Unidade Curricular: Laboratórios de Bioinformática, **Grupo 4**
Autores: Bruno Ferreira (pg58814), Joana Maia (pg58816), Sara Sousa (pg59766)

Análise Proteica

O presente documento realiza a análise proteica exaustiva de três proteínas chave do *Staphylococcus phage 88*. A análise foi conduzida através de uma pipeline bioinformática sequencial, integrando algoritmos de predição de homologia, topologia e modelação estrutural para inferir e validar a função biológica de cada uma das proteínas de interesse.

Endolisina (Gene ID: 5133735)

https://www.uniprot.org/uniprotkb/Q4ZAM7/entry

1. Identificação e Contexto

A proteína em estudo, identificada com o código de acesso UniProt (Bateman et al., 2025) Q4ZAM7 é uma N-acetylmuramoyl-L-alanine amidase codificada pelo *Staphylococcus phage 88*. Trata-se de uma proteína viral de 481 aminoácidos, atualmente classificada no TrEMBL como "Unreviewed". Dada a ausência da revisão manual por curadores, a presente análise bioinformática é fundamental para validar a anotação funcional, que é feita de forma automática, e caracterizar as suas propriedades estruturais e catalíticas.

2. Caracterização Físico-Química

A análise efetuada com o Expasy ProtParam (M.Walker, 2005) revela uma proteína com peso molecular de cerca de 54.1 kDa (54119.15 Da) e um ponto isoelétrico teórico de 8.73. O índice de instabilidade de 39.17 classifica a proteina como estável com confiança moderada, uma vez que o valor está próximo do limiar de instabilidade. Sendo assim, este valor que sugere que esta deverá manter a sua estrutura original em condições de teste realizadas *in vitro* por um período razoável.

3. Análise de Domínios e Famílias

A análise da sequência da proteína Q4ZAM7 - através das ferramentas UniProt, InterPro (Blum et al., 2025) e NCBI CDD (J. Wang et al., 2023) - permitiu identificar uma arquitetuta modular, característica de endolisinas com duplo domínio catalítico.
A anotação funcional identificou três domínios conservados funcionais distintos:

- Domínio Catalítico N-Terminal (~7-148 aa): contém domínio da família CHAP. Este domínio corresponde a uma endopeptidase, da família Peptidase 51, sendo responsável pela clivagem de pontes cruzadas de péptidos no peptidoglicano de *Staphylococcus*;
- Região Linker Flexível (~156-177 aa): região de baixa complexidade rica em aminoácidos polares, que impede a formação de estruturas secundárias rígidas. Este segmento atua como um conector (linker) flexível, conferindo mobilidade aos domínios catalíticos nas suas extremidades;
- Domínio Catalítico Central (~198-322 aa): representa o segundo domínio catalítico da proteína, tendo sido identificado como Ami-2 (N-acetilmuramoil-L-alanina amidase) pertencente à família PGRP. Foi revelada a presença de resíduos conservados de ligação ao zinco através do NCBI CDD, caracterizando este domínio como uma metaloenzima, que cliva a ligação entre açúcar e péptido do peptidoglicano;

- Domínio C-Terminal (~398-466 aa): Identificado como SH3b, representa o domínio de ligação à parede celular. Este domínio não possui atividade enzimática, mas confere especificidade e afinidade ao peptidoglicano presente na parede celular do *Staphylococcus*.

Name	Accession	Description	Interval	E-value
+ SH3_5	pfam08460	Bacterial SH3 domain	395-460	6.60×10^{-26}
+ Ami_2	smart00644	Ami_2 domain	198-322	1.11×10^{-24}
+ CHAP	pfam05257	CHAP domain	31-113	3.30×10^{-8}

Figura 1: Análise in silico da arquitetura modular da endolisina Q4ZAM7. **(A)** Visão geral da organização dos domínios funcionais obtida por InterPro, evidenciando os domínios N-terminal (CHAP), central (Amidase_2/PGRP) e C-terminal (SH3). **(B)** Detalhe dos domínios conservados (NCBI CDD), mostrando "hits" específicos (CHAP, Ami_2 e SH3_5). Os triângulos de cor ciano no topo indicam entre outras informações, os resíduos de ligação ao Zinco.

4. Estrutura Secundária e Terciária

Com a utilização das ferramentas PSIPRED (Buchan et al., 2024) e AlphaFold (Fleming et al., 2025) foram analisadas as estruturas secundárias e terciárias da proteína, respetivamente.

A predição da estrutura 2D apresenta uma elevada correlação com a análise de domínios. As regiões catalíticas e de ligação (CHAP, Amidase e SH3b) são ricas em hélices-α e folhas-β. Contudo, é destacada a região na posição 156-177 aa, prevista maioritariamente como Coil, que atua como conector flexível como referido anteriormente. A predição da estrutura 3D com índice de confiança (pLDDT) de 83.12 suporta esta topologia, demonstrando os dois domínios catalíticos separados pelo conector, permitindo o acesso simultâneo a diferentes locais de corte na parede celular.

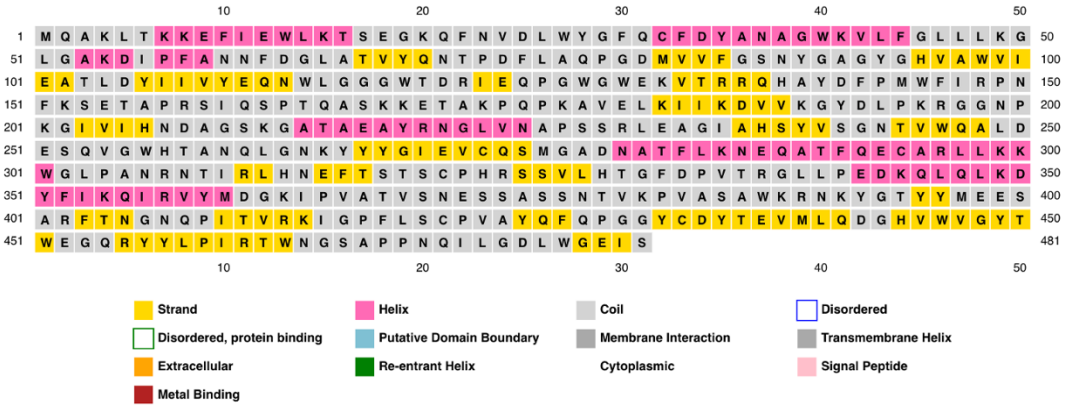


Figura 2: Predição da estrutura secundária da endolisina Q4ZAM7 (PSIPRED). A região N-terminal exibe uma predominância de hélices α (blocos rosa), características do domínio catalítico CHAP. Em contraste, a região C-terminal apresenta uma elevada densidade de folhas-β (blocos amarelos), consistentes com a topologia dos domínios de ligação SH3. As regiões a cinzento (Coil) representam linkers flexíveis que conectam estes módulos estruturados.

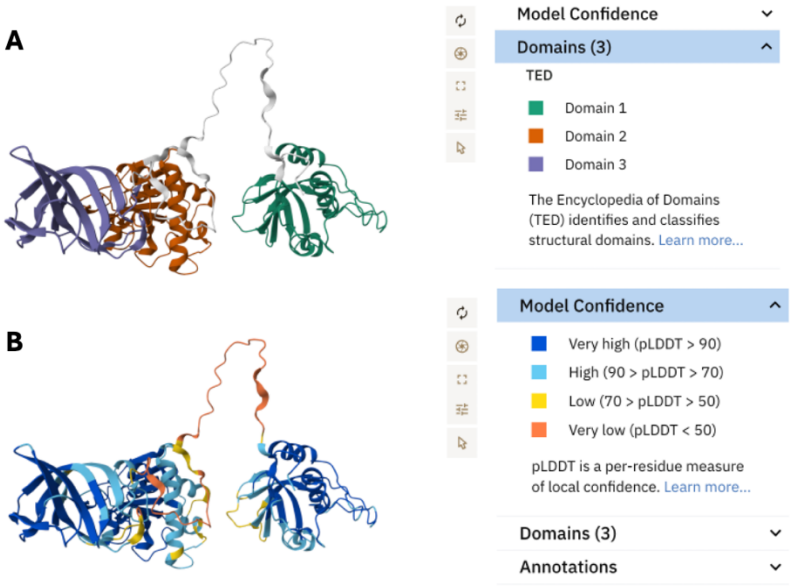


Figura 3: Análise estrutural e validação do modelo 3D da endolisina Q4ZAM7 (AlphaFold). **(A)** Representação da arquitetura modular da proteína, evidenciando os domínios funcionais: domínio catalítico N-terminal CHAP (roxo), domínio catalítico central Amidase (laranja) e domínio de ligação SH3b (verde). **(B)** O mesmo modelo colorido pelo pLDDT. Observa-se uma elevada confiança (azul escuro, pLDDT >90) nas regiões estruturadas correspondentes aos domínios funcionais.

Para validar funcionalmente os modelos previstos e ser feita a identificação de co-fatores, foi realizada uma pesquisa de homólogos estruturais utilizando a ferramenta BLAST (Altschul et al., 1990) contra a base de dados PDB (Protein Data Bank). O melhor alinhamento obtido foi da endolisina do *Staphylococcus phage G15* (Acession: 4OLS_A), que partilha uma boa conservação estrutural com a proteína em estudo.

A comparação entre ambas as proteínas revelou um facto evolutivo interessante: enquanto que a endolisina do *Staphylococcus phage G15* é uma enzima dependente de iões de Cálcio (Gu et al., 2014), a análise anterior da endolisina do nosso fago de interesse identificou a presença de resíduos conservados de ligação ao Zinco (visível na figura 1B). Esta diferença nos co-fatores sugere que, apesar da semelhança estrutural, a endolisina do *Staphylococcus phage 88* preserva

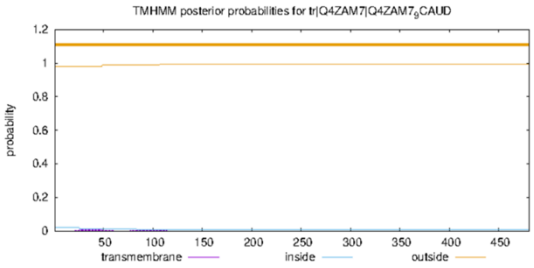
o mecanismo catalítico característico das amidases, atuando como uma metaloenzima dependente de Zinco.

5. Localização Subcelular e Topologia

A análise da topologia membranar realizada com a ferramenta TMHMM (Krogh et al., 2001) não detetou hélices transmembranares. O PSORTb (Yu et al., 2010) prevê uma localização extracelular (devido à homologia com lisostafinas secretadas), contudo o SignalP (Nielsen et al., 2024) indica uma probabilidade de 97.6% de ausência de péptido sinal. Esta divergência de resultados confirma o modelo de lise "holina-endolisina", na qual a proteína acumula-se no citoplasma, até que haja formação de poros pela holina, só aí permitindo a libertação da endolisina.

A

```
# tr|Q4ZAM7|Q4ZAM7_9CAUD Length: 481
# tr|Q4ZAM7|Q4ZAM7_9CAUD Number of predicted TMHs:  0
# tr|Q4ZAM7|Q4ZAM7_9CAUD Exp number of AAs in TMHs: 0.33761
# tr|Q4ZAM7|Q4ZAM7_9CAUD Exp number, first 60 AAs:  0.24468
# tr|Q4ZAM7|Q4ZAM7_9CAUD Total prob of N-in:        0.02127
tr|Q4ZAM7|Q4ZAM7_9CAUD  TMHMM2.0        outside     1   481
```



B

PSORTb Results (Click here for an explanation of the output formats)

```
SeqID: tr|Q4ZAM7|Q4ZAM7_9CAUD N-acetylmuramoyl-L-alanine amidase OS=Staphylococcus phage 88 OX=2908115 PE=3
  Analysis Report:
    CMSVM+              Unknown                 [No details]
    CWSVM+              Unknown                 [No details]
    CytoSVM+            Cytoplasmic             [No details]
    ECSVM+              Unknown                 [No details]
    ModHMM+             Unknown                 [No internal helices found]
    Motif+              Unknown                 [No motifs found]
    Profile+            Unknown                 [No matches to profiles found]
    SCL-BLAST+          Extracellular           [matched 3287967: Lysostaphin precursor (EC 3.4.24.75)
    SCL-BLASTe+         Unknown                 [No matches against database]
    Signal+             Unknown                 [No signal peptide detected]
  Localization Scores:
    Cytoplasmic             0.10
    CytoplasmicMembrane     0.14
    Cellwall                0.15
    Extracellular           9.60
  Final Prediction:
    Extracellular           9.60
```

C

Protein type	Signal peptide (Sec/SPI)	TAT signal peptide (Tat/SPI)	Lipoprotein signal peptide (Sec/SPII)	Other
Likelihood	0.0223	0.0004	0.0009	0.9764

Figura 4: (A) Análise da topologia transmembranar da endolisina Q4ZAM7 (TMHMM). O gráfico evidencia a ausência total de hélices transmembranares (Number of predicted TMHs: 0). **(B)** Predição da localização subcelular da endolisina Q4ZAM7 (PSORTb). A análise determinou uma localização final extracelular com confiança elevada. **(C)** Predição da presença de péptido sinal na endolisina Q4ZAM7 (SignalP). A análise classifica como "Other" com probabilidade de 97,64%, o que indica a ausência de péptido sinal.

6. Modificações Pós-Traducionais (PTMs)

A consulta de base de dados UniProt e a análise da sequência não revelaram anotações de modificações pós-traducionais complexas (glicosilação ou fosforilação regulatória) para esta proteína. Contudo, de forma a cumprir os requisitos da análise de modificações pós-traducionais foram utilizadas as ferramentas NetNGlyc 1.0 (Gupta & Brunak, 2002)e NetPhos 3.1(Blom et al., 1999).

A análise com NetNGlyc 1.0 revelou a presença de três potenciais locais de N-glicosilação, contudo apenas o resíduo com sequência NATF na posição 280 ultrapassou o valor limiar de 0.5, apresentando um score de 0.5764 e um resultado positivo "+". Os outros dois locais foram identificados como negativos "-", uma vez que o seu score é inferior ao *threshold*.
Relativamente à análise realizada pelo NetPhos 3.1, foram previstos múltiplos locais de fosforilação ao longo da sequência proteica. O gráfico resultante desta predição demonstra inúmeros picos acima do valor de *threshold*, a maioria deles pertencendo a resíduos de Thr e Ser.

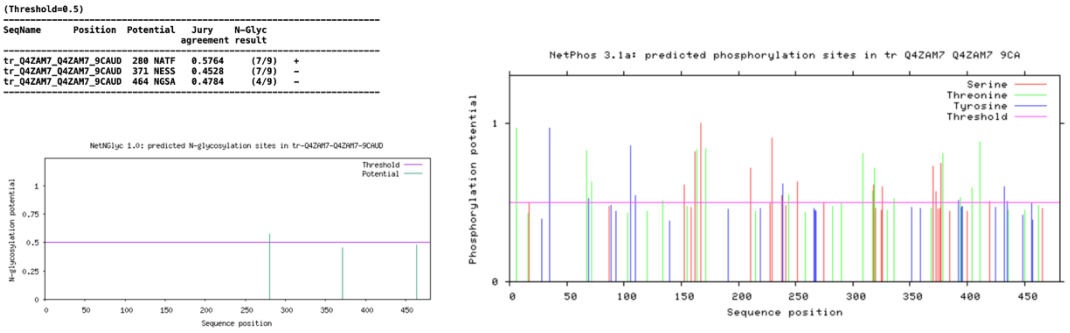


Figura 5: Predição de locais de glicosilação com NetNGlyci 1.0 (à esquerda) e de locais de fosforilação com NetPhos 3.1 (à direita) da endolisina Q4ZAM7.

Apesar dos resultados positivos obtidos em ambas as ferramentas, estes devem ser interpretados cuidadosamente, pois a sua ocorrência biológica tem uma probabilidade reduzida associada. Estas ferramentas (Blom et al., 1999; Gupta & Brunak, 2002) baseiam-se em dados de organismos eucariotas, que possuem muitas vezes maquinaria celular distinta.

Holina (Gene ID: 5133736)

https://www.uniprot.org/uniprotkb/Q4ZAM8/entry

1. Identificação e Contexto

A segunda proteína em estudo, identificada com o código de acesso UniProt Q4ZAM8 (anotada como ORF033) é a holina do *Staphylococcus phage 88*.
Trata-se de uma proteína se pequena dimensão com 145 aminoácidos, também classificada no TrEMBL como “Unreviewed”. A análise desta proteína é crucial para compreender o sistema de lise de dois componentes: enquanto a endolisina degrada a parede, a holina controla o tempo de lise, determinando o momento extao da morte da célula hospedeira através da permeabilização da membrana.

2. Caracterização Físico-Química

A caracterização físico-química efetuada com o Expasy ProtParam revela uma proteína com peso molecular de aproximadamente 15.7 kDa (15661.46 Da), significativamente menor que a endolisina. Apresenta um ponto isoelétrico teórico de 4.84, o que confere um carácter ácido.
O índice de instabilidade de 31.14 classifica-a como uma proteina estável *in vitro*.

3. Análise de Domínios e Famílias

A análise da sequência proteica Q4ZAM8 (UniProt) permitiu concluir que tem uma arquitetura bipartida com regiões funcionais distintas:

- Domínio N-Terminal (~3-80 aa): A região N-terminal contém o único domínio conservado detetado, pertencente à família Phage_holin_1. Corresponde à parte da proteína que se insere na membrana bacteriana, resultando na formação de poros;
- Domínio C-Terminal (~80-145 aa): Identificada como uma região intrinsecamente desordenada e rica em resíduos polares. Esta “cauda” citoplasmática não possui uma estrutura rígida e desempenha um papel crítico na regulação do tempo de lise.

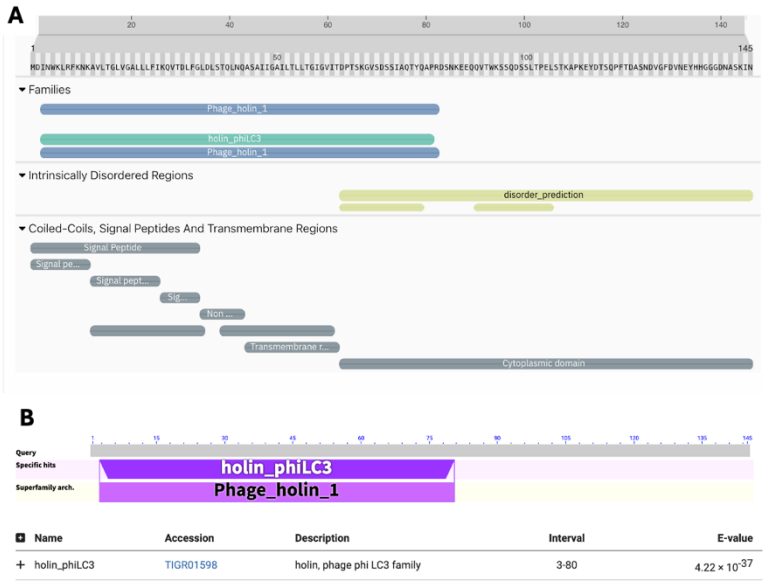


Figura 6: Análise in silico da arquitetura de domínios da holina Q4ZAM8. **(A)** Visão geral via InterPro, evidenciando a localização do domínio conservado Phage_holin_1 na região N-terminal. **(B)** Detalhe dos domínios conservados (NCBI CDD) identificando a subfamília específica holin_phiLC3 (superfamília Phage_holin_1), coincidente com a região transmembranar.

4. Estrutura Secundária e Terciária

Em consenso com o que foi descrito anteriormente, a estrutura secundária evidencia uma bipartição observada nos domínios: a região N-terminal é rica em em hélices-α, que são consistentes com domínios transmembranares, enquanto que a região C-terminal é dominada por *Coils*. O modelo 3D obtido apresenta um índice de confiança global de 63.84, considerado baixo. Através da análise da estrutura tridimensional da proteína completa, é possível a identificação de um único domínio funcional, sendo que a baixa confiança na região C-terminal confirma a previsão de desordem intrínseca, característica essencial para a função reguladora da cauda citoplasmática.

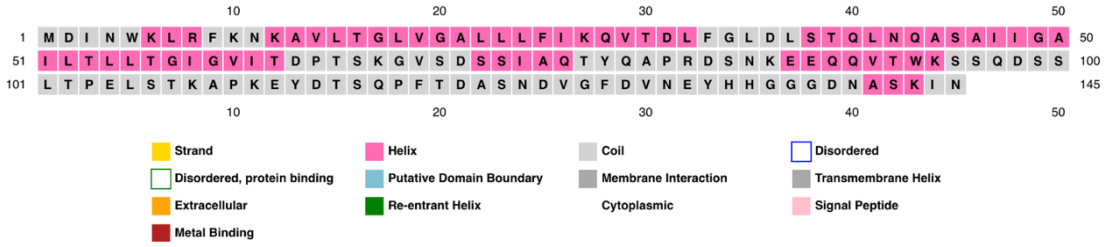


Figura 7: Predição da estrutura secundária da holina Q4ZAM8 (PSIPRED). Evidencia-se que a região N-terminal é dominada por elementos de hélices-α (blocos a rosa), que correspondem aos segmentos transmembranares. Já a região C-terminal é composta quase exclusivamente por regiões de Coil (blocos a cinzento), o que valida a natureza desordenada da cauda citoplasmática.

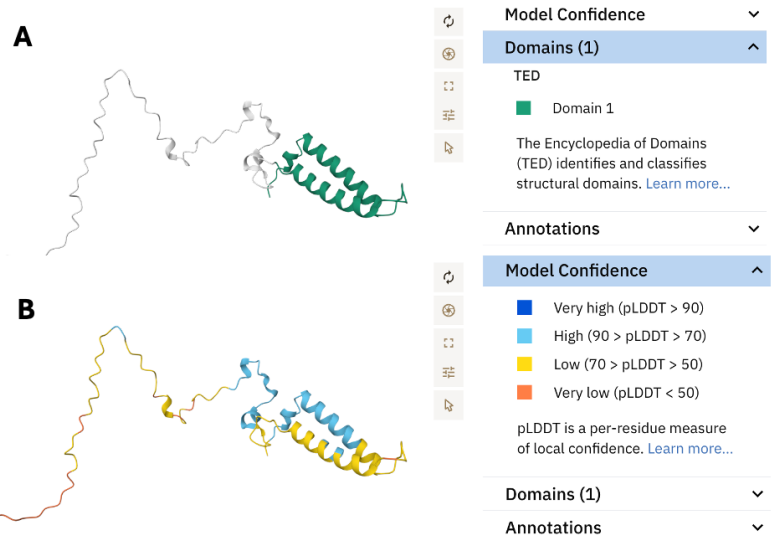
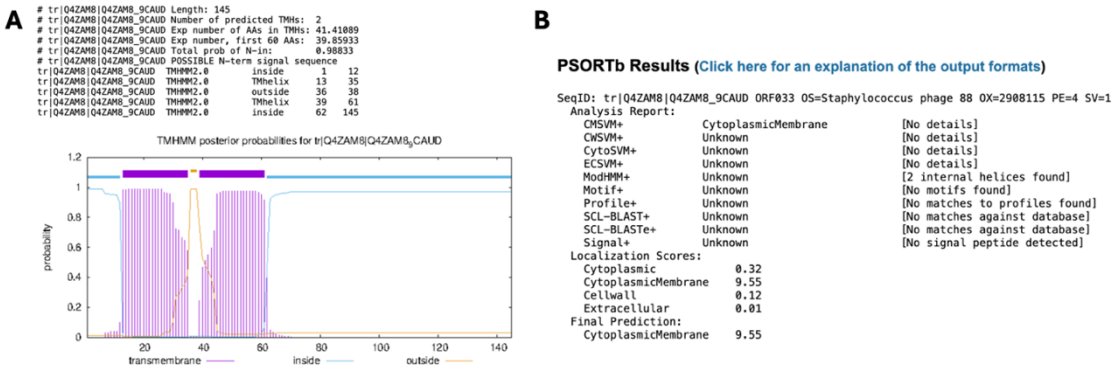


Figura 8: Análise estrutural e validação do modelo 3D da holina Q4ZAM8 (AlphaFold). **(A)** Representação da arquitetura da proteína, evidenciando o domínio N-terminal: Phage_holin_1 (verde) e a ausência de domínios na região C-terminal (cinza). **(B)** O mesmo modelo colorido pelo pLDDT. Observa-se uma elevada confiança (azul ciano, pLDDT >70) nas hélices-α que formam o poro, em contraste com a baixa confiança (amarelo/laranja, pLDDT <50) na cauda C-terminal.

A pesquisa de homólogos através do blastp contra a base de dados PDB não devolveu resultados significativos, não sendo possível identificar co-fatores por homologia direta. Este resultado é consistente com o mecanismo das holinas de Classe II, que normalmente dispensam co-fatores enzimáticos para a sua ativação.

5. Topologia e Localização

Com base na análise da topologia (TMHMM), foram identificadas duas hélices transmembranares, indicadas pelos picos de cor roxa na Figura 9A.
Esta arquitetura permite classificá-la como uma holina de classe II (I. N. Wang et al., 2000), o que sugere que o seu mecanismo de ação envolve a inserção na membrana e posterior oligomerização para a formação de poros, essenciais para a libertação da endolisina.
A ausência de péptido de sinal de secreção (SignalP: “Other”, 90.2%) e a previsão do PSORTb confirmam a sua localização final na Membrana Citoplasmática.



Protein type	Signal peptide (Sec/SPI)	TAT signal peptide (Tat/SPI)	Lipoprotein signal peptide (Sec/SPII)	Other
Likelihood	0.0907	0.0008	0.0056	0.9029

Figura 9: (A) Análise da topologia transmembranar da holina Q4ZAM8 (TMHMM). O gráfico identifica a presença de duas hélices transmembranares (Number of predicted TMHs: 2). **(B)** Predição da localização subcelular da holina Q4ZAM8 (PSORTb). A análise determinou uma localização final na membarana citoplasmática com confiança elevada. **(C)** Predição da presença de péptido sinal na holina Q4ZAM8 (SignalP). A análise classifica como “Other” com probabilidade de 90,29%, o que indica a ausência de péptido sinal.

6. Modificações Pós-Traducionais

À semelhança da análise realizada anteriormente na proteína endolisina, a consulta de base de dados UniProt não indicou modificações pós-traducionais confirmadas. As ferramentas NetNGlyc 1.0 e NetPhos 3.1 foram utilizadas para a realização de uma análise adicional.

A análise do NetNGlyc identificou um único potencial local de N-glicosilação na região C-terminal. O resíduo com a sequência NASK foi identificado na posição 140, com um score ligeiramente acima do *threshold* e com um resultado positivo "+" associado.
Relativamente ao NetPhos 3.1, o gráfico apresenta uma densidade elevada de locais de fosforilação concentrados na segunda metade da proteína, destacando os múltiplos resíduos de Serina (Ser), que ultrapassam o *threshold*.

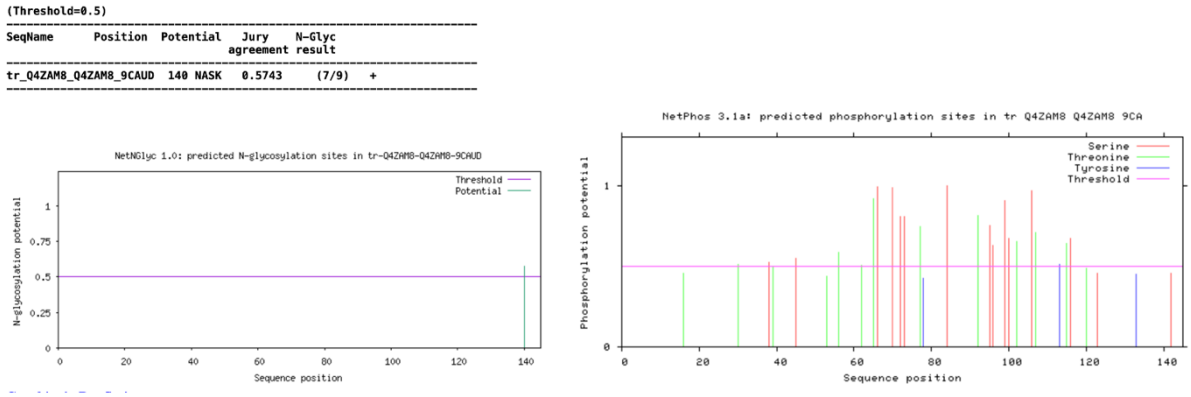


Figura 10: Predição de locais de glicosilação com NetNGlyci 1.0 (à esquerda) e de locais de fosforilação com NetPhos 3.1 (à direita) da holina Q4ZAM8.

O resultado positivo para N-glicosilação pode ser interpretado como um falso positivo, visto que a glicosilação apesar de poder ocorrer na zona citoplasmática, acontece predominantemente fora deste compartimento celular (Gupta & Brunak, 2002).

Fibra Caudal (Gene ID: 5133742)

https://www.uniprot.org/uniprotkb/Q4ZAN4/entry

1. Identificação e Contexto

A terceira proteína alvo do estudo, identificada no UniProt com o código de acesso Q4ZAN4 (anotada como ORF041) corresponde a uma fibra caudal do *Staphylococcus phage 88*.
Trata-se de uma proteína viral de 125 aminoácidos, atualmente classificada no TrEMBL como "Unreviewed". A sua análise é essencial para compreender a fase inicial do ciclo viral: a adsorção. As fibras da cauda funcionam como os sensores do fago, responsáveis pelo reconhecimento específico de recetores na superfície da bactéria hospedeira, desencadeando o processo de infeção.

2. Caracterização Físico-Química

A análise das propriedades físico-químicas (Expasy ProtParam) revela uma proteína de pequena dimensão com peso molecular de cerca de 14.2 kDa (14171.84 Da). É de destacar o ponto isoelétrico (pI) de 4.54, que indica um caráter bastante acídico, o que contrasta com a basicidade da endolisina. Outro dado relevante é o índice de instabilidade de 20.22, um valor extremamente baixo que classifica a proteína como sendo muito estável. Esta robustez é

biologicamente esperada para proteínas estruturais do virião, que necessitam de manter a sua integridade nos mais diversos tipos de ambientes extracelulares.

3. Análise de Domínios e Famílias

Ao contrário das proteínas analisadas anteriormente, a análise da sequência da proteína Q4ZAN4 não permitiu atribuir uma função enzimática direta pois não está descrita. A arquitetura da proteína revela-se bipartida:

- Domínio N-terminal (~1-60 aa): esta região contém o único domínio funcional anotado, pertencente à família DUF2977. A designação "DUF" indica que, embora esta região seja conservada entre vários fagos e bactérias, a sua função biológica ainda não é conhecida;
- Região C-terminal (~75-125 aa): região sem domínios conservados.

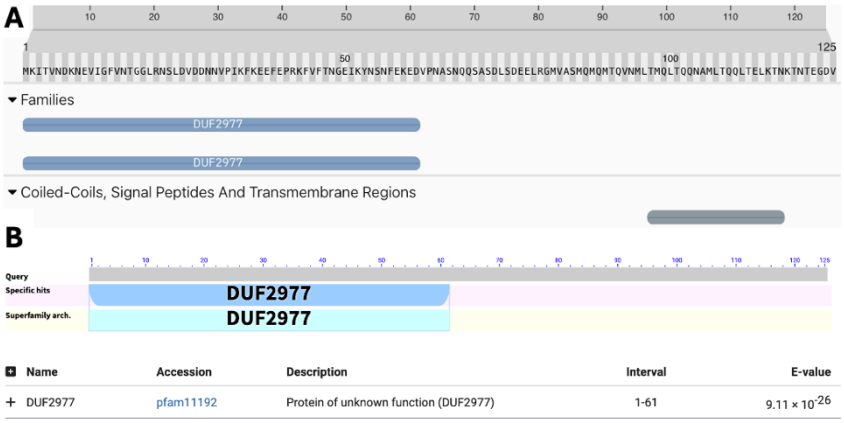


Figura 11: Análise in silico da arquitetura de domínios da fibra caudal Q4ZAN4. **(A)** Visão geral via InterPro, indicando a presença do domínio DUF2977 na região N-terminal. Adicionalmente, foi detetada uma região de Coiled-Coil na zona C-terminal. **(B)** Análise dos domínios conservados (NCBI CDD), que confirma a classificação da proteína na família DUF2977.

4. Estrutura Secundária e Terciária

A análise da estrutura secundaria com o PSIPRED permitiu concluir que esta proteína possui uma estrutura bipartida. O domínio N-terminal apresenta predominância de folhas-β, o que coincide com o posicionamento do domínio conservado DUF2977, sugerindo que a sua composição é rica em folhas-β.
Já o domínio C-terminal é caracterizado por uma extensa região de hélices-α, que sugerem a formação de estruturas coiled-coil como visto anteriormente, para formar a fibra da cauda.
O modelo 3D obtido apresenta um score de confiança (pLDDT) de 75.06, que é categorizado como alto. Este modelo AlphaFold permite a identificação de dois domínios funcionais, o que valida a predição da arquitetura bipartida em cabeça e cauda.

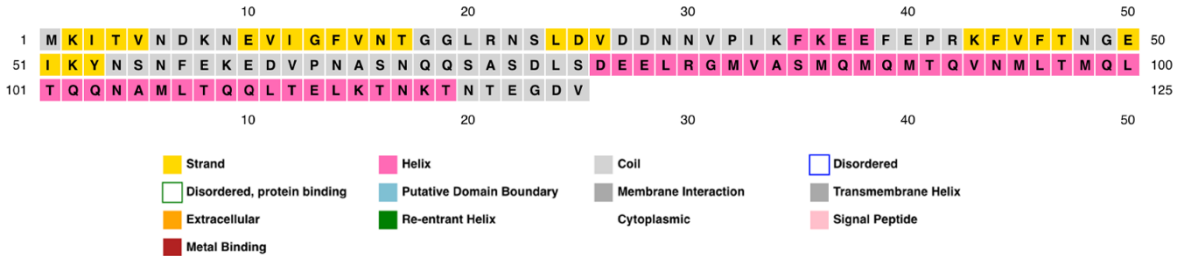


Figura 12: Predição da estrutura secundária da fibra caudal Q4ZAN4 (PSIPRED). A região N-terminal apresenta segmentos de folhas-β (blocos amarelos) intercalados por coils. A região C-terminal é dominada por região extensa de hélices-α (blocos rosa).

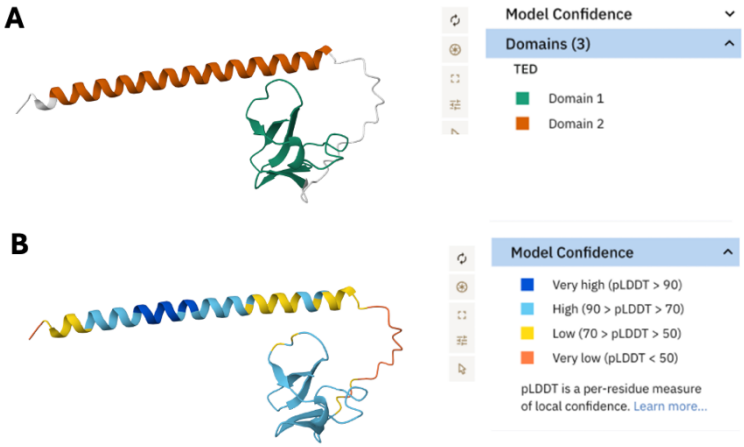
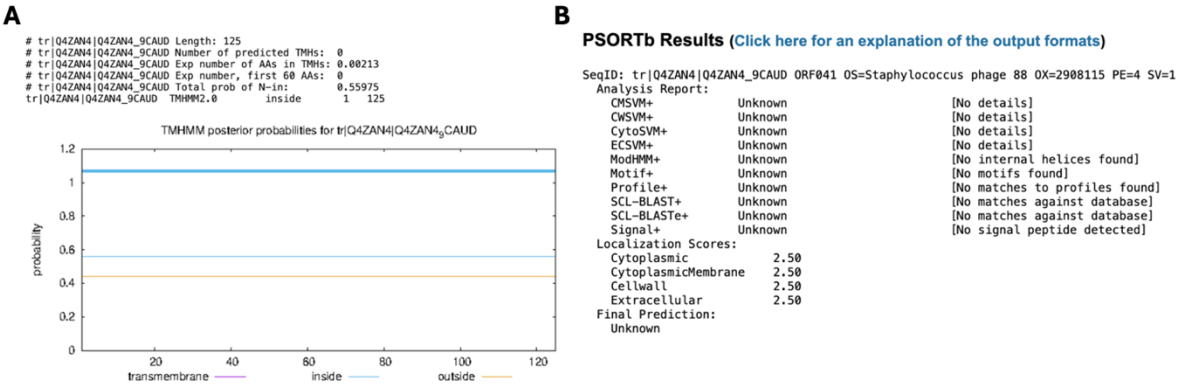


Figura 13: Análise estrutural e validação do modelo 3D da proteína fibra caudal Q4ZAN4 (AlphaFold). **(A)** Representação da arquitetura da proteína colorida por domínios, com evidência de arquitetura bipartida: um domínio na região N-terminal (verde) correspondente ao domínio DUF2977 e um domínio helicoidal na região C-terminal (laranja). **(B)** O mesmo modelo colorido pelo pLDDT. Observa-se uma elevada confiança (azul ciano, pLDDT >70) na estrutura do domínio N-terminal e numa das regiões da hélice.

Relativamente à fibra caudal, a pesquisa na base de dados PDB também não resultou na identificação de homólogos estruturais. Esta inexistência de homologia impede a identificação do recetor específico na superfície da bactéria.

5. Topologia e Localização

A análise da topologia transmembranar (TMHMM) confirma que a proteína não possui hélices transmembranares, o que se reflete numa probabilidade nula de inserção na membrana. É prevista ausência de secreção (SignalP), com uma probabilidade de 98.3% de não ter péptido sinal. A predição de localização subcelular (PSORTb) devolveu uma classificação de "Unknown", o que significa que o software não consegue atribuir um compartimento bacteriano, atribuindo pontuações baixas e idênticas de 2.50 a todos os compartimentos celulares possíveis. A análise conjunta destes dados possibilita chegar à conclusão de que esta proteína tem localização citoplasmática.



Protein type	Signal peptide (Sec/SPI)	TAT signal peptide (Tat/SPI)	Lipoprotein signal peptide (Sec/SPII)	Other
Likelihood	0.0126	0.0008	0.0035	0.9831

Figura 14: (A) Análise da topologia transmembranar da fibra caudal Q4ZAN4 (TMHMM). O gráfico evidencia a ausência de hélices transmembranares (Number of predicted TMHs:0). **(B)** Predição da localização subcelular da fibra caudal Q4ZAN4 (PSORTb). A análise determinou uma localização final desconhecida. **(C)** Predição da presença de péptido sinal na fibra caudal Q4ZAN4 (SignalP). A análise classifica como "Other" com probabilidade de 98,31%, o que indica a ausência de péptido sinal.

6. Modificações Pós-Traducionais

A análise da fibra caudal seguiu a mesma metodologia aplicada às proteínas anteriormente analisadas, sendo que a consulta do UniProt não indicou modificações experimentais confirmadas.

Relativamente à N-glicosilação foram previstos dois locais de potencial, contudo apenas um deles obteve um resultado positivo "+" e um score de 0.6587, que é superior ao valor do *threshold*. Este local encontra-se localizado na região C-terminal na posição 177, com sequência NKTN.
A análise da fosforilação com o NetPhos 3.1 revelou um gráfico com menos densidade de potenciais locais relativamente às análises anteriores, com um cluster de picos elevados localizados no centro da proteína e correspondentes à Serina (Ser).

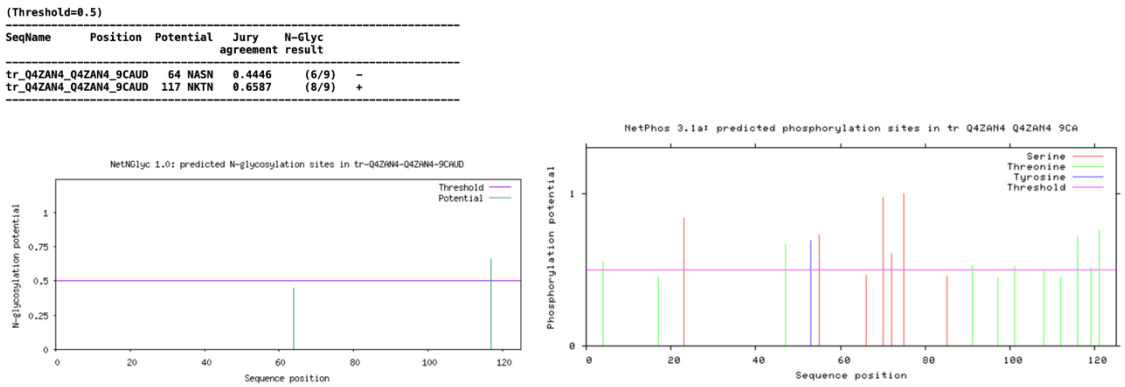


Figura 15: Predição de locais de glicosilação com NetNGlyci 1.0 (à esquerda) e de locais de fosforilação com NetPhos 3.1 (à direita) da fibra caudal Q4ZAN4.

Estas modificações estruturais são pouco prováveis de acontecer devido a restrições estruturais, sendo que o local de N-glicosilação e os picos de fosforilação coincidem com a região C-terminal identificada como Coiled-Coil (Figura 11A).

Referências Bibliográficas

1. Altschul, S. F., Gish, W., Miller, W., Myers, E. W., & Lipman, D. J. (1990). Basic local alignment search tool. *Journal of Molecular Biology*, *215*(3), 403–410. https://doi.org/10.1016/S0022-2836(05)80360-2

2. Bateman, A., Martin, M. J., Orchard, S., Magrane, M., Adesina, A., Ahmad, S., Bowler-Barnett, E. H., Bye-A-Jee, H., Carpentier, D., Denny, P., Fan, J., Garmiri, P., da Costa Gonzales, L. J., Hussein, A., Ignatchenko, A., Insana, G., Ishtiaq, R., Joshi, V., Jyothi, D., … Zhang, J. (2025). UniProt: the Universal Protein Knowledgebase in 2025. *Nucleic Acids Research*, *53*(D1), D609–D617. https://doi.org/10.1093/nar/gkae1010

3. Blom, N., Gammeltoft, S., & Brunak, S. (1999). Sequence and structure-based prediction of eukaryotic protein phosphorylation sites. *Journal of Molecular Biology*, *294*(5), 1351–1362. https://doi.org/10.1006/jmbi.1999.3310

4. Blum, M., Andreeva, A., Florentino, L. C., Chuguransky, S. R., Grego, T., Hobbs, E., Pinto, B. L., Orr, A., Paysan-Lafosse, T., Ponamareva, I., Salazar, G. A., Bordin, N., Bork, P., Bridge, A., Colwell, L., Gough, J., Haft, D. H., Letunic, I., Llinares-López, F., … Bateman, A. (2025). InterPro: the protein sequence classification resource in 2025. *Nucleic Acids Research*, *53*(D1), D444–D456. https://doi.org/10.1093/NAR/GKAE1082

5. Buchan, D. W. A., Moffat, L., Lau, A., Kandathil, S. M., & Jones, D. T. (2024). Deep learning for the PSIPRED Protein Analysis Workbench. *Nucleic Acids Research*, *52*(W1), W287–W293. https://doi.org/10.1093/NAR/GKAE328

6. Fleming, J., Magana, P., Nair, S., Tsenkov, M., Bertoni, D., Pidruchna, I., Lima Afonso, M. Q., Midlik, A., Paramval, U., Žídek, A., Laydon, A., Kovalevskiy, O., Pan, J., Cheng, J., Avsec, Ž., Bycroft, C., Wong, L. H., Last, M., Mirdita, M., … Velankar, S. (2025). AlphaFold Protein Structure Database and 3D-Beacons: New Data and Capabilities. *Journal of Molecular Biology*, *437*(15). https://doi.org/10.1016/j.jmb.2025.168967

7. Gu, J., Feng, Y., Feng, X., Sun, C., Lei, L., Ding, W., Niu, F., Jiao, L., Yang, M., Li, Y., Liu, X., Song, J., Cui, Z., Han, D., Du, C., Yang, Y., Ouyang, S., Liu, Z. J., & Han, W. (2014). Structural and Biochemical Characterization Reveals LysGH15 as an Unprecedented "EF-Hand-Like" Calcium-Binding Phage Lysin. *PLoS Pathogens*, *10*(5), e1004109. https://doi.org/10.1371/JOURNAL.PPAT.1004109

8. Gupta, R., & Brunak, S. (2002). Prediction of glycosylation across the human proteome and the correlation to protein function.

9. Krogh, A., Larsson, B., Von Heijne, G., & Sonnhammer, E. L. 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

10. M.Walker, J. (2005). The Proteomics Protocols Handbook. *The Proteomics Protocols Handbook*. https://doi.org/10.1385/1592598900

11. Nielsen, H., Teufel, F., Brunak, S., & von Heijne, G. (2024). SignalP: The Evolution of a Web Server. *Methods in Molecular Biology*, *2836*, 331–367. https://doi.org/10.1007/978-1-0716-4007-4_17

12. Wang, I. N., Smith, D. L., & Young, R. (2000). Holins: the protein clocks of bacteriophage infections. *Annual Review of Microbiology*, *54*, 799–825. https://doi.org/10.1146/ANNUREV.MICRO.54.1.799

13. Wang, J., Chitsaz, F., Derbyshire, M. K., Gonzales, N. R., Gwadz, M., Lu, S., Marchler, G. H., Song, J. S., Thanki, N., Yamashita, R. A., Yang, M., Zhang, D., Zheng, C., Lanczycki, C. J., & Marchler-Bauer, A. (2023). The conserved domain database in 2023. *Nucleic Acids Research*, *51*(D1), D384–D388. https://doi.org/10.1093/NAR/GKAC1096

14. Yu, N. Y., Wagner, J. R., Laird, M. R., Melli, G., Rey, S., Lo, R., Dao, P., Cenk Sahinalp, S., Ester, M., Foster, L. J., & Brinkman, F. S. L. (2010). PSORTb 3.0: improved protein subcellular localization prediction with refined localization subcategories and predictive capabilities for all prokaryotes. *Bioinformatics*, *26*(13), 1608–1615. https://doi.org/10.1093/BIOINFORMATICS/BTQ249