Vol.2011-BIO-25 No.40
2011/6/24

The Analysis of DNA Shuffling by nMDS

Ryota Doi[†1] and Y-h. Taguchi[†1]

DNA shuffling is widely used for optimizing complex properties contained within DNA and proteins. However, success rate of it is deeply dependent upon which pair of DNAs is employed for DNA shuffling. In this paper, we have used non-metric multidimensional scaling (nMDS) to select best pair of DNAs for it. It turns out that nMDS can sometimes choose better pairs of DNAs than hierarchical clustering which is frequently used to select the suitable pair of DNAs.

1. Introduction

DNA shuffling[1)] is one of the method to recombine two DNA sequences to generate new proteins to have new and better functionalities than original two proteins. During this process, some of domains in one protein are connected to other domains taken from another protein and result in newly obtained proteins. However, there is a problem on this. Usually, Polymerase Chain Reaction (PCR) technique is used to recombine two DNA fragments. This means, two fragments which are supposed to be joined must have some common sequences. If not, PCR cannot merge two fragments into one.

Thus, the situation is a little bit controversal. If two proteins are too far from each other, we cannot have newly generated sequence. On the other hand, if they are very same, recombination does not give us nothing new, since recombination of almost same sequence can generate something which is very close to both of its parent sequence. What we have to do is the following. First, we have to prepare the set of proteins whose sequences are more or less different from each other. Then, we have to estimate which pair of proteins is easy to be recombined.

Recently, Montera et al[2)] has proposed mutation based-measure, i.e., that based upon evolutionary process can outperform other simple measure base upon informatics. Hierarchical clustering based upon former gives us better pair of proteins than the later can. These pairs of proteins achieve better performance of recombination due to numerical simulation called eshuffle[3)]. However, they did not try any other method to employ better pairs than clustering.

In this paper, we have shown that replacement of method can sometimes give us better pairs of proteins. It turns out that not only measure of similarity is important, but also clustering method is.

2. Materials and Methods

2.1 Sequences

Luciana Montera kindly provided us the sequence they used in their paper. These are the sequences of 37 DNA gene sequences codifying to snake venom metallopeptidases.

2.2 Similarity between sequences

We have employed

$$d(X,Y) = 1 - \frac{GenCompress(X|\varepsilon) - GenCompress(X|Y)}{GenCompress(XY|\varepsilon)}$$

to compute distance $d(X,Y)$ between sequences X and Y using $GenCompress$ program[4)]. Details and definition of these measures can be obtained in original paper[2)].

2.3 Alignment

Alignment of pair of sequence has been done ClustalW2[5)].

2.4 nMDS

We have used nmds module in labdsv package in R[6)].

3. Results

Fig. 1 shows the embedding of 37 DNA sequences onto two dimensional space. It is clear that some of pairs which are closed to each other in the original study[2)] are not nearby pairs in Fig. 2. For example, EoMP06 and TSVDM are closed in Fig. 2 in the original study, but not in Fig. 1. In contrast to this, some of pairs which are far from each other in Fig. 2 in the original study are closed to each other in Fig. 1. For example, EoMP06 and hemor are closed to each other. Thus, it is obvious that nMDS can provide us different candidates for DNA shuffling.

[†1] Department of Physics, Chuo University, Tokyo, Japan

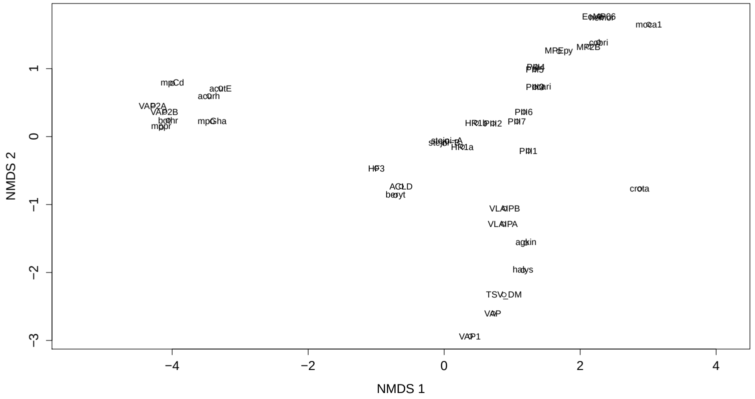

Fig. 1 Embedding of 37 DNA sequences onto two dimensional space by nMDS

Table 1 Comparison between pairs listed in Table 1 in the original work[2] and pairs newly obtained in the present study. The left most column is target gene and paired genes are listed next to this. nMDS is the present study, M_{mutatoin} is computed using newly obtained alignment[5], and M'_{mutatoin} is in origical work[2].

	$T_a = 50°C$			$T_a = 50°C$		
	$F_{\text{size}} = 45bp$			$F_{\text{size}} = 35bp$		
Target gene	Paired genes					
	nMDS	M_{mutatoin}	M'_{mutatoin}	nMDS	M_{mutatoin}	M'_{mutatoin}
EoMP06	hemor	TSVDM		hemor	TSVDM	
	2.091	3.850	3.39	2.344	4.768	4.63
cobri	MP2b	PIII2		MP2b	PIII2	
	3.267	4.983	5.99	4.424	5.931	5.80
PIII3	ecari	HR1b		ecari	HR1b	
	6.135	6.087	6.11	5.843	6.238	6.28
PIII1	PIII7	MPEpy		PIII7	MPEpy	
	6.259	5.883	4.94	6.241	6.084	5.15

In Table 1, we have shown that the comparison between pairs listed in Table 1 in the original work[2] and pairs newly obtained in the present study. Although those obtained in this study do not always outperform the previous study, at least, some of they can achieve better performance than previous ones.

4. Conclusion

In this paper, we have shown that nMDS sometimes outperforms conventional clustering method to infer the best pair of proteins for DNA shuffling. We propose to employ both hierarchical clustering and nMDS in order not to miss better pairs of proteins for DNA shuffling.

5. Acknowledgement

We would like to acknowledge Luciana Montera for kindly providing us DNA sequences which they used in the original research[2]. This work was supported by KAKENHI 23300357.

References

1) Kristian, M.M., Sabine, C.S., Susanne, K., Gregor, Z., Hubert, S.B., and Katja, M.A.: Nucleotide exchange and excision technology (NExT) DNA shuffling: a robust method for DNA fragmentation and directed evolution, *Nuc. Acid. Res.*, Vol.33, e117 (2005); Speck, J., Stebel., S.C,, Arndt, K.M., M ller, K.M.; Nucleotide exchange and excision technology DNA shuffling and directed evolution, *Methods Mol Biol.*, Vol.687, pp.333–44, (2011)

2) Montera, L, Nicoletti, M.C., da Siliva, F.H., and Moscato, P.: An effective mutation-based measure for evaluating the suitability of parental sequences to undergo DNA shuffling experiments, *2008 IEEE Congress on Evolutionary Computation (CEC2008)* , vol.1, pp.765–772, (2008).

3) Moore, G.L., Maranas, C.D., Lutz, S., and Benkovic, S.J.: Predicting Crossover Generation in DNA Shuffling, *Proc. Natl. Acad. Sci.* vol.98, pp.3226–3231, (2001).

4) Zhao, H., Arnold, F.H.: Functional and nonfunctional mutations distinguished by random recombination of homologous genes, *Proc. Natl. Acad. Sci.*, vol.94, pp.7997-8000, (1997).

5) http://www.ebi.ac.uk/Tools/msa/clustalw2/

6) R Development Core Team, R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org (2009).