Directed evolution of trypsin inhibiting peptides using a genetic algorithm

Yohei Yokobayashi, Kazunori Ikebukuro, Scott McNiven and Isao Karube *

Research Center for Advanced Science and Technology, The University of Tokyo, 4-6-1 Komaba, Meguro-ku, Tokyo 153, Japan

A new strategy which employs a process of directed evolution in the search for molecules exhibiting certain desirable properties is reported. By repeating a cycle of peptide synthesis, evaluation of the trypsin inhibitory activities of these peptides and subsequent selection and transformation based on a genetic algorithm, it is possible artificially to induce the evolution of a family of peptides and improve their biological activities. Commencing with a set of 24 randomly generated hexapeptides, a progressive improvement from 16 to 50% average inhibitory activity over six generations, with maximum activities of 80–90% is observed. The emergence of consensus sequences which concur with those previously generated using peptide libraries is also observed.

Introduction

The identification of compounds which exhibit desired biological or chemical activities has attracted much attention.[1] Synthetic libraries of peptides,[2] peptide derivatives[3] and other low molecular weight organic compounds[4] have been prepared and screened for a variety of applications. However, the construction of these libraries requires an enormous amount of time and effort as every possible combination is synthesised, screened for activity and then sequenced if need be. A more logical approach would substantially reduce the time, labour and cost involved in these operations.

Searching for bioactive peptides (or DNA, RNA or other oligomeric or multi-component compounds) is analogous to searching the multidimensional sequence space. For a molecule comprising n components, that space is n-dimensional. For example, if the biological activities of an array of dipeptides XY are mapped onto two-dimensional space, then it is easy to see that we want to know the peptide sequence corresponding to the point in the sequence space which shows the maximum biological activity. Mathematically, the maximum can be found by optimizing a function over two-dimensional space to find the combination of x and y which gives the maximum output. The important difference between these two problems is that for a mathematical function, the output can simply be calculated, whereas we physically have to synthesize, isolate and purify a peptide sequence and then evaluate its activity (output) experimentally.

Genetic algorithms (GAs)[5] are search methods which have been successfully applied to the optimisation of complex mathematical functions. Based on the process of genetic evolution observed in biological systems, three successive operations; *selection*, *crossover* and *mutation* are performed on a set of strings. These strings are a series of characters representing the individual components of the macromolecule. In this case, the strings were the six letters identifying the peptide sequences (Fig. 1). After these GA operations are performed, a new set of 24 peptide sequences is obtained (the second generation) and the process is repeated. In this way, we expected that the GA would effectively search through the possible peptide sequences and each successive generation would acquire higher inhibitory activities than the previous generation.

Implementation of the genetic algorithm

Previous work[6] using partially defined peptide libraries identified Ac-TTKIFT-NH_2 as a hexapeptide having high trypsin inhibitory activity. Consequently, in order to be able to compare the utility of the two selection methods, we generated a set of 24 random hexapeptide sequences comprising phenylalanine (F), isoleucine (I), lysine (K) and threonine (T). These first generation peptides were duly synthesized and a trypsin

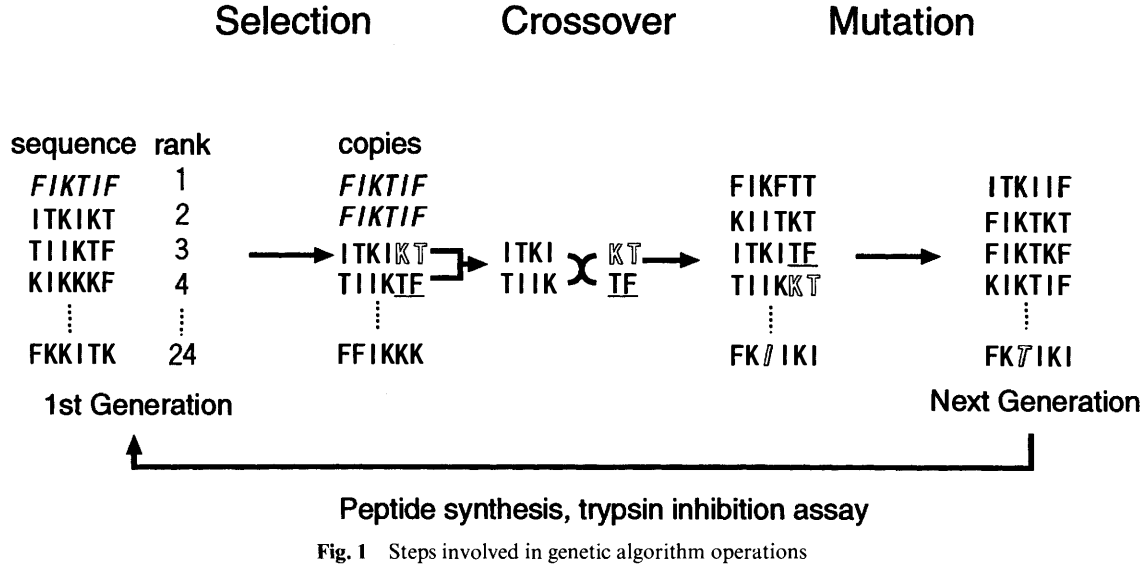

Fig. 1 Steps involved in genetic algorithm operations

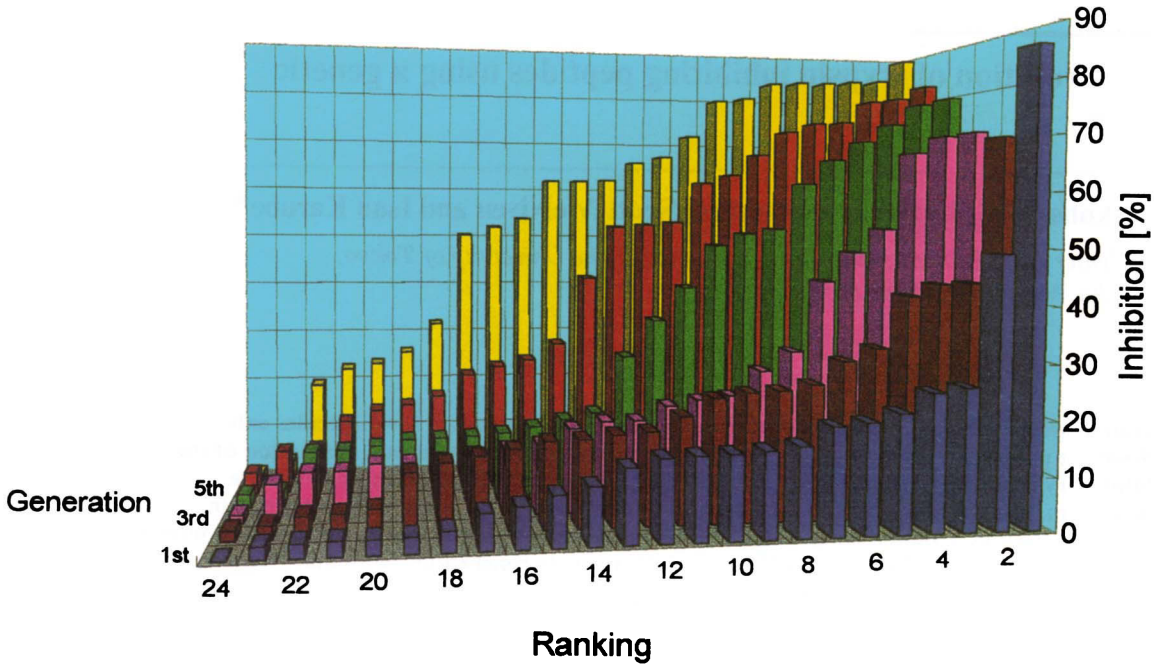

Fig. 2 Trypsin inhibitory activities of peptides according to their generations. For each generation, peptides were sorted according to their inhibitory activities.

inhibition assay was performed for each peptide. The data, consisting of peptide sequences and their inhibitory activities, were then fed into the GA program and the evolution initiated. In the *selection* operation, peptides were first ordered according to their inhibitory activities. Peptides ranking from 1st to 6th were each assigned two copies, 7th to 18th were each assigned one copy and those ranking below 19th were discarded. Thus a total of 24 sequences was kept for subsequent operations. The peptides were then randomly divided into 12 pairs and a *crossover* operation was performed for each pair whereby a random number of amino acids were exchanged between the pair. Finally the peptides were subjected to a *mutation* operation. Each amino acid in the sequence was allowed to be replaced by one of the four residues F, I, K or T with a probability of 3%. The 24 new peptide sequences obtained after these GA operations, the second generation, were synthesized and their inhibitory activities were evaluated. A total of six generations (144 peptides) were synthesized in order to evaluate the performance of the strategy.

Results and discussion

The inhibitory activities of the peptides are shown in Fig. 2. As expected, a progressive improvement was observed as the peptides evolved. While the average activity of the randomly generated first generation peptides was only 16%, the average for the sixth generation rose to 50%. The most effective peptides in each generation also showed a gradual increase in inhibitory activities.

The peptide sequences which exhibited high inhibitory activities (Table 1) bear several consensus sequences. Remarkably, every peptide with an activity of greater than 48% belongs to one of the two sequence patterns previously reported;[6] Ac-XKIXXX-NH_2 or Ac-XXKIXX-NH_2 (where X represents any of the four amino acids). In the sixth generation, 17 out of 22 discrete peptides (two peptides were duplicated) belonged to one of these patterns while only two peptides belonged to these patterns in the first generation, indicating that the GA was converging in the later generations.

Moreover, 13 of the 25 most active peptides contain the Ac-XXXXKI-NH_2 motif, eight of which possess the Ac-XXKIKI-NH_2 sequence. The -KI- unit undeniably confers activity on the hexapeptide, and it is tempting to conclude that any bioactivity is simply due to its presence. However, it must be noted that no peptide with an activity of greater than 11% has the sequence Ac-XXXKIX-NH_2, indicating that the position of the -KI- motif is quite important.

Not surprisingly, the two predominant patterns were identified by Eichler and Houghten[6] who used peptide libraries to identify sequences with trypsin inhibitory activity. Starting with an -XXKIXX- library and subsequently constraining the undefined amino acids, they identified Ac-TTKIFT-NH_2 as the optimal sequence. Although we thought our strategy might find a sequence which inhibits trypsin more strongly than this, it

Table 1 Peptide sequences which exhibited trypsin inhibitory activities of 60% or greater. The first digit in the peptide IDs indicates the generation, the second is an arbitrary number assigned in the crossover operation. Inhibitory activities were averaged for those peptides which appeared in more than one generation

Rank	Sequence	Peptide ID	Inhibition (%)
1	TTKIFT	6–02	89.0
2	KKIFIF	1–18	85.9
3	TKIFKI	6–12	84.3
4	FTKIKI	6–24	84.1
5	KTKIKI	6–04, 6–13	83.6
6	FKKIFI	6–22	83.2
7	TTKIKI	5–18	82.4
8	FKKIFT	6–08	79.9
9	TKIIKI	5–10, 4–19	79.3
10	TTKIFI	5–08	78.8
11	ITKIKI	6–16, 4–13	78.4
12	TKKIKI	5–14	74.4
13	FKKIKI	5–16	74.0
14	KTKIKT	5–21	72.2
15	TTKIKT	6–20, 4–10, 3–14	72.1
16	KKIIKI	3–09	70.6
17	KKIITK	2–22	70.1
18	FKKIFF	4–03	69.9
19	ITKIKT	5–11	67.4
20	TKIKKI	3–24	67.3
21	TKIIKT	6–15	66.9
22	KFKIKI	4–04	66.1
23	TFKIFT	6–14	65.7
24	KKIKKI	6–18, 5–09, 4–20	62.2
25	KKKIKI	6–11, 6–19	61.7
26	KKIKFT	5–12	61.5

appeared in the sixth generation and possessed the highest activity (89%) against trypsin of any peptide assayed. The non-emergence of a more potent inhibitor may be because an insufficient number of generations were examined in this experiment. More optimistically, it may be that Ac-TTKIFT-NH_2 is the optimal sequence for these amino acids.

Directed evolution using our GA searching strategy is widely applicable and independent of the synthetic methods and assay systems employed. Importantly, it is not limited to solid phase syntheses or to affinity-based assays. The strategy becomes even more efficient as the sequence becomes longer or as the molecule involves more components; conditions under which combinatorial library approaches become physically unfeasible. This method is thus capable of searching more efficiently through a greater diversity of compounds. In this case, we synthesized about 140 peptides out of some 4000 possible permutations (*ca.* 3%) and achieved 90% inhibition of trypsin activity! Recent success of DNA shuffling[7] for the *in vitro* selection of biological molecules confirms the potential of evolutionary approaches for searching multidimensional sequence spaces. Simpler organic compounds may also be examined using our method. Our strategy not only demonstrates a new application of GAs[8] but also provides an effective, more efficient approach to the exploration and exploitation of the diversity of combinatorial chemistry.

Experimental

Fluoren-9-ylmethoxycarbonyl (Fmoc)-protected amino acids were purchased from Novabiochem. All other reagents were of the highest available quality. Peptides were synthesized on a multiple peptide synthesizer (Shimadzu PSSM-8) using standard Fmoc chemistry. Rink amide MBHA resin (Novabiochem) was used as the solid support in order to obtain *C*-terminal amides. Peptides were acetylated at the *N*-terminus by treatment with acetic anhydride–DMF 1:3 (v/v) before being cleaved from the resin with a TFA–anisole–ethanedithiol 95:4:1 (v/v/v) mixture. Peptides were analysed by matrix assisted laser desorption ionization time of flight mass spectrometry (MALDI–TOF-MS) (PerSeptive Biosystems Voyager™-RP) and reversed-phase high performance liquid chromatography (RP-HPLC) (Waters 600 series). Peptides were further purified by HPLC if impurities were detected. Trypsin inhibition was assayed by monitoring the cleavage of N^{α}-benzoyl-DL-arginine-*p*-nitroanilide (BAPA) in the presence of the peptides. In a disposable cuvette, 200 μl of a 1 mM peptide solution in 0.1 M Tris–HCl buffer containing 0.025 M $CaCl_2$ (pH 7.8) and 400 μl of a 2.3 mM BAPA solution in Me_2SO–H_2O 1:9 (v/v) were mixed. Trypsin solution in 0.02 M HCl was prepared and diluted with Tris buffer in a ratio of 3:20. This trypsin solution (460 μl) was added to the cuvette and the absorbance at 410 nm was monitored using a Beckman DU-7400 spectrophotometer at 5 s intervals for 3 min at room temperature. The absorbance values from 1 to 3 min were fitted to a linear equation using a least squares method and the slope of the line calculated. The inhibitory activity (I_A) of the peptide was defined as in eqn. (1). In the control reaction, Tris buffer not

$$I_A = 100 \times (\text{slope of control} - \text{slope of peptide})/(\text{slope of control}) \quad (1)$$

containing peptide is added to the cuvette. Typically, trypsin solutions were prepared at concentrations of 0.100 to 0.125 mg ml^{-1} 0.02 M HCl. Concentrations were determined so that approximately equal reaction rates were obtained for the control in different assay sessions. All GA operations were implemented by programs written in Pascal and computations were performed on an IBM personal computer.

Acknowledgements

We thank Dr Shuichi Kojima of Gakushuin University for technical advice on the trypsin inhibition assay. S. M. thanks the Japan Society for the Promotion of Science for a fellowship.

References

1 For reviews on combinatorial chemistry, see (*a*) G. Lowe, *Chem. Soc. Rev.*, 1996, 309; (*b*) K. D. Janda, *Proc. Natl. Acad. Sci. USA*, 1994, **91**, 19 779; (*c*) *Acc. Chem. Res.*, 1996, **29**, pp. 111–170. Special Issue on Combinatorial Chemistry.

2 (*a*) K. S. Lam, S. E. Salmon, E. M. Hersh, V. J. Hruby, W. M. Kazmierski and R. J. Knapp, *Nature*, 1991, **354**, 82; (*b*) R. A. Houghten, C. Pinilla, S. E. Blondelle, J. R. Appel, C. T. Dooley and J. H. Cuervo, *Nature*, 1991, **354**, 84.

3 (*a*) Y. Cheng, T. Suenaga and W. C. Still, *J. Am. Chem. Soc.*, 1996. **118**, 1813; (*b*) R. Boyce, G. Li, H. P. Nestler, T. Suenaga and W. C. Still, *J. Am. Chem. Soc.*, 1994, **116**, 7955; (*c*) D. A. Campbell, J. C. Bermak, T. S. Burkoth and D. V. Patel, *J. Am. Chem. Soc.*, 1995, **117**, 5381.

4 B. A. Bunin, M. J. Plunkett and J. A. Ellman, *Proc. Natl. Acad. Sci. USA*, 1994, **91**, 4708.

5 (*a*) D. E. Goldberg, *Genetic Algorithms in Search Optimization and Machine Learning*, Addison-Wesley, Reading, MA, 1989; (*b*) S. Forrest, *Science*, 1993, **261**, 872.

6 J Eichler and R. A. Houghten, *Biochemistry*, 1993, **32**, 11 035.

7 (*a*) W. P. C. Stemmer, *Nature*, 1994, **370**, 389; (*b*) W. P. C. Stemmer, *Proc. Natl. Acad. Sci. USA*, 1994, **91**, 10 747.

8 While this work was in progress, two reports of similar work were published. The first uses GAs to explore the chemistry of Ugi reaction products while the second investigates the substrate specificity of stromelysin. (*a*) L. Weber, S. Wallbaum, C. Broger and K. Gubernator, *Angew. Chem., Int. Ed. Engl.*, 1995, **34**, 2280; (*b*) J. Singh, M. A. Ator, E. P. Jaeger, M. P. Allen, D. A. Whipple, J. E. Soloweij, S. Chowdhary and A. M. Treasurywala, *J. Am. Chem. Soc.* 1996, **118**, 1669.

Paper 6/02435A
Received 9*th April* 1996
Accepted 8*th July* 1996