Can large scale structural modeling be applied to peptide screening and deorphanization?

Shaima Hashem[a], Jérôme Leprince[b], Pierre Tuffery[a]

[a]BFA, Université Paris Cite, CNRS UMR 8251, Inserm U1133, Paris, France
[b]University Rouen Normandie, Inserm, NorDiC UMR 1239, F-76000 Rouen, France

Contact: pierre.tuffery@u-paris.fr

https://doi.org/10.17952/37EPS.2024.P2280

Background and objectives

– Peptides are ubiquitous in all living organisms and associated with varied fundamental physiological and biochemical processes. They are also becoming valuable therapeutic candidates, especially to modulate targeted protein-protein interactions.
– Thanks to the progress of techniques such as mass spectrometry, or large-scale genome analysis, more and more endogenous peptides are identified but frequently, their targets remain unknown. Their identification is referred to as the ***peptide deorphanization problem (1 peptide, n targets)***. It has motivated efforts in the past years, either from the experimental [1] or theoretical side [2].
– Conversely, for peptide therapeutic development, the target is usually known and one wishes to identify the best peptide targeting it among a collection of candidates. This is referred to as the ***peptide screening problem (1 target, n peptides)***.
– Recent advances in deep learning-based protein structure prediction [3,4] suggest it is possible to predict the effective interaction of protein (and peptides ?) thanks to their associated confidence scores, which motivates the development for in silico protocols to address peptide screening and deorphanization (e.g. [2]).

Here, our objectives are:
1: assess Alphafold2 performance for peptide deorphanization and screening for known binders of G Protein Coupled Receptors (GPCRs).
2: assess Alphafold2 performance for the ranking of peptides identified *in vitro* using PEPscan.

Materials

– **GPCR test set (GPCRts)**: It consists of 366 GPCR sequences corresponding to sequences extracted from the IUPHAR resource [5] on date of September 22nd, 2022.
– **GPCR peptide binders (GPCRpb)**: a set of 14 peptides known as GPCR binders, where the GPCR belongs to the GPCR test set (see Table 1)
– **PEPscan dataset**: a collection of peptides identified in vitro using PEPscan, scanning 7 protein complexes of known structures, as in [6].

Methods

To evaluate if a peptide interacts with a protein, we generate 3D models using the Alphafold2 framework [3], and analyze the scores associated with the complexes. We assume that the best scores will make it possible to identify the best pairs of partners.

– **Peptide-protein complex generation**: We have used Colabfold [7] version 1.5, using *multimer_v3* weights, enabling dropout, using 5 recycles and 5 seeds (resulting in 25 models). For the PEPScan analyses, the number of seed was increased to 20 (100 models). Given the size of some sequences, calculation have been performed using A100 GPUs thanks to the RPBS platform [8].

– **Peptide-protein complex scoring:** For each model we have then extracted the alphafold's **ipTM** (observed best score in [2]), the **pDockQ** and **pDockQ2** scores.

– **Deorphaning**: For each peptide of the **GPCRpb** set, we have considered the models generated with each member of the GPCRts collection (25 x 366 = 9150 models), and ranked them according to several confidence indexes (Alphafold2' ipTM, pDockQ and pDockQ2 [9]). The **Comb** score corresponds to a combination of the rankings obtained using ipTM and pDockQ2, with the assumption that if both scores rank a GPCR well, then it is more likely to truly interact with a peptide. In practice, we average the ranks (between 1 and 366) obtained using the ipTM and pDockQ2, but cases for which at least one score ranks the GPCR at a position greater than a value (200 here) are considered as too unlikely to be a binder, and the GPCR is removed from the list.

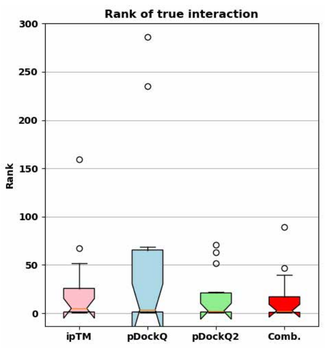
– **Screening**: For each peptide of the **GPCRpb** set, we have considered the models generated with each of their true binders in the GPCRts collection (25 x 14 = 350 models), and ranked them as for the deorphaning experiment.

– **PEPscan candidate validation**: Assess the ranking of the candidate peptide at protein-protein interface among all candidates identified using PEPscan for the target protein.

Name	Sequence	Target
alphaMSH	SYSMEHFRWGKPV	Q01727
angiotensin II	FDPRHVIY* / DRVYIHPF (*: scramble)	P35351
apelin	KFRRQRPRLSHKGPMPF	Q9JHG3
bradykinin	RPPGFSPFR	P25023/ P97583
Cholecystokinin	DYMGWMDF	P30551
CRH	SEEPPISLDLTFHLLREVLEMARA EQLAQQAHSNRKLMEII	P47866
Galanin	GWTLNSAGYLLGPHAIDNHRSFS DKHGLT	O08726
Ghrelin	GSSFLSPEHQKAQQRKESKKPP AKLQPR	O08725
glucagon	HSQGTFTSDYSKYLDSRRAQDF VQWLMNT	P30082
NeuromedinB	GNLWATGHFM	P24053
NeuromedinU	FFLFRPRN	Q9JJI5
NeuropeptideS	SFRNGVGSGV / SFRNGVGSGVKKTSFRRAKQ	P0C0L6
NeuropeptideW	WYKHVASPRYHTVGRASGLLMG LRRSPYLW	P48146
Oxytocin	CYIQNCPLG	P70536

Table 1: *GPCR binders (GPCRpb), with their sequences and the Uniprot Id of their true receptor*

Results (1) Peptide deorphanization



	ipTM	pDockQ	pDockQ2	Comb
mean	24.8	55.2	17.6	16.0
median	5	3.5	2.0	2.5
top10	9	9	9	9

Table 2: *Mean, median ranks of the true binders. Top10: number of targets for which the true binder is ranked 10 or less (over 366).*

Figure 1: *Distribution of ranks for different scores.*

Main observations:
- The true binders are in the best 10% for all cases but 2 (86%)
- The protocol ranks in the top10 the true receptor for 9 out of 14 cases (64%), independently of the score.
- The pDockQ2 improves largely over the pDockQ

- The different scores rank poorly the same targets. For alphaMSH and NeuropeptideS, the rank of the true receptor is of more than 40. For Ghrelin and NeuromedinU, the scores vary the most, one of the score only ranking the true target in the best 10%.
- Consequently, combining the rankings obtained using the ipTM and the pDockQ2 (Comb) improves only marginally the performance.

Results (2) Peptide screening

Name	Rank			Target
	ipTM	pDockQ2	Comb	
alphaMSH	3	3	2	Q01727
angiotensin II	2	2	2	P35351
apelin	2	1	1	Q9JHG3
bradykinin	1	1	1	P25023
Cholecystokinin	1	1	1	P30551
CRH	1	1	1	P47866
Galanin	1	1	1	O08726
Ghrelin	**12**	**4**	**9**	O08725
glucagon	1	1	1	P30082
NeuromedinB	1	1	1	P24053
NeuromedinU	2	4	2	Q9JJI5
NeuropeptideS	**4**	**8**	**5**	P0C0L6
NeuropeptideW	2	1	1	P48146
Oxytocin	1	2	1	P70536

Table 3: *Rank of the true binder according to different scores (see methods)*

Main observations:
- The natural binders are ranked in the top 3 for all cases but 2.
- Ranking heterogeneity depending on the scores is observed for the 2 poorly predicted targets.
- The pDockQ2 values are usually very low, suggesting pDockQ2 is irrelevant for peptides/protein interactions.
- The number of cases is however too limited to draw general conclusions.

Results (3) *In-Silico* PEPscan validation

Motivation: PEPscan is an *in vitro* approach to identify candidate interfering peptides to modulate protein-protein interactions. In previous studies, we have shown that it returns a limited number of candidates among which some are located in the vicinity of protein-protein interface. Here, we assess if Alphafold2 is able to discriminate the best interfering candidate among all the candidates.

PDB	Name	#ip	Rank		#if	Name	#ip	Rank		#if
			ipTM	pDQ2				ipTM	pDQ2	
1e96	NCF2	7	N/A	N/A	N/A	Rac1	4	**1**	**1**	4
1ktz	TGFβ3	12	12	12	9	TGFR2	8	**1**	**1**	0
1m27	SAP	4	3	3	0	FynSH3	4	2	**1**	0
1lfd	RalGDS	2	2	**1**	2	Ras	8	3	3	2
2j0t	MMP-1	5	3	4	3	TIMP1	2	**1**	**1**	2
3bx7	NGAL	8	2	2	6	CTLA-4	2	**1**	**1**	2
1eer	EPO	1	N/A	N/A	N/A	EPOR	8	4	3	8

Table 4: *Rank of the peptide at interface according to different scores among candidate peptides identified using PEPscan. Lines present the results for the each partner of the 3D complex of the PDB entry. N/A: no peptide identified at the interface. #ip: number of peptides identified using PEPscan #if: number of peptides contacting the experimental interface over #ip*

Main observations:
- Scores best rank peptides at complex interface in only 6 cases over 12 (50%)
- For 9 cases out of 12, the poses overlap the experimental binding site.
- All peptides tend to bind to the same position on protein, suggesting a very poor specificity to identify the binding interface.

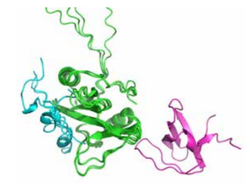


Figure 3: SAP/FynSH3.
All peptides (cyan) target the same spot, not at the experimental interface

Discussion/Conclusions

Present analyses rely on the assumption that the scoring of models generated using Alphafold2 is able to discriminate models containing different entities. Note that peptides are free to bind to any place on the protein surface.
- For the GPCR deorphanization, varied receptors are confronted with a given peptide, and the results are rather satisfactory, consistently with the results in [2].
- For the GPCR peptide screening, different peptides are confronted with a given receptor. Here again, the results are encouraging.
- For the PEPscan candidate validation, the results are more ambiguous. In some cases, no pose of the peptide targets the protein-protein interface. Furthermore, all peptides tend to bind the same location on protein surface (see Figure 3), which is possibly not the desired one.

Overall, the specificity of the protein-peptide modeling is at stake, and our results suggest that it is really poor.
- When peptides are confronted with the same protein, it is possible the scores lose effectiveness due to the smaller size of the peptides compared to their target protein.
- Ranking heterogeneity is possibly an indicator of the poor performance of the scoring.

Why such a difference between PEPscan candidate validation and GPCR peptide screening ? Different hypotheses under investigation are:
- All peptides confronted with GPCRs are naturally occurring peptides whereas PEPscan candidates might exhibit different properties and correspond to a class of peptides not learnt by AlphaFold2.
- PEPscan peptides that correspond to protein fragments could be more hydrophobic than natural peptides selected during evolution, and thus be more sticky. It is however surprizing that in some cases such peptides do not target the protein-protein interface which should be more hydrophobic.

Overall our results suggest the need to develop peptide-protein specific scores to discriminate the true binders from the incorrect ones. Another challenge is that of protein-peptide binding specificity, but this supposes to go back to AlphaFold2 internal prediction scheme, especially to increase the diversity of the poses.

Acknowledgements/Funding

The authors thank R. Guerois and J. Andreani for useful discussions.

This project is supported by the MSD Avenir foundation, MSD being a trade name of Merck & Co.

References

[1] Foster, S. R., Hauser, A. S., Vedel, L., Strachan, R. T., Huang, X. P., Gavin, A. C., ... & Gloriam, D. E. (2019). Discovery of human signaling systems: pairing peptides to G protein-coupled receptors. Cell, 179(4), 895-908.
[2] Teufel, F., Refsgaard, J. C., Kasimova, M. A., Deibler, K., Madsen, C. T., Stahlhut, C., ... & Madsen, D. (2023). Deorphanizing peptides using structure prediction. Journal of chemical information and modeling, 63(9), 2651-2655.
[3] Jumper J, Evans R, Pritzel A. et al. (2021) Highly accurate protein structure prediction with AlphaFold. *Nature* 596, 583–589
[4] Bryant, P., Pozzati, G., & Elofsson, A. (2022). Improved prediction of protein-protein interactions using AlphaFold2. Nature communications, 13(1), 1265.
[5] Harding SD, Armstrong JF, Faccenda E, Southan C, Alexander SPH, Davenport AP, Spedding M, Davies JA. (2023) The IUPHAR/BPS Guide to PHARMACOLOGY in 2024. Nucl. Acids Res. 2024; 52(D1):D1438-D1449.
[6] Rebollo, A., Fliedel, L., & Tuffery, P. (2022). PEPscan: A Broad Spectrum Approach for the Characterization of Protein-Binder Interactions?. Biomolecules, 12(2), 178.
[7] Mirdita M, Schütze K, Moriwaki Y, Heo L, Ovchinnikov S and Steinegger M. (2022) ColabFold: Making protein folding accessible to all. *Nature Methods*. doi: 10.1038/s41592-022-01488-1.
[8] Alland, C., Moreews, F., Boens, D., Carpentier, M., Chiusa, S., Lonquety, M., ... & Tufféry, P. (2005). RPBS: a web resource for structural bioinformatics. Nucleic acids research, 33(suppl_2), W44-W49.
[9] Zhu, W., Shenoy, A., Kundrotas, P., & Elofsson, A. (2023). Evaluation of AlphaFold-Multimer prediction on multi-chain protein complexes. Bioinformatics, 39(7), btad424.