Title	The bio molecule interaction prediction and design method only with primary structure
Sub Title	
Author	小川, 隆(Ogawa, Ryu)
Publisher	慶應義塾大学湘南藤沢学会
Publication year	2015
Jtitle	生命と情報 No.22 (2015.) ,p.1- 10
JaLC DOI	
Abstract	
Notes	慶應義塾大学湘南藤沢キャンパス先端生命科学研究会 2015年度学生論文集 博士論文ダイジェスト
Genre	Technical Report
URL	https://koara.lib.keio.ac.jp/xoonips/modules/xoonips/detail.php?koara_id=KO92001004-00000022-0001

The bio molecule interaction prediction and design method only with primary structure

Ryu Ogawa

An Abridged Edition of Doctoral Dissertation

Graduate School of Media and Governance

Systems Biology Program

KEIO UNIVERSITY

2015

Summary of this dissertation

Chapter 1

General introduction

Proteins are a crucial structural component of all cells in the body and act as one of the principal parts of life like pitching physical parameters and catalyzing chemical processes. Since 1983, after Ulmer proposed "Protein engineering," we still cannot design proteins longer than 100 amino acid residues (Ulmer, 1983).

Presently, protein engineering has the following two main approaches: rational design and directed evolution. The rational protein design approach is based on the knowledge of protein physics, structures and functions. This design aims at modifying the existing protein by changing its function via site-directed mutagenesis method (mutagenesis approach for rational design). By using the computational prediction (molecular simulation), some protein can be designed from the whole amino acid residue of

partial amino acid residue of short protein (computational approach for rational design) (Kuhlman *et al.*, 2003; Wu *et al.*, 2010). On the other hand, in the directed evolution approach appropriate protein is screened from the random protein library. However, this approach requires huge diversity of libraries and high-throughput screening technologies (random library screening) (Smith, 1985; Yang *et al.*, 2010).

Each existing approach has its own pros and cons. Both mutagenesis and computational approach for rational design have to consider the tertiary structure, which is difficult to be solved but is important. The random library screening approach for evolutional design remains as screening result, which causes difficulty in gaining background knowledge from this approach. In addition, sometimes the result is very different from the expectation, because these are black box processes and the scientist can only rebuild the screening library or change the parameters like washing buffer, washing time, target molecule preparation, and so on (Lutz, 2010).

To solve these problems, we propose “Primary structure driven computational approach.” This approach aims to gain

knowledge from primary structure of interacting molecules for protein design. To design bio molecule, this approach focused on collecting primary structure information and analyzing those dataset via computational algorithm. The process of collecting primary structure like DNA, RNA or protein sequences as source information for the protein design is easier than collecting tertiary structure information. In addition, calculation cost for primary structure analysis is lower than the cost incurred in molecular dynamics simulation. Compared with the random library screening, this method is able to have knowledge information through design processes.

In this dissertation, we propose our novel protein designing method with "primary structure driven approach." This approach focuses on gaining knowledge from primary structure from binding bio molecule and designing via that knowledge. To verify this approach, we set two milestones. The first milestone was DNA-histone interaction analysis as proof-of-concept and the second milestone was antibody design based on antibody-antigen interaction analysis as practical experiment (Figure 1).

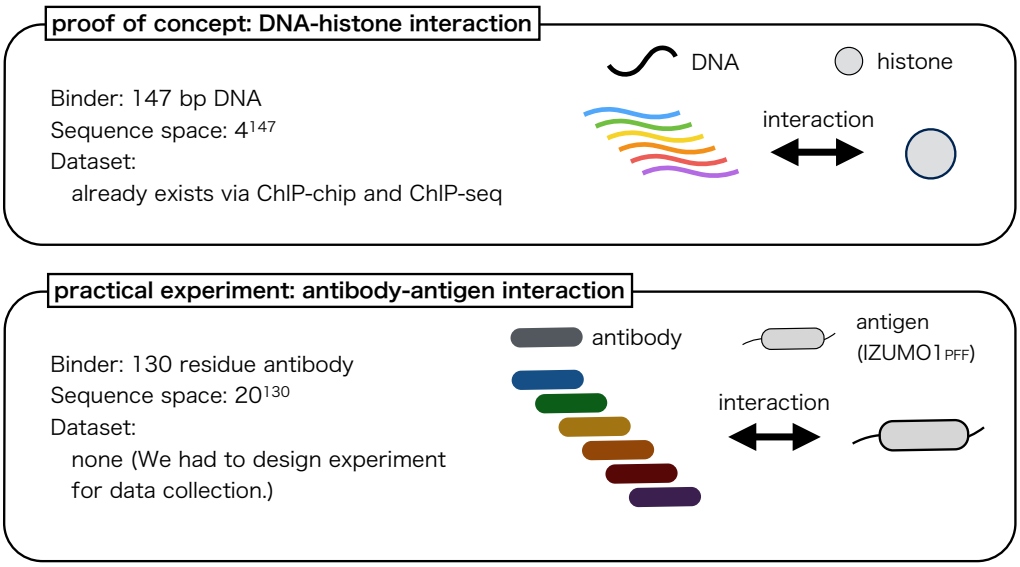


Figure 1 Schematic of two milestones of this study.

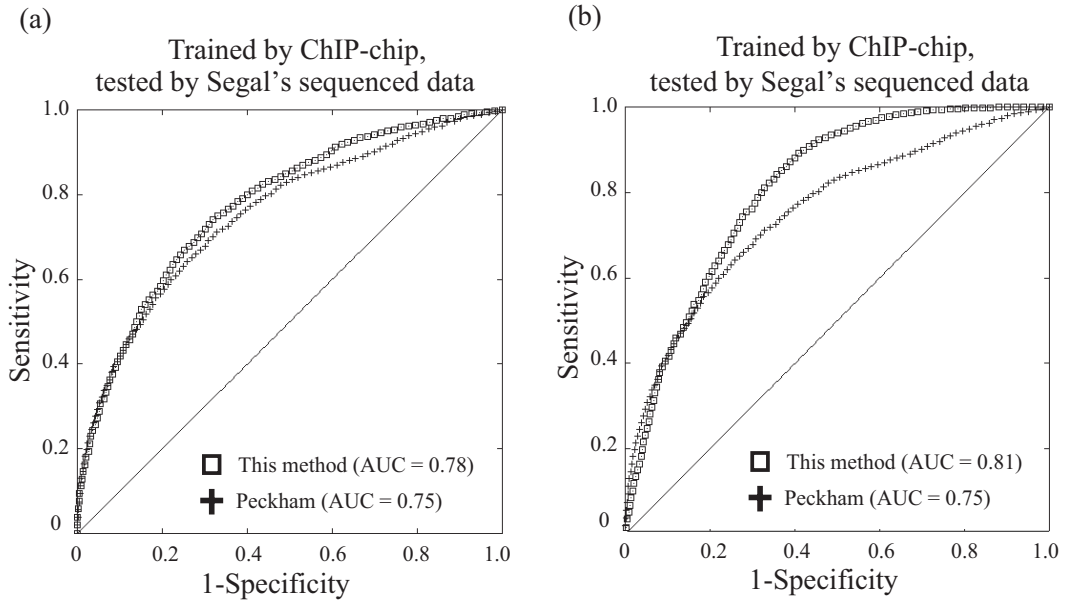
Top panel shows schematic of DNA-histone interaction and bottom panel shows schematic of antibody-antigen interaction.

Chapter 2

Computational prediction of DNA-histone interaction

For the first milestone theme, we selected histone-DNA interaction as a proof-of-concept. Histone protein binding partner is 147 bp DNA sequence and histone binding position on genomic DNA depends on the DNA sequences. The objective of this milestone was to detect knowledge from histone binding partner and to predict

histone-DNA binding with high accuracy. This is appropriate to the first milestone because gaining knowledge about DNA sequence with histone binding partner is easier than protein sequence as protein binding partner. This is because the sequence space of histone protein binding partner is much smaller than the sequence space of protein binding partner. As previously described, histone protein binding partner is 147 bp DNA sequence; therefore, size of sequence space is 4^{147} (3.18e+88). In contrast to histone binding partner, sequence space size of protein binding partner protein is at least bigger than 20^{100} (1.27e+130) because protein length distribution is approximated gamma distribution (alpha is 1 to 3) and mean length is 270 residue (Zhang, 2000; Brocchieri and Karlin, 2005). Furthermore, dataset of histone binding DNA sequences also exists. In Chapter 2, we showed our method via machine learning technology to predict histone binding DNA sequence on genomic DNA. Here, we used the relative fragment frequency index (we developed) and a support vector machine to screen for histone binding and linker DNA sequences.



Comparison of accuracy between Peckham's method (1- to 6-mer) and RFFI (1- to 6-mer)

Comparison of accuracy between Peckham's method (1- to 6-mer) and RFFI (1- to 8-mer)

Figure 2 Comparison of ROC curves of the SVM that was trained by our method and Peckham's method (trained by ChIP-chip data, tested by Segal's sequenced data).

(a) Comparison of accuracy between Peckham's method (1- to 6-mer) and RFFI (1- to 6-mer). AUC of Peckham's method was 0.75, AUC of RFFI was 0.78.(b) Comparison of accuracy between Peckham's method (1- to 6-mer) and RFFI (1- to 8-mer). AUC of Peckham's method was 0.75, AUC of RFFI was 0.81.

Our 2-fold cross-validation revealed that the accuracy of our method based on area under the receiver operating characteristic curve was 81%, whereas, that of previous method was 75% (Figure 2). This was revealed when both ChIP-chip and ChIP-seq data were used for validation. Thus, we suggest that our method is more effective in predicting histone binding DNA on genomic DNA.

Chapter 3

Antibody sequence design method combined with phage display library screening and next generation sequencing

For the second milestone, antigen-antibody interaction was selected for the practical experiment. Antibody is one of the most practical proteins used for clinical purposes. Antibodies are important protein of the immune system and considered as a new medical molecule. In Chapter 3, we compared antibody binding affinity and specificity both via our method and as well as by the conventional method. We developed highly sensitive antibody detecting method according to “Primary structure driven approach” by combining phage display screening method with next generation

sequencing. We detected two antibody clusters that were not detected with the existing methods. We gained antibodies with 9% library content rate after the panning process with the existing method. However, using our new method we found two new antibody clusters that contained only 0.2% and 1.1% of library content rate. All above facts supporting this method have highly sensitive antibody detecting functions than the existing methods (Figure 3).

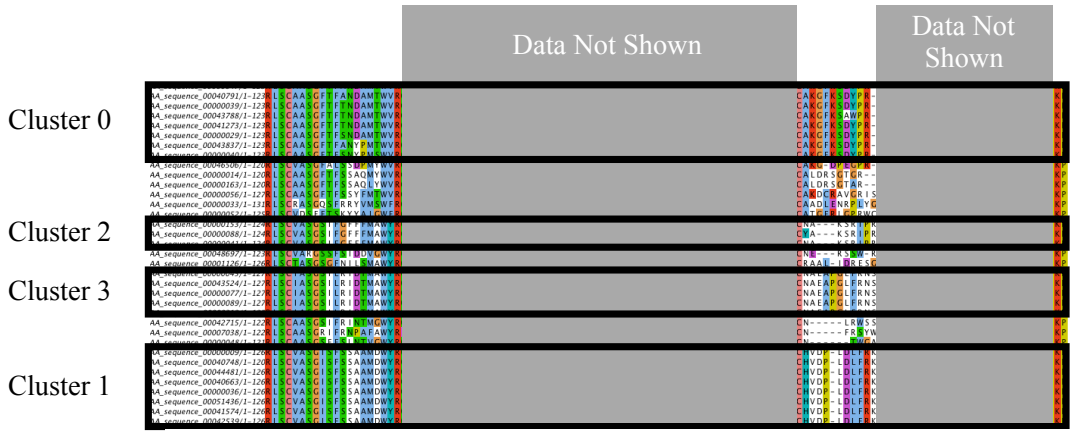


Figure 3 Clustering of top 100 amplification ratio VHH antibody sequences (partially shown).

This figure shows the clustering result via clustalW2. Our method detected new antibody clusters including Cluster 0, which was the only cluster found by the conventional method.

The amplification ratio of all sequences elucidated by NGS was examined. We detected 13 new antibody clusters and Cluster 0. We chose the three sequences without Cluster 0 as candidates for functional sequences to find the functional VHH sequences. We named them Clusters 1, 2, and 3, and examined them with different amplification ratios. Cluster 1 and 2 VHH phages displayed a clear binding ability to IZUMO1$_{PFF}$; however, Cluster 3 did not. Moreover, these VHHs and Cluster 0 were expressed in *E. coli* and complied to the affinity analysis on SPR. Clusters 0, 1 and 2 VHHs indicated high affinity to IZUMO1$_{PFF}$ with K_D8.5 nM, 6.8 nM and 13.6 nM.

Chapter 4.

Concluding remarks

Lastly, in Chapter 4, we have discussed my novel approach and its contributions and future prospects. These results indicated that our approach had higher ability to narrow down sequence space than the conventional approach. In conclusion, primary sequence driven approach could design high spec antibodies, which otherwise could not be developed via conventional approach.

References

Brocchieri, L. and Karlin, S. (2005) "Protein length in eukaryotic and prokaryotic proteomes." *Nucleic Acids Res.***33**, p.p. 3390-3400.

Kuhlman, B., Dantas, G., Ireton, G.C., Varani, G., Stoddard, B.L. and Baker, D. (2003) "Design of a novel globular protein fold with atomic-level accuracy." *Science***302**, p.p. 1364-1368.

Lutz, S. (2010) "Beyond directed evolution--semi-rational protein engineering and design." *Curr. Opin. Biotechnol.***21**, p.p. 734-43.

Smith, G.P. (1985) "Filamentous fusion phage: novel expression vectors that display cloned antigens on the virion surface." *Science***228**, p.p. 1315-1317.

Ulmer, K.M. (1983) "Protein engineering." *Science***219**, p.p. 666-671.

Wu, X., Yang, Z.Y., Li, Y., Hogerkorp, C.M., Schief, W.R., Seaman, M.S., Yuan, G.C., Liu, Y.J., Dion, M.F., Slack, M.D., Wu, L.F., Altschuler, S.J. and Rando, O.J. (2005) "Genome-scale identification of nucleosome positions in *S.cerevisiae*."*Science***309**, p.p. 626-630.

Zhang, J. (2000) "Protein-length distributions for the three domains of life." *Trends Genet.***16**, p.p. 107-109.

Zhou, T., Schmidt, S.D., Wu, L., Xu, L. and *et al*. (2010) "Rational design of envelope identifies broadly neutralizing human monoclonal antibodies to HIV-1." *Science***329**, 856-861.