Translating L-peptides into non-canonical linear and macrocyclic peptides

Anonymous Author(s)
Affiliation
Address
email

1 Introduction

Protein-protein interactions (PPIs) are critical to several biological functions, from regulating metabolic activities to several disease-causing pathways [1]. With limited success in using small molecules and large biologics to disrupt clinically relevant PPIs, peptides continue to garner further therapeutic interest [2, 3]. As of 2021, around 80 peptides were approved for clinical use worldwide, apart from more than 160 in various stages of clinical trials [4].

Peptides constitute of amino acid monomers, arranged usually in a linear fashion, but also found as macrocycles and other non-linear topologies [5, 6]. The chemical diversity arising from both the amino acid composition and their spatial arrangement [7], synthetic accessibility [8, 9], and potential for cell penetration [10, 11], have made them increasingly sought after modalities in drug discovery.

High-throughput screening for peptides that bind to desired target proteins is often carried out by using libraries made by genetically-encoded or chemical synthesis approaches, such as phage display [12, 13], mRNA display [14, 15], or affinity selection-mass spectrometry [16, 17]. A key challenge in genetically-encoded libraries for screening lies in the limitation of this approach to generate L-peptides, since cells can produce peptides with naturally occurring amino acids. Recent approaches have resulted in libraries with non-canonical amino acids, however, a majority of drug discovery efforts still uses L-peptide libraries [18, 19].

While L-peptide-based libraries are the workhorse of high throughput screening, peptides with non-canonical or unnatural amino acids have been shown to have similar or higher activity and proteolytically more stable [20, 21]. Thus, it is highly desirable to obtain non-canonical variants for L-peptides. Unfortunately, flipping the stereochemistry for each amino acid, and obtaining a D-peptide analogue of the L-peptide, does not work for most cases, necessitating further investigation into developing chemically similar non-canonical variants.

In this work, we developed a method to translate L-peptides into non-canonical linear and macrocylic peptides. We used a genetic algorithm with a hierarchical dual-objective function, mutating the seed peptide sequence to be chemically similar to the reference L-peptide, while ensuring that the binding affinity remains same or is higher than the seed sequence. Using DRD2 kinase inhibitor as the target protein, we obtained both linear and macrocyclic peptides with non-canonical amino acids with higher activity than observed by random sampling of L-peptides.

2 Results and Discussion

2.1 L-peptides data set generation

To simulate a data set of L-peptides, as the one that can be obtained from macrophage display, we randomly sampled individual L-amino acids, dipeptides, and tripeptides (Figure 1A). We docked these peptides against DRD2 kinase inhibitor to calculate their respective binding affinities characterized

Submitted to 36th Conference on Neural Information Processing Systems (NeurIPS 2022). Do not distribute.

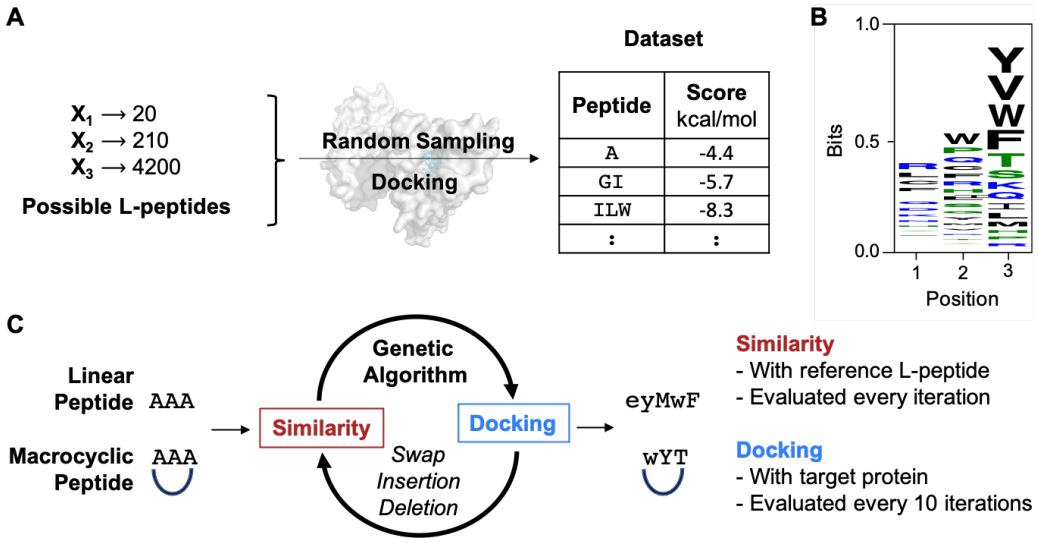

Figure 1: **Outline of the approach for translating L-peptides into non-canonical linear and macrocyclic peptides. A.** The data set was generated by random sampling, and docking of L-peptides to DRD2 kinase inhibitor. **B.** Amino acid positional frequency plot, resulting from the weighted multiple-sequence alignment of peptides in the data set was computed for the peptides in the data set. **C.** An overview of the genetic algorithm involving hierarchical similarity and docking for optimization of the seed sequence is shown.

by the docking scores. From the data set, we selected the tripeptide `YWY`, with a docking score of -10.3 kcal/mol as the reference peptide (or, the hit compound) for further optimization. We used the Python-based DOCKSTRING interface, with AutoDock Vina, for the generation of our initial data set, and for later evaluation during the genetic algorithm experiments [22, 23].

As a method of evaluating how these peptide binders compared with their small molecule analogues, we computed a weighted multiple-sequence alignment of the sequences in the data set [24]. The weights were based on the docking scores, with a higher docking score resulting in a directly proportional weight, and vice-versa. Visualizing the results using an amino acid positional frequency plot, we noted that amino acids with aryl groups, such as `Y`, tyrosine, and `W`, tryptophan, were in abundance (Figure 1B). The results are aligned to the observation of the interacting phenyl groups in small molecules binding to DRD2 kinase [25, 26].

2.2 Representation of peptide

Peptides were converted to molecular structures, and represented using circular fingerprints. Molecular structure representations for both linear and macrocyclic peptides, with L- and D-amino acids, were obtained by converting amino acid sequences to simplified molecular-input line-entry system (SMILES) strings using RDKit [27]. Bit-based and count-based circular fingerprints, with radius 3, size 128 and active chirality, were calculated using RDKit [28].

2.3 Genetic algorithm and optimization

Genetic algorithm was set up with a hierarchical objective of (1) achieving high chemical similarity to the reference peptide, and (2) increasing binding affinity with the target protein (Figure 1C). For mutations, random insertion, deletion, and swapping of amino acids in the seed sequence were performed. The ability to mutate in the discrete monomer space, using a pool of L- and D-amino acids, helped in increasing the diversity of the peptides, without any worry of synthetic accessibility. In the case of macrocyclic peptides, we mutated on the seed peptide string, and then defined a SMILES arbitrary target specification (SMARTS) based intra-molecular cyclization reaction to cyclize the linear peptides by a head-to-tail amide bond.

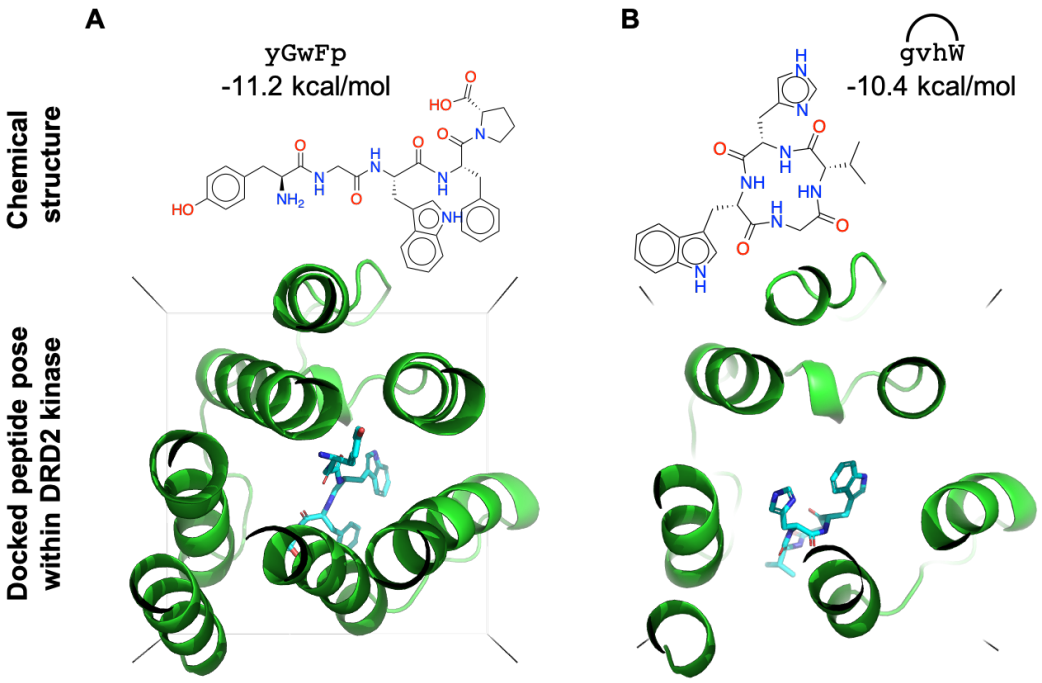

Figure 2: **Optimization for non-canonical linear and macrocyclic peptides.** Sequence information, chemical structure, docking score against DRD2 kinase, and the docked pose of the peptide in the protein are shown for **A.** linear `yGwFp`, and **B.** macrocyclic `gvhW` peptides.

In our experiments, we used a hierarchical approach of optimizing for similarity, followed by optimization for docking scores, motivated by the hypothesis that higher similarity to reference peptide will lead to better docking. Briefly, we approached the problem with the idea that if we could match the chemical similarity of the reference peptide with the mutated sequence, we could optimize further for a better docking score. We often started with simpler peptides, such as `AAA` and `GGG`, since initial experiments seeding with the reference peptide and other peptides with high docking scores did not result in any improvement beyond their initial scores. Additionally, we observed that while keeping the computational capacity the same (Dual-core Intel i5, with 8GB memory), chemical similarity between two peptides could be computed in microseconds, while docking took 1-3 minutes. This observation supported our hypothesis from a computational cost/time standpoint.

2.4 Optimizing for non-canonical linear and macrocyclic peptides

To benchmark different approaches, we set up two distinct experiments, first, optimizing for similarity and docking score, and another optimizing for docking score alone. In the first experiment, the chemical similarity for the mutated peptide to the reference peptide was computed, and the selection for further docking was based on the higher similarity between the seed or mutated peptide to the reference. Similar to the similarity evaluation, if the mutated peptide had a higher docking score, it was carried over to the next iteration, or else, the seed peptide was mutated again. In the second experiment, there was no similarity optimization done. In both cases, the seed peptide was mutated for at least 10 times, before being evaluated for docking. The genetic algorithm was run for 50 iterations, thus, a minimum of 500 mutations of the seed sequence was done for each experiment.

The chemical similarity was computed by evaluating Tanimoto similarity between the fingerprints of the reference and mutated peptides [29]. We noted that similarity between the bit-based fingerprints, with active and inactive bits for different molecular fragments, led to similarity score of 1.0 within 10-20 iterations. Such behavior arises from the bit-based fingerprints capturing individual fragments alone, irrespective of their position and number of occurrences [30]. As an alternative, we used the count-based fingerprints to compute chemical similarity. However, we noted that the docking scores of the resulting peptides at the end of experiments based on different fingerprint types were similar.

We optimized the seed sequences for both linear and macrocyclic peptides with non-canonical amino acids to be similar to the reference peptide YWY with a score of -10.3 kcal/mol, and dock against DRD2 kinase protein. The best linear peptide was yGwFp with a docking score of -11.2 kcal/mol, while the best macrocyclic peptide was gvhW with a docking score of -10.4 kcal/mol (Figure 2). In the optimization run, only the sequence composition could change, while the topology, linear or cyclic, had to remain constant for all iterations. We pursued the optimization for both linear and macrocyclic peptides through the similarity-docking and docking-only objective functions, and benchmarking both bit-based and count-based fingerprints.

For a particular experiment invovling non-canonical linear peptides, we set the seed to tripeptide aaa with a docking score of -7.0 kcal/mol, and the seed of the random generator to 0, and noted how the genetic algorithm proceeded. With the bit-based fingerprints, at 21 iterations, the mutated sequence, yMwFp, with a docking score of -11.2 kcal/mol, surpassed the docking score of the reference peptide sequence, and a Tanimoto similarity of 0.73. In further iterations, the sequence was mutated to yGwFp, although the docking score remained unchanged. With the count-based fingerprints, we noted that the docking score was surpassed at 3 iterations, with sequence yaNFY having a docking score of -10.7 kcal/mol and a Tanimoto similarity of 0.54. Ultimately, the sequence got mutated to yaVFY with a docking score of -11.1 kcal/mol. In multiple experiments, with different seed sequences and different numbers for the random seed generator, we observed similar trends, with the docking score being surpassed sooner when the similarity was computed using count-based fingerprints. We attribute this observation to the finer granularity of count-based fingerprints, and the Tanimoto similarity thereof.

Using the docking-only objective function, without any reference peptide sequence, the seed sequence, aaa, could mutate to RFEaa with a docking score of -9.2 kcal/mol, which is worse than that of the reference sequence. Thus, this experiment underscores the need of a similarity objective function, to guide the evolution of the sequence towards a high-affinity sequence, and then optimize the docking even further.

3 Limitations, Future Work and Conclusion

Our work discussed the development of a genetic algorithm-based method to translate linear L-peptides, commonly obtained through high-throughput library screening, to non-canonical linear and macrocylic peptides. With a dual objective function of similarity matching to a high-affinity reference peptide and maximization of the docking score against the target protein, we were able to obtain non-canonical peptides with better docking scores in less than 30 iterations. In the current state, the work is limited in its applicability to other target proteins. We aim to to include more proteins and evaluate our approach in the near future. Additionally, we look forward to working on experimental data sets to translate L-peptides therein or optimize the non-canonical peptides further. We strongly believe that this approach can accelerate drug discovery efforts, by enabling the development of more stable peptides with similar or higher activity, which can ultimately benefit patients.

References

[1] Irene MA Nooren and Janet M Thornton. Diversity of protein–protein interactions. *The EMBO journal*, 22(14):3486–3492, 2003.

[2] Natia Tsomaia. Peptide therapeutics: targeting the undruggable space. *European journal of medicinal chemistry*, 94:459–470, 2015.

[3] Ziqing Qian, Patrick G Dougherty, and Dehua Pei. Targeting intracellular protein–protein interactions with cell-permeable cyclic peptides. *Current opinion in chemical biology*, 38: 80–86, 2017.

[4] Lei Wang, Nanxi Wang, Wenping Zhang, Xurui Cheng, Zhibin Yan, Gang Shao, Xi Wang, Rui Wang, and Caiyun Fu. Therapeutic peptides: current applications and future directions. *Signal Transduction and Targeted Therapy*, 7(1):1–27, 2022.

[5] Norbert Sewald and Hans-Dieter Jakubke. *Peptides: chemistry and biology*. John Wiley & Sons, 2015.

[6] Sang-Hoon Joo. Cyclic peptides as therapeutic agents and biochemical tools. *Biomolecules & therapeutics*, 20(1):19–26, 2012.

[7] Knut Adermann, Harald John, Ludger Ständker, and Wolf-Georg Forssmann. Exploiting natural peptide diversity: novel research tools and drug leads. *Current opinion in biotechnology*, 15(6): 599–606, 2004.

[8] Somesh Mohapatra, Nina Hartrampf, Mackenzie Poskus, Andrei Loas, Rafael Gomez-Bombarelli, and Bradley L Pentelute. Deep learning for prediction and optimization of fast-flow peptide synthesis. *ACS central science*, 6(12):2277–2286, 2020.

[9] Alexander J Mijalis, Dale A Thomas, Mark D Simon, Andrea Adamo, Ryan Beaumont, Klavs F Jensen, and Bradley L Pentelute. A fully automated flow-based approach for accelerated peptide synthesis. *Nature chemical biology*, 13(5):464–466, 2017.

[10] Eva M López-Vidal, Carly K Schissel, Somesh Mohapatra, Kamela Bellovoda, Chia-Ling Wu, Jenna A Wood, Annika B Malmberg, Andrei Loas, Rafael Gómez-Bombarelli, and Bradley L Pentelute. Deep learning enables discovery of a short nuclear targeting peptide for efficient delivery of antisense oligomers. *JACS Au*, 1(11):2009–2020, 2021.

[11] Carly K Schissel, Somesh Mohapatra, Justin M Wolfe, Colin M Fadzen, Kamela Bellovoda, Chia-Ling Wu, Jenna A Wood, Annika B Malmberg, Andrei Loas, Rafael Gómez-Bombarelli, et al. Deep learning to design nuclear-targeting abiotic miniproteins. *Nature chemistry*, 13(10): 992–1000, 2021.

[12] Chien-Hsun Wu, I-Ju Liu, Ruei-Min Lu, and Han-Chung Wu. Advancement and applications of peptide phage display technology in biomedical science. *Journal of biomedical science*, 23 (1):1–14, 2016.

[13] Gianni Cesareni, Luisa Castagnoli, and Gianluca Cestra. Phage displayed peptide libraries. *Combinatorial Chemistry and High Throughput Screening*, 2:1–18, 1999.

[14] Kristopher Josephson, Alonso Ricardo, and Jack W Szostak. mrna display: from basic principles to macrocycle drug discovery. *Drug Discovery Today*, 19(4):388–399, 2014.

[15] Terry T Takahashi, Ryan J Austin, and Richard W Roberts. mrna display: ligand discovery, interaction analysis and beyond. *Trends in biochemical sciences*, 28(3):159–165, 2003.

[16] Renaud Prudent, D Allen Annis, Peter J Dandliker, Jean-Yves Ortholand, and Didier Roche. Exploring new targets and chemical space with affinity selection-mass spectrometry. *Nature Reviews Chemistry*, 5(1):62–71, 2021.

[17] Anthony J Quartararo, Zachary P Gates, Bente A Somsen, Nina Hartrampf, Xiyun Ye, Arisa Shimada, Yasuhiro Kajihara, Christian Ottmann, and Bradley L Pentelute. Ultra-large chemical libraries for the discovery of high-affinity peptide binders. *Nature communications*, 11(1):1–11, 2020.

[18] Benjamí Oller-Salvia and Jason W Chin. Efficient phage display with multiple distinct non-canonical amino acids using orthogonal ribosome-mediated genetic code expansion. *Angewandte Chemie*, 131(32):10960–10964, 2019.

[19] Titia Rixt Oppewal, Ivar D Jansen, Johan Hekelaar, and Clemens Mayer. A strategy to select macrocyclic peptides featuring asymmetric molecular scaffolds as cyclization units by phage display. *Journal of the American Chemical Society*, 144(8):3644–3652, 2022.

[20] Richard Obexer, Louise J Walport, and Hiroaki Suga. Exploring sequence space: harnessing chemical and biological diversity towards new peptide leads. *Current opinion in chemical biology*, 38:52–61, 2017.

[21] Jr Nestor et al. The medicinal chemistry of peptides. *Current medicinal chemistry*, 16(33): 4399–4418, 2009.

[22] Miguel García-Ortegón, Gregor NC Simm, Austin J Tripp, José Miguel Hernández-Lobato, Andreas Bender, and Sergio Bacallado. Dockstring: easy molecular docking yields better benchmarks for ligand design. *Journal of chemical information and modeling*, 62(15):3486–3502, 2022.

[23] Oleg Trott and Arthur J Olson. Autodock vina: improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading. *Journal of computational chemistry*, 31(2):455–461, 2010.

[24] Robert C Edgar and Serafim Batzoglou. Multiple sequence alignment. *Current opinion in structural biology*, 16(3):368–373, 2006.

[25] Kostas Papadopoulos, Kathryn A Giblin, Jon Paul Janet, Atanas Patronov, and Ola Engkvist. De novo design with deep generative models based on 3d similarity scoring. *Bioorganic & Medicinal Chemistry*, 44:116308, 2021.

[26] Didier Rognan. Binding site similarity search to identify novel target–ligand complexes. *Computational Chemogenomics*, page 171, 2013.

[27] Greg Landrum et al. Rdkit: A software suite for cheminformatics, computational chemistry, and predictive modeling. *Greg Landrum*, 2013.

[28] David Rogers and Mathew Hahn. Extended-connectivity fingerprints. *Journal of chemical information and modeling*, 50(5):742–754, 2010.

[29] Dávid Bajusz, Anita Rácz, and Károly Héberger. Why is tanimoto index an appropriate choice for fingerprint-based similarity calculations? *Journal of cheminformatics*, 7(1):1–13, 2015.

[30] Yuan Wang. *Molecular Complexity Effects and Fingerprint-Based Similarity Search Strategies*. PhD thesis, Universitäts-und Landesbibliothek Bonn, 2009.

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes]
   (c) Did you discuss any potential negative societal impacts of your work? [N/A]
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [N/A]
   (b) Did you include complete proofs of all theoretical results? [N/A]
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [No] Code can be shared upon the final publication of the work. Data and instructions have been provided to reproduce the work.
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes]
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [N/A]
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [Yes]
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [N/A]
   (b) Did you mention the license of the assets? [N/A]
   (c) Did you include any new assets either in the supplemental material or as a URL? [N/A]
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [N/A]
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A]
5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]