A New Biclustering Algorithm with Exclusive Random Selection of Columns for Predicting Recognition Spots on Protein Molecular Surfaces

Nishimura, Hiroto

Ohkawa, Takenao

(Citation)
International journal of bioscience, biochemistry, bioinformatics (IJBBB),8(1):11-19

(Issue Date)
2018-01

(Resource Type)
journal article

(Version)
Version of Record

(URL)
https://hdl.handle.net/20.500.14094/90004843

A New Biclustering Algorithm with Exclusive Random Selection of Columns for Predicting Recognition Spots on Protein Molecular Surfaces

Hiroto Nishimura, Takenao Ohkawa*
Graduate School of System Informatics, Kobe University, 1-1, Rokkodai, Nada, Kobe, 657-8501 Japan.

* Corresponding author. Tel.: +81-78-803-6213; email: ohkawa@kobe-u.ac.jp
Manuscript submitted July 9, 2017; accepted September 3, 2017.
doi: 10.17706/ijbbb.2018.8.1.11-19

Abstract: A protein and a ligand have a process of recognizing each other when they are remote before approaching for binding. In this process, there should be a local portion corresponding to a target of the recognition (a recognition spot) on the protein molecular surface. In this paper, we proposed a new biclustering algorithm BISERS for predicting recognition spots on protein molecular surfaces, in which a portion of the molecular surface of a query protein that frequently shows the similarity to other specific proteins binding to the similar ligand is extracted by biclustering. In BISERS, the similarity between rows as well as the similarity between columns is introduced into the evaluation function for updating the biclusters. In addition, the random sampling is applied to the process of exclusive selection of the column for reducing the computational cost effectively. Experimental results for 60 proteins show the effectiveness of our BISERS algorithm.

Key words: Biclustering, data mining, protein, recognition spot.

1. Introduction

Since protein is one of the important biological molecules for life phenomenon, understanding the functions of proteins is fundamental to biomedical science. Proteins often show their functions by binding to other compounds (ligands) on protein molecular surfaces. Since hotspot residues play an important role on binding process, so far, many computational methods for predicting the hotspots have been proposed [1]-[4]. On the other hand, a protein and a ligand have a process of recognizing each other when they are remote before approaching for binding. In this process, there should be a local portion corresponding to a target of the recognition (a recognition spot), which is not necessarily the same as a hotspot, on the protein molecular surface.

We have proposed a method of predicting a protein recognition spot based on the idea that the recognition spots observed from the proteins binding to similar compounds have some frequently shared features [5]. This method consists of three steps. The first step is extraction of feature points from protein molecular surfaces. In this process, feature points are extracted based on curvature-based approach [6] in the form of point feature histograms [7] with physical properties such as electrostatic potential and hydrophobicity. In the second step, structures of proteins, namely sets of feature points located on the 3D space, are compared and a list of matched feature points is derived from the results of pattern matching between a query protein, which is one whose recognition spot is unknown, and each of the reference

proteins, which are proteins whose binding ligands have been analyzed. A matched point matrix, which is a binary matrix that shows correspondences between feature points of the query protein and the reference proteins, is generated by aggregating the lists of matched feature points for all of reference proteins. In the final step, the recognition spot is extracted by biclustering of the matched point matrix. In this process, we have to consider the similarity between columns, since columns in the matched point matrix correspond to reference proteins, which can be classified by prior knowledge. In addition, the exclusive selection of column is required, because more than one columns correspond to the same reference protein. In other words, the bicluster that is composed of columns from the same reference protein is meaningless. Therefore, we have proposed a new biclustering algorithm BISES ((BIclustering based on Similarity and Exclusive Selection of column) [5], in which the similarity between columns and the exclusive selection of a column are considered.

However, our BISES has the following two drawbacks. The first is that the feature points located distantly each other are included in a bicluster generated by BISES. In general, the recognition spot seldom consists of distantly located portions. The second is the computational issue. In BISES, the concept of the group of columns is introduced and one column is selected from each group in updating the bicluster. Although only the combination of selected columns with the highest evaluation value is proceeded to the next step in the greedy manner, the number of combination becomes huge as the size of the data set scales up, which requires large computational resources.

In this paper, we propose a new biclustering algorithm BISERS (BIclustering based on Similarity and Exclusive Random Selection of column), which is a modified version of BISES. In BISERS, the similarity between rows, which correspond to feature points in the query protein, as well as the similarity between columns is introduced into the evaluation function for updating the biclusters. In addition, the random sampling is applied to the process of exclusive selection of the column for reducing the computational cost effectively.

The rest of this paper is organized as follows. Section 2 briefly describes the method for predicting the protein recognition spots and gives an example of the matched point matrix that is a target of biclustering. In section 3, we present the algorithm BISERS. Section 4 gives experimental results and discussions.

2. Overview of Method for Recognition Spots Prediction

Fig. 1 shows a general flow of a method for predicting recognition spots. The first step is extraction of feature points from a protein molecular surface represented as a set of polygon data consisting of the points in 3D space along with physical properties such as hydrophobicity and electrostatic potential. Fig. 2 illustrates an example of the original protein molecular surface and a set of feature points extracted from it.

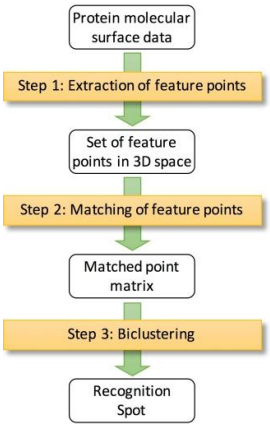
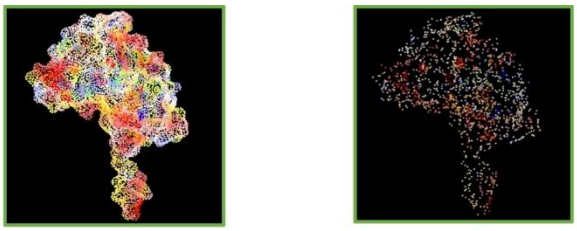


Fig. 1. General flow.



(a) Original molecular surface (b) Extracted feature points

Fig. 2. Examples of extracted feature points.

The second step is matching feature points between a query protein and each of reference proteins. In this step, a set of feature points of the query protein and a set of feature points of the reference protein are given, then a list of corresponding feature points in 3D space between a subset from the query feature points and a subset from the reference feature points is generated. An example of the result of matching is shown in Fig. 3, where the numbers are IDs of the feature points. In this figure, more than one correspondences are shown as 'TOP *x*' in descending order of the number of corresponding feature points. Since we have many reference proteins, a matched point matrix is generated as a final result of this step by composing of all of the list. Fig. 4 shows an example of the matched point matrix, in which '1' indicates that there is a correspondence between a feature point in the query and the reference proteins.

In the final step, important feature points are extracted as a bicluster. In other words, feature points in the query protein that have common correspondences to the feature points from the specific reference proteins are significant and regarded as a candidate of the recognition spot.

Query	Ref. A TOP 1
4726	4677
5120	8234
3463	7556
6523	2345
4345	7534
2534	3434
3425	6533
2355	4432
1424	4344
6346	3223
4412	6433
1255	7323
6312	1275

Query	Ref. A TOP 2
6332	5633
7252	3523
2334	4332
3463	7556
2632	2462
2534	2123
8453	4252
5120	2334
1424	3642
5356	6252
6346	4235

Query	Ref. A TOP 3
6335	7373
5120	3623
4726	3322
6523	4636
4345	1135
3425	7346
2342	2453
8453	5725

Query	Ref. B TOP 1
6523	5673
1255	8436
5424	2523
2521	6343
4345	6322
2534	3235
4122	4652
6346	2463
2334	3457
1431	3462
4412	2364
6312	2632

Query	Ref. B TOP 2
4242	4622
6523	2463
1255	3415
5424	6234
2521	2634
2534	1631
2632	6316
4223	6134
1413	1631
6332	2634
2211	2436

Query	Ref. B TOP 3
6523	2436
1255	2463
2521	1613
2534	3752
2632	2674
6324	6223
2211	2622

Fig. 3. An example of matching results of feature points.

Query protein	Reference protein A TOP 1	TOP 2	TOP 3	Reference protein B TOP 1	TOP 2	TOP 3
1255	1			1	1	1
1413					1	
1424	1	1				
1431				1		
2211					1	1
2334		1		1		
2342			1			
2355	1					
2521				1	1	1
2534	1	1		1	1	1
2632					1	1
3425	1		1			
3463	1	1				
4122				1		
4223					1	
4242					1	
4345	1		1	1		
4412	1			1		
4726	1		1			
5120	1	1	1			
5356		1				
5424				1	1	
6312	1			1		
6324						1
6332		1			1	
6335			1			
6346	1	1		1		
6523	1		1	1	1	1
7252		1				
8453		1	1			

Fig. 4. An example of matched point matrix.

In this paper, we focus on the final step and propose a method of generating an appropriate bicluster from a matched point matrix.

3. Biclustering Algorithm with Exclusive Random Selection of Columns

3.1. BISES

The matched point matrix has the following characteristics.

- Each column represents a reference protein, which can be classified based on various viewpoints.
- Each row corresponds to a feature point in a query protein, which means that the row has spatial information in 3D space.
- One or more columns correspond to the same reference protein.

BISES, a previous version of biclustering algorithm we have proposed, is based on the algorithm PDNS (Pattern-Driven Neighborhood Search) proposed by Ayadi *et al*. [8], in which biclusters are evaluated by correlations of rows and columns and are incrementally updated, namely adding or removing rows and columns, by neighborhood search. BISES introduces the inter-column similarity and the exclusive selection from 'group of columns' in order to deal with the first and the last points mentioned above, but does not consider second point. In addition, BISES encounters computational problem derived from combinatorial explosion for exclusive selection of a column.

3.2. BISERS

3.2.1. Evaluation function

To cope with the above-mentioned problems of BISES, we propose a new biclustering algorithm BISERS, which introduces the inter-row similarity as well as inter-column similarity. A row in the matched point matrix corresponds to a feature point that has coordinate values in 3D space. Since the feature points composing recognition spots tend to be closely located, biclusters that include locally arranged feature points should be obtain high evaluation values. Therefore, we define the inter-row similarity $SIM_r(i,j)$ based on the distance in 3D space between feature points as follow.

$$SIM_r(i,j) = \frac{D(i,j)}{\max_{i,j}(D(i,j))}, \tag{1}$$

where $D(i,j)$ is the distance in 3D space between two feature points p_i and p_j associated with rows indices i and j in the matched point matrix. On the other hand, each column represents a reference protein. We evaluate the similarity between columns based on the similarity between ligands that the corresponding reference proteins bind to. We use the ligand similarity that has been presented in COMPLIG algorithm proposed by Saito *et al*. [9] for evaluating inter-column similarity $SIM_c(i,j)$ as follows.

$$SIM_c(i,j) = \frac{Atom_{MN}+Bond_{MN}}{max(Atom_M+Bond_M, Atom_N+Bond_N)}, \tag{2}$$

where $Atom_M$ and $Atom_N$ are the numbers of atoms in ligands M and N respectively, and $Bond_M$ and $Bond_N$ are the numbers of bonds in ligands M and N respectively. $Atom_{MN}$ and $Bond_{MN}$ are the numbers of matched atoms and matched bonds that are calculated by COMPLIG.

In PDNS algorithm [8], the bicluster $B(I, J)$ (I is a set of rows and J is a set of columns included in the bicluster) is evaluated by Average Spearman's Rho (ASR) defined as follows.

$$ASR(B) = 2\max\left(\frac{\sum_{i\in I}\sum_{j\in I, j\geq i+1}\rho_{ij}}{|I|(|I|-1)}, \frac{\sum_{k\in J}\sum_{l\in J, l\geq k+1}\rho_{kl}}{|J|(|J|-1)}\right), \tag{3}$$

where $\rho_{ij}(i \neq j)$ is the Spearman's rank correlation between the rows indices i and j and $\rho_{kl}(k \neq l)$ is the Spearman's rank correlation between the columns indices k and l. In BISERS, since the bicluster that includes many rows with high inter-row similarity or many columns with high inter-column similarity is expected to be generated, the modified evaluation function ASR-S (Average Spearman's Rho with Similarity), in which weighted correlation values of rows and columns are introduced into ASR, is defined as follows.

$$ASR\text{-}S(B) = \frac{\sum_{i\in I}\sum_{j\in I, j\geq i+1}\rho_{ij}SIM_r(i,j)}{|I|(|I|-1)} \cdot \frac{\sum_{k\in J}\sum_{l\in J, l\geq k+1}\rho_{kl}SIM_c(k,l)}{|J|(|J|-1)}. \quad (4)$$

3.2.2. Efficient exclusive selection of column

As a result of updating the current bicluster in terms of the column, the updated bicluster might include more than one columns generated from the same reference protein. Since such a bicluster is meaningless, the columns derived from the same protein are grouped and one column at most has to be selected from the group and combined with other columns. In other words, multiple combinations of selected columns have to be considered. For each of them, rows are updated and the value of *ASR-S* is evaluated. Ideally, the bicluster with the highest *ASR-S* value is chosen from the all combinations for the next updating step, and these processes are iterated. In this iteration process, however, combinatorial explosion is serious, which prevents practical execution.

Therefore, in BISERS, only one column is randomly selected from each group without enumerating all combinations in order to generate an updated bicluster. Instead of considering exhaustive combinations, this random sampling process is repeated to get various biclusters. After each generated bicluster is updated in terms of rows, the best bicluster with the highest *ASR-S* value is proceeded to the next step. The number of random sampling S is defined as follows.

$$S = S_0 \sum_{i=1}^{n}(c_i - 1) + 1, \quad (5)$$

where n is the number of groups, c_i is the number of columns in group i, and S_0 is a constant value. This process is illustrated in Fig. 5 and the algorithm of BISERS is described in Fig. 6.

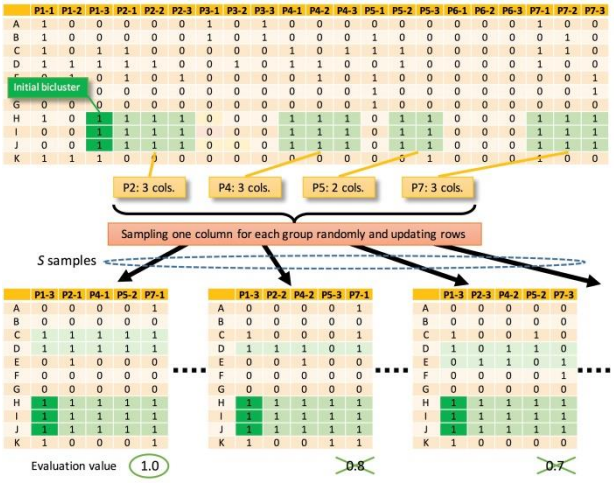


Fig. 5. Process for updating biclusters.

4. Experimental Results

4.1. Data Set

We made experiment with data set including 60 proteins shown in Table 1, four types of proteins binding to each of 15 types of ligands in order to confirm the effectiveness of BISERS. In this table, protein is specified with four letter PDB-ID with one letter chain ID and the number in the parentheses indicates the number of feature points on the molecular surface.

Algorithm of BISERS
Input: M: matched point matrix, B_0: initial bicluster, α, β: threshold,
S: sampling frequency, Y, Z: maximum number of iterations, γ: perturbation rate
Output: B^*: best bicluster

$C \leftarrow \emptyset$
$b \leftarrow B_0$
while the number of iterations $\leq Z$ **do**
  while the number of iterations $\leq Y$ **do**
    $b' \leftarrow$ update$_c(b)$: columns of (# of '1' / # of rows) > α are added
    remove $c \in \{C\}$ from b'
    divide b' into $\{g_1, g_2, \ldots, g_n\}$ each of which includes the columns in the same group
    generate $D = \{d_1, d_2, \ldots, d_s\}$ each of which consists of columns
      exclusively selected from each of $\{g_1, g_2, \ldots, g_n\}$ by random sampling
    for all $d_i \in \{D\}$ **do**
      $d' \leftarrow$ update$_r$(d): rows of (# of '1' / # of columns) > β are added
      calculate evaluation value *ASR-S*(d')
    end for
    $b'' \leftarrow d_i$ having the maximum *ASR-S*(d') in D
    $C \leftarrow$ columns of $b' \notin b''$
    if *ASR-S*(b'') > *ASR-S*(b) **then** $B^* \leftarrow b''$ **end if**
    $b \leftarrow b''$
  end while
  generate a new bicluster b by perturbing randomly γ% of best bicluster B^*
end while

Fig. 6. Algorithm of BISERS.

Table 1. Data Set for Experiments

Ligand	Protein
FMN	1bwk-A(143), 1huv-A(228), 3txz-A(182), 3gfx-A(93)
GLC	2osy-A(99), 2zid-A(129), 4gi6-A(148), 4hoz-A(88)
NAP	1s1p-A(147), 3r7m-A(125), 4lau-A(85), 4gq0-A(80)
SAM	4njk-A(123), 3cb8-A(110), 3t7v-A(249), 4njg-A(139)
PEP	1rzm-A(74), 1xuz-A(89), 3fyo-D(108), 3tfc-A(67)
C2E	4foj-A(76), 3pjt-A(88), 1w25-A(235), 3ign-A(38)
FBP	1w8s-A(98), 3srd-A(75), 3d1r-A(125), 2r8t-A(131)
G2F	4ptw-B(151), 3air-A(127), 2jie-A(164), 2pb1-A(90)
NGT	1hp5-A(61), 3sur-A(49), 2chn-A(51), 1np0-A(34)
PGE	4ogz-A(84), 4kie-A(122), 1xqi-A(112), 2esa-A(21)
ASD	3s79-A(196), 4qdc-A(166), 3nbr-A(40), 2vct-A(135)
B12	3whp-A(76), 5c8a-A(94), 3kox-A(44), 4hut-A(118)
MGR	3hti-A(96), 3bqz-A(168), 3btc-A(147), 3br0-A(168)
STU	4u97-A(99), 2yn8-B(73), 3p86-A(106), 1wvy-A(161)
ACR	1agm-A(54), 1kxh-A(70), 1mxg-A(44), 4uac-A(127)

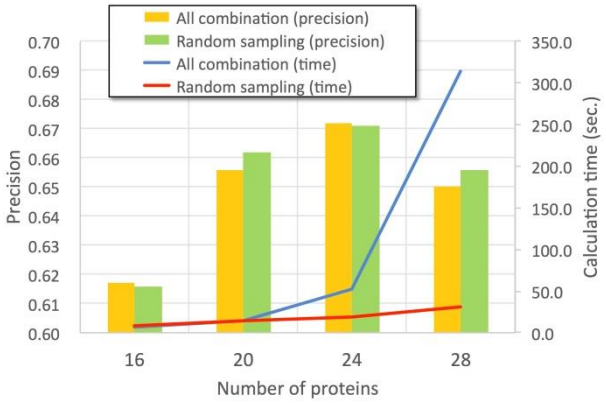


Fig. 7. Comparison between all combination and sampling.

4.2. Effectiveness of Random Sampling

First, we confirm the effectiveness of random sampling for the exclusive selection of columns from the viewpoint of the execution time and the accuracy of the results. The accuracy of the results is evaluated by calculating *precision*, which is the rate of extracted feature points that correspond to residues known to associate with binding in the Protein Data Bank (PDB) [10]. If the residue that is the closest to the extracted feature point is consistent with the residue in actual binding site specified in the PDB data, we judge that the feature point is extracted as the actual binding successfully.

BISERS was compared with the method without random sampling, namely all combinations of exclusive selection of columns from each group. Since the latter method takes much execution time for the original dataset mentioned above, we used subsets for this experiment. Fig. 7 and Table 2 show the average precision and calculation time based on ten times of trials for each method. In this table, the average number of the extracted feature points is also shown. While we can observe no significant differences in terms of the precision as well as the number of extracted feature points between two methods, the computational cost of the proposed method with random sampling is extremely reduced, which implies the effectiveness of the proposed method.

Table 2. All Combination vs. Random Sampling

Data set size (# of prteins)		16	20	24	28
Calculation time	All combination	6.2	14.1	52.9	314.0
	Random sampling	8.4	14.4	19.2	30.7
Precision	All combination	0.617	0.656	0.672	0.650
	Random sampling	0.616	0.662	0.671	0.656
# of extracted feature points	All combination	17.2	17.3	14.3	14.0
	Random sampling	17.6	17.2	14.4	13.7

Table 3. Comparison with Other Biclustering Methods

Method	TOP 1 Only					TOP 1 to 3	
	Bimax	Bibit	PDNS	BISES	BISERS	BISES	BISERS
Precision	0.548	0.572	0.583	0.644	0.655	0.689	0.724
Size of bicluster	53.8	44.1	61.0	54.9	54.6	56.2	55.7

4.3. Comparison with Other Biclustering Methods

A query protein is selected out of 60 proteins in the data set and the rest are treated as reference proteins. We prepare two types of matched point matrix from the results of matching feature points. One is the matrix consisting of only TOP 1 column for each of the reference proteins. The other consists of not only TOP 1 but also TOP 2 and 3 columns. Bimax [11], BiBit [12], PDNS [8], BISES [5] and BISERS are applied to the former matrix, which requires no exclusive selection of a column. BISES and BISERS are applied to the latter matrix. In BISERS, the initial bicluster B_0 has four rows and four columns at least, and other parameters are $\alpha = \beta = 0.6, Y = Z = 10, \gamma = 30$ and $S_0 = 100$.

The average precision and the average size of the generated bicluster for 60 query proteins are shown in Table 3. The size of the bicluster is defined as the product of the number of the feature points and the number of reference proteins in the bicluster. For the matched point matrix containing only TOP 1 columns, BISERS shows the highest average precision, which tells that BISERS considering both the inter-row similarity and the inter-column similarity is effective than some existing biclustering methods. In addition, BISERS considering the exclusive selection from TOP 1, 2, and 3 columns gives the highest precision, which suggests the use of multiple matching results is very significant for predicting the recognition spot from the matched point matrices.

5. Conclusions

This paper proposed BISERS, an efficient biclustering method for extracting important feature points from the matched point matrices to predict recognition sites on a protein molecular surface. Experimentally, we showed that the random sampling for the exclusive selection of column in BISERS reduces the computational cost drastically without sacrificing the prediction accuracy. In addition, we clarified the effectiveness of both the inter-row similarity and the inter-column similarity.

In this paper, we evaluated the prediction results by comparing with the known residues related to binding. However, the recognition spots do not always correspond to the binding residues. Therefore, in future work, we will evaluate the extracted feature points by experiments in vitro such as alanine substitution cooperating with a wet laboratory.

Acknowledgment

We wish to thank Katsumi Inoue, Masato Sasuga, Kenji Tominaga and Toshio Murayama for providing a device for matching feature points in the 3D space. We also wish to thank Dr. Kohei Tsumoto at the University of Tokyo for his relevant advice on interpreting the significance of the results of this study.

References

[1] Li, Z., & Li, J. (2010). Identifying protein binding hot spots by using deeply atomic contacts. *Proceedings of the Fourth International Conference on Computational Systems Biology* (pp. 155-167).

[2] Chen, Y., & Min, X. (2016). Predicting hot spots in protein interfaces based on feature selection using mRMR combining with SVM Forward. *Rev. Téc. Ing. Univ. Zulia*, *39(5)*, 8-13.

[3] Morrow, J. K., & Zhang, S. (2013). Computational prediction of hot spot residues. *Current Pharmaceutical Design*, *18(9)*, 1255–1265.

[4] Lise, S., Archambeau, C., Pontil, M., & Jones, D. T. (2009). Prediction of hot spot residues at protein-protein interfaces by combining machine learning and energy-based methods. *BMC Bioinformatics*, *10(365)*, DOI: 10.1186/1471-2105-10-365.

[5] Nishimura, H., Sakaue, K., & Ohkawa, T. (2016). Extraction of protein recognition spots by biclustering considering exclusive selection of column. *Proceedings of Biotechnology and Bioinformatics Symposium 2016*.

[6] Ho, H. T., & Gibbins, D. (2009). A curvature-based approach for multi-scale feature extraction from 3D meshes and unstructured point clouds. *IET Computer Vision*, *3(4)*, 201-212

[7] Rusu, R. B, Blodow, N., Marton, Z., & Beetz, M. (2008). Aligning point cloud views using persistent feature histograms. *Proceedings of IEEE/RSJ International Conference on Intelligent Robots and Systems* (pp. 3384-3391).

[8] Ayadi, W., Elloumi, M., & Hao, J. K. (2012). Pattern-driven neighborhood search for biclustering of microarray data. *BMC Bioinformatics*, *13(Suppl 7)*, DOI: 10.1186/1471-2105-13-S7-S11.

[9] Saito, M., Takemura, N., & Shirai, T. (2012). Classification of ligand molecules in PDB with fast heuristic graph match algorithm COMPLIG. *Journal of Molecular Biology*, *424(5)*, 379-390.

[10] Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., Shindyalov, I. N., & Bourne, P. E. (2000). The protein data bank. *Nucleic Acids Research*, *28(1)*, 235-242.

[11] Preli, A., Bleuler, S., Zimmermann, P., Wille, A., Buhlmann, P., Gruissem, W., Hennig, L., Thiele, L., & Zitzler, E. (2006). A systematic comparison and evaluation of biclustering methods for gene expression data. *Bioinformatics*, *22(9)*, 1122-1129.

[12] Rodriguez-Baena, D. S., Perez-Pulido, A. J., & Aguilar-Ruiz, J. S. (2011). A biclustering algorithm for extracting bit-patterns from binary datasets. *Bioinformatics*, *27(19)*, 2738-2745.

Hiroto Nishimura received his bachelor and master's degrees from Kobe University in 2015 and 2017, respectively. His major research interests include data mining and bioinformatics.

Takenao Ohkawa received his B.E, M.E., and Ph.D. degrees from Osaka University in 1986, 1988, and 1992, respectively. He is currently a professor in the Department of Information Science, Graduate School of System Informatics, Kobe University. His research interests include intelligent data processing and smart agriculture. He is a member of the IEEE, the IPSJ, the IEICE, the IEEJ, and the JSAI.