Korean J. Chem. Eng., 35(3), 717-724 (2018)
DOI: 10.1007/s11814-017-0350-4

pISSN: 0256-1115
eISSN: 1975-7220

Computational screening of potential non-immunoglobulin scaffolds using overlapped conserved residues (OCR)-based fingerprints

Ganapathiraman Munussami[‡], Sriram Sokalingam[†,‡], Selvakumar Edwardraja, Jung Rae Kim, Sungwook Chung, and Sun-Gu Lee[†]

Department of Chemical and Biomolecular Engineering, Pusan National University, Busan 46241, Korea
(*Received 25 September 2017 • accepted 20 November 2017*)

Abstract−Cystatins and lipocalins have attracted considerable interest for their potential applications in non-immunoglobulin protein scaffold engineering. In the present study, their potential homologs were screened computationally from non-redundant protein sequence database based on the overlapped conserved residues (OCR)-fingerprints, which can detect the protein family with low sequence identity, such as cystatins and lipocalins. Two types of OCR-fingerprints for each family were designed and showed very high detection efficiency (>90%). The protein sequence database was scanned by the fingerprints, which yielded the hypothetical sequences for cystatins and lipocalins. The hypothetical sequences were validated further based on their sequence motifs and structural models, which allowed an identification of the potential homologs of cystatins and lipocalins.

Keywords: Protein Scaffold, Cystatins, Lipocalins, OCR-fingerprints

INTRODUCTION

Non-immunoglobulin (non-Ig) protein scaffolds, which are expected to overcome the size and stability limitations of monoclonal antibodies, are becoming a growing trend in pharmaceutical biotechnology. In general, a protein is considered to be an ideal candidate for a non-Ig scaffold when it is small and a single polypeptide chain with high stability and solubility, and preferably contains native binding sites [11]. A good protein scaffold is not restricted to just the above-mentioned characteristics; it also has the potential to withstand various selection strategies and protein engineering techniques for target applications. A range of protein folds have been analyzed using combinatorial approaches to assess their potential as protein scaffolds in therapeutic studies [1]. Over the past two decades, approximately 50 different non-Ig protein scaffolds have been proposed as alternative binding protein candidates [13].

Adhiron and Anticalin proteins are representative commercialized non-Ig protein scaffolds that are considered to be an alternative monoclonal antibody (mAb) platform [14]. These proteins are structurally stable and bind the target molecules with similar specificity and affinity to that of antibodies. They can be engineered to have a variety of molecular recognition properties because of their flexible and structurally non-conserved binding sites. In addition, they are much smaller and have better tissue penetration properties than monoclonal antibodies. Identifying homologous proteins of Adhiron and Anticalin proteins in nature is important because their homologs can be a template for the design and engineering of new scaffold proteins.

[†]To whom correspondence should be addressed.
E-mail: biosriram@gmail.com, sungulee@pusan.ac.kr
[‡]These authors contributed equally to this work.

With the exponential growth of the protein sequence database, the approach to find novel proteins using computational screening has become a general tool. One representative method of computational screening is to scan the protein sequence database using the conserved sequence patterns or fingerprints for the homologs of the target protein. On the other hand, as the sequence identity of the target protein homologs is lower, it is becoming increasingly difficult to devise efficient fingerprints for target homologs. Although the structures of Adhiron or Anticalin homologs are highly conserved, their sequence identities are as low as 15%. This low sequence identity limits the design of efficient fingerprints for their homologous proteins, which impedes the computational screening of new potential scaffolds from the protein sequence database.

We previously developed a method to devise an efficient fingerprint for sequentially distant homologous proteins [6]. The approach, called the overlapped conserved residues (OCR) method, designs a protein fingerprint by considering the commonly conserved residues of the protein family in three attributes: sequence, structure, and intramolecular interaction. Technically, the commonly conserved residues are selected by performing three individual sequence alignments of known homologs: multiple sequence alignment (MSA), structure based alignment (SBA) and super-secondary structure (SSS) alignment methods, and are used in the design of fingerprints. The devised approach was implemented on Immunoglobulin V-set domain (IgV) as a model system to present the detailed procedure and efficiency of the fingerprint extracted using the proposed strategy. The approach was also benchmarked by applying on various protein folds, such as beta-strand rich, alpha+beta, and alpha/beta protein folds with a range of sequence similarities. The OCR-based approach was expected to be able to detect distant homologs, which might allow us to identify novel homologous proteins that might have specific functions in various organisms.

In this study, an attempt was made to devise fingerprints for the

homologous proteins of Adhiron and Anticalin using the OCR-based approach, and their potential homologs were screened from the protein sequence database. The first part of this paper describes how the OCR-based approach can be used to design fingerprints for target homologs and test their detection efficiencies. Second, candidates of target homologs are identified by scanning the non-redundant (nr) protein sequence database using the fingerprints. Finally, the identified candidates are validated at the sequence and structural levels using the conserved motifs of the targets and computational structural modeling of their sequences.

METHODS

1. Design of OCR-based Fingerprints

The OCR fingerprint was generated by selecting the commonly conserved residue positions identified by the multiple sequence alignment (MSA), structure-based alignment (SBA), and super-secondary structure-based alignment (SSS) methods. To perform the individual alignments, the representative distant proteins were taken from the same protein family listed in the SCOPe database with a sequence identity less than 90% between any two members. Five proteins (1cewI, 3nx0A, 1rn7A, 1eqkA, 3kfqC) and 9 proteins (1kt3A, 1iw2A, 1epaA, 1b0oA, 1gkaA, 1dfvA, 1jzuA, 1bbpA, 1i06A) were selected as the distant proteins for the cystatin homologs and lipocalin homologs, respectively. SSS-based alignment was conducted, as described elsewhere [9]. MSA was performed on the Clustal Omega server using the default parameters [12]. SBA was performed using the DALI server by selecting the structural neighbors [7]. The OCR^s fingerprint was generated by considering only the selected residues in the beta-strands of OCR. Details of the procedures to design the OCR-based fingerprints and the selection criteria for the representative protein sequences/structures are described in our previous report [6].

2. Measurement of Detection Efficiency of the Fingerprint

The fold detection efficiency of the individual alignment method was deduced by employing the generated fingerprints in the motif search tool from the GenomeNet to scan the PDB. The effectiveness of an OCR-based pattern was determined in terms of the 'specificity', 'sensitivity', and 'efficiency'. The 'specificity' is calculated as the ratio of 'true positive (TP)' hits to the total of 'true positive' and 'false positive (FP)'. The 'sensitivity' is calculated as the ratio of 'true positive' hits to the total number of homologs, i.e., true positives and false negatives (FN), in the PDB database. The 'efficiency' is calculated as the ratio of 'true positive' hits to the total number of hits.

3. Identification and Validation of Potential Cystatin and Lipocalin Homologs

To detect the novel homologs of cystatin and the lipocalin protein fold, the fingerprints were used as the input in the perl program ps_scan from prosite to scan the NCBI nr database [4]. The sequence hits identified were analyzed as true positive hits, false positive hits, and hypothetical hits. The identified hypothetical proteins were validated at the structure and sequence levels. InterProScan was used to analyze the novel structural homologs at the sequence level [2]. The consortium of databases that contribute to protein signature identification in the InterProScan application assists in identifying the novel structural homologs of cystatins and lipocalins and to place them in their respective protein family. Validation at the structural level was performed by modeling the structure of the hypothetical proteins using PHYRE2 (**P**rotein **H**omology/analog**Y** **R**ecognition **E**ngine V 2.0) [8]. In the fold prediction server, a one-to-one threading method was used to model and identify the fold of the resulting structural model. The predicted models were analyzed further in the PROFUNC and COFACTOR servers to assess the fold and structural analogs with the PDB database [10-16]. The PROFUNC server analyzes the modeled structure to retrieve the matching fold from the PDB database and the likely biochemical function of the hypothetical protein. The COFACTOR server also performs a structural analog search to present the proteins from the PDB database and its annotated biological function.

RESULTS

1. Generation of OCR Fingerprints for Adhiron and Anticalin Homologs

The detailed principle and procedure to devise an OCR-based fingerprint is described in our previous report [6]. Briefly, the representative sequences for a set of homologous proteins were selected and aligned by performing three individual sequence alignments: multiple sequence alignment (MSA), structure based alignment (SBA), and super-secondary structure (SSS) alignment methods. The commonly conserved positions were selected from the alignments, and used to generate a sequence pattern, designated as OCR-fingerprint. The sensitivity of the OCR-fingerprint can be improved further by using the conserved residues in the secondary structures and regarding the rest of the secondary structure elements as gaps. This fingerprint is designated as the OCR^s-fingerprint. Based on our previous study, OCR^s-fingerprint generally shows lower specificity sensitivity than OCR-fingerprint [5,6]. Thus, the set of proteins screened by OCR^s-fingerprint includes a higher ratio of false positives compared to that obtained by OCR-fingerprint. However, OCR^s-fingerprint generally shows higher sensitivity than OCR-fingerprint, which allows us to detect more candidates of homologs and can sometimes be advantageous in the screening of unknown homologs. Therefore, both the two types of fingerprints, OCR-fingerprint and OCR^s-fingerprint, were designed and used for the computational screening of the protein homologs in this study.

Adhiron proteins originate from phytocystatin, which belongs to the cystatins family in the cystatin/monellin superfamily under the cystatin-like fold and alpha and beta ($\alpha+\beta$) fold class of proteins [15]. The mean size of the proteins in the cystatin/monellin superfamily was approximately 100 amino acid residues, and it was defined structurally by the four-strand antiparallel β-sheet core with a central helix (Fig. 1(a)). To prepare the OCR-based fingerprints for the Adhiron homologs, the proteins in the cystatins family in the SCOPe database were determined to be a set of homologous proteins. The identity of the homologous protein sequences in the cystatins family was analyzed, and five distant representative protein templates were selected (Table 1(a)). The OCR- and OCR^s-fingerprints were designed based on the OCR approach, designated as OCR-cystatin and OCR^s-cystatin (Fig. 2(a)).

The Anticalin proteins are from the human lipocalins that belong

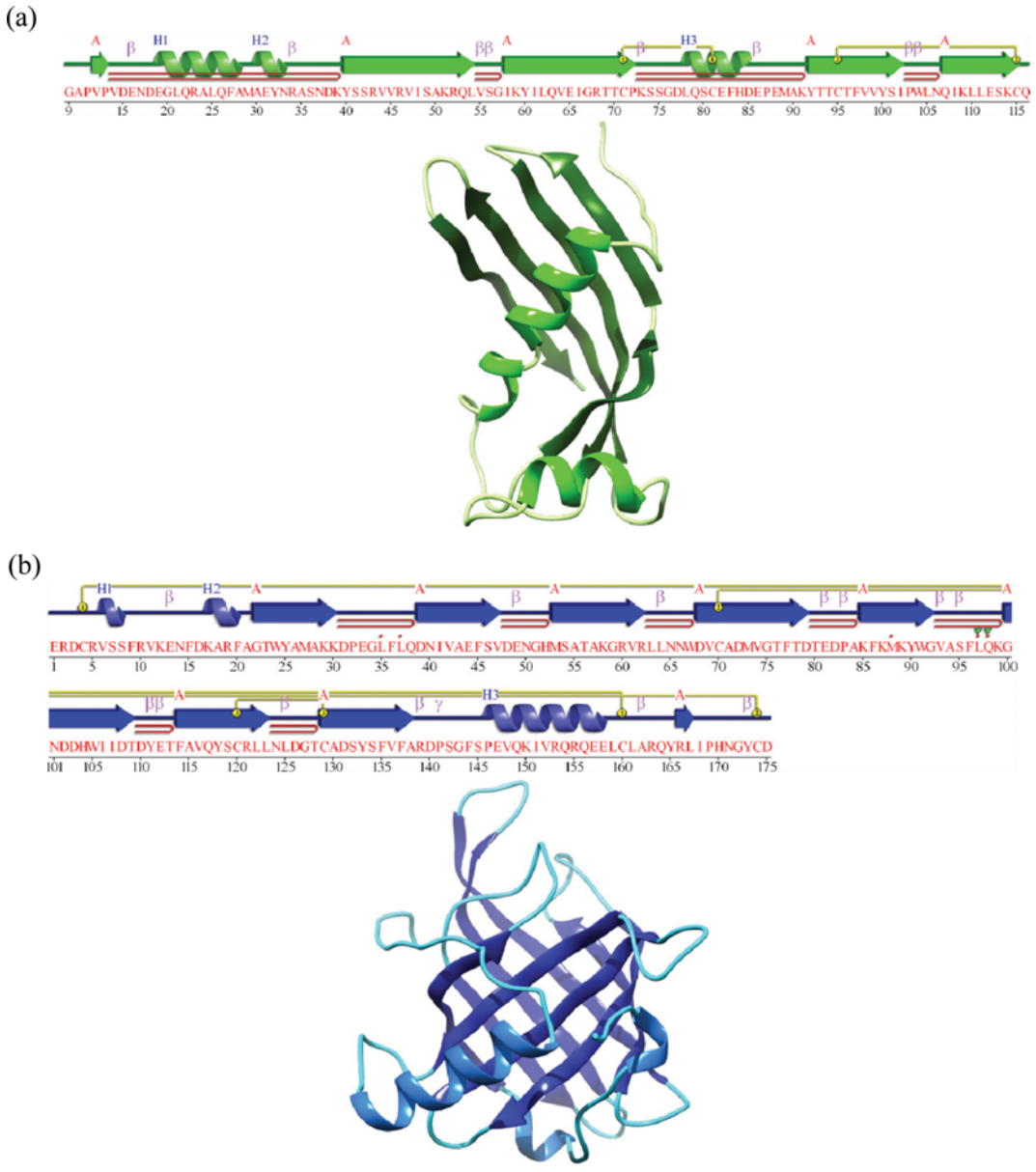

Fig. 1. Sequences, secondary structures and 3-dimensional structures of (a) a protein (PDB ID: 1cewI) in cystatins family, and (b) a protein (PDB ID: 1kt3A) in RBP-binding family.

to the retinol binding protein (RBP)-like family in the lipocalin superfamily under lipocalin fold and all-beta classes of protein fold in SCOPe. The proteins in the lipocalin superfamily have a single domain protein architecture with a β-barrel type of fold containing eight antiparallel β-strands and one or two geometrically outlier helices (Fig. 1(b)). Members of the lipocalin superfamily are divided into kernel and outlier lipocalins, which are defined by the conservation of sequence motifs. Kernel lipocalins have three highly conserved sequence motifs whereas outlier lipocalins are smaller subset of lipocalin proteins with only one conserved sequence motif [3]. Anticalin is from the kernel lipocalins, and the OCR-based fingerprints for Anticalin homologs were generated using the kernel lipocalins in the RBP-like family as a set of homologous proteins. Nine distant protein templates were selected from the homologous protein set (Table 1(b)). The OCR- and OCR^s-fingerprints were designed based on the templates, which were designated as OCR-lipocalin and OCR^s-lipocalin (Fig. 2(b)).

2. Detection Efficiencies of the Fingerprints

To test the efficiencies of the designed fingerprints, the PDB database, which includes 28 non-redundant proteins of the cystatins family and 148 non-redundant kernel type lipocalins of the RBP-like family, were scanned by the fingerprint. The protein hits were analyzed and classified as true positive (TP), false negative (FN), and false positive (FP) proteins. The effectiveness of the OCR-based pattern was determined in the terms of 'specificity', 'sensitivity', and 'efficiency'. The 'specificity' was calculated as the ratio of 'true positive (TP)' hits to the total 'true positives' and 'false positives (FP)'. The 'sensitivity' is calculated as the ratio of 'true posi-

Table 1. Representative protein templates of (a) cystatins and (b) lipocalins for OCR fingerprint generation

(a)

Sr. No.	PDB ID	Name	Sequence length	Sequence identity	rmsd	# Strands	# Helices
1	1cewI	Cystatin	108	100	0.0	4	2
2	3nx0A	Cystatin-C	109	39	1.5	4	2
3	1rn7A	Cystatin-D	112	37	1.9	5	1
4	1eqkA	Oryzacystatin-I	102	29	2.5	4	1
5	3kfqC	Cathepsin L2	98	15	2.6	4	1

(b)

Sr. No.	PDB ID	Name	Sequence length	Sequence identity	rmsd	# Strands
1	1kt3A	Plasma retinol-binding protein	175	100	0.0	8
2	1iw2A	Complement protein C8 Gamma	163	25	2.6	10
3	1epaA	Epididymal retinoic acid-binding protein	151	21	2.3	9
4	1b0oA	Beta-lactoglobulin	161	19	2.6	9
5	1gkaA	Crustacyanin A1 subunit	180	19	2.6	9
6	1dfvA	Human neutrophil gelatinase	173	16	2.7	8
7	1jzuA	Lipocalin Q83	157	18	3.8	8
8	1bbpA	Bilin binding protein	173	14	3.4	9
9	1i06A	Major urinary protein I	156	15	3.1	9

(a)

	LOOP 1			
OCR	x(8,19)-[DEPT]-x(1)-[EGHS]-x(2)-[DENR]-x(1)-[ALV]-[DHQR]-x(1,7)-[AV]-x(2)-[EQ]-x(1)-[EN]-[EKR]-x(2)-[KN]			
OCRs	x(37,49)			
	STRAND 1		LOOP 2	
OCR	x(9,13)-[AVY]-x(2)-[Q]		x(3,4)	
OCRs	[AVY]-x(2)-[Q]		x(3,4)	
	STRAND 2			
OCR	[G]-x(2)-[Y]-[FIY]-[FIL]-[DKNQT]-[IV]-[EKR]-[AFILV]-[GK]			
OCRs	[G]-x(2)-[Y]-[FIY]-[FIL]-[DKNQT]-[IV]-[EKR]-[AFILV]-[GK]			
	LOOP 3		STRAND 3	
OCR	x(6,29)		[AFL]	
OCRs	x(6,29)		[AFL]	
	LOOP 4			
OCR	x(19,31)			
OCRs	x(19,31)			

(b)

	LOOP 1			STRAND 1		
OCR	x(2,31)-[EKNQRS]-[FIVY]-x(1)			[G]-x(1)-[W]-x(2)-[FILMV]		
OCRs	x(5,34)			[G]-x(1)-[W]-x(2)-[FILMV]		
	LOOP 2	STRAND 2	LOOP 3	STRAND 3	LOOP 4	
OCR	x(40,56)					
OCRs	x(40,56)					
	STRAND 4	LOOP 5	STRAND 5	LOOP 6	STRAND 6	
OCR	[DEGKPS]	x(5,8)	[FILVY]	x(7,19)	[IV]-x(2)-[T]	
OCRs	[DEGKPS]	x(5,8)	[FILVY]	x(7,19)	[IV]-x(2)-[T]	
	LOOP 7	STRAND 7	LOOP 8	STRAND 8		
OCR	x(2,3)	[KNQST]-x(1)-[AIL]-[CILMV]	x(12,19)	[CILV]-[FLY]-x(1)-[R]		
OCRs	x(2,3)	[KNQST]-x(1)-[AIL]-[CILMV]	x(12,19)	[CILV]-[FLY]-x(1)-[R]		
	LOOP 9					
OCR	[DEST]-x(28,48)					
OCRs	x(29,49)					

Fig. 2. OCR and OCR^s fingerprint patterns for (a) cystatins and (b) lipocalins. The protein fingerprint is presented as PROSITE-like pattern. Here, in the expression of "x(d, r)," "d" indicates the minimum number of residues between two consecutive conserved positions and "r" indicates the maximum number of residues between two consecutive conserved positions. "x" is used if the minimum and maximum distance between two consecutive conserved positions is the same.

Table 2. Detection efficiency of cystatin and lipocalin fingerprints against PDB

Fingerprint	#Hits	TP	FN	FP	Specificity (%)	Sensitivity (%)	Efficiency (%)
OCR-cystatin	26	26	2	0	100	93	93
OCR^s-cystatin	26	26	2	0	100	93	93
OCR-lipocalin	141	141	7	0	100	95	95
OCR^s-lipocalin	141	141	7	0	100	95	95

Table 3. Detection of hypothetical homologs of cystatins and lipocalins in NR-database

Fingerprint	Total #hits	TP	FP	Unannotated	% Specificity
OCR-cystatin	81	81	0	0	100
OCR^s-cystatin	253	244	9	9	96
OCR-lipocalin	892	884	8	24	99
OCR^s-lipocalin	2180	1535	645	78	70

tive' hits to the total number of homologs, i.e., true positives and false negatives (FN), in the PDB database. The 'efficiency' is calculated as the ratio of 'true positive' hits to the total number of hits.

$$\text{Specificity (\%)} = TP/(TP+FP)*100 \quad (1)$$

$$\text{Sensitivity (\%)} = TP/(TP+FN)*100 \quad (2)$$

$$\text{Efficiency (\%)} = TP/(TP+FN+FP)*100 \quad (3)$$

The specificity, sensitivity, and overall efficiency of each fingerprint were estimated, as shown in Table 2. The OCR and OCR^s fingerprints for cystatins and lipocalins showed very high detection efficiency with 100% specificity and more than 90% selectivity. These results suggest that the fingerprints based on the OCR approach can be used efficiently to screen the homologs of the cystatins and kernel type lipocalin proteins in the RBP-like family protein.

3. Detection of Cystatin and Lipocalin Homologs in the NR Protein Sequence Database

The novel homologs of the target scaffolds were identified by scanning the NCBI NR database using the OCR-based fingerprints. The NR database is comprised of the non-identical sequences from GenBank CDS translations, PDB, Swiss-Prot, PIR, and PRF, which possess both annotated and unannotated protein sequences. Scanning was performed using the ps_scan - a PROSITE scanning perl program obtained from the expasy ftp server [4]. The program was executed with the OCR-based fingerprints to scan against the current NR database, which was downloaded to a local server computer.

Table 3 lists the results of NR database scanning. The OCR-cystatin fingerprint identified a total of 81 sequence hits. All 81 sequences were confirmed to be annotated as cystatin-like proteins. These results suggest that the OCR-cystatin is quite specific, but it cannot detect new lipocalin candidates, which can be included in the unannotated sequences. The OCR^s-cystatin fingerprint identified a total of 262 sequence hits. Among these, 244 sequences were annotated as lipocalin proteins; 9 sequences were annotated as non-lipocalin proteins and 9 sequences were identified as unannotated sequences. Based on the annotated protein hits, the specificity of the OCR^s-cystatin against the NR dataset was estimated to be approximately 96%. This suggests that potential cystatin homologs can be included in the nine unannotated sequences identified by OCR^s-cystatin.

Table 4. 3D protein structure prediction results using Phyre2 folds recognition server for the potential cystatins and lipocalins. "Confidence score" and "%ID" represent the probability (from 0 to 100) that the match with input sequence is a true homology and a template-query percentage sequence identity, respectively

Sr. no.	NCBI accession ID	Confidence (%)	% ID
Potential cystatins			
Query 1	EFB29599.1	96	23
Query 2	EPY77294.1	99	19
Query 3	KXG33568.1	100	32
Query 4	OQV17094.1	98	20
Potential lipocalins			
Query 1	JAT26890.1	100	21
Query 2	JAS45546.1	100	19
Query 3	EFB23711.1	100	23
Query 4	OBS79030.1	100	85
Query 5	OBS69293.1	100	15
Query 6	KXJ75307.1	100	17
Query 7	KXJ74931.1	100	22
Query 8	KXJ73040.1	100	18
Query 9	KXJ67955.1	100	20
Query 10	KTG35225.1	100	15
Query 11	KTG03214.1	100	17
Query 12	KTF91244.1	100	19
Query 13	EHH56965.1	100	19
Query 14	EHH23781.1	100	14
Query 15	EHH23656.1	100	27
Query 16	EHH19237.1	100	93
Query 17	ETN61002.1	100	20
Query 18	ETN61000.1	100	18
Query 19	XP_002593363.1	100	20

The OCR-lipocalin identified a total of 916 sequence hits. Among these, 884 sequences were annotated as lipocalin proteins; 8 were annotated as non-lipocalin proteins and 24 were identified as unannotated sequences. The OCR^s-lipocalin identified a total of 2258 sequence hits. Of these, 1535 sequences were annotated as lipocalin proteins; 645 were annotated as non-lipocalin proteins and 78 were identified as unannotated sequences. These results suggest that OCR^s-lipocalin can detect more lipocalin homologs in the NR database than OCR-lipocalin, but the detection specificity of OCR^s-lipocalin was much lower than that of OCR-lipocalin. Based on the annotated proteins, the specificity of OCR and OCR^s against the NR dataset was estimated to be approximately 99% and 70%, respectively. This suggests that the 24 unannotated sequences identified by OCR-lipocalin might include a relatively higher ratio of potential lipocalin homologs compared to the unannotated sequences identified by OCR^s-lipocalin.

4. Validation of the Potential Cystatin and Lipocalin Protein Homologs

As described above, the 9 and 24 unannotated sequences identified by OCR^s-cystatin and OCR-lipocalin, respectively, were expected to include potential target homologs with a high ratio. Further studies were performed for validation. The 78 unannotated sequences identified by OCR^s-lipocalin might also include the lipocalin proteins, but further validation was not performed for the sequences due to the relatively low specificity of OCR^s-lipocalin against the NR-database. To validate the identified proteins, their sequences were first validated through InterProScan, which contains a consortium of member databases and the results of sequence analysis from various databases [2]. The protein sequences were submitted to an InterProScan and the scan results were analyzed for the presence of the protein motif and conserved domains of the representative protein family. Sequence validation through InterProScan showed that 4 proteins from 9 cystatin candidates and 19 proteins from 24 lipocalin candidates had the respective motif and were found to belong to the cystatin and lipocalin homologs. Further validation was performed through computational structural modeling. The 4 and 19 proteins were modeled by a one-to-one threading method in the Phyre2 server [8]. In one-to-one threading, the sequence of interest was modelled based on the template submitted. For cystatins, the chicken egg white cystatin

Table 5. Potential biochemical functions for the identified homologs of cystatins and lipocalins predicted by ProFunc and COFACTOR server. ProFunc results are shown using Q-score, PDB-ID, and Family, which represent the quality function of an alignment, PDB-ID of the best matching fold, and family name, respectively. COFACTOR server results are shown using TM-score, PBD_ID and Family, which represent the quantitative assessment score, PDB-ID of the best matching fold and family name, respectively

Structural homolog and potential biochemical function with the PDB identified by ProFunc server				Biological function to the structural analog in PDB predicted by COFACTOR server		
Sr. No.	Q-score	PDB ID	Family	TM-score	PDB ID	Family
Potential cystatins						
1	0.444	1yvb	CY	0.94	1yvb	CY
2	0.769	1cew	CY	0.96	1yvb	CY
3	0.787	1cew	CY	0.96	1yvb	CY
4	0.759	1cew	CY	0.96	1yvb	CY
Potential lipocalins						
1	0.966	1kt3	LCN	0.98	1aqb	LCN
2	0.966	1kt3	LCN	0.98	1aqb	LCN
3	0.903	1kt3	LCN	0.98	1aqb	LCN
4	1	1kt3	LCN	0.98	1aqb	LCN
5	0.926	1kt3	LCN	0.98	1aqb	LCN
6	0.937	1kt3	LCN	0.98	1aqb	LCN
7	0.966	1kt3	LCN	0.98	1aqb	LCN
8	0.943	1kt3	LCN	0.98	1aqb	LCN
9	0.903	1kt3	LCN	0.98	1aqb	LCN
10	1	1kt3	LCN	0.98	1aqb	LCN
11	0.846	1kt3	LCN	0.98	1aqb	LCN
12	0.926	1kt3	LCN	0.98	1aqb	LCN
13	0.897	1kt3	LCN	0.98	1aqb	LCN
14	0.943	1kt3	LCN	0.98	1aqb	LCN
15	0.903	1kt3	LCN	0.98	1aqb	LCN
16	0.926	1kt3	LCN	0.98	1aqb	LCN
17	0.954	1kt3	LCN	0.98	1aqb	LCN
18	0.96	1kt3	LCN	0.98	1aqb	LCN
19	1	1kt3	LCN	0.98	1aqb	LCN

(PDB ID: 1cewI) was used as a template, and bovine retinol binding protein (PDB ID: 1kt3A) was used as a template for lipocalins. The Phyre2 server predicted that the 4 cystatin candidates and 19 lipocalin candidates have a cystatin-like fold and the fold of lipocalins (Table 4). The root mean square deviations between the structure templates and the one-to-one threading models were less than 1 Å. In addition, most of the predicted structure models showed 100% confidence, further validating the structural integrity of the modelled protein. The homology models of the hypothetical proteins were analyzed using the ProFunc [10] and COFACTOR [16] servers. The predicted models of cystatin and lipocalin homologs were estimated to be synonymous to function as the respective protein families (Table 5). These results support that the proteins listed in the Table 4 are potentially cystatin and lipocalin homologs.

DISCUSSION

In the present study, the OCR-based approach was used to identify potential cystatin and lipocalin homologs in the protein sequence database. The meaning of this study is two-fold. First, new potential homologs of cystatins and lipocalins were discovered. Despite the great potential of cystatin and lipocalin homologs as non-antibody scaffolds, it has been difficult to screen new cystatin and lipocalin homologs from the protein sequence database due to the low sequence identity. The new homologs identified can be a template for the engineering or design of cystatin- or lipocalin-based binding scaffolds. Second, this study showed that the OCR-based approach can be used to detect novel proteins from the sequence database. Although the protein sequence database has been recognized as a source of novel proteins, there has been a limitation in screening novel homologs from the sequence database if the sequence identity of the target homologs is low. The OCR-based approach might contribute to the discovery of novel homologs for many valuable proteins.

When the OCR and OCR^s fingerprints were used to scan the PDB (Table 2), the fingerprints showed 100% sensitivity. On the other hand, their selectivity decreased when the NR database was targeted (Table 3). This suggests that the efficiency of the fingerprint can be changed depending on the target database, which might be caused by the intrinsic information embedded in the designed fingerprint. The OCR fingerprints were designed using the proteins in the PDB as templates. Therefore, the designed fingerprint with information limited to the proteins in the PDB might not work efficiently if the protein homologs in the NR protein sequence database contain distant information from the homologs in PDB. As more distant homologs for a target fold are accumulated in the PDB, more information on the target homologs can be added to the fingerprint, which might increase the efficiency of the fingerprint against the protein databases other than PDB.

As shown in the Table 3, the OCR^s-fingerprint could detect more true positives and more false positives than the OCR-fingerprint. This suggests that the sensitivity of the OCR^s-fingerprint is lower, but its selectivity is higher than the OCR-fingerprint. The efficiency of a designed fingerprint is generally determined by the amounts of essential and nonessential information for the target homologs in the fingerprint. Theoretically, the selectivity and sensitivity of a fingerprint is 100% if the fingerprint contains all the essential information and has no nonessential information. On the other hand, the sensitivity decreases as essential information is lost, and the selectivity decreases as nonessential information is added. The OCR^s-fingerprint was designed by removing the conserved residues in the loop regions, which might include some essential and nonessential information. Therefore, the removal of the conserved residues in the loop can lead to the loss of essential and nonessential residues simultaneously, which might induce a decrease in sensitivity and an increase in selectivity. Hence, selecting only the essential information in the loops efficiently is crucial.

In the validation study of potential cystatin and lipocalin proteins, computational modeling was performed and the modeling results were found to be reliable. The possibility that the identified potential proteins belong to the target scaffold proteins is expected to be high because validation was also carried out using the motifs for the protein homologs. On the other hand, this should be confirmed by experimental approaches, such as X-ray crystallography or NMR methods.

ACKNOWLEDGEMENTS

This work was supported by a 2-Year Research Grant for Pusan National University.

REFERENCES

1. H. K. Binz, P. Amstutz and A. Pluckthun, *Nat. Biotechnol.*, **23**, 1257 (2005).
2. R. D. Finn, T. K. Attwood, P. C. Babbitt, A. Bateman, P. Bork, A. J. Bridge, H. Y. Chang, Z. Dosztanyi, S. El-Gebali, M. Fraser, J. Gough, D. Haft, G. L. Holliday, H. Z. Huang, X. S. Huang, I. Letunic, R. Lopez, S. N. Lu, A. Marchler-Bauer, H. Y. Mi, J. Mistry, D. A. Natale, M. Necci, G. Nuka, C. A. Orengo, Y. Park, S. Pesseat, D. Piovesan, S. C. Potter, N. D. Rawlings, N. Redaschi, L. Richardson, C. Rivoire, A. Sangrador-Vegas, C. Sigrist, I. Sillitoe, B. Smithers, S. Squizzato, G. Sutton, N. Thanki, P. D. Thomas, S. C. E. Tosatto, C. H. Wu, I. Xenarios, L. S. Yeh, S. Y. Young and A. L. Mitchell, *Nucleic. Acids Res.*, **45**, D190 (2017).
3. D. R. Flower, *Biochem. J.*, **318**(Pt 1), 1 (1996).
4. A. Gattiker, E. Gasteiger and A. M. Bairoch, *Appl. Bioinformatics*, **1**, 107 (2002).
5. A. Goyal, B. G. Kim, K. S. Hwang and S. G. Lee, *Biotechnol. Bioproc. E*, **20**, 431 (2015).
6. A. Goyal, S. Sokalingam, K. S. Hwang and S. G. Lee, *Sci. Rep-Uk*, **4**, 5643 (2014).
7. L. Holm and P. Rosenstrom, *Nucleic. Acids Res.*, **38**, W545 (2010).
8. L. A. Kelley, S. Mezulis, C. M. Yates, M. N. Wass and M. J. E. Sternberg, *Nat. Protoc.*, **10**, 845 (2015).
9. A. E. Kister and I. Gelfand, *P. Natl. Acad Sci. USA*, **106**, 18996 (2009).
10. R. A. Laskowski, J. D. Watson and J. M. Thornton, *Nucleic. Acids Res.*, **33**, W89 (2005).
11. P. A. Nygren and A. Skerra, *J. Immunol. Methods*, **290**, 3 (2004).
12. F. Sievers, A. Wilm, D. Dineen, T. J. Gibson, K. Karplus, W. Z. Li, R. Lopez, H. McWilliam, M. Remmert, J. Soding, J. D. Thompson

and D. G. Higgins, *Mol. Syst. Biol.*, **7**, 539 (2011).
13. A. Skerra and S. R. Schmidt, *Pharm. Bioprocess*, **3**, 383 (2015).
14. K. Skrlec, B. Strukelj and A. Berlec, *Trends Biotechnol.*, **33**, 408 (2015).
15. C. Tiede, A. A. Tang, S. E. Deacon, U. Mandal, J. E. Nettleship, R. L. Owen, S. E. George, D. J. Harrison, R. J. Owens, D. C. Tomlinson and M. J. McPherson, *Protein Eng. Des. Sel.*, **27**, 145 (2014).
16. C. X. Zhang, P. L. Freddolino and Y. Zhang, *Nucleic. Acids Res.*, **45**, W291 (2017).