# Binding Site Extraction by Detecting Optimal Graphs from Protein Molecular Surfaces

Takuma Mitsui and Takenao Ohkawa

***Abstract*—Proteins fulfill their functions by binding with molecular compounds called ligands. This research automatically extracts a binding site from the surface of a protein. A binding site candidate can be extracted as the local portion that satisfies the following two requirements. One is the structural similarity among proteins that bind the same kind of ligands. The other is the structural dissimilarity between a binding site in a protein and any local surfaces in the proteins that bind to any other ligands. By representing a protein molecular surface as a graph, the binding site extraction problem can be regarded as an optimal subgraph detection problem in which the best subgraph is extracted that satisfies the above requirements. However, if two ligands are different but have a partly similar structure, the binding sites of the proteins that bind these ligands often resemble each other. In such situations, an optimal graph may not present the binding site. Therefore, we introduce the concept of group integration, in which more than one group with similar ligands partners is regarded as a positive group. As a result of group integration, the number of proteins in a positive group and in a negative group is changed. Therefore, based on the distance from the virtual worst subgraph, an evaluation function is introduced to compare subgraphs with and without group integration. We clarified the effectiveness of binding site extraction with group integration through an experiment with 37 proteins.**

***Index Terms*—Binding site extraction, graph mining, group integration, protein surface.**

## I. Introduction

Proteins are the most essential and indispensable substance for living processes. Although some proteins fulfill their function alone, many show their faculty by binding with a small molecular compound called a ligand. A local structure that binds a ligand in a protein is called a binding site. The function of a protein and the binding site are related, in many cases.

This paper proposes a method for automatically extracting binding sites from the surface of proteins. Basically, a binding site candidate for extraction is defined as the local portion on the protein surface that satisfies the following two requirements. The first is the structural similarity among several proteins that bind the same kind of ligands, because binding sites, which are observed from different proteins but bind the same kind of ligands, often have similar structures. In other words, if the local portions from different proteins are structurally similar, they are probably binding sites. The second is the structural dissimilarity between a binding site in a protein and any local surfaces in the proteins that bind to any other ligands, because the structure of the binding site in a protein is scarcely observed in other proteins that bind to different ligands.

Recently, several methods for binding site prediction from protein surfaces have been proposed. Most of these methods use geometric or physical features of the binding sites themselves to identify them [1]-[3]. On the other hand, as mentioned above, we focus not only on the structural similarity among proteins in the same group but on the structural dissimilarity between the different groups [4].

In this research, we addressed the protein structure with protein molecular surfaces that are represented using triangular polygons. Since polygon data are a kind of a graph, the binding site extraction problem can be regarded as an optimal subgraph detection problem in which the best subgraph is extracted that satisfies the above requirements. In other words, a subgraph, which is frequently observed in one group (called a positive group) and is rarely observed in other groups (called negative groups), is extracted as a binding site, where a group is defined as a set of proteins that have the same kind of ligand partners.

An optimal subgraph can be effectively detected by applying the pruning method proposed by Morishita *et al.* [5]. However, when the optimal graph detection method is applied to the binding site extraction problem, we encounter the following problem. If two ligands are different but partly share a similar structure, the binding sites of the proteins that bind these ligands often resemble each other. In such a case, similar subgraph that actually correspond to the binding sites are observed in both positive and negative groups; such subgraphs are not detected as optimal graphs in many cases. Therefore, we introduce the concept of group integration, which is the idea that the label of the protein group with a similar ligand partner is changed from a negative to a positive group. After group integration, the positive group consists not only of a set of proteins with the same kind of ligand partner but also a set of proteins with similar ligand partners.

Since group integration is accompanied by conversion from a negative to a positive group, the number of proteins in the positive and negative groups is changed by group integration. Since the evaluation function proposed by Abe *et al.* [6] that compares and assesses the optimality of the extracted patterns assumes that the number of positive and negative instances are unchanged, it cannot directly compare an optimal graph obtained as a result of group integration with an optimal graph without group integration. Therefore, in this research, we introduce another evaluation function that focuses on the distance from the subgraph that scores the worst evaluation value. This evaluation function compares the graph with and without group integration because of relative evaluation using the distance from the virtual worst

Manuscript received July 12, 2013; revised September 25, 2013.

The authors are with the Graduate School of System Informatics, Kobe University, Japan (e-mail: mitsui@cs25.scitec.kobe-u.ac.jp, ohkawa@kobe-u.ac.jp).

DOI: 10.7763/IJBBB.2014.V4.305

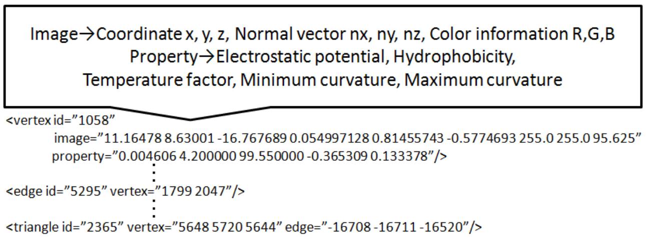
subgraph. In addition, the total processing time for enumerating subgraphs is shortened by simultaneously exploring both an optimal graph that considers group integration and an optimal subgraph without group integration.

The remaining sections are organized as follows. In Section II, we explain protein surface data. Section III presents our method of extracting binding sites. The experimental results of our method are shown in Section IV.

## II. Protein Data

### A. Protein Surface

We use the protein molecular surface data provided by the eF-site[1] database [7], where data about protein surfaces and physical properties are stored. The molecular surface is represented as a set of very small triangle polygons, and each vertex has numerical data on the curvature, the normal vector, the electrostatic potential, and the hydrophobicity of the amino acid that is located close to the vertex. An example of protein molecular surface data is shown in Fig. 1. The local part of the molecular surface data can be regarded as an undirected graph, which is constructed by linking adjacent vertices (Fig. 2).


Image→Coordinate x, y, z, Normal vector nx, ny, nz, Color information R,G,B
Property→Electrostatic potential, Hydrophobicity,
Temperature factor, Minimum curvature, Maximum curvature

```
<vertex id=”1058”
        image=”11.16478 8.63001 -16.767689 0.054997128 0.81455743 -0.5774693 255.0 255.0 95.625”
        property=”0.004606 4.200000 99.550000 -0.365309 0.133378”/>
                 ⋮
<edge id=”5295” vertex=”1799 2047”/>
                 ⋮
<triangle id=”2365” vertex=”5648 5720 5644” edge=”-16708 -16711 -16520”/>
```

Fig. 1. Protein surface data of eF-site.

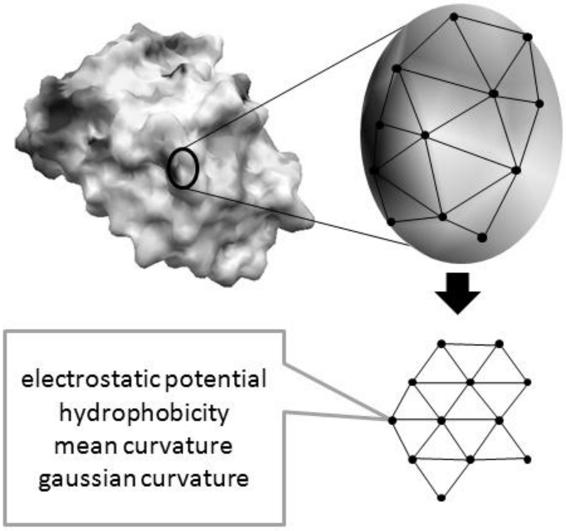


Fig. 2. Graph representing protein surface.

### B. Pocket

A binding site often forms a concave local portion in a protein molecular surface called a pocket. In our method, pockets are extracted from protein surface data in advance by the CASTp algorithm [8]. The CASTp server provides the information of atoms to compose a pocket. Then graphs are generated that only correspond to the pockets.

## III. Method

This section presents a method that predicts a binding site from the graph data of protein molecular surfaces. The extraction of binding sites can be formalized as the following optimal graph discovery problems. The input data are a set of graphs that represent each pocket of proteins, which consist of target protein $T$ for which the binding site is predicted and referential proteins $R=\{r_1,\ldots,r_n\}$. The target protein as well as each referential protein are grouped based on the kind of binding ligand, and each group is denoted by $G=\{g_1,\ldots,g_k\}$. These groups are classified into positive and negative groups by function $\xi : G \rightarrow$ {positive, negative} that labels (positive or negative) to each group based on its partner ligand. If $\xi(g_i)=$ positive, $g_i$ is a group of proteins that binds to the same ligand as a target protein. If $\xi(g_i)=$ negative, $g_i$ is a group of proteins that binds to a different ligand.

The output is a pocket containing the optimal subgraph of the target protein.

The following is an outline of our method:

1) For each pocket of target protein $T$, 2 and 3 are executed.
2) A subgraph of the pocket is explored, and a subgraph that satisfies a condition mentioned below are stored.
3) The stored subgraphs are evaluated by a function that considers graph size, and a subgraph given the best value is the score of the pocket.

All pockets of target protein $T$ are ranked by score, and the top ranked pocket is extracted as a binding site.

### A. Detecting Optimal Graphs

We must find a subgraph that is likely to be a binding site from the subgraphs of each pocket of target protein $T$. The following statements define intuitively desirable subgraph $I$ as a part of the binding site to be extracted from target protein $T$.

- Graphs similar to subgraph $I$ are observed in mnay proteins in positive groups.
- Graphs similar to subgraph $I$ are rarely observed from almost all the proteins in negative groups.

These intuitive statements are formalized as follows. Certain subgraph $I$ of a pocket in target protein $T$ divides a set of referential proteins $R$ into two subsets, $S_i=\{r\in R \mid I(r)=i\}$ and $(i=0,1)$, where $I(r)$ is a function that judges whether a subgraph resembles $I$ in protein $r$. If $I(r)=1$, a subgraph similar to $I$ is observed in protein $r$. On the other hand, if $I(r)=0$, no subgraph resembles $I$ in protein $r$. $N_i(I)$ denotes the number of proteins in subset $S_i$, and $M_i(I)$ denotes the number of proteins that belong to the positive groups in subset $S_i$, and $\theta_i(I)=M_i(I)/N_i(I)$ denotes the ratio of proteins belonging to the positive groups.

We introduce an evaluation function, where a subgraph satisfying the above requirements has better value, based on a

[1] http://ef-site.hgc.jp

method proposed by Abe *et al.* [6]:

$$\psi(\theta) = 2\theta(1-\theta) \tag{1}$$

$$G_\xi(I) = \psi(\theta_0)N_0 + \psi(\theta_1)N_1 . \tag{2}$$

Function $G_\xi(I)$ becomes smaller (better) when the proteins belonging to a positive group and a negative group can be divided well. That is, the ratio of the number of proteins belonging to a positive group is high (or low) in $S_1$, and the ratio of the number of proteins belonging to a positive group is low (or high) in $S_2$. However, since the subgraph that satisfies our goal, namely, the binding site extraction, can only be detected in the former situation, a subgraph, where the ratio of the number of proteins belonging to a negative group is high in $S_1$ is eliminated by applying the following condition:

$$M_1(I) \geq \frac{m}{n} N_1(I), \tag{3}$$

where $m$ denotes the number of proteins that belong to the positive groups and $n$ denotes the number of proteins that belong to the negative groups. We detect the subgraph that minimizes function $G_\xi(I)$ in (2) and fulfills (3).

To discover an optimal graph, a graph mining method called gApprox [9] is used for each pocket of the target protein. gApprox can search for all the subgraphs formed from the graph of the pocket of the target protein without omission and overlapping. However, if all the subgraphs are explored by gApprox, the calculation cost is extremely high. To cope with this problem, we applied an effective pruning method [1] when extending a subgraph by gApprox.

Whenever graph mining extends a graph, the number of similar graphs in the referential proteins decreases monotonically. $J$ denotes a graph that is generated by extending subgraph $I$ $(J \supseteq I)$. Then $M_1(I) \geq M_1(J)$ and $N_1(I) \geq N_1(J)$ hold. Effective pruning is realized using the monotonicity of the number of proteins with similar graphs:

$$\begin{aligned} &G_\xi(N_1(J), M_1(J)) \geq \\ &\quad \min\{G_\xi(M_1(I), M_1(I)),\ G_\xi(0, N_1(I) - M_1(I))\}, \end{aligned} \tag{4}$$

where function $G_\xi(I)$ in (2) is decided by variables $N_1(I)$ and $M_1(I)$, and therefore $G_\xi(N_1(I), M_1(I))$ is used instead of $G_\xi(I)$.

In comparison with a small subgraph, a large size graph should be evaluated highly as a meaningful structure. If we use the evaluation function that considered the graph size in graph mining, the soundness of the above pruning scheme is not guaranteed. Therefore, all subgraphs that are explored and fulfill (3) are stored and evaluated using the following evaluation function after completing the graph mining:

$$G_\xi^{size}(I) = (\psi(\theta_0)N_0 + \psi(\theta_1)N_1)\big/\sqrt{h}, \tag{5}$$

where $h$ represents the size of a graph defined as the number of vertices in it. The optimal graph is decided by minimizing function $G_\xi^{size}(I)$ in (5). The pocket including the optimal graph is extracted as the binding site of the target protein.

### *B. Group Integration*

If a ligand that binds a protein in a positive group resembles another ligand that binds a protein in a negative group, these proteins may share a structurally similar binding site. In this case, a graph similar to the subgraph of the binding site in the target protein is frequently observed in the proteins in the negative groups as well as in the positive groups and fails to extract a binding site as an optimal graph.

To solve this problem, we consider the concept of group integration, in which more than one group is assumed to be positive. Group integration introduces function $\eta : G \rightarrow \{\text{positive, negative}\}$. If $\eta(g_i) = \text{positive}$, $g_i$ is not only a group of proteins with the same kind of ligand partner as a target protein but also a group of proteins with a ligand partner that resembles a ligand partner of a target protein.

A subgraph is evaluated by two different functions: $\xi$ and $\eta$. We must compare the results of evaluation that assumes group integration and evaluation without group integration to determine which one is better. However, the values of evaluation functions $G_\xi(I)$ and $G_\eta(I)$ in (2), without and with group integration, cannot be directly compared, because the ratios of the number of proteins in the positive and negative groups is different in these two situations. Therefore, graph mining needs to be implemented with and without group integration. However, if graph mining is excecuted twice, its processing time becomes longer. Therefore, in the graph mining procedure, an explored subgraph is simultaneously evaluated both with and without group integration. In addition, if the pruning conditions in (4) both with and without group integration are satisfied, pruning is executed.

We introduce an evaluation function that can select an optimal one from a graph extracted from the with group integration and a graph identified without group integration, based on the distance from the virtual worst subgraph. If $M_1(I)/N_1(I)$ has the same value as $m/n$, the partition into $S_1$ and $S_0$ by subgraph $I$ is the worst partition. In other words, in the coordinates where the abscissa is $N_1(I)$ and the ordinate is $M_1(I)$, if subgraph $I$ is the worst partition, the point corresponding to subgraph $I$ is on line $M_1 = \frac{m}{n} N_1$. Subgraph $I$, which satisfies $M_1(I) = M_0(I)$ and $N_1(I) = N_0(I)$, has no ability to classify proteins into positive or negative groups. On the other hand, if point $(N_1(I), M_1(I))$ is far from line $M_1 = \frac{m}{n} N_1$, subgraph $I$ has high ability to classify proteins into positive or negative groups. Therefore, subgraph $I$ is evaluated based on the distance from the line to the point corresponding to $I$. This distance-based evaluation function is defined as follows:

$$D(N_1, M_1) = \frac{|\, mN_1 - nM_1 \,|}{\sqrt{n^2 + m^2}} \sqrt{h}\ , \tag{6}$$

where $h$ means the size of subgraph $I$ that is introduced for adequately evaluating the large subgraph.

## IV. Experiments and Discussions

We experimentally tested our method on extracting binding sites from the protein structural data in which the binding sites and the binding ligands are known. All experiments were conducted on a PC with a 3.40 GHz CPU and 16 GB main memory.

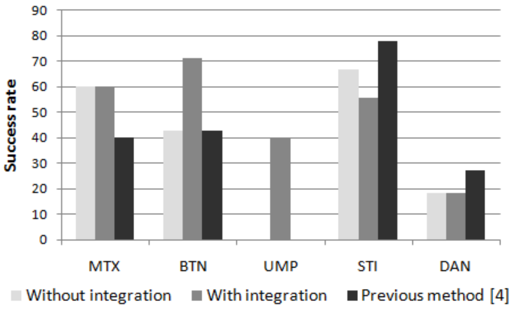


Fig. 3. Success rate.

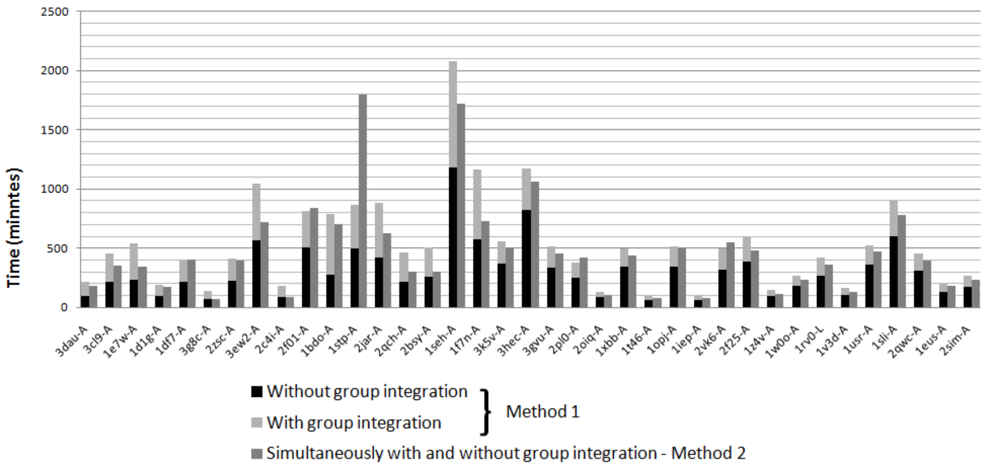


Fig. 4. Processing time.

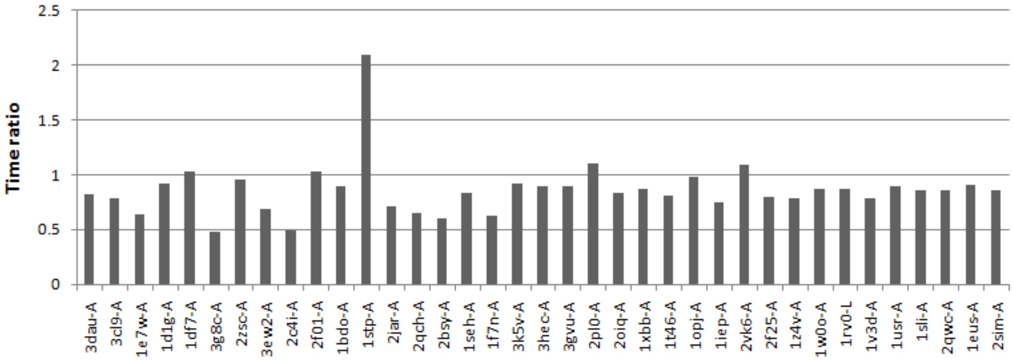


Fig. 5. Processing time ratio

The dataset for the experiments consists of 37 proteins, each of which has about 15 pockets on average. These proteins compose protein groups based on the binding partners (ligands), summarized in Table I. In group integration, the groups to be integrated are decided by the results of [4] and are shown in Table II.

TABLE I: Ligands and Protein Groups

| Ligand | Protein |
|---|---|
| MTX(METHOTREXAT) | 3dau,3cl9,1e7w,1d1g,1df7 |
| BTN(BIOTIN) | 3g8c,2zsc,3ew2,2c4i,2f01<br>1bdo 1stp |
| UMP[2] | 2jar,2qch,2bsy,1seh,1f7n |
| STI[3] | 3k5v,3hec,3gvu,2pl0,2oiq<br>1xbb,1t46,1opj,1iep |
| DAN[4] | 2vk6,2f25,1z4v,1w0o,1rv0<br>1v3d,1usr,1sli,2qwc,1eus,2sim |

TABLE II: Positive Groups in Group Integration

| Ligand of positive group without group integration | Ligands of positive group assuming group integration |
|---|---|
| MTX | MTX, BTN |
| BTN | BTN, MTX |
| UMP | UMP, BTN |
| STI | STI,DAN |
| DAN | DAN, MTX |

Each of the 37 proteins is considered a target protein, and the others are used as referential proteins. The binding sites of all the proteins are known, but the binding site extraction is conducted under the assumption that their binding sites are unknown. We ranked the optimal graphs extracted from each pocket based on the value of the evaluation function defined in (6). Fig. 3 shows the comparative results of the three methods, including our method without group integration, with group integration, and previous work [4]. The method that predicted binding sites by mining graphs representing protein surface was previously proposed by Kurumatani *et al.* [4]. Fig. 3 shows the success rate, which is defined as the

[2] 2'-DEOXYURIDINE 5'-MONOPHOSPHATE
[3] 4-(4-METHYL-PIPERAZIN-1-YLMETHYL)-N-[4-METHYL-3-(4-PYRIDIN-3-YL-PYRIMIDIN-2-YLAMINO)-PHENLY]-BENZAMIDE
[4] 2-DEOXY-2,3-DEHYDRO-N-ACETYL-NEURAMINIC ACID

percentage of proteins in which the correct binding sites were successfully extracted as the most optimal graph. The horizontal axis expresses each protein group with its ligand partner.

Fig. 3 shows that the success rate of group integration was the highest in three groups: MTX, BTN and UMP. However, for groups: STI and DAN, the success rate of group integration were lower than previous work.

In group UMP, the binding site of proteins 2qch and 1seh obtain considerably low evaluation values in the previous method, because subgraphs similar to these binding sites are also obtained in other proteins in a negative group. In such cases, group integration can successfully boost the success rate.

The binding sites have been successfully extracted in 17 proteins out of 37 proteins by the proposed method with group integration, whereas the number of proteins in which the binding sites have been accurately extracted is only 14 or 15 using other methods. These result show that optimal graph detection along with the mechanism of group integration is very effective for accurate binding site extraction.

The processing times for binding site extraction, in which the major part was spent in the graph search process, is evaluated in each of the following two methods:

- Method 1: Graph searches with and without group integration were separately (in series) executed. The total processing time was calculated by adding the processing time for each.
- Method 2 (proposed implementation): Graph searches with and without group integration were simultaneously (in parallel) executed.

The results are shown Fig. 4.

Method 2 is about one hour faster than Method 1 per protein on average. In other words, Method 2 reduces the processing time by 14% compared with Method 1 per protein on average. The number of proteins for which Method 2 is faster than Method 1 is 32, and the opposite is 5. Fig. 5 compares the relative processing speed of Methods 2 and 1. The processing time in Method 2 was successfully decreased or very slightly increased (almost the same) in almost all proteins. However, only in protein 1stp-A, the processing time of Method 2 was two times longer than Method 1. The main reason is that 1stp-A has more edges than the other proteins. In our proposed method, since the simultaneous processing of subgraph mining with and without group integration accelerates the relaxation of the pruning condition, the number of expanded subgraph patterns is greatly increased for large-scale graphs. In such cases, the total processing time would probably be reduced by introducing a mechanism in which either Method 1 and 2, were dynamically selected based on the number of edges in the graph.

## V. Conclusion

We proposed a method of extracting the binding sites of proteins using protein molecular surface data. Binding site were extracted by detecting an optimal graph. The success rate of predicting them is improved with group integration. Although group integration improved the prediction success rate, the calculation cost is high. For applying our proposed method to large-scale datasets, we must reduce its calculation cost.

A future challenge is abstracting a graph to reduce the calculation cost. In this idea, similar and neighboring vertices are unified. Then the number of vertices is decreased and graph mining is expected to be processed in a short time.

## Acknowledgment

This work was partly supported by Grant-in-Aid for Scientific Research (B) (24300056) from MEXT.

## References

[1] T. Dai, Q. Liu, J. Gao, Z. Cao, and R. Zhu, "A new protein-ligand binding sites prediction method based on the integration of protein sequence conservation information," *BMC Bioinformatics*, vol. 12, suppl. 14, pp. S9, 2011.

[2] S. Koizumi, K. Imada, T. Ozaki, and T. Ohkawa, "Extraction of binding sites in proteins by searching for similar local molecular surfaces," *Lecture Notes in Computer Science*, vol. 5265, pp. 87-97, 2008.

[3] M. Weisel, E. Proschak, and G. Schneider, "PocketPicker: analysis of ligand binding-sites with shape descriptors," *Chemistry Central Journal*, vol. 1, no. 7, pp. 1-17, 2007.

[4] N. Kurumatani, H. Monji, and T. Ohkawa, "Binding site extraction by similar subgraphs mining from protein molecular surface," in *Proc. 2012 IEEE 12th International Conference on,BioInformatics and BioEngineering (BIBE)*, 2012, pp. 255-259.

[5] S. Morishita and J. Sese, "Traversing itemset lattices with statistical metric pruning," in *Proc. the nineteenth ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems*, 2000, pp. 226-236.

[6] K. Abe, S. Kawasoe, T. Asai, H. Arimura, and S.Arikawa, "Optimized substructure discovery for semi-structured data," *Principles of Data Mining and Knowledge Discovering*, pp. 57-100, 2002.

[7] K. Kinoshita and H. Nakamura, "eF-site and PDBjViewer: database and viewer for protein functional sites," *Bioinformatics*, vol. 20, no. 8, pp. 1329-1330, 2004.

[8] T. A. Binkowski, S. Naghibzadeg, and J. Liang, "CASTp: Computed atlas of surface topography of proteins," *Nucleic Acid Research 31*, no. 13, pp. 3352-3355, 2003.

[9] C. Chen, X. Yan, F. Zhu, and J. Han, "gApprox: Mining frequent approximate patterns from a massive network," in *Proc. Seventh IEEE International Conference on Data Mining*, 2007, pp. 445-450.

**Takuma Mitsui** gained his bachelor of engineering from Kobe University in 2013. He is presently a student of the master course at Graduate School of System Informatics, Kobe University (Kobe, Japan). His research interests include graph 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 bioinformatics. He is a member of the IEEE, the Information Processing Society of Japan, the Japanese Society for Bioinformatics, the Institute of Electronics, Information, and Communication Engineers, the Institute of Electrical Engineers in Japan, and the Japanese Society for Artificial Intelligence.