Exploration of peptide structures and generation of therapeutic peptide candidates using Protein Language Model

Interdisciplinary Team Project

Mateusz Chojnacki, Younginn Park, Łukasz Milewski, Hubert Wąsiewicz
Team Warsaw 3

This project aims to explore therapeutic peptides targeting the SARS-CoV-2 nsp13 helicase, crucial for viral replication with use of ProtGPT2 to propose novel peptide sequences. Our findings lay the groundwork for further investigations into peptide-based therapies against SARS-CoV-2, providing insights into potential drug targets.

1 Methods

1.1 Peptide Base Expansion and Docking Analysis

Despite minimal studies on the subject, we found one relevant paper[1] focusing on peptide drugs targeting the SARS-CoV-2 helicase nsp13 providing 45 potential peptides (P1-P45). To expand our peptide base, we employed advanced search options in the RCSB PDB database[2] to identify peptides with structural similarity to P1-P45, what yielded an additional 27 unique peptides (K1-K27). Before conducting docking procedures, peptides with undetermined amino acids (X) at the C-end or N-end of their sequences were removed. To confirm the binding of peptides to helicase nsp13, we utilized the HPepDock[3] and CABS-dock[4] docking algorithms, which showed that each peptide bound to at least one of the three potential binding sites, with a focus on the first or/and second site.

The SARS-CoV-2 helicase nps13[5, 6, 7] contains 5 domains, starting from N-terminal: ZBD - zinc-binding domain, S - stalk domain, 1B - β-domain, 1A - catalytic "RecA1 like" helicase domain and 2A - catalytic "RecA2 like" helicase domain. Previous studies[5,6] indicate two crucial binding sites essential for the replication/transcription process. The first site, binding ATP, is situated between the 1A and 2A domains, while the second site, binding the 5'-end of the substrate RNA, resides in the pocket between the 1A, 2A, and 1B domains. These binding sites exhibit strong conservation across the coronavirus family[5], making them potential targets for therapeutic peptides. Additionally, a third potential target pocket between ZBD, S, and 1B domains was identified, often docked by peptides from our dataset and marked as a potential allosteric site.

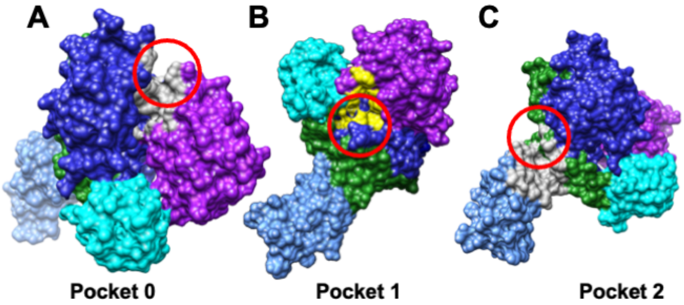

Figure 1: Possible peptide binding sites of nsp13 helicase. A) Pocket '0' involved in ATP binding, B) pocket '1' involved in binding 5 ′-end of the substrate RNA, C) pocket '2' which can be possible allosteric binding site

1.2 Sequence Generation with ProtGPT2

In recent years, there has been remarkable progress in natural language processing (NLP), largely driven by the emergence of large pre-trained language models. In our work, we used ProtGPT2[8], an autoregressive Transformer model with 738 million parameters, designed to generate *de novo* protein sequences at a high throughput. The model was trained on approximately 50 million non-annotated sequences spanning the entire known protein space[9]. We fed pre-trained model with P and K peptides, treated as a 'context' for the model to base its generation. These known peptides were appended with new amino acids being 'appropriate' for given 'context'. In evaluating the generated sequences, some key metrics were applied. These included hydrophobicity measurements, which were calculated using the grand average of hydropathy (GRAVY)[10], assessing the balance between hydrophobic and hydrophilic properties of the amino acids in the chain. Metrics like instability index[11] and isoelectric point (pI) also provided crucial insights for drug design. For instance, any value of instability index above 40 is said to imply instability in a test tube, while the isoelectric point informs about the pH of a solution at which the net charge of a peptide becomes zero[12].

Table 1: Sequences generated with ProtGPT2 with selected preliminary metrics (pI - isoelectric point, II - instability index, gravy - grand average of hydropathicity index, ppl - perplexity)

Alias	**pI**	**II**	**gravy**	**ppl**	**Sequence**
G1	6.46	21.70	-0.48	844.69	SLPYPFIWGNQMWMLTWPDHR
G2	6.74	32.63	-0.56	868.12	HMWPGDIKPAAVSRDLSQ
G3	6.92	27.30	0.21	904.70	IIVTQTMKSGDVSVILHQIHYKAD
G4	6.06	19.66	-0.38	1007.95	WNPADYGGIKPLLTETNIVGKY
G5	7.84	25.80	-0.41	1020.43	GCCSDPLCAWRCHAGRCGRD
G6	7.94	35.50	0.74	1063.46	CKFFWATYTSCCLSGGNLGIFVPS
G7	6.22	18.45	-0.37	1089.33	LSITENGEFKPLGFQFSQKSIEKV
G8	6.77	29.40	0.18	1100.54	LVGPTIWRAALLESAPRHAAE
G9	7.82	11.31	-0.03	1200.32	GCCSDPRCAWRCYGCLS
G10	6.80	35.61	0.29	1287.75	ALKIPISKIYIDSHSVLSPE
G11	6.75	35.87	0.02	1371.14	LHTPLPLTRRDKALLDDALSLFG
G12	6.21	39.74	-0.47	1400.99	GWLEPLLARPWLIVGRDQRGVMTRPYDEG
G13	6.91	14.87	-0.71	1567.13	HEGFTSDFRNPQHAFGSLMCRFNT
G14	7.02	27.67	-0.31	1689.99	LTFQHNFQTHRGHEVGSAQGFTAILW
G15	6.05	34.96	0.60	1731.80	YCKFEWATFAKSCAFPVDGLSFPFFGI
G16	6.00	33.07	-0.33	1800.79	QIPTVNNLKVSEPFTP
G17	6.12	6.10	-0.03	1831.41	GLDIQKVKDMEQLLTQVRLSI
G18	6.74	27.94	-0.04	1927.21	VLEKYKDVIMNSSSLLEHIATGIKKFE
G19	6.40	3.73	-0.26	1964.08	TLPFHSVIYVDSATGQTWTGNR
G20	6.21	37.61	-0.89	2220.56	GYDPETGTWGRRMTLFTPDSRAEVAAR

2 Docking Results

In comparing various docking methods, including HpepDock, CABS-dock, and Alphafold, it becomes evident that the choice of method significantly influences the outcome of peptide docking experiments. Building upon the previous analysis of peptide docking, particularly focusing on peptides from group G, we delve into the specifics of each method's performance. Both HpepDock and CABS-dock exhibited successful docking of G peptides to at least two binding sites (sites 1 and 2). For every G peptide, number of models docked to specific binding pockets is shown in Table 2. In Fig. 2 is shown example of CABS-dock docking result for G13 peptide, which part is in the binding pocket '0' (ATP binding pocket, as it is described in subsection 1.2). This outcome underscores the reliability and versatility of these methods in accommodating different peptide conformations and binding scenarios. However, the results took a different turn when employing Alphafold 2 Multimer for docking.

Despite its renowned capabilities in protein structure prediction, Alphafold encountered challenges in docking G peptides effectively. None of the peptides were docked to any of the three pockets, and the predicted plDDT values ranged disappointingly between 10-20%. This discrepancy in performance raises intriguing questions about the suitability of Alphafold for peptide docking tasks, especially when dealing with shorter peptide sequences.

The discrepancy in performance among the methods prompts further investigation into the underlying factors influencing their efficacy. While HpepDock and CABS-dock demonstrated proficiency in handling the peptides' structural complexities, Alphafold's limitations in accurately predicting peptide folding could be attributed to several factors. One plausible explanation is the inherent difficulty in simulating the folding dynamics of short peptides within the constraints of Alphafold's algorithms. Thus, while the outcomes of docking experiments may vary depending on the selected method, it's essential to consider the method's strengths and limitations in the context of the specific peptide sequences and binding scenarios under investigation. This comparative analysis highlights the importance of employing a diverse array of computational tools and methodologies to gain a comprehensive understanding of peptide-protein interactions.

Table 2: Number of G peptide models or clusters that docked to specific binding pockets (described in subsection 1.1) from docking runs from HPepDock and CABS-dock to helicase 6zsl from 10 results with the best binding score. Numbers in brackets next to results from Cabs-dock show total number of models from top 1000 supporting specific binding pockets.

Alias	HpepDock			CABS-dock		
	0	1	2	0	1	2
G1		1	6	3 (342)	5 (511)	1 (42)
G2	1	3	3	4 (382)	2 (255)	1 (93)
G3	1	1	8	7 (677)	1 (123)	
G4	1	1	4	1 (129)	3 (290)	4 (272)
G5	2	3	4	1 (123)	6 (511)	1 (134)
G6	2	1	7	2 (195)	3 (222)	5 (388)
G7		2	6	5 (425)	3 (374)	
G8		1	7	4 (308)	4 (404)	
G9	2	1	7	6 (459)	1 (98)	
G10		3	3	4 (308)	4 (481)	1 (149)
G11		1	8	4 (391)	4 (418)	1 (70)
G12		3	5	6 (617)	4 (383)	
G13	1	3	6	7 (690)		
G14	2	3	3	3 (266)	2 (200)	1 (147)
G15		2	6	3 (272)	5 (400)	1 (100)
G16	2	1	7	4 (329)	2 (263)	
G17	1	1	7	3 (338)	2 (144)	2 (243)
G18	1		4	4 (397)	3 (373)	1 (134)
G19		1	8	5 (438)	5 (562)	
G20	1	2	7	8 (699)	2 (201)	

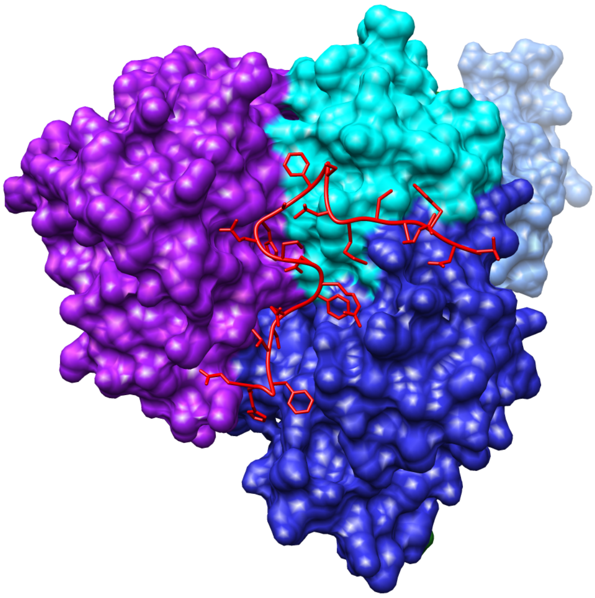

Figure 2: Visulization of G13 peptide docked to ATP binding pocket '0' in CABS-dock software.

3 Discussion

When assessed using Alphafold, our designed peptides exhibited challenges in effectively docking onto the target protein. Alphafold's predictions suggested that the designed peptides struggled to bind to the intended binding sites on the nsp13 helicase, indicating potential limitations in their structural compatibility. Although our designed peptides encountered difficulties in docking onto the target protein when assessed using Alphafold, the results obtained from CABS-dock were notably more promising. CABS-dock's ability to generate multiple docking clusters allowed for a more comprehensive exploration of peptide-protein interactions. This led to the identification of docking clusters indicating successful binding of our designed peptides to specific pockets. The dominance of clusters docking into a single pocket typically indicates that >600 out of 1000 simulations successfully matched a peptide there. Among all 20 G peptides, results from Cabs-dock suggest peptides G3, G13 and G20 as the most likely candidates for helicase nsp13 inhibitors, which could be base for future analysis. Furthermore, considering the inherent limitations of Hpepdock, which favors docking onto exposed regions of rigid proteins, it became evident that relying solely on a single docking tool may lead to biased results. These findings underscore the critical importance of employing a diverse range of computational tools for peptide design and evaluation.

References

[1] H. Pradeep, U. Najma, and H. S. Aparna, "Milk Peptides as Novel Multi-Targeted Therapeutic Candidates for SARS-CoV2," *Protein J*, vol. 40, no. 3, pp. 310–327, Jun. 2021. [Online]. Available: https://doi.org/10.1007/s10930-021-09983-8

[2] "RCSB PDB." [Online]. Available: https://www.rcsb.org/

[3] P. Zhou, B. Jin, H. Li, and S.-Y. Huang, "HPEPDOCK: a web server for blind peptide–protein docking based on a hierarchical algorithm," *Nucleic Acids Research*, vol. 46, no. W1, pp. W443–W450, Jul. 2018. [Online]. Available: https://doi.org/10.1093/nar/gky357

[4] M. Blaszczyk, M. P. Ciemny, A. Kolinski, M. Kurcinski, and S. Kmiecik, "Protein–peptide docking using CABS-dock and contact information," *Briefings in Bioinformatics*, vol. 20, no. 6, pp. 2299–2305, Nov. 2019. [Online]. Available: https://doi.org/10.1093/bib/bby080

[5] J. A. Newman, A. Douangamath, S. Yadzani, Y. Yosaatmadja, A. Aimon, J. Brandão-Neto, L. Dunnett, T. Gorrie-stone, R. Skyner, D. Fearon, M. Schapira, F. von Delft, and O. Gileadi, "Structure, mechanism and crystallographic fragment screening of the SARS-CoV-2 NSP13 helicase," *Nat Commun*, vol. 12, no. 1, p. 4848, Aug. 2021, number: 1 Publisher: Nature Publishing Group. [Online]. Available: https://www.nature.com/articles/s41467-021-25166-6

[6] J. Chen, Q. Wang, B. Malone, E. Llewellyn, Y. Pechersky, K. Maruthi, E. T. Eng, J. K. Perry, E. A. Campbell, D. E. Shaw, and S. A. Darst, "Ensemble cryo-EM reveals conformational states of the nsp13 helicase in the SARS-CoV-2 helicase replication–transcription complex," *Nat Struct Mol Biol*, vol. 29, no. 3, pp. 250–260, Mar. 2022, number: 3 Publisher: Nature Publishing Group. [Online]. Available: https://www.nature.com/articles/s41594-022-00734-6

[7] K. J. Mickolajczyk, P. M. M. Shelton, M. Grasso, X. Cao, S. E. Warrington, A. Aher, S. Liu, and T. M. Kapoor, "Force-dependent stimulation of RNA unwinding by SARS-CoV-2 nsp13 helicase," *Biophys J*, vol. 120, no. 6, pp. 1020–1030, Mar. 2021.

[8] N. Ferruz, S. Schmidt, and B. Höcker, "ProtGPT2 is a deep unsupervised language model for protein design," *Nat Commun*, vol. 13, no. 1, p. 4348, Jul. 2022, number: 1 Publisher: Nature Publishing Group. [Online]. Available: https://www.nature.com/articles/s41467-022-32007-7

[9] "nferruz/UR50_2021_04." [Online]. Available: https://huggingface.co/datasets/nferruz/UR50_2021_04

[10] J. Kyte and R. F. Doolittle, "A simple method for displaying the hydropathic character of a protein," *Journal of Molecular Biology*, vol. 157, no. 1, pp. 105–132, May 1982. [Online]. Available: https://www.sciencedirect.com/science/article/pii/0022283682905150

[11] K. Guruprasad, B. Reddy, and M. W. Pandit, "Correlation between stability of a protein and its dipeptide composition: a novel approach for predicting in vivo stability of a protein from its primary sequence," *Protein Engineering, Design and Selection*, vol. 4, no. 2, pp. 155–161, Dec. 1990. [Online]. Available: https://doi.org/10.1093/protein/4.2.155

[12] C.-H. Shen, "Chapter 8 - Extraction and purification of proteins," in *Diagnostic Molecular Biology (Second Edition)*, C.-H. Shen, Ed. Academic Press, Jan. 2023, pp. 209–229. [Online]. Available: https://www.sciencedirect.com/science/article/pii/B9780323917889000077