīvām *Microbacterium paraoxydans* baktērijām (no laboratorijas kolekcijas), izmantojot tikai augstāko enzīma koncentrāciju. Visas reakcijas tika veiktas trijos atkārtojumos, kā negatīvo kontroli izmantojot PBS, bet kā pozitīvo kontroli – vistas olu baltuma lizocīmu (Biochemica).

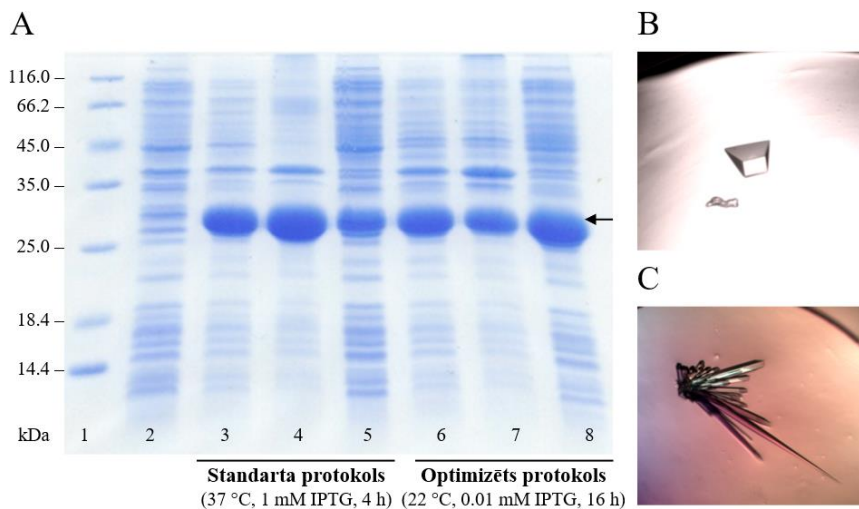
3. REZULTĀTI UN DISKUSIJA

Mans promocijas pētījums ir atainots trīs publicētos manuskriptos: (I) prezentē un raksturo metodes, kas tika izstrādātas Enc34 ORF6 proteīna izpētei un vēlāk tika piemērotas arī citiem mērķproteīniem no Enc34 proteoma, tostarp ORF39, (II) lielākoties analizē Enc34 ORF6 proteīna struktūrbioloģiju, un (III) atklāj jaunu lizocīma folda variāciju, kas pārstāvēta Enc34 ORF39 proteīnā. Šīs nodaļas struktūra ir veidota tā, lai iepazīstinātu ar galvenajiem minēto publikāciju rezultātiem (attiecīgi 3.1., 3.2. un 3.3. apakšnodaļā) un īsi apspriestu projekta nākotni (3.4. apakšnodaļā).

3.1. Vienpavediena DNS saistošā proteīna struktūras raksturošana: bakteriofāga Enc34 ORF6 proteīna gadījuma analīze

Lai labāk izprastu vienpavediena DNS saistošā proteīna (SSB) funkcijas molekulāros pamatus, ir jānosaka tā trīsdimensionālā struktūra, ja iespējams, kompleksā ar DNS. Šajā publikācijā ir aprakstīti visi soļi, kas tika veikti, lai strukturāli raksturotu bakteriofāga Enc34 SSB, ko kodē tā genoma ORF6. Mēs sniedzam sīkāku informāciju par galvenajām pētījumā izmantotajām procedūrām, sākot ar ekspresijas konstruktū izstrādi, proteīna producēšanu, attīrīšanu un kristalizēšanu, līdz pat trīsdimensionālās struktūras noteikšanai un analīzei. Papildus mēs dalāmies ar padomiem un ieteikumiem, kas ļāva veiksmīgi realizēt bakteriofāga Enc34 ORF6 SSB pētījumu un kuru viens no piemēriem ir zemāk ilustrētais rekombinantā proteīna producēšanas optimizēts protokols.

Producējot rekombinantus proteīnus *E. coli*, interesējošais proteīns gana bieži izrādās pilnībā vai daļēji nešķīstošs, kas tika novērots arī Enc34 ORF6 proteīna gadījumā. Tomēr, gadu gaitā producējot dažādus rekombinantus proteīnus, mūsu laboratorijā ir ticis izstrādāts alternatīvs protokols, kurā tiek izmantota zemāka kultivēšanas temperatūra (22 °C) un induktora koncentrācija (0,01 mM IPTG), un tas ir bijis ļoti efektīvs šķīstošu proteīnu ražošanai, vienlaikus saglabājot līdzīgu vai pat lielāku mērķproteīna iznākumu, nekā producējot standartapstākļos (2. attēls, A).

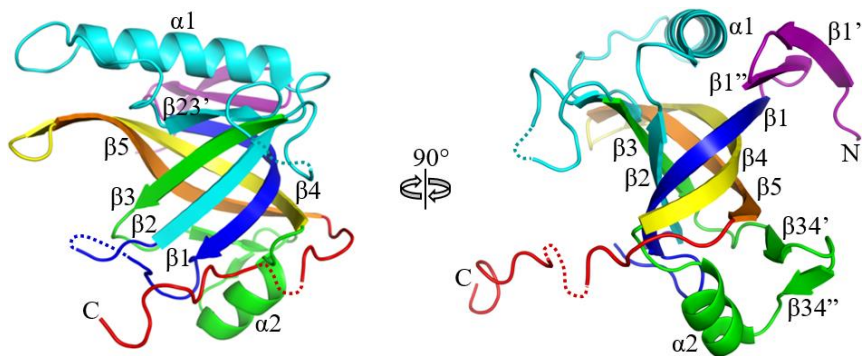


2. attēls. Enc34 ORF6 proteīna producēšana un kristalizēšana. SDS-PAGE analīze ORF6 proteīna produkcijai un šķīdībai, izmantojot standarta vai optimizētu producēšanas protokolu (A). 1 – molekulasmasas marķieris; 2 – neinducētu šūnu lizāts; 3 un 6 – šūnas proteīnu saturs ORF6 producēšanas beigās; 4 un 7 – nešķīstošo proteīnu frakcija; 5 un 8 – šķīstošo proteīnu frakcija. Bultiņa norāda uz ORF6 proteīna zonu. Pilna garuma ORF6 proteīna (B) un ORF6 Δ C-oligotimidīna kompleksa (C) kristāli.

Natīvais pilna garuma Enc34 ORF6 proteīns tika producēts saskaņā ar optimizēto protokolu, un tālāk tas tika attīrīts, izmantojot imobilizētu metāla jonu afinitātes hromatogrāfiju un anjonu apmaiņas hromatogrāfiju. Proteīns viegli kristalizējās vairākos apstākļos (2. attēls, B), tādēļ darbs tika turpināts ar pilna garuma ORF6 selenometionīna aizvietotā variantā (SeMet-ORF6) producēšanu, kas bija nepieciešams, lai iegūtu eksperimentālos fāžu datus struktūras noteikšanai. Pēc SeMet-ORF6 attīrīšanas un kristalizēšanas protokolu optimizēšanas iegūtie kristāli tika izmantoti rentgenstaru datu ievākšanai un struktūras atrisināšanai. Visbeidzot, tika izveidots C-termināli saīsināts ORF6 variants (ORF6 Δ C), pamatojoties uz to, ka proteīna nestrukturētais C-gals var gan kavēt kristalizāciju tā fleksibilitātes dēļ, gan konkurēt ar vienpavediena DNS par vienu un to pašu saistīšanās virsmu (Marintcheva et al., 2008). Visbeidzot, tika iegūta ORF6 Δ C struktūra kompleksā ar DNS, izmantojot rentgenstaru datus no proteīna-oligotimidīna ko-kristāla (2. attēls, C).

3.2. *Enterobacter* fāga Enc34 vienpavediena DNS saistošā proteīna DNS atpazīšanas strukturālie pamati

Enc34 ORF6 proteīna struktūra tika noteikta 1,50 Å izšķirtspējā, un, ja neskaita nestrukturēto C-galu, pārējam proteīnam ir kompakta, globulāra forma ar aptuveni 40 Å diametru un jauktu α/β arhitektūru, ko galvenokārt veido β -virtnes. Proteīnu var raksturot kā veidotu no ieliektas septiņu virkņu β -plāksnes vienā proteīna pusē, no kuras trīs centrālās virtnes kopā ar divām β -virknēm no pretējās puses veido nedaudz saplacinātu piecu pavedienu β -mucu proteīna kodolā, savukārt β -mucas abus galus aizklāj α -spirāles (3. attēls). Šī proteīna arhitektūra viennozīmīgi atgādina oligonukleotīdu/oligosaharīdu saistīšanas (OB) foldu, ko parasti definē kā piecu virkņu β -mucu ar galu nosedzošu α -spirāli (Theobald et al., 2003).



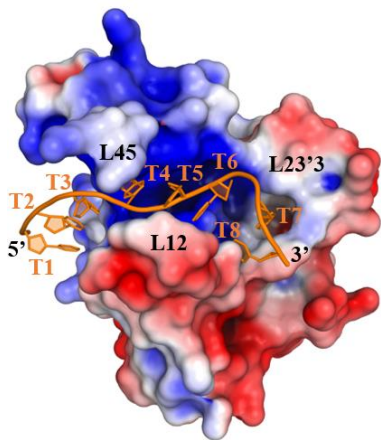
3. attēls. Fāga Enc34 ORF6 proteīna terciārā struktūra. Proteīna aminoskābju secība ir iekrāsota varavīksnes krāsās no violetas līdz sarkanai, attiecīgi no N līdz C-galam. Virtnes proteīna β -mucas kodolā ir numurētas saskaņā ar Murzin, 1993.

Lai arī sekvenču līmenī Enc34 ORF6 homologi ir tikai tuvradniecīgi fāgu proteīni no nezināmas funkcijas domēna (DUF) 2815 saimes, tā trīsdimensionālā struktūra ir līdzīga dažādiem OB folda proteīniem. Trīs struktūras, kas pašlaik atrodamas Proteīnu datu bankā (PDB), nepārprotami veido kopīgu strukturālās homoloģijas grupu ar Enc34 ORF6, un tajā ietilpst vēl divi līdz šim neraksturoti fāgu izcelsmes proteīni no *Bacillus cereus* (PDB ID: 4JG2) un *Enterococcus faecalis* (PDB ID: 4KLL), kā arī bakteriofāga T7 gēna 2.5 proteīns (PDB ID: 1JE5), kas ir plaši pētīts SSB (Hernandez, Richardson, 2019). Lai

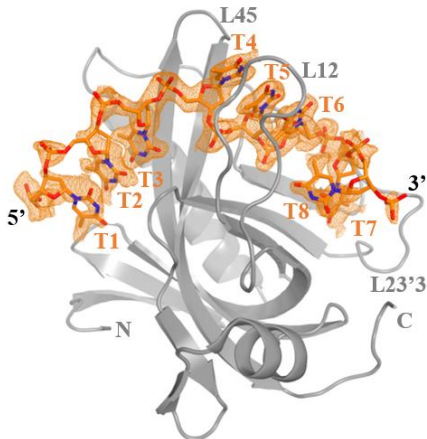
apstiprinātu, ka Enc34 ORF6 patiešām darbojas kā SSB, proteīns tika pārbaudīts agarozes gēla elektroforēzē dažāda veida DNS klātbūtnē. Tika skaidri novērota DNS kustības nobīde vpDNS gadījumā, bet ne proteīnam ar dpDNS. Pārsteidzošā kārtā tika atklāts, ka ko-kristalizācijas izmēģinājumiem izveidotais C-termināli saīsinātais proteīns saistīja kā vpDNS, tā dpDNS ar līdzīgu afinitāti.

Turpinot pētīt Enc34 ORF6 DNS saistīšanas mehānismu, mēs noteicām ORF6 Δ C 3D struktūru kompleksā ar vpDNS. ORF6 Δ C-(dT)₃₅ kristāli difragēja līdz 1,56 Å un saturēja vienu ORF6 Δ C molekulu asimetriskajā vienībā, kas bija saistījusi deviņu timidīna nukleotīdu sekvenci (4. attēls). Kompleksa struktūrā bija redzams, ka DNS pavediens izklāj visu OB folda ligandu saistīšanas virsmu, kas lielākoties ir pozitīvi lādēta un norobežota ar cilpām L12, L23'3 un L45, kuras savieno attiecīgi β 1 un β 2, β 23' un β 3 un β 4 un β 5. Proteīna DNS saistīšanas mehānisma pamatā ir divas aromātiskās aminoskābes un vairākas elektrostātiskās mijiedarbības, kuras ir konservatīvas un raksturo T7 tipa SSB saimi.

A



B



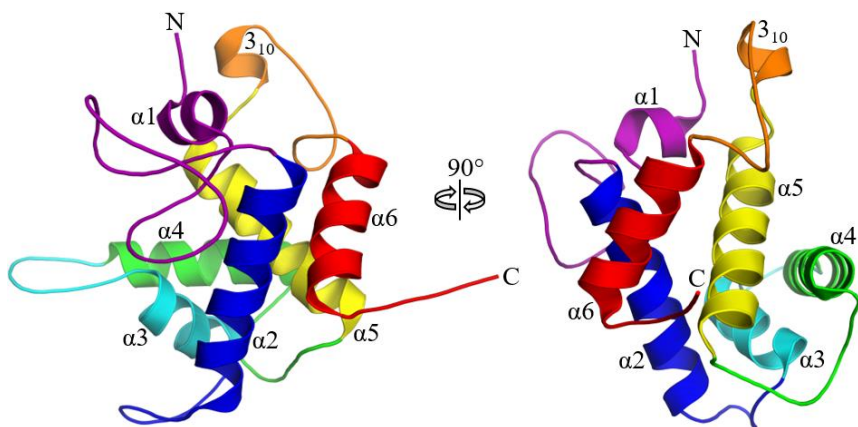
4. attēls. ORF6 Δ C-(dT)₃₅ kompleksa struktūra. DNS oligonukleotīds (oranžā krāsā) saistās pozitīvi lādētajā iedobē OB folda domēnā (A). Atšķirīga kompleksa orientācija, kurā DNS ir pārklāta ar 2Fo-Fc “omit” elektronblīvuma karti, konturētu pie 0,7 σ (B). Abos paneļos timidīni ir numurēti T1-T8 5' \rightarrow 3' virzienā un ir norādītas arī saistīšanās virsmu norobežojošās cilpas.

3.3. Lizocīma folda daudzveidība: neparasta endolizīna, ko kodē fāgs Enc34, katalītiskā domēna struktūra

Lai uzsāktu Enc34 hipotetizētā endolizīna funkcijas un struktūras pētījumus, ORF39 sekvence tika klonēta ekspresijas konstruktā un ar sešiem histidīniem N-galā iezīmēts ORF39 proteīns tika producēts *E. coli* šūnās. Plazmīdai, kas kodēja pilna garuma proteīnu, bija zema transformācijas efektivitāte, tika novērota arī samazināta šūnu augšana un pēc indukcijas nebija konstatējama mērķa proteīna produkcija, kas liecināja par ORF39 produkta toksicitāti baktēriju šūnām. Turpmāki eksperimenti atklāja, ka ORF39 proteīna saīsinātais variants, kas saturēja tikai N-terminālo PHA02564 domēnu, ir viegli producējams šķīstošā formā un ir šūnām nekaitīgs. Tādēļ visos turpmākajos pētījumos tika izmantots proteīna C-gala delēcijas variants ORF39 Δ C.

Lai pārbaudītu Enc34 endolizīna enzimatisko aktivitāti, attīrītam ORF39 Δ C proteīnam tika veikti optiskā blīvuma samazināšanas testi, kā substrātu izmantojot ārējās membrānas ($\bar{A}M$) permeabilizētas baktērijas. Enc34 endolizīns uzrādīja nepārprotamu muralītisko aktivitāti pret $\bar{A}M$ permeabilizētām *Escherichia coli* W3100, *Pseudomonas aeruginosa* PAO1 un *Hafnia alvei* šūnām, pierādot, ka Enc34 endolizīna N-terminālais domēns un līdz ar to arī citu bakteriofāgu genomos atrodamie PHA02564 saimes proteīni patiešām ir funkcionālu endolizīnu enzimatiski aktīvie domēni (EAD).

ORF39 EAD struktūra tika noteikta, izmantojot viena viļņa garuma anomālās difrakcijas metodi ar SeMet aizvietotiem ORF39 Δ C kristāliem, kas difragēja līdz 1,8 Å. Vēlākos eksperimentos tika izrēķināta cita struktūra ar izšķirtspēju līdz 1,6 Å, un tāpēc tā tika izvēlēta publicēšanai. ORF39 EAD kopējā struktūra sastāv tikai no spirālēm un cilpām, un to veido sešas α -spirāles ($\alpha 1$ līdz $\alpha 6$) un viena vienpagrieziena 3_{10} spirāle starp $\alpha 5$ un $\alpha 6$ (5. attēls). Šo struktūru var raksturot kā divu sapludinātu α -spirāļu saišķu kopu, no kuriem pirmo veido $\alpha 2$, $\alpha 3$, $\alpha 4$ un $\alpha 5$, bet otro – $\alpha 2$, $\alpha 5$ un $\alpha 6$. EAD ir globulāras formas ar lielu iedobi starp cilpām, kas savieno $\alpha 1$ - $\alpha 2$ un $\alpha 3$ - $\alpha 4$, un tās centrs ir apmēram $\alpha 3$ spirāles C-galā.

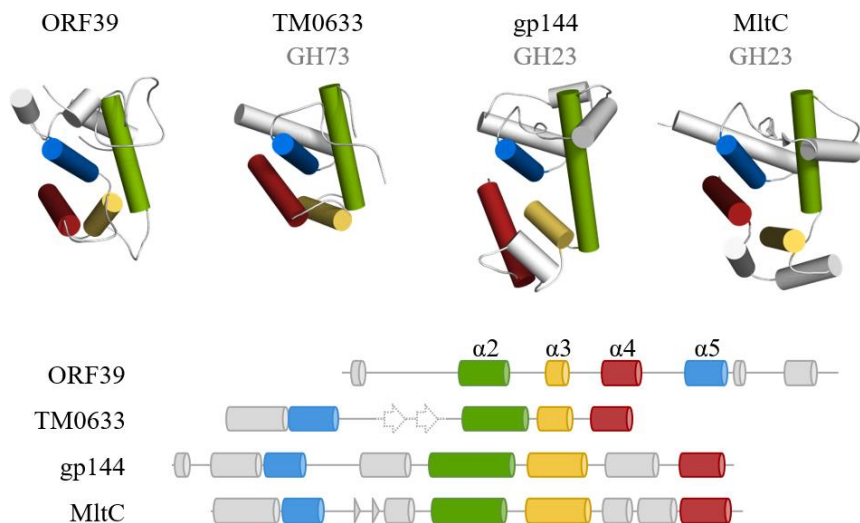


5. attēls. Enc34 endolizīna enzimatiski aktīvā domēna terciārā struktūra. Proteīna aminoskābju secība ir iekrāsota varavīksnes krāsās no violetas līdz sarkanai, attiecīgi no N līdz C-galam.

Strukturālās homoloģijas analīze ar Dali kā līdzīgus atklāja dažādus α -strukturētus proteīnus, tomēr nevienam no “top” līdziniekiem nebija nekādas acīmredzamas radniecības ar endolizīniem. Tikai trīs proteīni, proti, N-acetilglikozaminidāze no *Thermotoga maritima* (PDB ID: 4QDN; Z vērtība: 3,0), lītiskā transglikozilāze gp144 no bakteriofāga ϕ KZ (PDB ID: 3BKH; Z vērtība: 2,6) un *E. coli* lītiskā transglikozilāze MltC (PDB ID: 4C5F; Z vērtība: 2,5) liecināja par potenciālu homoloģiju, kaut arī sekvenču līmenī tie bija tik atšķirīgi no Enc34 ORF39, ka sekvenču salīdzinājumi nebija iespējami. Savukārt struktūras līmenī šie trīs homoloģiskie proteīni reprezentē dažādus α/β lizocīma folda variantus, kas visi sastāv no α un β -strukturētām daivām, starp kurām veidojas dziļa iedobe substrāta saistīšanai un šķelšanai. Enc34 endolizīna EAD ir līdzīgs lizocīma folda α -daivai, bet tajā nav β -strukturētu elementu. Tomēr jāatzīmē, ka lielā cilpa, kas ORF39 proteīnā savieno $\alpha 1$ un $\alpha 2$, būtībā atrodas tajā pašā vietā, kur TM0633 β -daiva, un tai varētu būt līdzīga funkcija substrātu saistošās iedobes aizsegšanā.

ORF39 EAD struktūrā Glu131 pozīcija atbilst katalītiskajām glutamāta atlikumiem TM0633, gp144 un MltC proteīnos, kas apstiprina, ka dziļā iedobe patiešām ir Enc34 endolizīna aktīvais centrs. Konservatīvi ir arī Trp79 un Tyr95, kas ir tikuši parādīti kā svarīgi katalīzei ORF39 strukturālajos homologos (Lipski et al., 2015; Chertkov et al., 2017). Spirāles, kas Enc34 endolizīnā atbilst $\alpha 2$, $\alpha 3$ un $\alpha 4$, šādā secībā ir

atrodamas arī pārējos trīs proteīnos, bet, kamēr $\alpha 5$ Enc34 endolizīnā seko tieši aiz $\alpha 4$, TM0633, gp144 un MltC proteīnos attiecīgā spirālē atrodas pirms $\alpha 2$ analoga (6. attēls). Zīmīgi, ka Enc34 endolizīna katalītiskais Glu131 ir tieši $\alpha 5$ spirālē, bet, neraugoties uz cirkulāro permutāciju, attiecīgais glutamāta atlikums TM0633, gp144 un MltC struktūrās atrodas tieši tajā pašā vietā, kur ORF39 proteīnam. Kopumā šīs konservatīvās strukturālās īpašības liecina par Enc34 endolizīna un TM0633, gp144 un MltC proteīnu evolucionāro radniecību, kā arī iezīmē peptidoglikāna šķelšanas molekulāro motīvu lielai enzīmu supergrupai, kas apvieno glikozīdu hidrolāzes (GH) 73 un GH23 proteīnus.



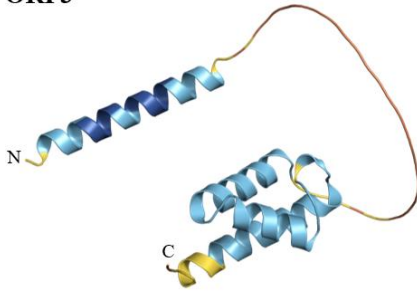
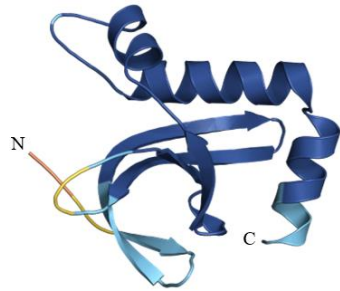
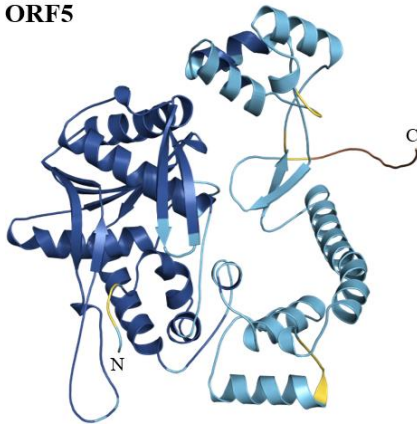
6. attēls. Enc34 endolizīna un GH73 un GH23 glikozīdu hidrolāžu folda līdzība. Četras α -spirāles, kas veido proteīna kodolu, ir izceltas dažādās krāsās. Šo proteīnu sekundārās struktūras elementu diagrammas ir parādītas zemāk, izmantojot to pašu krāsu shēmu. Gp144 un MltC proteīnu pārējie domēni ir noņemti, lai saglabātu attēla uzskatāmību.

Būtisks šī pētījuma trūkums ir tas, ka netika iegūta endolizīna-liganda kompleksa struktūra, kas ļautu sīkāk izpētīt šī Enc34 diverģentā enzīma peptidoglikāna saistīšanas un šķelšanas molekulāros principus. Tika veikti daudzi kristālu mērcēšanas un ko-kristalizācijas eksperimenti, izmantojot komerciāli pieejamo N-acetilglikozamīnu (GlcNAc), N-acetilmuramīnskābi (MurNAc), hitotetraozi un GlcNAc-MurNAc disaharīdu, bet diemžēl bez panākumiem. Lai gan endolizīna-liganda kompleksi ir tikuši iegūti, izmantojot proteīna dabisko formu (Pérez-

Dorado et al., 2007; Fokine et al., 2008), turpmāki kristālu mērcēšanas un ko-kristalizācijas eksperimenti, izmantojot neaktīvu Enc34 endolizīna variantu, varētu radīt labāku iespēju proteīna-liganda mijiedarbības notveršanai. Jāpiemin arī, ka strukturālais homologs TM0633 izrādījās enzimatiski neaktīvs uz peptīdus nesaturošiem glikānu pavedieniem un kristalizēšanas izmēģinājumos ar lineāriem peptidoglikāna analogiem, tieši tāpat kā ORF39 gadījumā, neizdevās iegūt nevienu ar ligandu saistītu TM0633 proteīna struktūru (Lipski et al., 2015). Tāpēc mums ir zināma interese par kristalogrāfijas eksperimentu turpināšanu, šoreiz izmantojot sazarotus peptidoglikāna analogus.

3.4. Enc34 ORF3, ORF4 un ORF5 proteīnu AlphaFold2 struktūras un turpmākā darba perspektīvas

“Proteīnu struktūras prognozēšanas kritiska novērtējuma” pasaules konkursa četrpadsmitā kārtā 2020. gadā iezīmēja nozīmīgu pavērsienu proteīnu struktūrbioģijas jomā. DeepMind izstrādātā dziļās mācīšanās mākslīgā intelekta programma AlphaFold2 (AF2) (Jumper et al., 2021) kļuva par šī ilggadējā izaicinājuma uzvarētāju, spējot prognozēt atsevišķu proteīnu molekulu foldingu ar teju eksperimentālu precizitāti (Pereira et al., 2021). Turpmākais zinātniskās sabiedrības izvērtējums ir tikai apstiprinājis AF2 modeļu uzticamību, un pētnieki ir vienprātis par šīs tehnoloģijas nozīmīgumu struktūrbioģijas un dabaszinātņu turpmākajā attīstībā (Cramer, 2021; Akdel et al., 2022). Taču iepriekš, pirms AF2, gan ORF6, gan ORF39 proteīnu trīsdimensionālās struktūras tika atrisinātas eksperimentāli, ar publikācijā I raksturotajām metodēm. Lai gan ilgais un darbietilpīgais *de novo* struktūras noteikšanas ceļš ir bijis veiksmīgs fāga Enc34 ORF6 un ORF39 produktu strukturālajai raksturošanai, ORF3, ORF4 un ORF5 hipotētisko proteīnu struktūras, kas arī savulaik tika uzstādītas kā darba mērķi, tā arī nekad nematerializējās, jo vienkārši neizdevās iegūt šo proteīnu kristālus. Savukārt AF2 spēj prognozēt visas trīs struktūras minūšu laikā (7. attēls), demonstrējot neironu tīklos balstītu skaitļošanas pieeju apbrīnojāmās un pat pārcilvēciskās spējas.

ORF3**ORF4****ORF5****Modeļa ticamība**

- Ļoti augsta (pLDDT > 90)
- Pārliecinoša (90 > pLDDT > 70)
- Zema (70 > pLDDT > 50)
- Ļoti zema (pLDDT < 50)

7. attēls. Fāga Enc34 ORF3, ORF4 un ORF5 proteīnu AlphaFold2 struktūras. Parādīts katra proteīna visaugstāk novērtētais modelis, un atlikumi ir iekrāsoti atbilstoši to prognozēto lokālo attālumu starpības testa (pLDDT) vērtībām, kas korelē ar strukturālās prognozes ticamību (nepublicēti dati).

Lai ilustrētu, kā mūsdienu sasniegumi mākslīgā intelekta un mašīnmācīšanās jomā var paātrināt proteomikas “tumšās matērijas” izpēti, kā piemēru es aplūkošu Enc34 ORF5 proteīnu. Šis proteīns satur DUF2800, kas Konservatīvo domēnu datubāzē pašlaik tiek anotēts kā CRISPR/Cas sistēmas asociētais proteīns Cas4. Tā prognozētās struktūras pārbaude ar Dali atklāja izteiktu homoloģiju ar vairākiem proteīniem, kam piemīt eksonukleāzes aktivitāte. Promocijas darba sagatavošanas laikā ORF5 Dali “top” līdzinieks ir mikobaktēriju heterodimēriskā helikāze-nukleāze AdnAB (PDB ID: 6PPR; Z-score: 13,2), kas ir atbildīga par DNS divpavedienu pārrāvumu (DSB) apstrādi, lai tos tālāk varētu labot ar homologās rekombinācijas palīdzību (Jia et al., 2019).

Enc34 ORF5 proteīna terciāro struktūru var raksturot kā veidotu no divām daivām ar dziļu iedobi starp tām. Lielākā N-terminālā daiva ir kompakta un īpaši konservatīva, un tā atbilst nukleāzes modulim AdnA subvienībā un citiem strukturāli raksturotiem eksonukleāzes domēniem. Turpretī C-terminālā daiva ir unikāla Enc34 ORF5 un radniecīgajiem fāgu proteīniem, un interesanti, ka prognozētajos AF2 modeļos tā ieņem vairākas alternatīvas konformācijas, atklājot struktūru dažādos atvērtības stāvokļos, no kuriem daži varētu būt bioloģiski relevanti. Atvērtākajās ORF5 proteīna konformācijās ir iespējams iemodelēt “dakšveida” DNS molekulu no AdnAB kompleksa struktūras tikai ar nebūtiskiem stēriskajiem traucējumiem, kas liecina, ka ORF5 proteīns, līdzīgi kā AdnA, potenciāli apstrādā Y formas DNS substrātu. Daudzi no identificētajiem ORF5 proteīna homoloģiem darbojas kopā ar helikāzēm, kas nodrošina DNS pavedienu atvīšanu un kompleksa pārvietošanu, un tamdēļ varētu izpētīt, vai arī ORF5 proteīns veido kompleksus ar paša fāga vai saimniekbaktērijas helikāzes domēniem. Jāatzīmē, ka fāgam Enc34, papildus replikatīvajai helikāzei, kas ir ietverta ORF1 bifunkcionālajā praimāzes-helikāzes proteīnā, ir arī papildus DNS helikāze, ko kodē ORF9 un kuras funkcija vēl nav izpētīta. Ja šādi daudzproteīnu kompleksi tiktu identificēti, tie būtu interesanti kriogēnās elektronmikroskopijas pētījumu objekti, jo īpaši ņemot vērā ORF5 radniecību ar DSB labošanas sistēmām un pat CRISPR asociētajiem proteīniem. Būtiska līdzība starp Enc34 ORF5 un Cas4 proteīniem patiešām ir konstatējama, tomēr bez papildus pierādījumiem ORF5 iespējamā iesaiste CRISPR speiseru apstrādē (Lee et al., 2019) vai jebkurā citā ar CRISPR saistītā procesā joprojām nav zināma, un tas ir vēl viens saistošs jautājums turpmākajiem pētījumiem.

SECINĀJUMI

1. Bakteriofāga Enc34 ORF6 produktu un tā saīsināto, kā arī selenometionīna aizvietoto variantu var rekombinanti producēt *E. coli*. Samazināta kultivēšanas temperatūra un induktora koncentrācija palielināja mērķproteīnu šķīdību.
2. Enc34 ORF6 proteīnu un tā variantus var attīrīt, kombinējot hromatogrāfijas metodes, un kristalizēt, izmantojot komerciālus šķīdumu komplektus. C-termināli saīsinātais ORF6 proteīns kristalizējās arī DNS oligonukleotīda klātbūtnē.
3. Struktūras līmenī Enc34 ORF6 proteīns sastāv no N-termināla oligonukleotīdu/oligosaharīdu saistīšanas folda domēna un C-termināla nestrukturēta segmenta, un tas ir līdzīgs fāga T7 vienpavediena DNS saistošajam proteīnam gp2.5.
4. Enc34 ORF6 proteīns saista vienpavediena DNS *in vitro*, un tā C-gals spēlē lomu vienpavediena un divpavedienu DNS izšķiršanā.
5. Bakteriofāga Enc34 ORF39 produkts ir toksisks *E. coli* šūnām, taču C-termināli saīsinātu proteīnu un tā selenometionīna aizvietoto variantu var iegūt, attīrīt un kristalizēt, izmantojot līdzīgas laboratorijas metodes kā ORF6 proteīna gadījumā.
6. Enc34 ORF39 proteīna N-terminālais domēns šķeļ Gram-negatīvu peptidoglikānu *in vitro* ar līdzīgu aktivitāti kā komerciāls vistas olu baltuma lizocīms.
7. Struktūras līmenī Enc34 ORF39 N-terminālais domēns sastāv no piecām α -spirālēm, kas sakārtotas divos sapludinātos spirāļu saišķos, un tā struktūra ir attāli līdzīga lizocīma folda α -strukturētajai daļai.
8. Enc34 ORF39 proteīns ir radniecīgs glikozīdu hidrolāzes GH73 un GH23 enzīmiem, kas ietver fāgu endolizīnus un baktēriju autolizīnus un kuriem visiem ir kopīgs ļoti konservatīvs trīs aminoskābju motīvs substrāta šķelšanai.
9. Proteīni, kas satur nezināmas funkcijas domēnu DUF2815, ir vienpavediena DNS saistošie proteīni, un fāgu proteīnu PHA02564 saimes pārstāvji ir Gram-negatīvu peptidoglikānu noārdoši endolizīni.

PATEICĪBAS

SIA “Mikrotīkls” stipendija un Eiropas Sociālā fonda projekts “Doktorantūras kapacitātes stiprināšana Latvijas Universitātē jaunā doktorantūras modeļa ietvaros” (granta nr. 8.2.2.0/20/I/006) mani ļoti būtiski finansiāli atbalstīja praktiski visā doktorantūras garumā. Esmu pagodināta un pateicīga, ka esmu bijusi šo grantu saņēmēja, bez kuriem mans promocijas darbs, visticamāk, nekad nebūtu tapis.

Esmu pateicīga savam darba vadītājam Dr. biol. Andrim Kazākam par to, ka iepazīstināja mani ar brīnumaino fāgu bioloģijas jomu un vadīja mani cauri visiem pētniecības projekta posmiem, ar pacietību, labsirdību un dāsnumu laika un uzmanības veltīšanā. Tas ir bijis prieks un privilēģija – strādāt plecu pie pleca kopā laboratorijā, mācīties no Jums un piedalīties visās diskusijās un humora pilnajās sarunās, kas mums ir bijušas šo gadu gaitā. Vēlos pateikties arī prof., Dr. biol. Kasparam Tāram par to, ka uzņēma mani BMC Struktūrbioloģijas grupā, par to, ka aicināja mani līdzī sinhrotrona braucienos un konferencēs, un par to, ka mudināja mani pieteikties grantiem, kad es biju sākusi šaubīties. Vēl vienu milzīgu paldies es veltu savam kolēģim Dr. biol. Jānim Rūmniekam par to, ka viņš principā bija mans mentors visdažādākajos struktūrbioloģijas un datu prezentēšanas jautājumos. No struktūru noteikšanas un datu analīzes līdz akadēmisko rakstu sagatavošanai, Tu man esi ļoti palīdzējis un daudz iemācījis, par ko es esmu tiešām neizmērojami pateicīga.

Vēlos teikt paldies arī citiem Struktūrbioloģijas grupas biedriem un BMC kolēģiem, it īpaši Inārai, Ņikitam, Annai, Evai un Mihailam, ar kuriem man ir bijis prieks sadarboties pētniecībā, pasniegšanā un citos projektos. Esmu pateicīga Ilzei un visiem BMC Studentu padomes locekļiem par kopā sanākšanu un foršo zinātnes komunikācijas pasākumu organizēšanu, kuros mēs ne vien demonstrējām zinātni plašām auditorijām, bet arī paši guvām iedvesmu turpmākam darbam. Paldies maniem draugiem un vienaudžiem Elzai, Aivaram, Kārlim, Karīnai un Vladimiram par to, ka bijāt kopā ar mani šajā procesā, padarījāt to daudz jautrāku un sniedzāt nepārtrauktu atbalstu. Visbeidzot, sirsnīgs paldies Rūtai par to, ka viņa ir absolūti vislabākais draugs, kāds vien var būt, un manai apbrīnojamajai mammai par to, ka viņa mani iedvesmo ar savu piemēru un ka nemitīgi lepojas ar mani, lai ko arī es nebūtu nolēmusi darīt un sasniegt.

University of Latvia
Faculty of Biology

Elīna Černooka

**STRUCTURAL AND FUNCTIONAL STUDIES OF
UNCHARACTERIZED CONSERVED PROTEINS
FROM BACTERIOPHAGE ENC34**

Summary of Doctoral Thesis

For the degree of Doctor of Science in Biology
Subfield of Molecular Biology

Rīga, 2022

The research for the thesis was conducted at the Latvian Biomedical Research and Study Centre from 2016 to 2022.



Latvian Biomedical
Research and Study Centre
research and education in biomedicine from genes to human

This work was supported by SIA “Mikrotīkls” (MikroTik) scholarship administered by the University of Latvia Foundation and the European Social Fund grant no. 8.2.2.0/20/I/006.



UNIVERSITY OF LATVIA
FOUNDATION

NATIONAL
DEVELOPMENT
PLAN 2020



EUROPEAN UNION
European Social
Fund

INVESTING IN YOUR FUTURE

Form of the thesis: Collection of publications in Molecular Biology

Supervisor: Dr. biol. Andris Kazāks

Reviewers:

- 1) Dr. biol. Kalvis Brangulis, Latvian Biomedical Research and Study Centre
- 2) Prof., Dr. chem. Kristaps Jaudzems, Latvian Institute of Organic Synthesis
- 3) Asst. prof. Emilio Parisini, PhD, Istituto Italiano di Tecnologia and Latvian Institute of Organic Synthesis

The thesis will be defended at the public session of the Doctoral Committee of Biology, University of Latvia, at 13:00 on May 12, 2023, at the Latvian Biomedical Research and Study Centre, Rātsupītes iela 1 k-1.

The thesis and its summary are available at the University of Latvia Library, Kalpaka bulvāris 4.

University of Latvia Doctoral Committee of Biology

Chairman _____ / Prof., Dr. biol. Nils Rostoks /

Secretary _____ / Dr. biol. Daina Eze /

© University of Latvia, 2022

© Elīna Černooka, 2022

ANNOTATION

A significant proportion of most phage genomes constitutes the “dark matter of the biosphere”, exhibiting low or undetectable similarity to characterized sequences and encompassing numerous fascinating protein targets for dedicated studies that could finally unearth their novel biology or translational potential. *Hafnia* virus Enc34, discovered by our laboratory, encodes several such proteins including the open reading frame (ORF) 6 hypothetical DNA replication protein and the ORF39 putative endolysin, both of which represent evolutionarily conserved yet uncharacterized protein families. To gain a better understanding about these proteins, their native, mutant and selenomethionine-substituted variants were produced, purified and crystallized, and their three-dimensional structures were experimentally determined. Structural homology analyses revealed that the Enc34 ORF6 protein is a diverged member of the phage T7-type single-stranded DNA-binding proteins (SSBs). Its DNA-binding properties were verified *in vitro*, and a crystal structure of the protein in complex with a DNA oligonucleotide was solved. The complex structure elucidated for the first time the molecular basis of DNA binding for a major SSB family, found in both phages and eukaryotic viruses. In the case of the Enc34 ORF39 protein, based on genomic context we had hypothesized it to represent a new type of phage endolysin, a peptidoglycan-degrading enzyme that digests the bacterial cell wall “from within”. The proposed enzymatic activity was verified by *in vitro* assays, and the three-dimensional structure of the protein’s enzymatically active domain unveiled a highly diverged variation of the lysozyme fold. Structural analysis uncovered that the Enc34 ORF39 protein is only distantly related to one other phage endolysin and two bacterial autolysins that, nevertheless, all share a conserved three amino acid scaffold for substrate cleavage. As a consequence of our work, the domain of unknown function (DUF) 2815 and phage protein family PHA02564 annotations are expected to become obsolete, having been illuminated from the proteomic dark matter as SSBs and endolysins, respectively.

TABLE OF CONTENTS

ABBREVIATIONS	5
INTRODUCTION	6
Aims of the thesis	6
Tasks for achieving the aims	7
Theses	7
APPROBATION OF RESEARCH	8
Publications	8
Conferences and seminars	8
Awards and scholarships	10
1. LITERATURE REVIEW	11
1.1. Phage biology	11
1.2. <i>Hafnia</i> virus Enc34.....	13
2. MATERIALS AND METHODS	16
2.1. Sequence analysis.....	16
2.2. Plasmid construction	16
2.3. Protein production	17
2.4. Protein purification.....	17
2.5. Crystallization and data collection	18
2.6. Structure determination	18
2.7. Structure analysis.....	19
2.8. Electrophoretic mobility shift assay	19
2.9. Turbidity reduction assay	20
3. RESULTS AND DISCUSSION	21
3.1. Structural characterization of a single-stranded DNA-binding protein: a case study of the ORF6 protein from phage Enc34.....	21
3.2. Structural basis for DNA recognition of a single-stranded DNA-binding protein from <i>Enterobacter</i> phage Enc34	23
3.3. Diversity of the lysozyme fold: structure of the catalytic domain from an unusual endolysin encoded by phage Enc34.....	25
3.4. AlphaFold2 structures of the Enc34 ORF3, ORF4 and ORF5 proteins and prospects for future work	28
CONCLUSIONS	31
ACKNOWLEDGEMENTS	32
REFERENCES	33

ABBREVIATIONS

Å	ångström
AF2	AlphaFold2
BMC	Latvian Biomedical Research and Study Centre
CDC	Centers for Disease Control and Prevention
CRISPR	clustered regularly interspaced short palindromic repeats
Dali	distance matrix alignment
DSB	double-strand break
dsDNA	double-stranded deoxyribonucleic acid
DTT	dithiothreitol
DUF	domain of unknown function
EAD	enzymatically active domain
ECDC	European Centre for Disease Prevention and Control
GH	glycoside hydrolase
GlcNAc	N-acetylglucosamine
IPTG	isopropyl β -D-1-thiogalactopyranoside
kbp	kilobase pairs
MurNAc	N-acetylmuramic acid
NCBI	National Center for Biotechnology Information
OB	oligonucleotide/oligosaccharide binding
OD600	optical density at 600 nanometers
OM	outer membrane
ORF	open reading frame
PBS	phosphate buffered saline
PCR	polymerase chain reaction
PDB	Protein Data Bank
PHA	phage protein
pLDDT	predicted local distance difference test
SeMet	selenomethionine
SSB	single-stranded deoxyribonucleic acid-binding protein
ssDNA	single-stranded deoxyribonucleic acid
ssRNA	single-stranded ribonucleic acid
TAE	Tris-acetate-ethylenediaminetetraacetic acid
TEV	tobacco etch virus
TN	Tris-NaCl (sodium chloride)
WHO	World Health Organization

INTRODUCTION

Bacteriophages are beloved model systems in natural sciences, and the continuous studies of the paradigm coliphages λ , T4, M13 and a few others, have generated many fundamental discoveries in biology and a myriad of tools for science and industry. Genomic and metagenomic sequencing efforts have revealed the occurrence of diverse phages in every part of the nature, where they exist as a major driving force for ecological and evolutionary processes, yet many aspects of phage biology are still thought to remain undiscovered. The global threat of antimicrobial resistance has further invigorated interest in phages as agents for combating bacterial infections and contamination. Many innovative antibacterials can be expected to emerge from phage-encoded cell lysis machineries, which have turned out to be more diverse than previously thought and still remain underexplored.

Within the growing record of bacteriophage sequences, most of the encoded proteins lack reliable functional annotations due to low or undetectable similarity to experimentally characterized gene products and, thus, constitute the proverbial dark matter of the biosphere. Some of these mysterious proteins, nevertheless, show notable conservation with other viral, prokaryotic or even eukaryotic homologs and, therefore, are excellent targets for structural and functional studies, having the potential to elucidate entire protein families with apparently essential biological roles. Our model bacteriophage Enc34, isolated at the BMC, encodes several such proteins in its genomic modules for DNA replication and host cell lysis, that upon closer inspection have attracted our interest for experimental investigation.

Aims of the thesis

The aims of my doctoral thesis were to gain new knowledge about the structure and function of uncharacterized proteins from the *Hafnia* virus Enc34 proteome, to shed light on understudied protein families and to contribute to functional annotation of biological sequences.

Tasks for achieving the aims

The specific tasks for achieving the aims were:

- To identify conserved uncharacterized proteins of interest in the Enc34 proteome using *in silico* sequence analysis tools;
- To develop production, purification and crystallization pipelines for the native and selenomethionine-substituted protein variants;
- To *de novo* determine and analyze the three-dimensional structure of the selected proteins and to initiate hypotheses regarding their function;
- To explore the proposed function of the target proteins by additional structural studies and *in vitro* assays.

Theses

The main theses of this work were:

1. Bacteriophage Enc34 ORF6 protein is a diverged component of the viral DNA replication machinery.
2. Bacteriophage Enc34 ORF39 encodes an endolysin with an atypical architecture and a highly diverged enzymatically active domain.
3. Focused structural studies, complemented with *in vitro* assays, can illuminate previously uncharacterized proteins from the viral dark matter.
4. Investigation of conserved understudied proteins can provide insight into larger protein families and further improve functional annotation of related sequences.

APPROBATION OF RESEARCH

Publications

- I. **Černooka, E.**, Rūmnieks, J., & Kazāks, A. (2021). Structural characterization of a single-stranded DNA-binding protein: a case study of the ORF6 protein from bacteriophage Enc34. *Methods in molecular biology* (Clifton, N.J.), 2281, 343–373. https://doi.org/10.1007/978-1-0716-1290-3_23
- II. **Černooka, E.**, Rūmnieks, J., Tārs, K., & Kazāks, A. (2017). Structural basis for DNA recognition of a single-stranded DNA-binding protein from *Enterobacter* phage Enc34. *Scientific reports*, 7(1), 15529. <https://doi.org/10.1038/s41598-017-15774-y>
- III. **Černooka, E.**, Rūmnieks, J., Zrelavs, Ņ., Tārs, K., & Kazāks, A. (2022). Diversity of the lysozyme fold: structure of the catalytic domain from an unusual endolysin encoded by phage Enc34. *Scientific reports*, 12(1), 5005. <https://doi.org/10.1038/s41598-022-08765-1>

Conferences and seminars

1. **Černooka, E.**, Rūmnieks, J., Zrelavs, Ņ., Tārs, K., Kazāks, A. Structural analysis of a divergent endolysin from *Hafnia* bacteriophage Enc34. Instruct Biennial Structural Biology Conference. Poster presentation. 2022, Utrecht, The Netherlands
2. **Černooka, E.**, Rūmnieks, J., Zrelavs, Ņ., Kazāks, A. Structure and function of proteins from a bacteriophage model system. The 78th International Scientific Conference of the University of Latvia, Scientific session of scholarship recipients supported by the patrons of the University of Latvia Foundation. Oral presentation. 2020, Riga, Latvia

3. Zrelavs, N., Černooka, E., Dišlers, A., Kazāks, A. Genomic and proteomic studies of novel bacteriophages from Latvian Biomedical Research and Study Centre's phage collection. The 9th International Conference 'Bioresources and Viruses'. Oral presentation. 2019, Kyiv, Ukraine
4. Černooka, E., Rūmnieks, J., Zrelavs, N., Tārs, K., Kazāks, A. Structural studies of conserved uncharacterized proteins in the genomes of novel DNA phages. FEBS3+ Conference. Oral presentation. 2019, Riga, Latvia
5. Černooka, E., Rūmnieks, J., Zrelavs, N., Tārs, K., Kazāks, A. Characterization of newly isolated DNA phages and structural and functional studies of their proteins. The 77th International Scientific Conference of the University of Latvia, Scientific session of scholarship recipients supported by the patrons of the University of Latvia Foundation. Oral presentation. 2019, Riga, Latvia
6. Černooka, E., Kazāks, A. Studies of non-structural proteins encoded by DNA bacteriophages. Instruct-ERIC seminar. Oral presentation. 2019, Riga, Latvia.
7. Černooka, E., Kazāks, A. Genomic characterization of newly discovered DNA bacteriophages and structural and functional studies of their encoded proteins. The 76th International Scientific Conference of the University of Latvia, Molecular Biology Section. Oral presentation. 2018, Riga, Latvia
8. Černooka, E., Tārs, K., Kazāks, A. Structural and functional studies of a hypothetical replication protein from *Enterobacter* phage Enc34. The 75th International Scientific Conference of the University of Latvia, Molecular Biology Section. Oral presentation. 2017, Riga, Latvia

Awards and scholarships

- Certificate of Appreciation to Zrelovs, N., **Černooka, E.**, Rūmnieks, J., Kazāks, A., Dišlers, A., from the Latvian Academy of Sciences for “Establishing a new research direction in Latvia that is dedicated to DNA bacteriophage discovery and investigation”, 2021
- **Černooka E.** SIA “Mikrotīkls” scholarship for doctoral students in natural and medical sciences, 2017-2020
- **Černooka E.** Mundheim family excellence scholarship, 2014-2015

1. LITERATURE REVIEW

From proving that DNA is the molecule of inheritance to the discovery of clustered regularly interspaced short palindromic repeats (CRISPR), studies involving bacteriophages have always been in the forefront of molecular biology, delivering important insights and tools now widely used in science and industry. Many of the most impactful discoveries and applications originated from curiosity about both phages themselves and biological processes as they manifested in bacteriophage model systems (Salmond & Fineran, 2015). Still, these achievements may only be the proverbial tip of the iceberg, since only a tiny fraction of all phages in nature has been studied, and merely a small section of the viral gene pool has been mapped and explored (Aevarsson et al., 2021). As natural antagonists of bacteria, phages can also serve as biocontrol agents and their lytic mechanisms can be exploited to create new pharmaceuticals for bacterial infections. This line of research has recently reactivated in light of mounting antimicrobial resistance among common bacterial pathogens and prominent public health agencies sounding the alarm on the imminent “post-antibiotic era” as well as calling for the development of new antibacterial treatments (WHO, 2015; CDC, 2019; ECDC & WHO, 2022). Undoubtedly, many more discoveries and innovations will be derived from both basic and translational phage research, which will steadily continue its contribution to science, technology and medicine.

1.1. Phage biology

Bacteriophages infect the ubiquitous bacteria and, thus, are among the smallest and most simple, yet abundant entities in nature with their own genetics, evolutionary histories and some self-governing capacities (Yolles & Frieden, 2022). Now it is recognized that phages are present in all prokaryotic habitats, from aquatic to terrestrial, to plant and animal-associated, and that they outnumber all organisms in the biosphere combined (Cobián Güemes et al., 2016). As the most numerous biological entities on Earth, phages have major roles to play in biogeochemical cycling (Suttle, 2007), shaping of bacterial ecosystems (Chevallereau et al., 2021) and driving microbial evolution (Hampton et al., 2020).

Phages differ not only in their virion structures, which can be tailed, non-tailed, filamentous or enveloped, but also in the nucleic acid type comprising the viral genome, as single-stranded (ss) and double-stranded (ds) DNA and RNA phages are currently known (Ofir & Sorek, 2018). With the advancements in sequencing technologies, genomics became an increasingly large and important part of phage research (Brüssow & Hendrix, 2002) and still remains a cornerstone of new phage discovery and characterization. The expansion of phage genomics immediately uncovered remarkable variation in genome size and sequence, and further revealed that phage genomes are mosaic in structure, i.e., composed of gene blocks or modules acquired from other phages, due to extensive genetic exchange within the population (Hatfull, 2008). As of now, the GenBank genetic sequence database holds more than 16,000 unique phage sequences (curated according to Cook et al., 2021) and this collection appears to be doubling every two to three years, with no apparent end in sight. However, only a small minority of bacteria and their associated phages are readily cultivable under laboratory conditions; therefore, the arrival and recent surge of viral metagenomics has allowed for an even more comprehensive assessment of phage distribution in different natural habitats (Paez-Espino et al., 2016; Gregory et al., 2019) and has further illuminated the vastness of genetic diversity within the virus world (Dion et al., 2020). To date, a large section of the phage genomic and metagenomic data still constitutes the so-called dark matter of the biosphere, that exhibits low or undetectable similarity to any known sequences and, thus, encompasses entire genome modules and enigmatic individual genes yet to be explored (Hatfull, 2015; Santiago-Rodriguez & Hollister, 2022).

Tailed phages with DNA genomes around 25-100 kilobase pairs (kbp) are ostensibly the most common (Zreløvs et al., 2020a) and typically encode tens of phage proteins in a strictly modular arrangement, likely due to selective pressures for maintaining an efficient compact genome (Berg & Roux, 2021). The functional modules generally cluster proteins pertaining to lysogeny, DNA replication, genome packaging, phage head and tail morphogenesis, and host lysis (Brüssow & Hendrix, 2002) which in the larger genomes are often punctuated with hyperplastic regions encoding novel proteins with potentially adaptive functions (Comeau et al., 2008).

1.2. *Hafnia* virus Enc34

The Latvian Biomedical Research and Study Centre (BMC) has had a prevalent interest in phage research, historically and still directed towards ssRNA bacteriophages and their structural proteins (Rūmnieks & Tārs, 2011; Rūmnieks & Tārs, 2014; Rūmnieks & Tārs, 2017; Rūmnieks et al., 2020). During previous efforts to find new RNA phages for study from different environmental sources we, in fact, had accumulated a collection of more than 40 isolates of tailed DNA viruses. For a while these phages were not further explored, until DNA sequencing became much more available to us. That instigated a renewed interest in characterizing the viruses we had discovered and ultimately resulted in a new direction of tailed phage research at BMC and on a national level. At the time of writing, the BMC phage collection has grown by tens of additional isolates from insect, wastewater and even Antarctic soil samples, among other sources. A total of 18 phage whole genome sequences have been deposited in GenBank as of now, and our findings include the first cultured phages of *Virgibacillus halotolerans* (Zreløvs et al., 2020b) and *Erwinia persicina* (Zreløvs et al., 2020c), a highly diverged *Lactococcus* bacteriophage (Zreløvs et al., 2021), a *Morganella* phage with a distinct genome packaging mode (Zreløvs et al., 2022a), and sequences of novel *Psychrobacillus* viruses originating from Antarctic soil (Zreløvs et al., 2022b) all with unique and fascinating features.

The very first fully sequenced isolate from the collection, however, was the bacteriophage Enc34 (Kazāks et al., 2012), then thought to infect *Enterobacter cancerogenus* based on biochemical tests, but more recent molecular analyses showed its host to be an unusual strain of *Hafnia alvei* (see Publication III). *H. alvei* is a Gram-negative bacterium of the family *Hafniaceae* within the order *Enterobacterales* (Adeolu et al., 2016) and while it is typically regarded as a commensal of animal gut microbiota (Song et al., 2017) the species has also been identified as the causative agent of opportunistic infections in humans (Ionescu et al., 2022) and various diseases in bees, fish and other animals (Padilla et al., 2014). The virion of *Hafnia* virus Enc34 has a head-tail structure with a long non-contractile tail (Figure 1, left) and the viral genome is a linear dsDNA molecule approximately 60 kbp in length with 5' overhangs, known as

cohesive ends or cos, that indicate the 5' cos genome packaging strategy (Casjens & Gilcrease, 2009). Formally named *Enchivirus Enc34*, the phage is currently the only member of its genus, which further belongs to the family *Casjensviridae* (Figure 1, right). At present, this family comprises a large cluster of *Salmonella* viruses of the genus *Chivirus* and many small genera of one or a few isolated phages. The latter infect an array of different Gram-negative bacteria, including *Burkholderia* (Lynch et al., 2012), *Providencia* (Onmus-Leone et al., 2013), *Xylella* (Ahern et al., 2014), *Proteus* (Nguyen et al., 2019), *Pectobacterium* (Zaczek-Moczyłowska et al., 2020) and others, while still exhibiting similar genome size and notable synteny of genes.

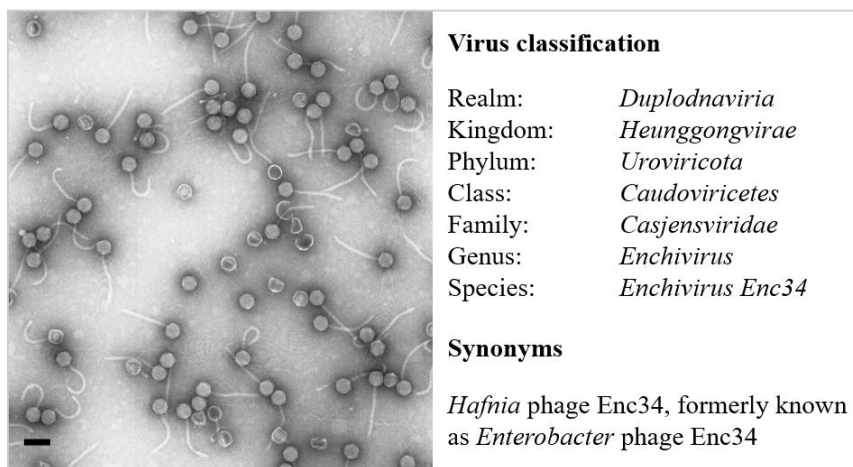


Figure 1. *Hafnia* phage Enc34. Transmission electron microscopy of virions (left) and phage taxonomy and nomenclature (right). Scale bar, 100 nm (unpublished data).

The Enc34 genome contains 80 open reading frames (ORFs) organized in functional modules responsible for viral DNA replication, genome packaging, virion morphogenesis, and host cell lysis, however, 45 ORFs had no recognizable similarity to known gene products and thus were annotated as hypothetical (Kazāks et al., 2012) (GenBank ID: JQ340774.2). In my work, I narrowed the focus to uncharacterized proteins in the phage Enc34 DNA replication module, in the middle of which there appeared to be a gene block encoding four enigmatic hypothetical proteins (ORFs 3 to 6), and to the cell lysis cassette, seemingly containing a new type of bacterial cell wall-degrading enzyme

or endolysin (ORF39). The latter protein was predicted to have an unusual architecture containing a C-terminal transmembrane domain and a highly diverged enzymatically active domain. The selected proteins were notably conserved, however, none of the identifiable sequence homologs, at the time, had been investigated experimentally. Thus, studies of such protein targets, in addition to uncovering essential structural and functional aspects of the protein itself, would also illuminate entire protein families, bringing them forth from the “viral dark matter” (Galperin & Koonin, 2004). My doctoral thesis details *de novo* structure determination and function verification of the ORF6 hypothetical DNA replication protein and the ORF39 putative endolysin. The computationally solved structures of the ORF3, ORF4 and ORF5 products are also presented below with a brief discussion on future work prospects.

2. MATERIALS AND METHODS

To study the structure and function of proteins from our model phage Enc34, I employed a range of molecular, microbiological and computational techniques. Here I present a summary of the main methods and procedures used to study Enc34 ORF6 and ORF39 proteins. More specific details can be found in the respective papers, especially Publication I, which in itself is a comprehensive guide for a structural biology study.

2.1. Sequence analysis

Enc34 protein sequences were examined with BLASTP (Boratyn et al., 2013) using the National Center for Biotechnology Information (NCBI) non-redundant protein sequence database, as well as HHpred (Gabler et al., 2020) and the PDB_mmCIF70 database. Transmembrane helices were identified with TMHMM (Krogh et al., 2001) and DeepTMHMM (Hallgren et al., 2022) and signal peptides were predicted with SignalP (Teufel et al., 2022). Multiple sequence alignments for further homology and phylogenetic analyses were generated using Clustal Omega (Sievers & Higgins, 2021).

2.2. Plasmid construction

The coding sequences for proteins of interest were PCR-amplified from the Enc34 genomic DNA (GenBank ID: JQ340774). ORF6 was amplified using primers that contained the sequence of the tobacco etch virus (TEV) protease cleavage site and restriction sites for molecular cloning and clone selection. The amplified DNA was cloned into a pETDuet-1 vector (Novagen) in-frame with an N-terminal hexa histidine-tag (His-Tag). The resultant vector containing the affinity tag, followed by the TEV site and a blunt-end restriction site for insert excision was used to construct all subsequent plasmids, including that encoding ORF39. To facilitate blunt-end cloning, forward primers were phosphorylated by the T4 polynucleotide kinase (Thermo Scientific) prior to PCR. Plasmids for the C-terminally truncated protein variants were created using appropriate reverse primers and the same pETDuet-1-derived vector backbone.

2.3. Protein production

Native proteins and their truncated variants were produced in *E. coli* BL21(DE3) cells using an optimized protocol that we have developed to increase solubility and yield of target proteins. Briefly, cells transformed with the protein-encoding plasmid were grown in 2xTY medium with 50 µg/mL ampicillin at 25 °C to an OD600 of 0.6-0.8. Then the cultivation temperature was reduced to 22 °C and target protein production was induced with 0.01 mM isopropyl β-D-1-thiogalactopyranoside (IPTG) overnight. The selenomethionine-substituted (SeMet) proteins were produced in *E. coli* B834 using a reagent kit (Molecular Dimensions) and a slightly modified protocol that in our studies improved target protein solubility. First, the cells were grown 2xTY medium with 50 µg/mL ampicillin at 25 °C to an OD600 of 0.8-1.0, upon which they were transferred to a dedicated commercial medium with nutrients and incubated at 25 °C for 2 h. Then, SeMet solution was added and overproduction was induced with 0.1-1.0 mM IPTG at 25 °C overnight. The cells were harvested by centrifugation and stored at -20 °C until further use.

2.4. Protein purification

To enable target protein isolation, the cell mass was thawed and resuspended in TN buffer (20 mM Tris-HCl (pH 8.0), 300 mM NaCl) and disrupted by sonication. The lysate was clarified by centrifugation and applied onto a HisTrap affinity chromatography column (GE Healthcare). The column was washed with TN buffer containing 20 mM imidazole, after which the bound protein was eluted in TN buffer containing 300 mM imidazole. The His-Tag was removed by TEV protease, after which another round of affinity chromatography was performed to separate the cleaved protein. To obtain sufficient purity for crystallization, ORF6 protein required an additional step high-resolution ion-exchange chromatography, for which I used the MonoQ column (GE Healthcare). The SeMet proteins were isolated and purified following the same protocol as for the respective native protein, except that all buffer solutions were supplemented with dithiothreitol (DTT) to reduce protein aggregation.

2.5. Crystallization and data collection

Purified protein preparations were buffer exchanged into 20 mM Tris-HCl (pH 8.0), 100 mM NaCl (with 1 mM DTT for the SeMet proteins) and concentrated to ~10 mg/mL using appropriate molecular weight cut-off filters (Millipore). Then, sitting-drop vapor-diffusion crystallization trials were set up at room temperature, using commercially available crystallization screens, such as JCSG-plus, Structure screen 1 + 2 and PACT premier (Molecular Dimensions) and a liquid handling system (Tecan). The crystal-growing conditions were replicated and manually optimized to obtain more crystals, which were then transferred to cryoprotectant solution and flash-frozen in liquid nitrogen, awaiting X-ray data collection. To produce protein-ligand complexes, I primarily used co-crystallization, where the target protein was mixed with the ligand in varying molar ratios, just prior to screening and/or subsequent crystallization experiments. In the case of Enc34 endolysin, crystal soaking was also attempted, however, it together with co-crystallization failed to reveal recognizable ligand binding in the obtained structures. Diffraction data from all the crystal samples were collected on several occasions at MAX-lab and MAX IV (Lund University) and BESSY II (Helmholtz-Zentrum Berlin).

2.6. Structure determination

To solve the ORF6 protein structure, diffraction images from a SeMet crystal were processed with MOSFLM (Battye et al., 2011) and scaled using SCALA (Evans, 2006) from the CCP4 software suite (Potterton et al., 2018). Phasing was done using the SHELX C/D/E pipeline (Usón & Sheldrick, 2018), followed by density modification in DM (Cowtan, 2010). An initial model was generated with BUCCANEER (Cowtan, 2006), which was then manually re-built in COOT (Casañal et al., 2020) and refined with REFMAC5 (Kovalevskiy et al., 2018). Structure of the C-terminally truncated ORF6 protein was solved by molecular replacement using MOLREP (Vagin & Teplyakov, 2010) and the protein-DNA complex structure was later solved using PHASER (McCoy et al., 2007). For the ORF39 protein, diffraction images were processed with XDS (Kabsch, 2010a), after which multiple SeMet datasets were scaled together using XSCALE (Kabsch, 2010b), through

the XDSAPP interface (Sparta et al., 2016). The structure was solved by AutoSol Wizard (Terwilliger et al., 2009) from the Phenix suite (Liebschner et al., 2019), and the model was used to phase a slightly higher-resolution native dataset. In later co-crystallization and crystal soaking experiments, many more high-resolution structures were solved with PHASER, from which the most complete model was further improved with COOT and REFMAC5 and used for publishing. All ORF6 and ORF39 reported structures underwent validation in COOT and MolProbity (Williams et al., 2018).

2.7. Structure analysis

Model inspection and close-up analyses were done in COOT and PyMOL (Schrödinger & DeLano, 2020). Structural homologs of the ORF6 and ORF39 proteins were identified using the distance matrix alignment (Dali) server (Holm, 2022). Intermolecular interactions within the crystals were examined with PDBePISA (Krissinel & Henrick, 2007). For the ORF6 protein, electrostatic surface was calculated with PyMOL's APBS Tools plug-in (Jurrus et al., 2018), and an omit map was created with REFMAC and FFT (Murshudov et al., 1999) in CCP4. For the ORF39 protein, ConSurf (Ashkenazy et al., 2016) was used for amino acid conservation analysis and mapping. All visualization of the models and figure preparation was performed with PyMOL.

2.8. Electrophoretic mobility shift assay

The DNA-binding properties of the ORF6 protein were studied in a series of agarose gel electrophoresis experiments. Both full-length and truncated protein variants were tested with either ssDNA or dsDNA fragments derived from the *Enc34* genome. To set up the reactions, aliquots of DNA were mixed with varying concentrations of the ORF6 protein (molar ratios ranging from 1:156 to 1:2500) and incubated at room temperature for 20 min. Then, the mixtures were loaded onto a 1% native agarose gel containing ethidium bromide in Tris-acetate-EDTA (TAE) buffer and run for 20 min at 10 V/cm, after which the DNA was visualized under ultraviolet light.

2.9. Turbidity reduction assay

The peptidoglycan-degrading activity of the ORF39 protein was determined spectrophotometrically by assaying the protein against bacterial cell substrate with externally exposed peptidoglycan layer. To produce such substrate, Gram-negative *Escherichia coli* W3100, *Pseudomonas aeruginosa* PAO1 and Enc34-sensitive *Hafnia alvei* (all from laboratory collection) were treated with chloroform-saturated 50 mM Tris-HCl (pH 7.7), washed and suspended in phosphate buffered saline (PBS) at an OD600 of 0.6-1.0. The reactions were set up in a 96-well plate by adding varying amounts of the ORF39 protein to the membrane-permeabilized cells, and then absorbance at 600 nm was measured on a microplate reader (BioTek) for 3 h at 3 min intervals. The enzymatic activity was then calculated with the ActivityCalculator from <https://www.biw.kuleuven.be/logt/ActivityCalculator.htm> (Briers et al., 2007). The ORF39 protein was also tested on unpermeabilized Gram-negative cells and Gram-positive *Microbacterium paraoxydans* (from laboratory collection) using only the highest concentration of the enzyme. All reactions were performed in triplicate, with PBS used as negative controls and chicken egg white lysozyme (Biochemica) as positive controls.

3. RESULTS AND DISCUSSION

My doctoral research is presented in three published manuscripts: (I) details the methods developed for the investigation of the Enc34 ORF6 protein, which were later applied to other target proteins from the Enc34 proteome, including ORF39, (II) focuses on the structural biology of the Enc34 ORF6 protein, and (III) unveils the novel lysozyme fold variation of the Enc34 ORF39 protein. This section is structured to present the main findings of the three publications (subsections 3.1., 3.2. and 3.3., respectively) and to briefly discuss the future of this project (subsection 3.4.).

3.1. Structural characterization of a single-stranded DNA-binding protein: a case study of the ORF6 protein from phage Enc34

In the quest to understand how single-stranded DNA-binding proteins (SSBs) function at a molecular level, determination of their three-dimensional structure, especially in complex with DNA, is indispensable. This publication is a walk-through of all the steps that were taken to structurally characterize the SSB of the bacteriophage Enc34, encoded by ORF6 in its genome. We provide details on all major procedures involved in the study, from designing expression constructs through to protein production, purification, and crystallization, to determination and analysis of the protein's three-dimensional structure. We further share tips and tricks that proved successful in the study of bacteriophage Enc34 ORF6 SSB, e.g., an optimized protocol for recombinant protein production, illustrated below.

When producing recombinant proteins in *E. coli*, a not too uncommon issue is a complete or partial insolubility of the target protein, which was the observation also for the Enc34 ORF6 protein. However, during our efforts to produce a wide variety of different recombinant proteins over the years, we have developed an alternative protocol in our laboratory which uses lower cultivation temperature (22 °C) and inducer concentration (0.01 mM IPTG) and has a high success rate in producing soluble protein while maintaining comparable or better protein yields than the standard protocol (Figure 2A).

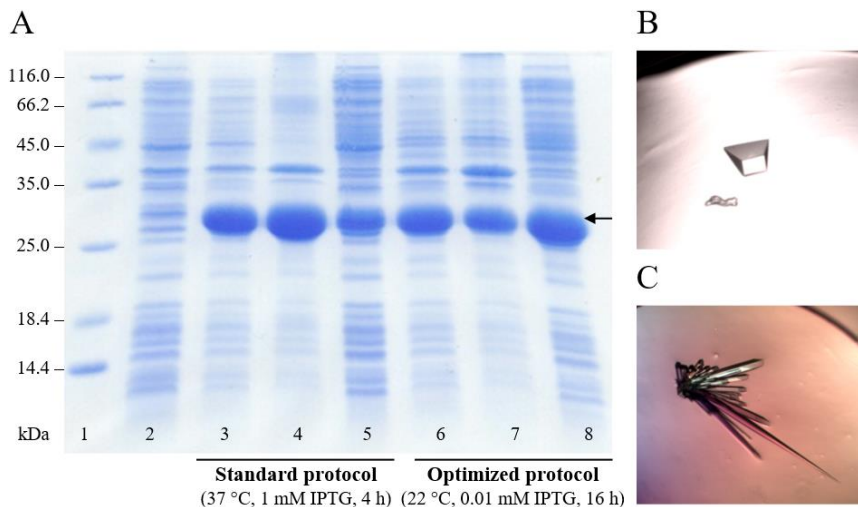


Figure 2. Production and crystallization of the Enc34 ORF6 protein. SDS-PAGE analysis of production and solubility of the ORF6 protein using standard or optimized production protocols (A). Lane 1, molecular weight marker; lane 2, lysate of uninduced cells; lanes 3 and 6, total cellular protein at the end of protein production; lanes 4 and 7, insoluble protein fraction; lanes 5 and 8, soluble protein fraction. The arrow points to the ORF6 protein. Crystals of the full-length ORF6 (B) and ORF6 Δ C-oligothymidine complex (C).

The native full-length Enc34 ORF6 protein was produced according to the optimized protocol and purified by immobilized metal ion affinity chromatography and anion exchange chromatography. The protein readily crystallized under several conditions (Figure 2B), so we proceeded to produce a selenomethionine-substituted variant of the full-length ORF6 (SeMet-ORF6) which is required to provide experimental phasing data for structure solution. After optimization of purification and crystallization protocols for SeMet-ORF6, the obtained crystals were used for X-ray data collection and structure determination. Finally, a C-terminally truncated variant of ORF6 (ORF6 Δ C) was constructed with the rationale that the unstructured tail might both impede crystallization due to its flexibility and compete for the same binding surface as the single-stranded DNA (Marintcheva et al., 2008). Finally, a structure of the ORF6 Δ C in complex with DNA was obtained, using X-ray data from a protein-oligothymidine co-crystal (Figure 2C).

3.2. Structural basis for DNA recognition of a single-stranded DNA-binding protein from *Enterobacter* phage Enc34

The structure of the Enc34 ORF6 protein was solved to 1.50 Å, which apart from the disordered C-terminus adopts a compact, roughly globular shape with a diameter of approximately 40 Å and has a mixed α/β architecture comprised predominantly of β -strands. The protein can be regarded as built up from a concave seven-stranded β -sheet at one side of the protein, of which the three central strands together with the two β -strands from the opposite side form a slightly flattened five-stranded β -barrel at the protein core, and two α -helices cap the β -barrel at both ends (Figure 3). The core architecture of the protein clearly resembles the oligonucleotide/oligosaccharide-binding (OB) fold, which is commonly defined as a five-stranded β -barrel with a capping α -helix (Theobald et al., 2003).

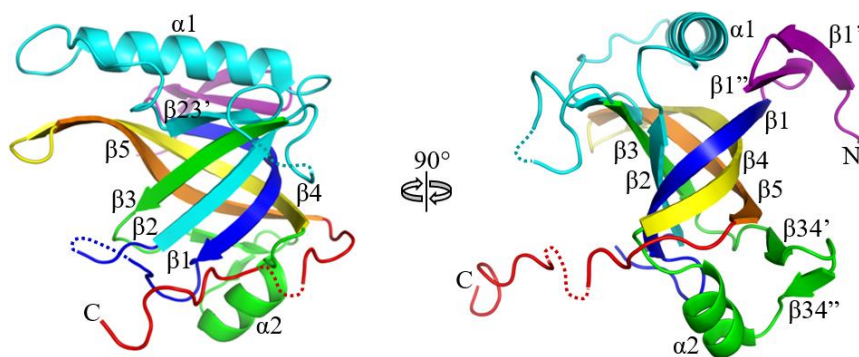


Figure 3. Overall structure of the phage Enc34 ORF6 protein. The protein chain is rainbow-colored purple to red from the N to the C-terminus. The conserved strands of the core β -barrel are named in accordance to Murzin, 1993.

While at the sequence level Enc34 ORF6 shows conservation with a number of recognizably homologous phage proteins from the domain of unknown function (DUF) 2815 family, its three-dimensional structure is similar to a variety of different OB fold proteins. Three structures currently in the Protein Data Bank (PDB) clearly form a high-similarity group with Enc34 ORF6, which includes two uncharacterized phage-related proteins from *Bacillus cereus* (PDB ID: 4JG2) and *Enterococcus*

faecalis (PDB ID: 4KLLK), and the bacteriophage T7 gene 2.5 protein (PDB ID: 1JE5), an extensively studied SSB (Hernandez & Richardson, 2019). To confirm that the Enc34 ORF6 indeed functions as an SSB, the protein was assayed in agarose gel electrophoresis in presence of different types of DNA. A clear mobility shift of the DNA was observed in presence of ssDNA, but not with dsDNA. Somewhat surprisingly, the C-terminally truncated protein constructed for co-crystallization trials, appeared to bind ssDNA and dsDNA with a similar affinity.

To further investigate the DNA-binding mechanism of the Enc34 ORF6, we determined the 3D structure of the ORF6 Δ C in complex with ssDNA. The ORF6 Δ C-(dT)₃₅ crystals diffracted to 1.56 Å and contained a single chain of ORF6 Δ C in the asymmetric unit bound to a stretch of nine thymidine nucleotides (Figure 4). The complex structure reveals the DNA strand stretching across the entire ligand-binding surface of the OB fold domain, which is largely positively charged and enclosed by loops L12, L23'3 and L45 that connect strands β 1 and β 2, β 23' and β 3, and β 4 and β 5, respectively. The DNA-binding mode of the protein involves two aromatic stacks and multiple electrostatic interactions, that are conserved among the T7-type SSBs.

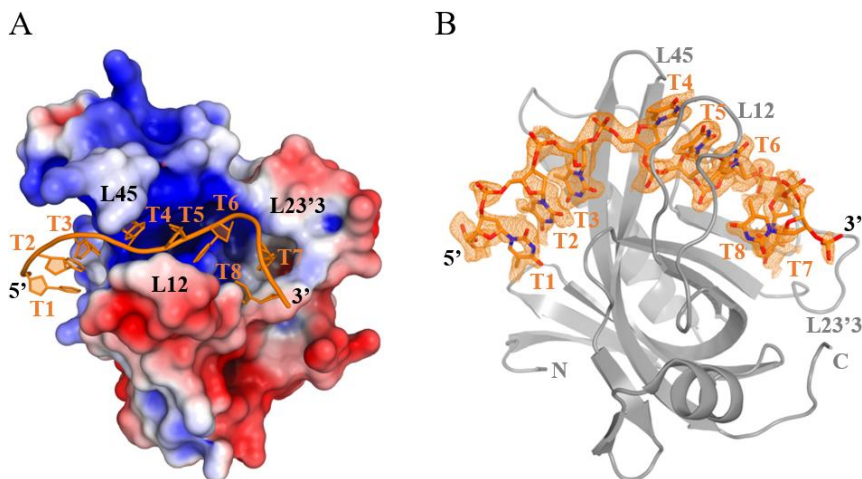


Figure 4. Structure of the ORF6 Δ C-(dT)₃₅ complex. The DNA oligonucleotide (orange) binds to the positively charged cleft along the core OB fold (A). A different orientation of the complex with an overlay of a 2Fo-Fc omit map of the DNA contoured at 0.7 σ (B). The thymidines are numbered T1-T8 in the 5' to 3' direction and the cleft-defining loops are indicated in both panels.

3.3. Diversity of the lysozyme fold: structure of the catalytic domain from an unusual endolysin encoded by phage Enc34

To enable functional and structural studies of the putative Enc34 endolysin, the coding sequence of the ORF39 protein with a cleavable N-terminal hexahistidine-tag was cloned and expressed in *E. coli*. A plasmid encoding the full-length protein showed low transformation efficiency, reduced cell growth and no detectable production of the target protein upon induction, indicative of toxicity of the ORF39 product to the bacterial cells. Further experimentation revealed that a truncated variant of the ORF39 protein containing only the N-terminal PHA02564 domain can be readily produced in a soluble form and without adverse effects to the cells. This C-terminal deletion variant of the protein, ORF39 Δ C, was therefore used for all subsequent studies.

To test for the enzymatic activity of the Enc34 endolysin, purified ORF39 Δ C protein was subjected to turbidity reduction assays using outer membrane (OM)-permeabilized bacteria as a substrate. The Enc34 endolysin showed clear muralytic activity towards OM-permeabilized *Escherichia coli* W3100, *Pseudomonas aeruginosa* PAO1 and *Hafnia alvei* cells, demonstrating that the N-terminal domain of the Enc34 endolysin, and hence the PHA02564 family proteins in other bacteriophage genomes, indeed encode an enzymatically active domain (EAD) of a functional endolysin.

The structure of the ORF39 EAD was determined using the single-wavelength anomalous diffraction method with SeMet-substituted ORF39 Δ C crystals diffracting to 1.8 Å. In later experiments another structure was solved to 1.6 Å and, thus, was selected for reporting here. The overall structure of the ORF39 EAD is all-helical and consists of six α helices (α 1 to α 6) and a single one-turn 3_{10} helix in-between α 5 and α 6 (Figure 5). The structure can be regarded as two overlapping helical bundles, the first of which is formed by α 2, α 3, α 4 and α 5 and the second consisting of α 2, α 5 and α 6. The EAD is roughly globular in shape with a large groove between loops connecting α 1– α 2 and α 3– α 4 and centered on the C-terminal end of helix α 3.

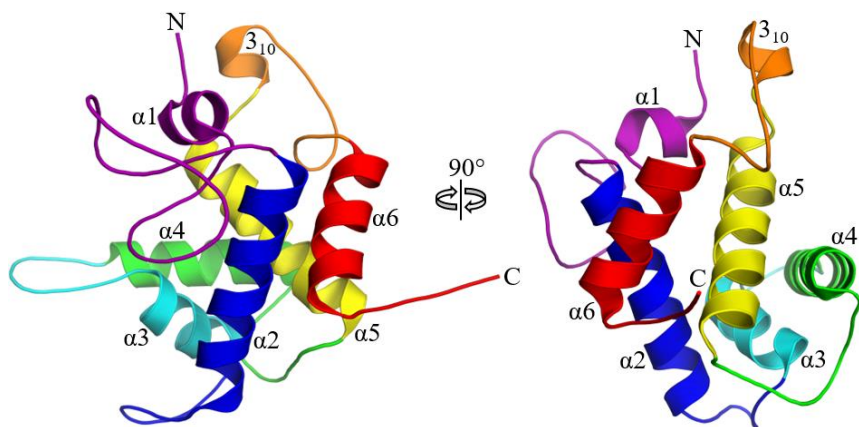


Figure 5. Overall structure of the Enc34 endolysin's enzymatically active domain. The protein chain is rainbow-colored purple to red from the N to the C-terminus.

Structural homology analysis with Dali uncovered similarities with a range of α -helical proteins; however, none of the top hits had any apparent relatedness to endolysins. Only three proteins, namely, *N*-acetylglucosaminidase from *Thermotoga maritima* (PDB ID: 4QDN; Z score: 3.0), lytic transglycosylase gp144 from bacteriophage ϕ KZ (PDB ID: 3BKH; Z score: 2.6), and the *E. coli* lytic transglycosylase MltC (PDB ID: 4C5F; Z score: 2.5) did indicate potential homology, but the sequence similarity of Enc34 ORF39 to any of these proteins was insufficient for a sequence alignment. Structurally, however, the homologous proteins all are variations of the α/β lysozyme fold which canonically consists of α and β -structured lobes arranged to form a deep cleft for substrate binding and cleavage. The EAD of the Enc34 endolysin bears resemblance to the α -lobe of the lysozyme fold but lacks any β -structured elements. It can be noted, however, that the large loop connecting $\alpha 1$ and $\alpha 2$ in the ORF39 protein is located at essentially the same position as the β -lobe in TM0633 and could likewise function as a lid over the substrate-binding groove.

The Glu131 in the ORF39 EAD corresponds to the catalytic glutamate residues in the superimposed TM0633, gp144 and MltC proteins which suggests that the groove is indeed the active site of the Enc34 endolysin. Furthermore, residues corresponding to Trp79 and Tyr95 are also conserved; these have been shown to be important for catalysis in the structural homologs (Lipski et al., 2015; Chertkov et al., 2017). Helices corresponding to $\alpha 2$, $\alpha 3$ and $\alpha 4$ in the Enc34 endolysin are

found in that particular order also in the other structural homologs but while the remaining helix $\alpha 5$ in the Enc34 endolysin directly follows $\alpha 4$, in the TM0633, gp144 and MltC proteins the corresponding helix is located before their $\alpha 2$ counterpart (Figure 6). Notably, the $\alpha 5$ holds the catalytic Glu131 of the Enc34 endolysin but despite the permuted core, the respective glutamate residue in the superimposed TM0633, gp144 and MltC structures is located at essentially the same position as in the ORF39 protein. Overall, these conserved features indicate a shared evolutionary history between the Enc34 endolysin and TM0633, gp144 and MltC proteins and highlight a molecular signature for peptidoglycan cleavage for this superclade of enzymes, bringing together glycoside hydrolase (GH) 73 and GH23 family proteins.

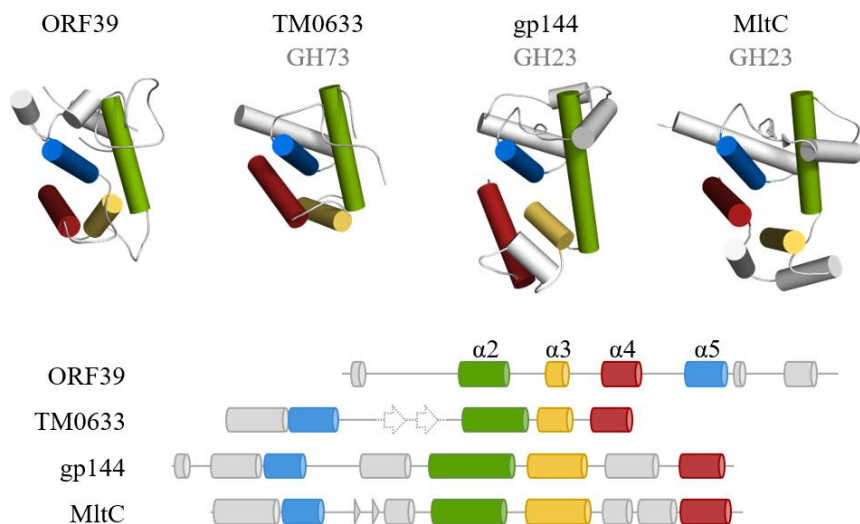


Figure 6. Similarities in the core fold of the Enc34 endolysin and GH73 and GH23 glycoside hydrolase enzymes. The four helices constituting the protein core are shown in different colors. Secondary structure diagrams of the proteins are presented below using the same color scheme. Additional domains of the gp144 and MltC proteins have been removed for clarity.

A notable shortcoming of this study is the lack of an endolysin-ligand complex structure that would allow to delve deeper into the molecular basis for peptidoglycan binding and cleavage of the divergent Enc34 enzyme. Considerable efforts of crystal soaking and co-crystallization, using commercially available N-acetylglucosamine

(GlcNAc), N-acetylmuramic acid (MurNAc), chitotetraose and GlcNAc-MurNAc disaccharide were undertaken, but sadly to no avail. Although endolysin-ligand complexes have been obtained using wild type proteins (Pérez-Dorado et al., 2007; Fokine et al., 2008), future efforts of crystal soaking and co-crystallization with an inactive variant of the Enc34 enzyme may provide a better chance for capturing the protein-ligand interaction. It is also worth noting that the structural homolog TM0633 was shown to be enzymatically inactive on peptide-free glycan strands and extensive crystallization trials with linear peptidoglycan analogs likewise failed to produce any ligand-bound protein structures (Lipski et al., 2015). Therefore, continued crystallography experiments using branched peptidoglycan analogs are also of considerable interest to us.

3.4. AlphaFold2 structures of the Enc34 ORF3, ORF4 and ORF5 proteins and prospects for future work

The fourteenth iteration of the Critical Assessment of protein Structure Prediction world-wide competition in 2020, marks a major development in the landscape of protein structural biology. The deep learning-based artificial intelligence program AlphaFold2 (AF2) developed by DeepMind (Jumper et al., 2021), emerged as the victor of this long-standing challenge, having reached the capacity to predict the folding of single protein chains with near-experimental accuracy (Pereira et al., 2021). Subsequent assessments by the wider scientific community have validated the reliability of AF2 models, and are in agreement about the transformative impact this technology will have on structural biology and life sciences as a whole (Cramer, 2021; Akdel et al., 2022). Back in the “pre-AF2” era, however, the novel three-dimensional structures of both ORF6 and ORF39 proteins were solved experimentally, as detailed in Publication I with the ORF6 SSB as a case study. While the long and laborious journey of *de novo* structure determination was successful for structural characterization of the phage Enc34 ORF6 and ORF39 products, the structures of ORF3, ORF4 and ORF5 hypothetical proteins that were also selected as targets, never materialized, since their crystals could not be obtained. Yet, AF2 can predict all three of the structures in minutes (Figure 7) showcasing the incredible superhuman capabilities of neural network-based computational approaches.

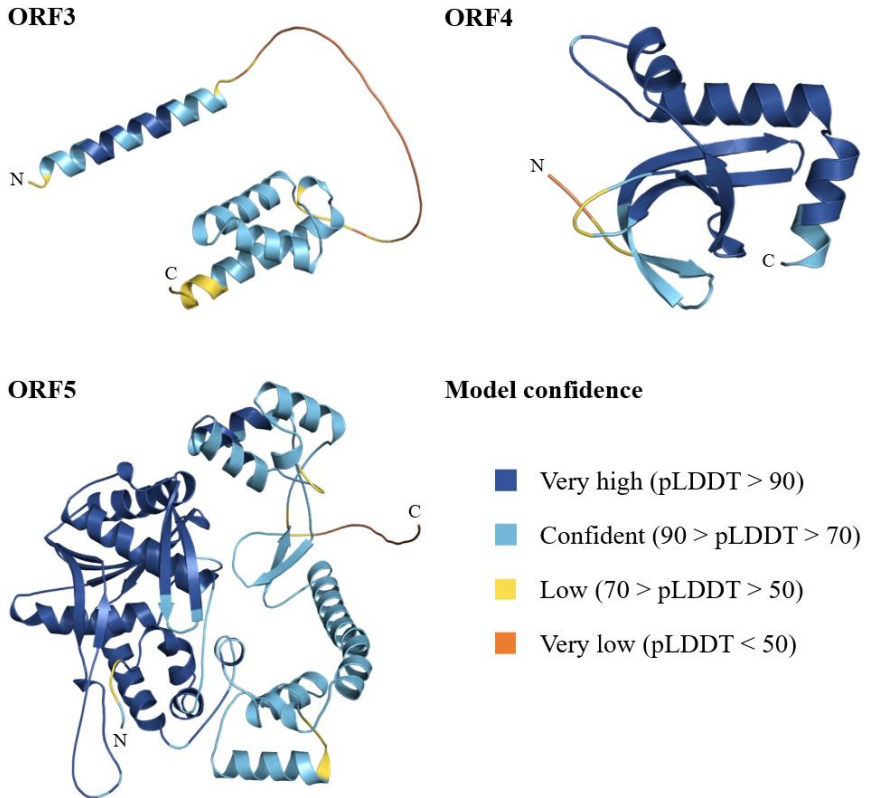


Figure 7. AlphaFold2 structures of the phage Enc34 ORF3, ORF4 and ORF5 proteins. The top-ranked model of each protein is presented, with residues colored according to their predicted local-distance difference test (pLDDT) scores that correlate to structural prediction confidence (unpublished data).

To illustrate how the present-day breakthroughs in artificial intelligence and machine learning can accelerate the study of the proteomic dark matter, I will expand on the case of the Enc34 ORF5 protein as an example. The protein contains a DUF2800, which in the Conserved Domain Database is currently annotated as CRISPR/Cas system-associated protein Cas4. Examination of the predicted structure with Dali reveals strong homology to various proteins with exonuclease activity. At the time of writing, the top Dali match is the mycobacterial heterodimeric helicase-nuclease AdnAB (PDB ID: 6PPR; Z-score: 13.2) which is responsible for processing DNA double-strand breaks (DSBs) to

enable repair by homologous recombination (Jia et al., 2019). The overall structure of the Enc34 ORF5 protein can be seen as composed of two lobes with a deep cleft between them. The larger N-terminal lobe is well-ordered and notably conserved, corresponding to the nuclease module of the AdnA subunit and other structurally characterized exonuclease domains. In contrast, the C-terminal lobe is a unique to Enc34 ORF5 and related phage proteins and, interestingly, it adopts several alternative conformations in the predicted AF2 models representing the structure in various states of openness, some of which could be biologically relevant. The more open ORF5 protein conformations accommodate the forked DNA moiety of the AdnAB complex structure with only minor steric hindrance, suggesting that the protein could likewise act upon Y-shaped DNA substrate. Many of the identified ORF5 protein homologs work in concert with helicases that perform DNA unwinding and complex translocation, and, thus, it could be investigated whether the ORF5 protein forms complexes with phage or host helicase domains. Notably, Enc34 is thought to have an accessory DNA helicase encoded by ORF9, the functional role of which has not yet been studied, in addition to its replicative helicase which is embodied in the bifunctional primase-helicase of ORF1. If identified, such multiprotein complexes would be interesting objects for cryogenic electron microscopy studies, especially given their similarity to DSB repair systems and even CRISPR-associated proteins. Indeed, significant similarities between Enc34 ORF5 and Cas4 proteins can be detected; however, in the absence of further evidence, the potential involvement of ORF5 in prespacer processing (Lee et al., 2019) or any other CRISPR-related process, remains unknown, and is yet another compelling question for future studies.

CONCLUSIONS

1. Bacteriophage Enc34 ORF6 product and its truncated and selenomethionine-substituted variants could be produced in *E. coli*. Lowered culturing temperature and inducer concentration increased solubility of the recombinant proteins.
2. Enc34 ORF6 protein and its variants could be purified by a combination of chromatography methods and crystallized using commercial screens. The C-terminally truncated ORF6 protein also crystallized in the presence of a DNA oligonucleotide.
3. Structurally, Enc34 ORF6 protein is composed of an N-terminal oligonucleotide/oligosaccharide-binding fold domain and a C-terminal disordered segment, and it resembles the phage T7 single-stranded DNA-binding protein gp2.5.
4. Enc34 ORF6 protein binds single-stranded DNA *in vitro* and its C-terminus contributes to discrimination between single-stranded and double-stranded DNA.
5. Bacteriophage Enc34 ORF39 product is toxic to *E. coli* cells; however, a C-terminally truncated protein and its selenomethionine-substituted variant could be produced, purified and crystallized, using similar laboratory methods as for the ORF6 protein.
6. The N-terminal domain of the Enc34 ORF39 protein readily digests Gram-negative peptidoglycan *in vitro* with activity comparable to that of commercial chicken egg white lysozyme.
7. Structurally, the N-terminal domain of Enc34 ORF39 is composed of five α -helices organized in two overlapping helical bundles, and it distantly resembles the α -structured lobe of the lysozyme fold.
8. Enc34 ORF39 protein is related to glycoside hydrolase GH73 and GH23 enzymes, which include phage endolysins and bacterial autolysins, all sharing a highly conserved three amino acid scaffold for substrate cleavage.
9. Proteins containing the domain of unknown function DUF2815 are single-stranded DNA-binding proteins, and members of the phage protein family PHA02564 are Gram-negative peptidoglycan-degrading endolysins.

ACKNOWLEDGEMENTS

SIA “Mikrotīkls” scholarship, together with the project “Strengthening of the Capacity of Doctoral Studies at the University of Latvia within the Framework of the New Doctoral Model” from the European Social Fund (grant no. 8.2.2.0/20/I/006) provided vital financial support for the most of my PhD journey. I am honored and grateful to have been a recipient of these grants, without which this PhD thesis would likely never come to exist.

I am thankful to my supervisor Dr. biol. Andris Kazāks for introducing me to the wondrous field of phage biology and for personally guiding me through all stages of a research project, with patience, goodwill and generosity of time and attention. It has been a joy and a privilege to work side-by-side in our lab, to learn from you and to engage in all the discussions and humor we have had throughout the years. I also want to thank prof., Dr. biol. Kaspars Tārs for welcoming me into the BMC’s Structural Biology lab, for inviting me along to synchrotron trips and conferences, and for nudging me to apply for grants, when I was starting to doubt myself. Another huge thank you is dedicated to my colleague Dr. biol. Jānis Rūmnieks for essentially becoming my mentor in all-things structural biology and data presentation. From structure determination, to data analysis, to academic writing, you have help me immensely and taught me so much, for which I am beyond grateful.

I would like to extend my thanks to all the Structural Biology lab members and colleagues at the BMC, especially Ināra, Ņikita, Anna, Eva and Mihails, with whom I have had the pleasure to work with in research, teaching and other projects. I am thankful to Ilze and everybody at the BMC’s Student Council for coming together and organizing all the cool science communication events, in which we brought science to large audiences and got inspired in the process. Thank you to my friends and peers Elza, Aivars, Kārlis, Karīna and Vladimirs for being together with me on this journey and providing continuous support and comic relief. Last but not least, a heartfelt thanks to Rūta for being just the absolute best friend there ever was, and to my amazing mom, for being such an icon and inspiration herself, and for always being proud of me no matter what I decided to do and achieve.

REFERENCES

- Adeolu, M., Alnajjar, S., Naushad, S., & S Gupta, R. (2016). Genome-based phylogeny and taxonomy of the '*Enterobacteriales*': proposal for *Enterobacterales* ord. nov. divided into the families *Enterobacteriaceae*, *Erwiniaceae* fam. nov., *Pectobacteriaceae* fam. nov., *Yersiniaceae* fam. nov., *Hafniaceae* fam. nov., *Morganellaceae* fam. nov., and *Budviciaceae* fam. nov. *International journal of systematic and evolutionary microbiology*, 66(12), 5575–5599. <https://doi.org/10.1099/ijsem.0.001485>
- Aevansson, A., Kaczorowska, A. K., Adalsteinsson, B. T., Ahlqvist, J., Al-Karadaghi, S., Altenbuchner, J., Arsin, H., Átlason, Ú. Á., Brandt, D., Cichowicz-Cieślak, M., Cornish, K., Courtin, J., Dabrowski, S., Dahle, H., Djefane, S., Dorawa, S., Dusaucy, J., Enault, F., Fedøy, A. E., Freitag-Pohl, S., ... Zhang, R. (2021). Going to extremes – a metagenomic journey into the dark matter of life. *FEMS microbiology letters*, 368(12), fnab067. <https://doi.org/10.1093/femsle/fnab067>
- Ahern, S. J., Das, M., Bhowmick, T. S., Young, R., & Gonzalez, C. F. (2014). Characterization of novel virulent broad-host-range phages of *Xylella fastidiosa* and *Xanthomonas*. *Journal of bacteriology*, 196(2), 459–471. <https://doi.org/10.1128/JB.01080-13>
- Akdel, M., Pires, D. E. V., Pardo, E. P., Jänes, J., Zalevsky, A. O., Mészáros, B., Bryant, P., Good, L. L., Laskowski, R. A., Pozzati, G., Shenoy, A., Zhu, W., Kundrotas, P., Serra, V. R., Rodrigues, C. H. M., Dunham, A. S., Burke, D., Borkakoti, N., Velankar, S., Frost, A., ... Beltrao, P. (2022). A structural biology community assessment of AlphaFold2 applications. *Nature structural & molecular biology*, 29(11), 1056–1067. <https://doi.org/10.1038/s41594-022-00849-w>
- Ashkenazy, H., Abadi, S., Martz, E., Chay, O., Mayrose, I., Pupko, T., & Ben-Tal, N. (2016). ConSurf 2016: an improved methodology to estimate and visualize evolutionary conservation in macromolecules. *Nucleic acids research*, 44(W1), W344–W350. <https://doi.org/10.1093/nar/gkw408>
- Battye, T. G., Kontogiannis, L., Johnson, O., Powell, H. R., & Leslie, A. G. (2011). iMOSFLM: a new graphical interface for diffraction-image processing with MOSFLM. *Acta crystallographica. Section D, Biological crystallography*, 67(Pt 4), 271–281. <https://doi.org/10.1107/S0907444910048675>
- Berg, M., & Roux, S. (2021). Extreme dimensions – how big (or small) can tailed phages be? *Nature reviews. Microbiology*, 19(7), 407. <https://doi.org/10.1038/s41579-021-00574-z>

- Boratyn, G. M., Camacho, C., Cooper, P. S., Coulouris, G., Fong, A., Ma, N., Madden, T. L., Matten, W. T., McGinnis, S. D., Merezuk, Y., Raytselis, Y., Sayers, E. W., Tao, T., Ye, J., & Zaretskaya, I. (2013). BLAST: a more efficient report with usability improvements. *Nucleic acids research*, 41(Web Server issue), W29–W33. <https://doi.org/10.1093/nar/gkt282>
- Briers, Y., Lavigne, R., Volckaert, G., & Hertveldt, K. (2007). A standardized approach for accurate quantification of murein hydrolase activity in high-throughput assays. *Journal of biochemical and biophysical methods*, 70(3), 531–533. <https://doi.org/10.1016/j.jbbm.2006.10.009>
- Brüssow, H., & Hendrix, R. W. (2002). Phage genomics: small is beautiful. *Cell*, 108(1), 13–16. [https://doi.org/10.1016/s0092-8674\(01\)00637-7](https://doi.org/10.1016/s0092-8674(01)00637-7)
- Casañal, A., Lohkamp, B., & Emsley, P. (2020). Current developments in Coot for macromolecular model building of electron cryo-microscopy and crystallographic data. *Protein science: a publication of the Protein Society*, 29(4), 1069–1078. <https://doi.org/10.1002/pro.3791>
- Casjens, S. R., & Gilcrease, E. B. (2009). Determining DNA packaging strategy by analysis of the termini of the chromosomes in tailed-bacteriophage virions. *Methods in molecular biology (Clifton, N.J.)*, 502, 91–111. https://doi.org/10.1007/978-1-60327-565-1_7
- CDC. (2019). Antibiotic resistance threats in the United States, 2019. US Department of Health and Human Services, Centres for Disease Control and Prevention. <https://www.cdc.gov/drugresistance/Biggest-Threats.html>
- Chertkov, O. V., Armeev, G. A., Uporov, I. V., Legotsky, S. A., Sykilinda, N. N., Shaytan, A. K., Klyachko, N. L., & Miroshnikov, K. A. (2017). Dual active site in the endolytic transglycosylase gp144 of pacteriophage phiKZ. *Acta naturae*, 9(1), 81–87.
- Chevallereau, A., Pons, B. J., van Houte, S., & Westra, E. R. (2022). Interactions between bacterial and phage communities in natural environments. *Nature reviews. Microbiology*, 20(1), 49–62. <https://doi.org/10.1038/s41579-021-00602-y>
- Cobián Güemes, A. G., Youle, M., Cantú, V. A., Felts, B., Nulton, J., & Rohwer, F. (2016). Viruses as winners in the game of life. *Annual review of virology*, 3(1), 197–214. <https://doi.org/10.1146/annurev-virology-100114-054952>
- Comeau, A. M., Hatfull, G. F., Krisch, H. M., Lindell, D., Mann, N. H., & Prangishvili, D. (2008). Exploring the prokaryotic virosphere. *Research in microbiology*, 159(5), 306–313. <https://doi.org/10.1016/j.resmic.2008.05.001>

- Cook, R., Brown, N., Redgwell, T., Rihtman, B., Barnes, M., Clokie, M., Stekel, D. J., Hobman, J., Jones, M. A., & Millard, A. (2021). Infrastructure for a PHAge REference Database: identification of large-scale biases in the current collection of cultured phage genomes. *PHAGE* (New Rochelle, N.Y.), 2(4), 214–223. <https://doi.org/10.1089/phage.2021.0007>
- Cowtan K. (2006). The Buccaneer software for automated model building. 1. Tracing protein chains. *Acta crystallographica. Section D, Biological crystallography*, 62(Pt 9), 1002–1011. <https://doi.org/10.1107/S0907444906022116>
- Cowtan K. (2010). Recent developments in classical density modification. *Acta crystallographica. Section D, Biological crystallography*, 66(Pt 4), 470–478. <https://doi.org/10.1107/S090744490903947X>
- Cramer P. (2021). AlphaFold2 and the future of structural biology. *Nature structural & molecular biology*, 28(9), 704–705. <https://doi.org/10.1038/s41594-021-00650-1>
- Dion, M. B., Oechslin, F., & Moineau, S. (2020). Phage diversity, genomics and phylogeny. *Nature reviews. Microbiology*, 18(3), 125–138. <https://doi.org/10.1038/s41579-019-0311-5>
- ECDC & WHO (2022). Antimicrobial resistance surveillance in Europe 2022 – 2020 data. World Health Organization. Regional Office for Europe. <https://apps.who.int/iris/handle/10665/351141>
- Evans P. (2006). Scaling and assessment of data quality. *Acta crystallographica. Section D, Biological crystallography*, 62(Pt 1), 72–82. <https://doi.org/10.1107/S0907444905036693>
- Fokine, A., Miroshnikov, K. A., Shneider, M. M., Mesyanzhinov, V. V., & Rossmann, M. G. (2008). Structure of the bacteriophage phiKZ lytic transglycosylase gp144. *The Journal of biological chemistry*, 283(11), 7242–7250. <https://doi.org/10.1074/jbc.M709398200>
- Gabler, F., Nam, S. Z., Till, S., Mirdita, M., Steinegger, M., Söding, J., Lupas, A. N., & Alva, V. (2020). Protein sequence analysis using the MPI Bioinformatics Toolkit. *Current protocols in bioinformatics*, 72(1), e108. <https://doi.org/10.1002/cpbi.108>
- Galperin, M. Y., & Koonin, E. V. (2004). 'Conserved hypothetical' proteins: prioritization of targets for experimental study. *Nucleic acids research*, 32(18), 5452–5463. <https://doi.org/10.1093/nar/gkh885>
- Gregory, A. C., Zayed, A. A., Conceição-Neto, N., Temperton, B., Bolduc, B., Alberti, A., Ardyna, M., Arkhipova, K., Carmichael, M., Cruaud, C., Dimier, C., Domínguez-Huerta, G., Ferland, J., Kandels, S., Liu, Y., Marec, C., Pesant, S., Picheral, M., Pisarev, S., Poulain, J., ... Sullivan, M. B. (2019). Marine DNA viral macro- and microdiversity from pole to

- pole. Cell, 177(5), 1109–1123.e14.
<https://doi.org/10.1016/j.cell.2019.03.040>
- Hallgren, J., Tsigos, K. D., Pedersen, M. D., Almagro Armenteros, J. J., Marcatili, P., Nielsen, H., Krogh, A., & Winther, O. (2022). DeepTMHMM predicts alpha and beta transmembrane proteins using deep neural networks. <https://doi.org/10.1101/2022.04.08.487609>
- Hampton, H. G., Watson, B., & Fineran, P. C. (2020). The arms race between bacteria and their phage foes. *Nature*, 577(7790), 327–336.
<https://doi.org/10.1038/s41586-019-1894-8>
- Hatfull G. F. (2008). Bacteriophage genomics. *Current opinion in microbiology*, 11(5), 447–453. <https://doi.org/10.1016/j.mib.2008.09.004>
- Hatfull G. F. (2015). Dark matter of the biosphere: the amazing world of bacteriophage diversity. *Journal of virology*, 89(16), 8107–8110.
<https://doi.org/10.1128/JVI.01340-15>
- Hernandez, A. J., & Richardson, C. C. (2019). Gp2.5, the multifunctional bacteriophage T7 single-stranded DNA binding protein. *Seminars in cell & developmental biology*, 86, 92–101.
<https://doi.org/10.1016/j.semcd.2018.03.018>
- Holm L. (2022). Dali server: structural unification of protein families. *Nucleic acids research*, 50(W1), W210–W215. Advance online publication.
<https://doi.org/10.1093/nar/gkac387>
- Ionescu, M. I., Neagoe, D. Ș., Crăciun, A. M., & Moldovan, O. T. (2022). The Gram-negative bacilli isolated from caves – *Sphingomonas paucimobilis* and *Hafnia alvei* and a review of their involvement in human infections. *International journal of environmental research and public health*, 19(4), 2324. <https://doi.org/10.3390/ijerph19042324>
- Jia, N., Unciuleac, M. C., Xue, C., Greene, E. C., Patel, D. J., & Shuman, S. (2019). Structures and single-molecule analysis of bacterial motor nuclease AdnAB illuminate the mechanism of DNA double-strand break resection. *Proceedings of the National Academy of Sciences of the United States of America*, 116(49), 24507–24516.
<https://doi.org/10.1073/pnas.1913546116>
- Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Židek, A., Potapenko, A., Bridgland, A., Meyer, C., Kohl, S. A. A., Ballard, A. J., Cowie, A., Romera-Paredes, B., Nikolov, S., Jain, R., Adler, J., Back, T., ... Hassabis, D. (2021). Highly accurate protein structure prediction with AlphaFold. *Nature*, 596(7873), 583–589. <https://doi.org/10.1038/s41586-021-03819-2>
- Jurrus, E., Engel, D., Star, K., Monson, K., Brandi, J., Felberg, L. E., Brookes, D. H., Wilson, L., Chen, J., Liles, K., Chun, M., Li, P., Gohara, D. W., Dolinsky, T., Konecny, R., Koes, D. R., Nielsen, J. E., Head-Gordon, T.,

- Geng, W., Krasny, R., ... Baker, N. A. (2018). Improvements to the APBS biomolecular solvation software suite. *Protein science: a publication of the Protein Society*, 27(1), 112–128. <https://doi.org/10.1002/pro.3280>
- Kabsch W. (2010a). XDS. *Acta crystallographica. Section D, Biological crystallography*, 66(Pt 2), 125–132. <https://doi.org/10.1107/S0907444909047337>
- Kabsch W. (2010b). Integration, scaling, space-group assignment and post-refinement. *Acta crystallographica. Section D, Biological crystallography*, 66(Pt 2), 133–144. <https://doi.org/10.1107/S0907444909047374>
- Kazāks, A., Dišlers, A., Lipowsky, G., Nikolajeva, V., & Tārs, K. (2012). Complete genome sequence of the *Enterobacter cancerogenus* bacteriophage Enc34. *Journal of virology*, 86(20), 11403–11404. <https://doi.org/10.1128/JVI.01954-12>
- Kovalevskiy, O., Nicholls, R. A., Long, F., Carlon, A., & Murshudov, G. N. (2018). Overview of refinement procedures within REFMAC5: utilizing data from different sources. *Acta crystallographica. Section D, Structural biology*, 74(Pt 3), 215–227. <https://doi.org/10.1107/S2059798318000979>
- Krissinel, E., & Henrick, K. (2007). Inference of macromolecular assemblies from crystalline state. *Journal of molecular biology*, 372(3), 774–797. <https://doi.org/10.1016/j.jmb.2007.05.022>
- Krogh, A., Larsson, B., von Heijne, G., & Sonnhammer, E. L. (2001). Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. *Journal of molecular biology*, 305(3), 567–580. <https://doi.org/10.1006/jmbi.2000.4315>
- Lee, H., Dhingra, Y., & Sashital, D. G. (2019). The Cas4-Cas1-Cas2 complex mediates precise prespacer processing during CRISPR adaptation. *eLife*, 8, e44248. <https://doi.org/10.7554/eLife.44248>
- Liebschner, D., Afonine, P. V., Baker, M. L., Bunkóczi, G., Chen, V. B., Croll, T. I., Hintze, B., Hung, L. W., Jain, S., McCoy, A. J., Moriarty, N. W., Oeffner, R. D., Poon, B. K., Prisant, M. G., Read, R. J., Richardson, J. S., Richardson, D. C., Sammito, M. D., Sobolev, O. V., Stockwell, D. H., ... Adams, P. D. (2019). Macromolecular structure determination using X-rays, neutrons and electrons: recent developments in Phenix. *Acta crystallographica. Section D, Structural biology*, 75(Pt 10), 861–877. <https://doi.org/10.1107/S2059798319011471>
- Lipski, A., Hervé, M., Lombard, V., Nurizzo, D., Mengin-Lecreux, D., Bourne, Y., & Vincent, F. (2015). Structural and biochemical characterization of the β -N-acetylglucosaminidase from *Thermotoga maritima*: toward rationalization of mechanistic knowledge in the GH73 family. *Glycobiology*, 25(3), 319–330. <https://doi.org/10.1093/glycob/cwu113>

- Lynch, K. H., Stothard, P., & Dennis, J. J. (2012). Comparative analysis of two phenotypically-similar but genomically-distinct *Burkholderia cenocepacia*-specific bacteriophages. *BMC genomics*, 13, 223. <https://doi.org/10.1186/1471-2164-13-223>
- Marintcheva, B., Marintchev, A., Wagner, G., & Richardson, C. C. (2008). Acidic C-terminal tail of the ssDNA-binding protein of bacteriophage T7 and ssDNA compete for the same binding surface. *Proceedings of the National Academy of Sciences of the United States of America*, 105(6), 1855–1860. <https://doi.org/10.1073/pnas.0711919105>
- McCoy, A. J., Grosse-Kunstleve, R. W., Adams, P. D., Winn, M. D., Storoni, L. C., & Read, R. J. (2007). Phaser crystallographic software. *Journal of applied crystallography*, 40(Pt 4), 658–674. <https://doi.org/10.1107/S0021889807021206>
- Murshudov, G. N., Vagin, A. A., Lebedev, A., Wilson, K. S., & Dodson, E. J. (1999). Efficient anisotropic refinement of macromolecular structures using FFT. *Acta crystallographica. Section D, Biological crystallography*, 55(Pt 1), 247–255. <https://doi.org/10.1107/S090744499801405X>
- Murzin A. G. (1993). OB (oligonucleotide/oligosaccharide binding)-fold: common structural and functional solution for non-homologous sequences. *The EMBO journal*, 12(3), 861–867. <https://doi.org/10.1002/j.1460-2075.1993.tb05726.x>
- Nguyen, J., Harb, L., Moreland, R., Liu, M., Gill, J. J., & Ramsey, J. (2019). Complete genome sequence of *Proteus mirabilis* siphophage Saba. *Microbiology resource announcements*, 8(41), e01094-19. <https://doi.org/10.1128/MRA.01094-19>
- Ofir, G., & Sorek, R. (2018). Contemporary phage biology: from classic models to new insights. *Cell*, 172(6), 1260–1270. <https://doi.org/10.1016/j.cell.2017.10.045>
- Onmus-Leone, F., Hang, J., Clifford, R. J., Yang, Y., Riley, M. C., Kuschner, R. A., Waterman, P. E., & Lesho, E. P. (2013). Enhanced *de novo* assembly of high throughput pyrosequencing data using whole genome mapping. *PloS one*, 8(4), e61762. <https://doi.org/10.1371/journal.pone.0061762>
- Padilla, D., Acosta, F., Ramos-Vivas, J., Grasso, V., Bravo, J., El Aamri, F., & Real, F. (2014). The pathogen *Hafnia alvei* in veterinary medicine: a review. *Journal of Applied Animal Research*, 43(2), 231–235. <https://doi.org/10.1080/09712119.2014.963086>
- Paez-Espino, D., Eloë-Fadrosch, E. A., Pavlopoulos, G. A., Thomas, A. D., Huntemann, M., Mikhailova, N., Rubin, E., Ivanova, N. N., & Kyrpides, N. C. (2016). Uncovering Earth's virome. *Nature*, 536(7617), 425–430. <https://doi.org/10.1038/nature19094>

- Pereira, J., Simpkin, A. J., Hartmann, M. D., Rigden, D. J., Keegan, R. M., & Lupas, A. N. (2021). High-accuracy protein structure prediction in CASP14. *Proteins*, 89(12), 1687–1699. <https://doi.org/10.1002/prot.26171>
- Pérez-Dorado, I., Campillo, N. E., Monterroso, B., Heseck, D., Lee, M., Páez, J. A., García, P., Martínez-Ripoll, M., García, J. L., Mobashery, S., Menéndez, M., & Hermoso, J. A. (2007). Elucidation of the molecular recognition of bacterial cell wall by modular pneumococcal phage endolysin CPL-1. *The Journal of biological chemistry*, 282(34), 24990–24999. <https://doi.org/10.1074/jbc.M704317200>
- Potterton, L., Agirre, J., Ballard, C., Cowtan, K., Dodson, E., Evans, P. R., Jenkins, H. T., Keegan, R., Krissinel, E., Stevenson, K., Lebedev, A., McNicholas, S. J., Nicholls, R. A., Noble, M., Pannu, N. S., Roth, C., Sheldrick, G., Skubak, P., Turkenburg, J., Uski, V., ... Wojdyr, M. (2018). CCP4i2: the new graphical user interface to the CCP4 program suite. *Acta crystallographica. Section D, Structural biology*, 74(Pt 2), 68–84. <https://doi.org/10.1107/S2059798317016035>
- Rūmnieks, J., Liekņiņa, I., Kalniņš, G., Šišovs, M., Akopjana, I., Bogans, J., & Tārs, K. (2020). Three-dimensional structure of 22 uncultured ssRNA bacteriophages: flexibility of the coat protein fold and variations in particle shapes. *Science advances*, 6(36), eabc0023. <https://doi.org/10.1126/sciadv.abc0023>
- Rūmnieks, J., & Tārs, K. (2011). Crystal structure of the read-through domain from bacteriophage Q β A1 protein. *Protein science: a publication of the Protein Society*, 20(10), 1707–1712. <https://doi.org/10.1002/pro.704>
- Rūmnieks, J., & Tārs, K. (2014). Crystal structure of the bacteriophage Q β coat protein in complex with the RNA operator of the replicase gene. *Journal of molecular biology*, 426(5), 1039–1049. <https://doi.org/10.1016/j.jmb.2013.08.025>
- Rūmnieks, J., & Tārs, K. (2017). Crystal structure of the maturation protein from bacteriophage Q β . *Journal of molecular biology*, 429(5), 688–696. <https://doi.org/10.1016/j.jmb.2017.01.012>
- Salmond, G. P., & Fineran, P. C. (2015). A century of the phage: past, present and future. *Nature reviews. Microbiology*, 13(12), 777–786. <https://doi.org/10.1038/nrmicro3564>
- Santiago-Rodriguez, T. M., & Hollister, E. B. (2022). Unraveling the viral dark matter through viral metagenomics. *Frontiers in immunology*, 13, 1005107. <https://doi.org/10.3389/fimmu.2022.1005107>
- Schrödinger, L., & DeLano, W. (2020). PyMOL. <http://www.pymol.org/pymol>

- Sievers, F., & Higgins, D. G. (2021). The Clustal Omega multiple alignment package. *Methods in molecular biology* (Clifton, N.J.), 2231, 3–16. https://doi.org/10.1007/978-1-0716-1036-7_1
- Song, H. S., Kim, J. Y., Kim, Y. B., Jeong, M. S., Kang, J., Rhee, J. K., Kwon, J., Kim, J. S., Choi, J. S., Choi, H. J., Nam, Y. D., & Roh, S. W. (2017). Complete genome sequence of a commensal bacterium, *Hafnia alvei* CBA7124, isolated from human feces. *Gut pathogens*, 9, 41. <https://doi.org/10.1186/s13099-017-0190-0>
- Sparta, K. M., Krug, M., Heinemann, U., Mueller, U., & Weiss, M. S. (2016). XDSAPP2.0. *Journal of Applied Crystallography*, 49(3), 1085–1092. <https://doi.org/10.1107/s1600576716004416>
- Suttle C. A. (2007). Marine viruses – major players in the global ecosystem. *Nature reviews. Microbiology*, 5(10), 801–812. <https://doi.org/10.1038/nrmicro1750>
- Terwilliger, T. C., Adams, P. D., Read, R. J., McCoy, A. J., Moriarty, N. W., Grosse-Kunstleve, R. W., Afonine, P. V., Zwart, P. H., & Hung, L. W. (2009). Decision-making in structure solution using Bayesian estimates of map quality: the PHENIX AutoSol wizard. *Acta crystallographica. Section D, Biological crystallography*, 65(Pt 6), 582–601. <https://doi.org/10.1107/S0907444909012098>
- Teufel, F., Almagro Armenteros, J. J., Johansen, A. R., Gislason, M. H., Pihl, S. I., Tsirigos, K. D., Winther, O., Brunak, S., von Heijne, G., & Nielsen, H. (2022). SignalP 6.0 predicts all five types of signal peptides using protein language models. *Nature biotechnology*, 40(7), 1023–1025. <https://doi.org/10.1038/s41587-021-01156-3>
- Theobald, D. L., Mitton-Fry, R. M., & Wuttke, D. S. (2003). Nucleic acid recognition by OB-fold proteins. *Annual review of biophysics and biomolecular structure*, 32, 115–133. <https://doi.org/10.1146/annurev.biophys.32.110601.142506>
- Usón, I., & Sheldrick, G. M. (2018). An introduction to experimental phasing of macromolecules illustrated by SHELX; new autotracing features. *Acta crystallographica. Section D, Structural biology*, 74(Pt 2), 106–116. <https://doi.org/10.1107/S2059798317015121>
- Vagin, A., & Teplyakov, A. (2010). Molecular replacement with MOLREP. *Acta crystallographica. Section D, Biological crystallography*, 66(Pt 1), 22–25. <https://doi.org/10.1107/S0907444909042589>
- WHO. (2015). Antibiotic resistance: multi-country public awareness survey. World Health Organization. <https://apps.who.int/iris/handle/10665/194460>
- Williams, C. J., Headd, J. J., Moriarty, N. W., Prisant, M. G., Videau, L. L., Deis, L. N., Verma, V., Keedy, D. A., Hintze, B. J., Chen, V. B., Jain, S., Lewis,

- S. M., Arendall, W. B., 3rd, Snoeyink, J., Adams, P. D., Lovell, S. C., Richardson, J. S., & Richardson, D. C. (2018). MolProbity: more and better reference data for improved all-atom structure validation. *Protein science: a publication of the Protein Society*, 27(1), 293–315. <https://doi.org/10.1002/pro.3330>
- Yolles, M. & Frieden, R. (2022). Viruses as living systems – a metacybernetic view. *Systems*, 10(3), 70. <https://doi.org/10.3390/systems10030070>
- Zaczek-Moczydłowska, M. A., Young, G. K., Trudgett, J., Fleming, C. C., Campbell, K., & O'Hanlon, R. (2020). Genomic characterization, formulation and efficacy *in planta* of a *Siphoviridae* and *Podoviridae* protection cocktail against the bacterial plant pathogens *Pectobacterium* spp. *Viruses*, 12(2), 150. <https://doi.org/10.3390/v12020150>
- Zrelovs, N., Dišlers, A., & Kazāks, A. (2020a). Motley crew: overview of the currently available phage diversity. *Frontiers in microbiology*, 11, 579452. <https://doi.org/10.3389/fmicb.2020.579452>
- Zrelovs, N., Černeoka, E., Dišlers, A., & Kazāks, A. (2020b). Isolation and characterization of the novel *Virgibacillus*-infecting bacteriophage Mimir87. *Archives of virology*, 165(3), 737–741. <https://doi.org/10.1007/s00705-019-04516-2>
- Zrelovs, N., Dišlers, A., & Kazāks, A. (2020c). Novel *Erwinia persicina* infecting phage Midgardsormr38 within the context of temperate *Erwinia* phages. *Frontiers in microbiology*, 11, 1245. <https://doi.org/10.3389/fmicb.2020.01245>
- Zrelovs, N., Dišlers, A., & Kazāks, A. (2021). Genome characterization of Nocturne116, novel *Lactococcus lactis*-infecting phage isolated from moth. *Microorganisms*, 9(7), 1540. <https://doi.org/10.3390/microorganisms9071540>
- Zrelovs, N., Jansons, J., Dišlers, A., & Kazāks, A. (2022a). *Morganella* phage Mecenats66 utilizes an evolutionarily distinct subtype of headful genome packaging with a preferred packaging initiation site. *Microorganisms*, 10(9), 1799. <https://doi.org/10.3390/microorganisms10091799>
- Zrelovs, N., Lamsters, K., Karušs, J., Krievāns, M., Dišlers, A., & Kazāks, A. (2022b). PVJ1 is not the first tailed temperate phage infecting bacteria from genus *Psychrobacillus*. Comment on Liu et al. Isolation and characterization of the first temperate virus infecting *Psychrobacillus* from marine sediments. *Viruses* 2022, 14, 108. *Viruses*, 14(3), 495. <https://doi.org/10.3390/v14030495>