Structure of the Potential Virulence Factor from Francisella tularensis Shows Unexpected Presence of the SHS2 Motif and Similarity to Other Bacterial Virulence Factors

Janette Chammas[1], Mallika Iyer[2], George Minasov[3], Ludmilla Shuvalova [3], Wayne Anderson[4], and Adam Godzik[5]

[1]University of California Riverside College of Natural and Agricultural Sciences
[2]Sanford Burnham Prebys Medical Discovery Institute
[3]Northwestern University Feinberg School of Medicine
[4]Northwestern University
[5]University of California Riverside School of Medicine

January 14, 2022

Abstract

Pathogenic bacteria attack their host by secreting virulence factors that in various ways interrupt host defenses and damage their cells. Functions of many virulence factors, even from well-studied pathogens, are still unknown. Francisella tularensis is a class A pathogen and a causative agent of tularemia, a disease that is lethal without proper treatment. Here we report the three-dimensional structure and preliminary analysis of the potential virulence factor identified by the transcriptomic analysis of the F. tularensis disease models that is encoded by the FTT_1539 gene. The structure of the FTT_1539 protein contains two sets of three stranded antiparallel beta sheets, with a helix placed between the first and the second beta strand in each sheet. This structural motif, previously seen in virulence factors from other pathogens, was named the SHS2 motif and identified to play a role in protein-protein interactions and small molecule recognition. Sequence and structure analysis identified FTT_1539 as a member of a large family of secreted proteins from a broad range of pathogenic bacteria, such as Helicobacter pylori and Mycobacterium tuberculosis. While the specific function of the proteins from this class is still unknown, their similarity to the H. pylori Tip-α protein that induces TNF-a and other chemokines through NF-kB activation suggests the existence of a common pathogen-host interference mechanism shared by multiple human pathogens.

Janette Chammas[1], Mallika Iyer[2], George Minasov[3,4], Ludmilla Shuvalova[3,4], Wayne F. Anderson[4], and Adam Godzik[5,6] *

[1] Undergraduate Research Project, College of Natural and Agricultural Sciences, University of California Riverside, Riverside, CA, 92521, USA

[2] Graduate School of Biomedical Sciences, Sanford Burnham Prebys Medical Discovery Institute, La Jolla, CA, 92037, USA

[3] Department of Microbiology-Immunology, Northwestern University Feinberg School of Medicine, Chicago, IL 60201

[4] Center for Structural Genomics of Infectious Diseases, Northwestern University Feinberg School of Medicine, Chicago, IL 60201

[5] Biosciences Division, University of California Riverside School of Medicine, Riverside, CA, 92521, USA

[6] Center for Structural Genomics of Infectious Diseases, University of California Riverside School of Medicine, Riverside, CA, 92521, USA

ABSTRACT:

Pathogenic bacteria attack their host by secreting virulence factors that, in various ways, interrupt host defenses and damage their cells. Specific functions of many putative virulence factors, even from well-studied pathogens, are still unknown. *Francisella tularensis*is a class A pathogen and the causative agent of tularemia, a rare but potentially lethal disease that can be treated only with specialized antibiotics. Here we report the three-dimensional structure and preliminary analyses of the potential virulence factor, identified by the transcriptomic analysis of the *F. tularensis* disease models, that is encoded by the FTT_1539 gene. The structure of the FTT_1539 protein contains two sets of three-stranded antiparallel beta sheets, with a helix placed between the first and the second beta strand in each sheet. This structural motif, previously seen in virulence factors from other pathogens, was named the SHS2 motif and was identified to play a role in protein-protein interactions and small molecule recognition. Sequence and structure analysis identifies FTT_1539 as a founding member of a novel family of secreted proteins from a broad range of pathogenic bacteria, including *Helicobacter pylori* and*Mycobacterium tuberculosis.* While the specific function of the proteins from this class is still unknown, their structural similarity to the *H. pylori* Tip-α protein that induces TNF-α and other chemokines through NF-ϰB activation suggests the existence of a common pathogen-host interference mechanism shared by multiple human pathogens.

INTRODUCTION:

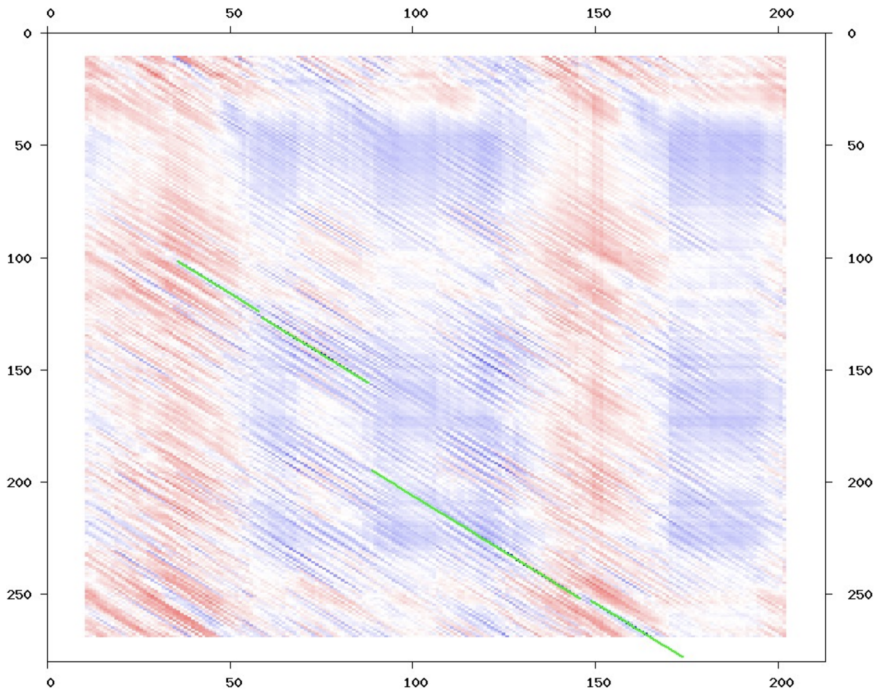
Francisella tularensis is a Gram-negative coccobacillus [1] that is the causative agent of tularemia (rabbit fever), a disease that results in extreme respiratory illness and distress that can be fatal without treatment [2]. Tularemia can be spread through bug bites or contracted through an airborne route. It was considered a viable biological weapon and its actual use in WWII was claimed, even though this claim was later disputed [3]. Despite extensive research on*F. tularensis* , we still do not fully understand its virulence mechanisms, and the functions of many of its proteins are unknown. However, *F. tularensis* studies on mouse infection models identified several proteins that were highly upregulated during infection and strongly reacted with the mouse immune system. [4]. While most of these proteins have a known role in *F. tularensis*pathogenicity, one of the standouts was an uncharacterized protein encoded by a gene in the locus FTT_1539. In another study in which mice were infected with *F. tularensis* type A strain FSC033 [5], the amounts of bacterial proteins isolated from infected mice were compared to that in bacteria that were cultured *in vitro* and it was found that the abundance of FTT_1539 increased by more than 4-fold [6, 7]. The protein FTT_1539 was also shown to co-purify with the membrane fraction and to be recognized by antibodies as playing a role in cell surface associations [8]. All these observations support the hypothesis that FTT_1539 may play a role in *F. tularensis*virulence, or more generally, in its interactions with the host. In addition, FTT_1539 is identified as an immunodominant antigen, which suggests its possible use as a tularemia vaccine candidate [9].

To further our understanding of the possible role of the FTT_1539 protein as the *F. tularensis* virulence factor, its structure was determined by the Center for Structural Genomics of Infectious Diseases (CSGID) and its coordinates were deposited to the Protein Data Bank (PDB) with the code 4QVR. In this paper we describe the structure and provide genomic analyses of FTT_1539.

RESULTS:

The locus FTT_1539 (also spelled FTT1539) of the *Francisella tularensis* strain Schu S4 genome encodes a protein with 511 amino acid residues, which we will refer to as the FTT_1539 protein. Initial sequence analysis, including Hidden Markov Model (HMM) analysis with Pfam [10], did not indicate homology to any well characterized proteins and thus in the genomics databases it is described as a "conserved hypothetical protein". Below we present the results of the sequence and structure analyses of this protein.

Sequence and Phylogenetic Analysis



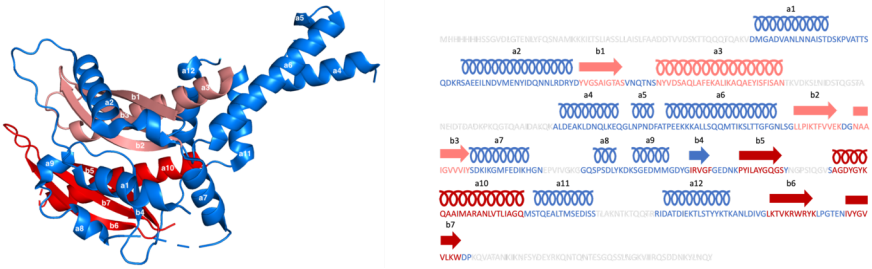
Francisella tularensis protein FTT_1539 defines a large novel protein family, with at least 532 homologs identified in the NCBI refseq database [11] with an iterative PSI-BLAST [12] search. Most of its homologs were found within the *Francisella* genus, with more distant ones found among several human pathogens, especially pathogens of the gastrointestinal tract such as *Vibrio parahaemolyticus* ,*Salmonella enterica* or *Campylobacter jejuni*. None of the homologs were experimentally characterized and most bore the name "hypothetical protein" suggesting the need for further research on this protein family. Sequence similarity within the *Francisella*genus was high (over 50% sequence identity) and it dropped to around 20% for the homologs from other genera, suggesting a genus specific expansion in *Francisella* species.

Analysis with the distant homology prediction algorithms FFAS [13] and HHpred [14] suggested that the C-terminus of FTT_1539 may contain a domain that could be a distant homolog of the Pfam LPP20 lipoprotein family (PF02169.16) such as Lpp20 (HP1456) from*Helicobacter pylori* (PDB code 5OK8). In addition, FFAS dot plot analysis (see Figure 1) also suggested the existence of an internal repeat with the C-terminal LPP20-like domain showing weak sequence similarity to the N-terminal part of the FTT_1539 protein.

Structure Determination

The structure of the FTT_1539 protein was determined using the high throughput structure determination pipeline at the Center for Structural Genomics of Infectious Diseases (CSGID). Data was collected to 2.30 Å resolution and a model with R-factor 0.174 (R-free of 0.224) was deposited to the PDB with the code 4QVR. The structure determination protocol is described in the Methods section.

Structure analysis

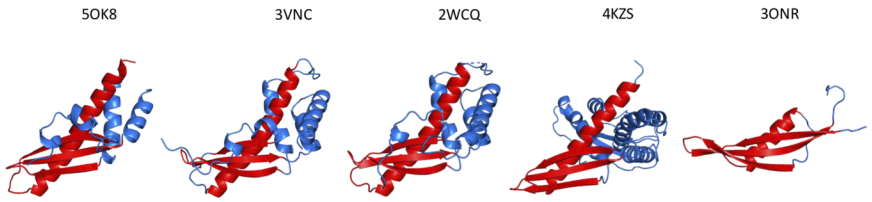


The experimentally determined structure presented a complex α/β fold consisting of 9 long and 3 short helices and two independent, antiparallel, twisted beta sheets with the 132 topology, shown in Figure 2 in mauve and red, respectively, together with a conserved helix inserted between the first and second strand. The second beta sheet has an additional short beta strand (b4), so its topology may be also described as 1243.

Structure comparison using the Dali [15] and FATCAT [16] algorithms confirmed the presence of an internal repeat (Figure 2 and 3) and its similarity to the LPP20 proteins. Database searches among other PDB structures identified hundreds of proteins with statistically significant structural similarity to the structure of FTT_1539. The top of the list contained several secreted or cell surface proteins from human pathogens such as HP1454 (PDB ID:4KZS), HP1456 (PDB ID: 5OK8), Tip-αN34 (PDB ID: 2WCQ) and Tip-αN25 (PDB ID: 3VNC) from*Helicobacter pylori* and calcium dodecin (Rv0397) from*Mycobacterium tuberculosis* (PDB ID: 3ONR). More distant similarities to other proteins from the SHS2 Pfam clan were also found in the search. The similarity to the Tip-α proteins is especially interesting, as they are known to induce human TNF-α and have carcinogenic effects [17].

The motif repeated in the FTT_1539 protein, consisting of the three-stranded antiparallel beta sheets, with a helix inserted between the first and the second beta strand, was previously identified in the bacterial ATPase FtsA (Pfam family PF14450), the archaea-eukaryotic RNA polymerase subunit Rpb7p (Pfam family PF03876), the GyrI-like small molecule binding domain (Pfam family PF06445), and the archease protein family MTH1598/Tm1083-like (Pfam family PF01951) [18], together forming an SHS2 clan. Interestingly, proteins from the GyrI family also contain a tandem repeat of this motif. The N-terminus half of the tandem repeat in the FTT_1539 protein is more similar to the "classical" SHS2 motif. We hypothesize that after duplication, one repeat, the N-terminal one, retained its structure (and possibly function) while the second repeat located at the C-terminus diverged and possibly lost its function.

The protein with the structure most similar to that of FTT_1539 identified by the FATCAT algorithm was the *Helicobacter pylori*protein HP1454 (PDB ID:4KZS) [19], with an RMSD (Root Mean Square Deviation) of 2.58 Å over the length of 157 amino acids in the pairwise structural alignment. The protein HP1454 contains 3 distinct domains, but the region similar to the structure of FTT_1539 consists of the N-terminal Domain I containing a classical SHS2 motif – a three-stranded antiparallel β-sheet with a single α-helix inserted between the first and second beta strand [19]. This SHS2 motif is highlighted in Figure 3. The N-terminal domain of HP1454 is extracellular and has structural and potential functional similarities to Tip-α proteins, which are classified as carcinogenic factors. Although the significance of this functional similarity is still unclear, it has been suggested that this motif is involved in protein-protein interactions due to its cellular localization [19]. The second most similar structure belongs to the *H. pylori*protein LPP20 (HP1456) (PDB ID: 5OK8), with an RMSD of 3.03 Å over the length of 125 amino acids [20]. The LPP20 protein was the founding member of the Pfam LPP20 protein family which was initially characterized as a non-essential class of lipoproteins [21]. Other members of this family are virulence factors that are bound to the outer membrane of the bacteria and secreted and transported via vesicles [20]. *H. pylori* LPP20 was also found to play a role in cancer suppression by reducing the expression of E-cadherin in gastric cancer cells [20].



The most interesting similarity found was to the structure of the*Helicobacter pylori* protein Tip-α (Tumor necrosis factor α (TNF-α) inducing protein) (PDB IDs: 2WCQ, 3VNC) with an RMSD of 3.08 Å over the length of 136 amino acids. Tip-α is an unusual virulence factor, being able to penetrate gastric cells and stimulate the production of the TNF- α and simultaneously activate the NF-ϰB pathway [17], finally resulting in a strong carcinogenic effect. Tip-α is highly expressed during bacterial infection [22] and forms homodimers in its active state.

Finally, another pairwise alignment between the structure of calcium dodecin (Rv0397) from *Mycobacterium tuberculosis* (PDB ID: 3ONR) and the structure of FTT_1539 had an RMSD of 2.19 Å over the length of 137 amino acids. Calcium dodecin is also classified as a member of the SHS2 clan and is involved in secretion and signaling of proteins that help regulate the life cycle of bacteria [23]. The structural fold of calcium dodecin allows it to form a pocket-like motif which binds calcium and is similar to metal or flavin binding sites [23].

CONCLUSION/DISCUSSION:

The analysis presented here highlights the presence of a common structural motif in a large group of diverse secreted proteins from a broad range of human pathogens such as *Franciscella tularensis* ,*Helicobacter pylori and Mycobacterium tuberculosis* . Although the sequence identity between these proteins is low, their overall structure contains the signature SHS2 motif and they all are secreted or located on the bacterial cell surface during infection and were broadly classified as virulence factors. Interestingly, most of these proteins also form homodimers. This suggest that the tandem duplication in FTT_1539 could be important for its function, mimicking the dimer arrangement of other proteins from this clan. Of these proteins only 3VNC has been functionally and structurally characterized and the location of matched regions in the structures of the FTT_1539 and the homologs discussed here suggests a possible similarity in their function and pathogenicity.

The genome of *F. tularensis* contains components homologous to Type I, IV and VI secretion systems [24], suggesting a mechanism for how the FTT_1539 protein can be secreted. Analogies with the function of Tip-α and connecting fragmentary knowledge of the functions of other proteins from this group allows us to propose that FTT_1539 may be involved in interactions with the host inflammatory pathways, and the suppression of the inflammatory immune response activation through intracellular molecules such as TNF-α [24].

METHODS:

Cloning, expression, purification, crystallization, data collection, structure solution and refinement

The FTT_1539 encoding gene from *Francisella tularensis* subsp. tularensis SCHU S4 strain (gi: 56708571; residues 1-476 was cloned [25], expressed [26], purified [27], set up for crystallization, screened, data was collected and structure was solved using the standard CSGID protocol, as described before [28] (see Supplementary Tables 1-5 with data on cloning and expression (Table S1), purification (Table S2), crystallization (Table S3), data collection (Table S4) and structure refinements (Table S5)). The structure was deposited to the Protein Data Bank (https://www.rcsb.org/) with the PDB code 4QVR.

Sequence and structure analysis

Sequence similarity was identified using Position-Specific Iterated BLAST (PSI-BLAST) [12] on the set of "Refseq" dataset, and the algorithm went through 12 iterations until convergence.

Protein with similar structured we identifies by structure comparison programs DALI [15] and FATCAT [16] searching the PDB database. The structure visualization was done with PyMOL [29] and Chimera [30].

A multiple protein structure alignment was performed on the POSA (Partial Order Structure Alignment) server [31], which provided a superimposed PDB file of all studied proteins to view and individual structurers were visualized using PyMOL.

The FATCAT server was used [16] to calculate pairwise alignments to obtain RMSD values and database similarity searches.

The clan and the family of the FTT_1539 protein we identified using the Pfam database of protein family Hidden Markov Models (HMM) [10] and the local installation of the HMMER [32]. Pfam HMM identified the SHS2 domain found in the N-terminus of the FTT_1539 protein, and the Pfam database provided details of the nine other members of the SHS2 clan discussed here, such as GyrI-Like, Tip-alpha, and Dodecin type proteins.

ACKNOWLEDGMENTS:

This project has been funded in whole or in part with Federal funds from the National Institute of Allergy and Infectious Diseases, under Contracts No. HHSN272201200026C and HHSN272201700060C and National Institute of General Medical Sciences grant GM118187, both at the National Institutes of Health, Department of Health and Human Services. This research used resources of the Advanced Photon Source, a U.S. Department of Energy (DOE) Office of Science User Facility operated for the DOE Office of Science by Argonne National Laboratory under Contract No. DE-AC02-06CH11357. Use of the LS-CAT Sector 21 was supported by the Michigan Economic Development Corporation and the Michigan Technology Tri-Corridor (Grant 085P1000817).

We would also like to acknowledge Drs. Nicole L. Inniss and Karla J.F. Satchell, both at Northwestern University Feinberg School of Medicine for help in editing and collecting data for this manuscript and Drs. Ievgeniia Dubrovska and Kristin Flores for technical assistance with the structure determination.

CONFLICT OF INTEREST

Authors declare no conflicts of interest related to this publication.

REFERENCES:

1. Larsson, P., et al., *The complete genome sequence of Francisella tularensis, the causative agent of tularemia.* Nat Genet, 2005. **37** (2): p. 153-9.

2. Prevention, C.f.D.C.a. *Tularemia* . 2020; Available from: https://www.cdc.gov/tularemia/index.html.

3. Dennis, D.T., et al., *Tularemia as a biological weapon: medical and public health management.* JAMA, 2001. **285** (21): p. 2763-73.

4. Andersson, H., et al., *Transcriptional profiling of host responses in mouse lungs following aerosol infection with type A Francisella tularensis.* J Med Microbiol, 2006. **55** (Pt 3): p. 263-271.

5. Keim, P., A. Johansson, and D.M. Wagner, *Molecular epidemiology, evolution, and ecology of Francisella.* Ann N Y Acad Sci, 2007. **1105** : p. 30-66.

6. Ellis, J., et al., *Tularemia.* Clin Microbiol Rev, 2002.**15** (4): p. 631-46.

7. Heras, B., et al., *DSB proteins and bacterial pathogenicity.*Nat Rev Microbiol, 2009. **7** (3): p. 215-25.

8. Chandler, J.C., *The Surface proteome of Francisella tularensis* , in *Department of Microbiology, Immunology, and Pathology* . 2011, Colorado State University, Fort Collins.

9. Chu, P., et al., *Live attenuated Francisella novicida vaccine protects against Francisella tularensis pulmonary challenge in rats and non-human primates.* PLoS Pathog, 2014. **10** (10): p. e1004439.

10. El-Gebali, S., et al., *The Pfam protein families database in 2019.* Nucleic Acids Res, 2019. **47** (D1): p. D427-D432.

11. O'Leary, N.A., et al., *Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation.*Nucleic Acids Res, 2016. **44** (D1): p. D733-45.

12. Altschul, S.F., et al., *Gapped BLAST and PSI-BLAST: a new generation of protein database search programs.* Nucleic Acids Res, 1997. **25** (17): p. 3389-402.

13. Jaroszewski, L., et al., *FFAS server: novel features and applications.* Nucleic Acids Res, 2011. **39** (Web Server issue): p. W38-44.

14. Zimmermann, L., et al., *A Completely Reimplemented MPI Bioinformatics Toolkit with a New HHpred Server at its Core.* J Mol Biol, 2018. **430** (15): p. 2237-2243.

15. Holm, L., *Using Dali for Protein Structure Comparison.*Methods Mol Biol, 2020. **2112** : p. 29-42.

16. Ye, Y. and A. Godzik, *FATCAT: a web server for flexible structure comparison and structure similarity searching.* Nucleic Acids Res, 2004. **32** (Web Server issue): p. W582-5.

17. Tosi, T., et al., *Structures of the tumor necrosis factor alpha inducing protein Tipalpha: a novel virulence factor from Helicobacter pylori.* FEBS Lett, 2009. **583** (10): p. 1581-5.

18. Anantharaman, V. and L. Aravind, *The SHS2 module is a common structural theme in functionally diverse protein groups, like Rpb7p, FtsA, GyrI, and MTH1598/TM1083 superfamilies.* Proteins, 2004.**56** (4): p. 795-807.

19. Quarantini, S., L. Cendron, and G. Zanotti, *Crystal structure of the secreted protein HP1454 from the human pathogen Helicobacter pylori.* Proteins, 2014. **82** (10): p. 2868-73.

20. Vallese, F., et al., *Helicobacter pylori antigenic Lpp20 is a structural homologue of Tipalpha and promotes epithelial-mesenchymal transition.* Biochim Biophys Acta Gen Subj, 2017. **1861** (12): p. 3263-3271.

21. Kostrzynska, M., et al., *Molecular characterization of a conserved 20-kilodalton membrane-associated lipoprotein antigen of Helicobacter pylori.* J Bacteriol, 1994. **176** (19): p. 5938-48.

22. Gao, M., et al., *Crystal structure of TNF-alpha-inducing protein from Helicobacter pylori in active form reveals the intrinsic molecular flexibility for unique DNA-binding.* PLoS One, 2012.**7** (7): p. e41871.

23. Arockiasamy, A., et al., *Crystal structure of calcium dodecin (Rv0379), from Mycobacterium tuberculosis with a unique calcium-binding site.* Protein Sci, 2011. **20** (5): p. 827-33.

24. Walters, K.A., et al., *Francisella tularensis subsp. tularensis induces a unique pulmonary inflammatory response: role of bacterial gene expression in temporal regulation of host defense responses.* PLoS One, 2013. **8** (5): p. e62412.

25. Kwon, K. and S.N. Peterson, *High-Throughput Cloning for Biophysical Applications* in *Structural Genomics and Drug Discovery: Methods and Protocols* , W.F. Anderson, Editor. 2014, Springer: New York.

26. Millard, C.S., et al., *A less laborious approach to the high-throughput production of recombinant proteins in Escherichia coli using 2-liter plastic bottles.* Protein Expr Purif, 2003.**29** (2): p. 311-20.

27. Shuvalova, L., *Parallel protein purification.* Methods Mol Biol, 2014. **1140** : p. 137-43.

28. Klancher, C.A., et al., *The ChiS-Family DNA-Binding Domain Contains a Cryptic Helix-Turn-Helix Variant.* mBio, 2021.**12** (2).

29. Schrödinger, L., *The PyMOL Molecular Graphics System, Version 2.0 Schrödinger, LLC.* 2020.

30. Pettersen, E.F., et al., *UCSF Chimera–a visualization system for exploratory research and analysis.* J Comput Chem, 2004.**25** (13): p. 1605-12.

31. Li, Z., et al., *POSA: a user-driven, interactive multiple protein structure alignment server.* Nucleic Acids Res, 2014.**42** (Web Server issue): p. W240-5.

32. Potter, S.C., et al., *HMMER web server: 2018 update.* Nucleic Acids Res, 2018. **46** (W1): p. W200-W204.