Journal of Bioscience and Bioengineering
VOL. xx No. xx, 1–9, 2016

www.elsevier.com/locate/jbiosc

Exploring high-affinity binding properties of octamer peptides by principal component analysis of tetramer peptides

Akiko Kume,[1] Shun Kawai,[2] Ryuji Kato,[2] Shinmei Iwata,[1] Kazunori Shimizu,[1] and Hiroyuki Honda[1,*]

Department of Biotechnology, Graduate School of Engineering, Nagoya University, Furo-cho, Chikusa-ku, Nagoya 464-8603, Japan[1] *and Department of Basic Medicinal Sciences, Graduate School of Pharmaceutical Sciences, Nagoya University, Furo-cho, Chikusa-ku, Nagoya 464-8601, Japan*[2]

Received 31 March 2016; accepted 17 August 2016
Available online xxx

To investigate the binding properties of a peptide sequence, we conducted principal component analysis (PCA) of the physicochemical features of a tetramer peptide library comprised of 512 peptides, and the variables were reduced to two principal components. We selected IL-2 and IgG as model proteins and the binding affinity to these proteins was assayed using the 512 peptides mentioned above. PCA of binding affinity data showed that 16 and 18 variables were suitable for localizing IL-2 and IgG high-affinity binding peptides, respectively, into a restricted region of the PCA plot. We then investigated whether the binding affinity of octamer peptide libraries could be predicted using the identified region in the tetramer PCA. The results show that octamer high-affinity binding peptides were also concentrated in the tetramer high-affinity binding region of both IL-2 and IgG. The average fluorescence intensity of high-affinity binding peptides was 3.3- and 2.1-fold higher than that of low-affinity binding peptides for IL-2 and IgG, respectively. We conclude that PCA may be used to identify octamer peptides with high- or low-affinity binding properties from data from a tetramer peptide library.

[**Key words:** Peptide; Binding analysis; Peptide array; Principal component analysis; Prediction]

Peptides are among the most important of biological molecules, being able to bind to various intracellular molecules such as enzymes, receptors, cytokines, polysaccharides, and nucleotides, and thereby regulate complex biological mechanisms. A variety of peptides specifically bound to target molecules has been described, and the information obtained may be applied to many biological research areas such as molecular drug targeting, bio-imaging and diagnostics (1–11). Such functional peptides need to be rationally explored, and its variants should be designed with precision.

Recently, the construction of peptide libraries for screening and development of functional peptides has attracted much attention. For example, phage display (12) (a biological strategy), combinatorial split synthesis (13) (a chemical strategy), and beads display (14) (a chemical strategy) are the most popular and convenient methods of obtaining peptide libraries and are also powerful tools for the identification of new ligands. Such methods identify functional peptides based simply on positive screening (15–18). In these methods, however, the rational analysis of the characteristics of functional peptides is difficult.

A larger number of peptides are included in libraries of peptides with a longer sequence. Therefore, longer peptides have a higher chance of interacting with the target protein than shorter peptides. For example, the peptides binding to antibody have 13–16 residues (19) and the peptides binding to MHC class I have 9–10 residues (20). However, it is very difficult to exhaustively explore the functional properties of peptides from such large libraries. An octamer peptide library, for instance, consists of 25.6 billion different peptides. In contrast, a tetramer peptide library consists of only 160,000 peptides. Thus, exploring peptide libraries in an efficient and rational manner is crucial when screening for functional peptides in a large number of candidate molecules.

In the context of screening a peptide library efficiently, many researchers have demonstrated that the physicochemical properties of longer peptides are different depending on the target protein (21–26). Recently, we developed a screening method for peptides with binding affinity to bile acid and to the death receptor using both binding data from solid-phase peptide arrays and physicochemical characteristics from computational analysis (27,28). In the method described in these articles, we used random peptide libraries containing 2212 peptides for bile acid-binding peptides and 643 peptides for receptor-binding peptides. The larger the size of the random library, the higher the number of high-affinity binding peptides we found. The minimum size requirements for screening of peptide libraries still needed to be determined. Thus, we proposed the use of a small, nonrandom, predesigned library for the screening of high-affinity binding peptides. In the predesigned library, the 20 amino acids were categorized into four groups according to the physicochemical properties of high/low hydrophobicity and positive/negative charge, and a tetramer peptide library (512 peptides, $4^4 \times 2$ versions) covering all four physicochemical properties was constructed (29). The size of this library was the minimum size for a tetramer library. Using the predesigned library, we found the physicochemical properties underlying the binding of peptides to the target protein, and succeeded in

* Corresponding author. Tel.: +81 52 789 3215; fax: +81 52 789 3214.
E-mail address: honda@nubio.nagoya-u.ac.jp (H. Honda).

1389-1723/$ – see front matter
http://dx.doi.org/10.1016/j.jbiosc.2016.08.005

developing a screening method for functional peptides using both binding affinity data and physicochemical properties of peptides in an efficient and rational manner.

However, our previously proposed screening method (29) was adapted for the use of only short tetramer peptides. In addition, if the peptide length were extended from tetramer to octamer in our proposed method, the size of the peptide library would need to increase 256-fold ($256 = 4^4$ to $65{,}536 = 4^8$). Therefore, a new concept for predesigned peptide libraries of longer peptides was necessary.

Principal component analysis (PCA) is a mathematical algorithm that reduces dimensionality of the data without loss of information, and recently it has been incorporated into biotechnological research such as genome-wide expression studies (30). The reduction is accomplished by identifying variables, called principal components (PCs), along which the variation in the data is maximal. PCA identifies new variables, such as PC1 or PC2, which are linear combinations of the original variables. Assuming that the physicochemical properties of peptides are used as original variables, the reduction of dimensionality obtained by PCA would result in two peptides with similar characteristics to be placed at points roughly close to each other in the PCA plot. Many different properties, such as polarity, charge, hydrophobicity, and molecular weight of each amino acid residue, and the average, maximum, and minimum values of each in the peptide sequence, can be used as original variables. It is precisely the multi-dimensionality of the properties of a peptide sequence that makes the analysis of peptide characteristics difficult.

Indeed, analysis of peptides with high-affinity binding to a target protein should be performed by selecting relevant variables for peptide binding affinity first, and then the combinatory model of the relevant variables should be constructed to predict the binding strength. However, it is not possible to guarantee that the correct relevant variables will be selected from restricted data. In addition, selection of variables by supervised learning is more difficult when only a small sample is used compared to a very large original population. The reduction of dimensionality without loss of information by PCA is very attractive because only two PCs allow a clear understanding of localized characteristics of high-affinity binding peptides.

In the present study, the following methodology was employed. A PCA of peptide sequence properties was conducted against a predesigned short peptide library and only two PCs were identified. Next, binding affinity of the peptide library with the target protein was investigated. After obtaining the area of the high-affinity binding peptides in a score plot of the two PCs, the area characteristics were used for searching the larger peptide library, and any resulting correlation between sequence and function was investigated. For instance, we conducted PCA of a tetramer peptide library consisting of 512 peptides, and the variables corresponding to the physicochemical properties of peptides were reduced to two PCs. We selected IL-2 and IgG as model proteins and the binding affinity for each protein was assayed using a peptide array consisting of the 512 peptides. When the binding affinity data was laid over PCA plots, PCAs with 16 and 18 variables were suitable for localizing IL-2 and IgG high-affinity binding peptides, respectively, into the restricted region of the PCA plot. We then tested whether the binding properties of an octamer peptide library to IL-2 and IgG could be predicted using the identified region from the PCA of the tetramer library. The binding affinities of octamer peptide libraries using this method are discussed.

MATERIALS AND METHODS

Preparation of peptide array The peptide arrays were prepared as described previously (29). A cellulose membrane (grade 542; Whatman, Maidstone, UK) was activated using β-alanine as the basal spacer. The activated 9-fluorenylmethoxycarbonyl amino acids (at 0.25 M) were then spotted using a peptide auto-spotter (ASP222, Intavis AG, Köln, Germany). The membrane was washed with *N*,*N*′-dimethyl formamide after spotting, and deprotected with 20% piperidine. Additional washes were performed with *N*,*N*′-dimethyl formamide, followed by washing with methanol. These steps were repeated at every elongation step. After elongation, the side-chain-protecting groups were removed by incubation in *m*-cresol:thioanisole:ethanedithiol:trifluoroacetic acid (1:6:3:40) for 3 h, and the membrane was washed with diethyl ether and methanol.

Preparation of peptide library A tetramer peptide library consisting of 512 peptides covering the four important physicochemical properties for protein–protein interactions was obtained as previously described (29) and is shown in Table S1.

An ordinary octamer peptide library was obtained from sequences of the extracellular domain of the human IL-2 receptor. The library consisted of 234 peptides designed as octamer peptides with two amino acid shifts (Table S2). The sequences of the human IL-2 receptor alpha (P01589) and beta (P14784) chains were obtained from Universal Protein Resource (http://www.uniprot.org/).

A random octamer peptide library was designed by combining two peptides randomly selected from the tetramer peptide library. The random octamer library consisted of 640 peptides (Table S3).

Binding assay of peptide arrays to IL-2 or IgG The binding assay of the peptide arrays to the target proteins was performed as described previously (29). After removing side-chain-protecting groups by incubation for 16 h, the peptide array membranes were washed several times with phosphate-buffered saline (PBS, pH 7.0) to remove traces of the reagent. The membranes were then soaked in 1% Block Ace Powder (DS Pharma Biomedical, Osaka, Japan) in PBS for 1 h to block unspecific adsorption. After washing three times with PBS containing 0.05% Tween-20 (T-PBS) for 5 min, the arrays were incubated with each of the IL-2 solutions in PBS at a final concentration of 130 nM for 2 h at 37°C with rotation at 55 rpm using a small-size shaker (NR-3, Taitec Corporation, Saitama, Japan). Unbound cytokine was removed by washing the array with 0.01% T-PBS three times for 5 min. The peptide arrays were then incubated with primary rabbit antibodies against the target (ab9618 for IL-2, Abcam, Cambridge, UK) in PBS containing 0.25% Block Ace Powder for 2 h at 37°C with rotation at 55 rpm. Unbound primary antibody was removed by washing the array three times for 5 min with 0.01% T-PBS. Next, the peptide arrays were incubated with secondary antibody (anti-mouse goat antibody) conjugated with Alexa 488 (ab150117, Abcam) in PBS containing 0.25% Block Ace Powder for 2 h at 37°C with rotation at 55 rpm. Unbound secondary antibody was removed by washing the array three times for 10 min with 0.01% T-PBS.

The fluorescence intensity of each peptide spot from the tetramer peptide library and from the ordinary octamer peptide library derived from the extracellular domains of the IL-2 receptor was determined using a Typhoon FLA-9500 (GE Healthcare Life Sciences, Buckinghamshire, UK). Fluorescence intensity from the random octamer peptide library was determined using an FLA-7000 (GE Healthcare Life Sciences) at 494/519 nm and ImageQuantTL (GE Healthcare Life Sciences).

The amount of protein bound to a peptide spot was calculated from the fluorescence intensity as follows: (fluorescence intensity of protein bound to a peptide spot) = (fluorescence intensity of the peptide spot with cytokine) − (fluorescence intensity of the peptide spot without cytokine).

Principal component analysis A principal component analysis (PCA) was performed using R (R Development Core Team, https://www.r-project.org/). Principal component 1 and 2 (PC1 and 2) were composed of physicochemical properties which were included isoelectric point (31), hydropathy index (32), polarity (31), side-chain contribution to protein stability (33), partition coefficient (Log P), molecular weight (34), Pk1 and Pk2 (35) of the amino acid residue, and the number of the amino acid group in peptides (29) (Table S4).

To determine the suitable number of input variables for tetramer peptide libraries, we performed several PCAs (numbered 1–16) with two components. Variable decreasing method was conducted using 49 to 9 variables (Table S4). We first selected PCAs number 10, 11, 14, and 16 as candidates for fitting variables by cumulative proportion of PC1 and PC2. We then calculated the correlation coefficient between fluorescence intensity and PC1 or PC2, and found that PCA 11 (16 variables) for IL-2 and PCA 10 (18 variables) for IgG were the most fitting variables with the highest coefficients. The proportions of variance and cumulative proportions were calculated from the PCA with R.

Analysis of high- and low-affinity binding peptide rules for IL-2 and IgG We determined the high- and low-affinity binding rules for the tetramer, ordinary octamer, and random octamer peptide libraries from the PC scores of the tetramer peptide library. First, we selected the top 20 and bottom 20 peptides in fluorescence intensity from the tetramer peptide library. Second, we performed PCA of the top 20 and bottom 20 peptides using 16 variables for IL-2 and 18 variables for IgG, and calculated the averages and standard deviations (SDs) of the PC scores for the top 20 and bottom 20 peptides. Third, we determined the high-binder region from average ± SD of the top 20 peptides in a score plot of PC1 and PC2. In addition, we determined the low-binder region from average ± of the bottom 20 peptides in a score plot of PC1 and PC2. The area excluded by these regions was referred to as others.

Next, we calculated the averages and SDs of the PC scores for the ordinary and random octamer peptide libraries. Peptides from these octamer libraries localized

Please cite this article in press as: Kume, A., et al., Exploring high-affinity binding properties of octamer peptides by principal component analysis of tetramer peptides, J. Biosci. Bioeng., (2016), http://dx.doi.org/10.1016/j.jbiosc.2016.08.005