Design of bioactive peptides by machine learning using neural network

Naoya Ozawa & Jaekyun Hwang

Authors

Naoya Ozawa He is a member of Suga Laboratory in the Department of Chemistry. He is developing functional peptides by screening of peptides containing unnatural amino acids. In this research, he performed screening and evaluation of peptides.

Jaekyun Hwang He is a member of Watanabe / Minamitani Laboratory in the department of materials engineering. He is developing a novel material searching methodology using first principles calculations and machine learning. In this research, he performed the training of machine learning model with amino acid sequences and candidate screening.

Introduction

Macrocyclic peptides are appealing class of molecules for drug discovery[1,2]. Many drugs work by binding to a specific target protein involved in a disease and thereby modulating its function. Currently, small molecules and antibodies are the major classes of drugs. While small molecules have low production cost, antibodies have high selectivity to various target proteins. Macrocyclic peptides can potentially combine these advantages.

For discovery of bioactive macrocyclic peptides, a screening methodology called the RaPID system had been developed[3]. This methodology allows for construction of a library of macrocyclic peptides with the diversity of 10^{13} in one tube and selective amplification of molecules that bind to a target protein. However, the amplification procedure must be repeated 5–10 times for identification of strong binders, and only a small subset of theoretically possible amino acid sequences (4×10^{16} sequences for 13 residues consisting of 19 kinds of amino acids) can be tested. Moreover, although next-generation sequencing enabled us to read 10^5 sequences from an enriched library, the enormous information from the sequencing has not been fully utilized.

On the other hand, thanks to the rapid growth of calculation speed and understanding of artificial neural network (ANN) models, it had become possible to use powerful deep learning techniques in various fields. Although various approaches had been studied for organic materials, no effective methods had been reported for learning a huge amount of amino acid sequences whose three-dimensional structure was unknown. In this research, each amino acid was described by nine characteristic numerical properties suggested by Kidera *et al.*[4], which were derived by dimensionality reduction of 188 physiochemical properties. For the ANN model, we decided to use a long short-term memory (LSTM) model[5], which was known to be powerful for the arbitrary length of inputs.

In this research, we aimed at designing peptides with high expected affinity while shortening the screening process by utilizing a machine learning model trained with the big data obtained from next-generation sequencing. As the model target protein, we employed the extracellular region of a receptor, Met, which was expressed on cell surface. Because a macrocyclic peptide which bound to the protein had been reported[6], we expected that the protein would be suitable for validation of the methodology developed in this study.

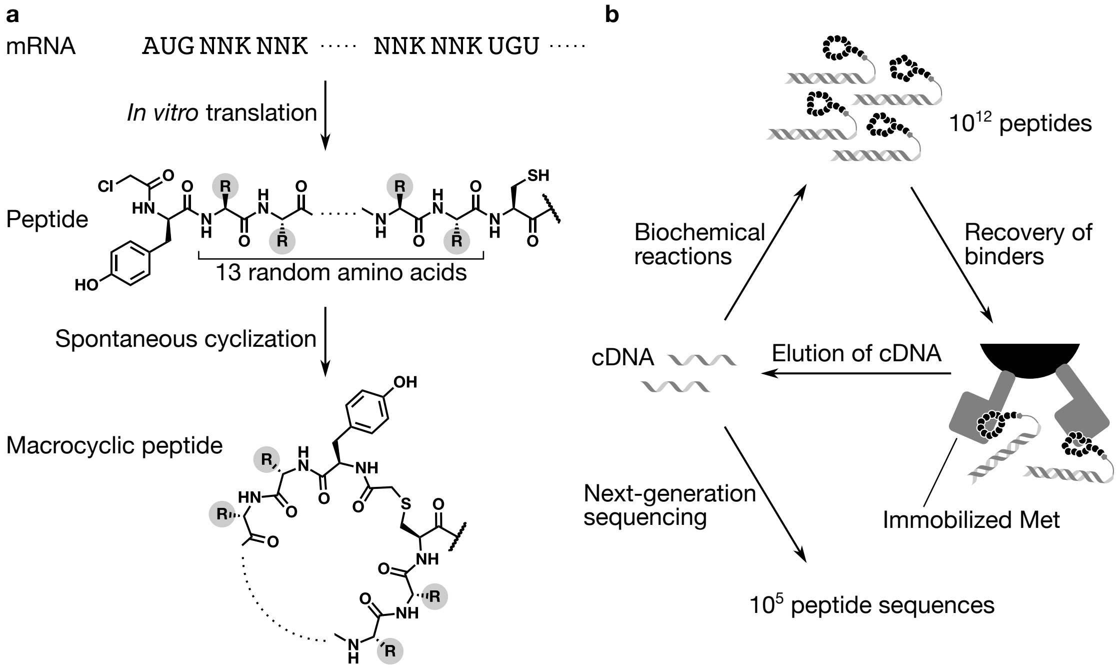

Figure 1. Screening of macrocyclic peptides
(a) Preparation of the macrocyclic peptide library. Macrocyclic peptides with a random sequence of 13 residues were synthesized by *in vitro* translation. (b) Scheme of screening. Peptides binding to immobilized Met were recovered, and their cognate cDNA was analyzed by next-generation sequencing to obtain the sequence data.

Results and discussion

Screening

To prepare a library of macrocyclic peptides, a mixture of mRNA having a random region was subjected to *in vitro* translation (Figure 1a). By using genetic code reprogramming technique, chloroacetyl-D-tyrosine was introduced at the N-terminus of the peptides so that the chloroacetyl group would react with a downstream cysteine to form a macrocyclic structure[7]. The mRNA was designed so that a random sequence of 13 amino acid residues would appear between the N-terminal amino acid and the cysteine. 19 kinds of natural amino acids (excluding methionine) was used for the random sequence. The peptides were fused with mRNA coding for their sequences by means of mRNA display technique[8], and the mRNA was reverse transcribed to make a library of macrocyclic peptides fused with their cognate mRNA/cDNA.

Using this macrocyclic peptide library, screening was performed (Figure 1b). Met immobilized on resin beads via Fc-tag was mixed with a solution of the peptide library containing ~10^{12} molecules. Then, the solution was removed, and the beads were washed to recover peptides binding to beads carrying Met. The cDNA of the binding peptides was eluted by heat to obtain ~10^7 molecules (Sample P1). Because the proportion of the binding peptides in the initial library should be very low, we anticipated that Sample P1 was contaminated with large amount of non-binding peptides (noise). To reduce the proportion of the non-binding peptides, a peptide library was constructed using the eluted cDNA, and another round of screening was performed (Sample P2). We also anticipated that some of the peptides that could bind to the beads carrying Met might bind to the Fc-tag or the resin instead of Met. To obtain data for such peptides, we also performed screening using beads carrying only Fc-tag (Sample N1 and N2). Then, to obtain sequence data of the recovered peptides, the cDNA samples

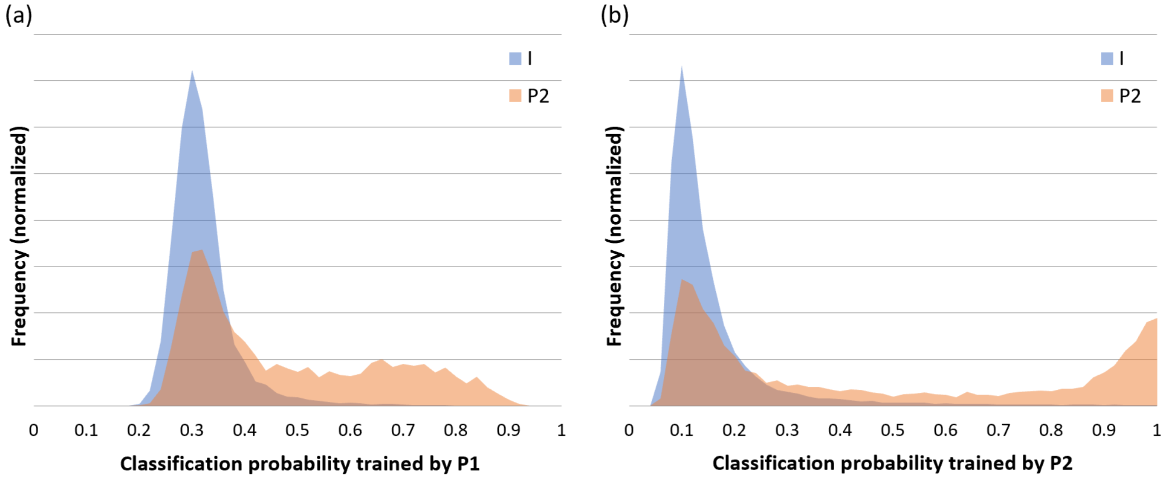

Figure 2. Classification probability distribution of Initial state (I) and 2nd screened state (P2)
The horizontal axis represents the classification probability. The value smaller than 0.5 would be predicted as peptides from I, and larger than 0.5 would be predicted as peptides from P2. Actual peptides from I are shown in blue, and P2 in orange. Each distribution was trained by the 1st screening results (a) or the 2nd screening results (b).

(P1, P2, N1 and N2) were subjected to next-generation sequencing. The initial library (I) was also sequenced. From each sample, ~10^5 sequences were obtained.

Machine learning and design of peptides

Machine learning model by LSTM was designed to accept input data taken from amino acid sequences. Each amino acid was converted into an array of length nine (nine characteristic properties) at each timestep, and the timestep was repeated until the end of amino acid sequence. Note that the number of timesteps corresponds to the length of amino acid sequence. The objective function was set as a three categorical classification problem among before the screening (I), and after the screening (P2, N2). Training was done by minimizing the categorical cross entropy.

Although it was desirable to evaluate the binding affinity of each peptide, since only the sequence information was available, the probability to be classified into P2 was used as a "score" instead of binding affinity. This method would also have a merit that if the P2 score was high, the probability of being classified as N2 should be low, meaning that binding to the resin would be automatically eliminated.

We tried various combinations of training models, but the accuracy of training (guess whether the amino acid sequence is before or after the screening) was no better than 80%. We examined the distribution of the classification probability to find out the reason. Figure 2 shows the predicted probability distribution to be classified as I and P2 by the 10% of test data (not used for model training) of I and P2. The left side of each figure represents the class I, and the right side represents the class P2. However, more than half of the actual class P2 peptides were indistinguishable from class I. In general, if the classification training fails for some reason, the probability distribution is smooth, and the central part is almost overlapped. Since the two peaks appeared in P2, two causes are possible. The first possibility is that the LSTM model could describe only some partial features of the true objective function. The better separation of the two peaks in Figure 2b implies that the training after 2nd screening had captured more features. The second possibility is that half of the screened peptides were initial peptides. Further investigations are needed on this problem.

In the next step, we tried to figure out the score of aMD5, which was reported in the previous study[6] but not used for training the LSTM model. On average, one in 330 random sequences had higher score value than aMD5. This rarity was widely distributed from one in 10 to one in 100,000 depending on the random initial values of the LSTM model. Meanwhile, the training accuracy as shown in Figure 2 was roughly the same for each random initial value. The possible reason for this situation is that the characteristics of the strong binding affinity did not clearly appear in the 2nd screening results. Then the LSTM model randomly filled the missing part of the information, where some of them gave the correct answer (or near the aMD5 answer) and some wrong.

Based on the above results, we tested two strategies for candidate screening. Strategy 1 used the minimum score of 20 LSTM model ensemble that gave a high score on aMD5 from randomly trained LSTMs. This strategy increased the rarity of aMD5 to one in 14,000,000. The disadvantage of this strategy is that it cannot be used without experimental results. In strategy 2, the average value of the scores from randomly trained 50 LSTMs without any constraint conditions was used as the overall score.

Candidate searching was done in two steps. First, we generated and evaluated a billion of random peptides, and recorded top 100,000 sequences. Then, we generated the mutants for the second step. We substituted one amino acid for another while rotating the sequences. This method created 3,054 mutants from each parent. The second step of screening was done by generating all possible mutants from top 1,000 peptides in the previous generation (using random search results at first generation), and the scores were evaluated and recorded if mutants were newly suggested peptides. Then, new top 1,000 peptides are used again as new parents, and this process were repeated until no more parents were found. The top five peptides are shown in Table 1.

For the experimental evaluation, we also tried to search other families of peptides from the champion peptides. We estimated the score barrier when substitute one amino acid for another to find other families. Because it was difficult to set up a clear family boundary, we used a provisional value to suggest peptides for evaluation in Table 2.

Evaluation of designed peptides

To test whether the designed peptides could bind to Met, the peptide-mRNA/cDNA fusion for each sequence was prepared in a similar manner as in the screening, and the amount of binding fraction on the beads bearing Met was quantified. While aMD5 showed a binding fraction of 90%, the designed peptides showed <1%, suggesting that their affinity was lower than that of aMD5.

Table 1. Top rank of peptides by LSTM models trained by different two strategies

Rank	Strategy 1		Strategy 2	
	Sequence	Score	Sequence	Score
1st	**WYYYGAKWQRLLP**	0.988106	**YYYYYAKQRWLLP**	0.985776
2nd	**WYYYGAKWRQLLP**	0.987411	**YYYYYAKQRWLLA**	0.985612
3rd	**YYYYANFKQLYLP**	0.987244	**YYYYYARQRWLLP**	0.985422
4th	**WYYYSAKWQKLLP**	0.987196	**YYYYYAKQRFLLP**	0.985402
5th	**YYYYANFKLQYLP**	0.987173	**YYYYYAQKRWLLP**	0.985400

Table 2. Peptides for experimental verification suggested by LSTM models

	Strategy 1		Strategy 2	
	Sequence	Score	Sequence	Score
1	**WYYYGAKWQRLLP**	0.988106	**YYYYYAKQRWLLP**	0.985776
2	**YYYYAKWGKLLLP**	0.983291	**YYYYLKCKLRLLL**	0.958273
3	**FYYPYCFELRLLL**	0.947954	**LLKLKWCWLKLLE**	0.927284
4			**LERLRWCWLKLAL**	0.870536

Conclusion

We performed machine learning using the data from screening of macrocyclic peptides and constructed models for prediction of the binding affinity from peptide sequences. Although the model seemed to have succeeded in learning the screening results, attempts at design of sequences with high affinity suggested the necessity of improved prediction accuracy. We expect that detailed analysis of the screening results and improved quality and quantity of data would solve this problem. For example, investigating the cause of the change in the probability distribution observed in Figure 2, increasing the number of screening cycles to reduce the noise, and scaling up the sequencing might be effective.

Acknowledgements

We would like to thank our supervisors, Prof. Hiroaki Suga and Prof. Satoshi Watanabe, and our assistant supervisors, Prof. Masatoshi Imada and Prof. Atsuo Yamada, for the kind supports and encouragement. We also thank the MERIT program for providing us the opportunity for this joint research.

参考文献

1. Leenheer, D., ten Dijke, P. & Hipolito, C. J. A current perspective on applications of macrocyclic-peptide-based high-affinity ligands. *Biopolymers* **106**, 889–900 (2016).
2. Zorzi, A., Deyle, K. & Heinis, C. Cyclic peptide therapeutics: past, present and future. *Curr. Opin. Chem. Biol.* **38**, 24–29 (2017).
3. Yamagishi, Y. *et al.* Natural product-like macrocyclic N-methyl-peptide inhibitors against a ubiquitin ligase uncovered from a ribosome-expressed de novo library. *Chem. Biol.* **18**, 1562–1570 (2011).
4. Kidera, A., Konishi, Y., Oka, M., Ooi, T. & Scheraga, H. A. Statistical analysis of the physical properties of the 20 naturally occurring amino acids. *J. Protein Chem.* **4**, 23–55 (1985).
5. Gers, F. A., Schmidhuber, J. & Cummins, F. Learning to Forget: Continual Prediction with LSTM. *Neural Comput.* **12**, 2451–2471 (2000).
6. Ito, K. *et al.* Artificial human Met agonists based on macrocycle scaffolds. *Nat. Commun.* **6**, 6373 (2015).
7. Goto, Y. *et al.* Reprogramming the translation initiation for the synthesis of physiologically stable cyclic peptides. *ACS Chem. Biol.* **3**, 120–129 (2008).
8. Roberts, R. W. & Szostak, J. W. RNA-peptide fusions for the in vitro selection of peptides and proteins. *Proc. Natl. Acad. Sci. U. S. A.* **94**, 12297–12302 (1997).