79. Artificial evolution. Algorithms for searching for amino acid sequences.

Beniamin Dziurdza, [1] **Jacek Błażewicz** [1,2] **and Wojciech T. Markiewicz** [2]

Keywords: artificial evolution, randomisation, genetic hyperspace, phage display, SELEX

1 Introduction

Artificial evolution can be viewed as an exploration of *genetic hyperspace*. One can say that a *genetic hyperspace* is a "space" of DNA sequences, where DNA sequences encode peptides and RNAs of one organism. In natural evolution of living organisms mutations are random and a survival of a given organism (and its genotype) depends on environment's pressure. In an artificial evolution mutations are designed from outside of a system by an investigator who decides what the "best qualities of evoluating organism" are.

¿From a technical point of view, mutations are realised by cloning DNA with *randomisation*. *Randomisation* of DNA gives a possibility to obtain DNA sequences that belong into a product of sets, where each set is a subset of $\{A, G, C, T\}$.

In practice, it is not possible to obtain all organisms coded by generated DNA sequences. The reason is that a number of possible DNA sequences grows exponentially with a length of a sequence. In case when a number of sequences becomes too large, there is no technology that enables an "implementation" of all organisms coded by DNA sequences. For that reason the artificial evolution methods such as *phage display* [1] or *SELEX* [2] exploit only a minute and rather unknown part of genetic hyperspace [3].

Because of the reasons mentioned above, nowadays we can search for very short sequences only. Good example is searching for one peptide, where only few (6-10) positions of amino acid sequence are randomised (it gives 18-30 bp) [3].

In this poster a new approach to this problem is presented, however the obtained results can be in simple way applied to instances with many peptides.

2 New concept of artificial evolution

The most popular randomisation's pattern is $(NNS)_x$, where $N = \{A, G, C, T\}$, $S = \{G, C\}$ and x is the number of codons [3]. Unfortunately, although $(NNS)_x$ pattern generates all amino acid sequences, in fact using it does not ensure obtaining all amino acid sequences due to technological constraints. Investigator can only choose a best peptide, find its primary structure and optionally repeat experiment. It is not possible to get the knowledge about other sequences, because it is not known exactly which of them were obtained [3].

New approach to artificial evolution was proposed in [3]. Instead of $(NNS)_x$ pattern, partial randomisation patterns are applied. These patterns generate just a part of molecular variability but one can obtain really all required amino acid sequences and test them. The result gives a significant knowledge about the explored part of genetic hyperspace.

[1]Institute of Computing Science, Poznań University of Technology, Piotrowo 3a, PL–60965, Poznań, Poland. E-mail: `Beniamin.Dziurdza@cs.put.poznan.pl`

[2]Institute of Bioorganic Chemistry, Polish Academy of Sciences, Noskowskiego 12, PL–61704, Poznań, Poland. E-mail: `markwt@ibch.poznan.pl`

This knowledge is applied for navigation in genetic hyperspace. In next iterations of experiment randomisations (mutations) are chosen in such a way that one tries to obtain new amino acid sequences that have not been tested yet. Thus, currently chosen randomisation should generate a new set of sequences laying in the neighbourhood of the best known so far sequence. It is based on a reasonable assumption that peptides with similar primary structure have similar properties. If one cannot find a better sequence in a current iteration of the experiment then the randomisation (explored neighbourhood) should be changed or another amino acid sequence should be chosen in order to explore its neighbourhood in genetic hyperspace.

3 Model of artificial evolution

Because we search for a peptide, it is useful to analyse randomisation of DNA sequence as a sequence of randomised codons. It is easy to observe that not all codon randomisation patterns are as good as others, e.g. $\{A\} \times \{T\} \times \{G, C\}$ induces *Tyr* only and 2 DNA sequences but $\{A\} \times \{A, T\} \times \{G\}$ induces *Tyr*, *Phe* and also 2 DNA sequences.

This observation is the reason for an introduction of efficient codon patterns where efficient codon pattern is defined as such a randomisation pattern that there's no pattern which generates more (the same) different amino acid sequences and generates no more (less) DNA sequences. Naturally, using these patterns is always more efficient than using non-efficient codon patterns.

In the poster an algorithm for finding efficient codon patterns is presented. In addition the list of found patterns is attached.

A mathematical model of this experiment is proposed. It results in a new concept of artificial evolution. Experiment iterations are presented as a collection of the following randomisations. Each randomisation is presented as a sequence of efficient codon patterns. The goal is to find the new randomisation such that it generates as many new amino acid sequences as possible and fulfils constraints on a maximal number of DNA sequences. Moreover, it should generate neighbourhood of the required sequence.

The above model makes an efficient exploration of genetic hyperspace possible. Unfortunately, the corresponding optimisation problem called Optimal Randomisation is NP–hard. An algorithm for the optimal randomisation problem is described in the poster.

4 ArtEvol

The computational approach presented in the poster has led to the computer program called ArtEvol (Artificial Evolution). ArtEvol for Windows OSes can be downloaded from `http://www.cs.put.poznan.pl/bdziurdza`.

References

[1] Houghten, R. A., Pinilla, C., Blondelle, S. E., Appel, J. R., Dooley, C. T. and Cuervo, J. H. (1991) Generation and use of synthetic peptide combinatorial libraries for basic research and drug discovery. *Nature*, **354**, 84–86.

[2] Tuerk, C., Gold, L. (1990). Systematic evolution of ligands by expotential enrichment: RNA ligands to bacteriophage T4 DNA polymerase. *Science*, 249: 505–510.

[3] (a) Markiewicz, W. T., Kwaśnikowski, P., Talarek, J., a manuscript in preparation and *The Polish Patent Application*; (b) Kwaśnikowski, P., Ph.D. Thesis: Semi–synthetic combinatorial phage libraries (in Polish). Inst. Bioorg. Chem., Poznań, 2000.