Proceedings of the 2012 IEEE 12th International Conference on Bioinformatics
& Bioengineering (BIBE), Larnaca, Cyprus, 11-13 November 2012

Binding site extraction by similar subgraphs mining from protein molecular surfaces

Natsumi Kurumatani, Hiroyuki Monji, and Takenao Ohkawa
Graduate School of System Informatics,
Kobe University,
1-1, Rokkodai, Nada, Kobe 657-8501, Japan
Email: kurumatani@cs25.scitec.kobe-u.ac.jp, ohkawa@kobe-u.ac.jp

***Abstract*—Most proteins express their functions by binding with other proteins or molecular compounds called ligands. The local portion involved in binding is called a binding site. The characteristics of the binding site often determine the function of the protein, so clarifying the location of the binding site of the protein helps analyze the function of proteins. Binding sites that bind to similar ligands often have common surface structures. Such common structures are called surface motifs. Therefore, extracting the surface motifs among several proteins with similar functions improves binding site prediction.**

We propose a method of predicting binding sites by extracting the surface motifs that are frequently observed in only a specific group, which means a set of proteins that bind to the same ligand. Since most binding sites have concave structures called pockets, the pockets are compared and common structures are searched for to extract the surface motifs by applying similar graph mining to the pocket data, which are represented as graphs, to find the frequent subgraphs among the pockets of several proteins. In addition, the common binding sites across several groups can be predicted in such a way to integrate more than one group.

Applying our proposed method to a set of 37 proteins of five groups, we achieved success rates of binding site prediction over 40% and 50% for more than half of the groups without group integration and using integration, respectively.

***Index Terms*—protein surface comparison, binding site extraction, graph mining.**

I. INTRODUCTION

The functional analysis of proteins is important for revealing the mechanisms of living bodies. Most proteins express their functions by binding other proteins or molecular compounds (ligands). The local portion involved in binding is called a binding site that has specific features of 3D structure and properties [1]–[4]. The features on binding sites often determine the function of the protein. Clarifying what part is the binding site in a protein and what are its features has improved the analysis of the function of proteins.

The binding sites of proteins that bind to the same ligand often have specifically common surface structures. Such surface structures, which are called surface motifs [5], [6], are considered candidates for binding sites. We propose a method of predicting binding sites by extracting the surface motifs that are frequently observed in a group, which means a set of proteins that bind to the same ligand.

Most binding sites have concave structures, which we call pockets. In our method, the pockets are compared and common structures are extracted by applying similar graph mining to the pocket data represented as graphs and finding the frequent subgraphs among the pockets of several proteins. Even if similar structures are frequently observed in the proteins in a particular group and also frequently in the proteins in other groups, we cannot regard them as significant surface motifs. We introduce a score function for evaluating the specificity of the extracted subgraphs to a particular protein group in which both the inter-group frequency and the intra-group frequency of similar subgraphs are considered.

Most existing methods compare the structures of proteins by focusing on the residues or the atoms of the protein's structure [7] to find motifs by searching for commonly frequent local structures [5], [6]. Recently, a number of methods for predicting binding sites from protein surfaces have been proposed. In these methods, geometric and/or physical features of binding sites themselves are often employed to identify the binding sites [8]–[11]. In our method, however, we try to find the local surfaces that not only characterize a group of proteins binding to the similar ligands but also distinguish this group from other groups.

Our method basically assumes that the similar structures regarded as binding sites in a particular group are rarely observed in other groups. However, the structures of the binding site are not always common only in a single group but may be similar in other groups. From this point of view, several groups with similar binding sites must be integrated to improve the accuracy of the binding site prediction. In our method, several groups, which are selected from the combination of all groups, are integrated and regarded as one group, if the variance of the values of the score function increases.

The objectives of our work are

- introducing a framework for representing shape and properties of protein surfaces by using graphs,
- formulating a graph mining algorithm for discovering subgraphs that distinguish a part of graphs from a number of graphs, and hereby,
- developing a method for predicting binding sites with high accuracy.

Additionally, we try to improve its prediction accuracy by introducing group integration.

978-1-4673-4358-9/12/$31.00 ©2012 IEEE

II. Method

A. Molecular surfaces and their graph representation

We use the protein molecular surface data provided by the eF-site[1] database. In eF-site, protein surface data are constructed from a number of triangular polygons. Each vertex of the polygons corresponds to a very small local portion of the molecular surface that has such structural and physical attributes as position (3D coordinates), normal vectors, electrostatic potential, and hydrophobicity. Neighboring vertices are connected by edges. Therefore, the molecular surface data in an eF-site can be regarded as an undirected graph with attributed vertices, suggesting that we can employ graph mining algorithms to find frequently observed local surfaces.

B. Outline of binding site extraction

Since the structure of the binding site tends to be conserved on the surfaces of the proteins, it may be similar among proteins that have the same ligand partner. The surface motifs refer to the local portion of the protein surface that is frequently observed among the proteins with the same ligand partner but is rarely observed among proteins with different ligand partners. These motifs are candidates for the binding sites.

Most binding sites appear on the protein's surface, especially in a pocket. To reduce the computation cost, in our method, we extract them not from the entire surface but only from the pockets.

In binding site prediction, the inputs are the surface data of a target protein, from which the binding site should be extracted, the surface data of the proteins that bind to the same ligand as the target protein binds, and the surface data of the proteins that bind to different ligands. These proteins have been classified into several groups based on the types of ligands. The proteins except the target protein, are called referential proteins. The outputs are the pockets that have surface motifs.

Binding site extraction consists of the following three steps,

1) Extract the pockets from the protein surface data.
2) Search for surface motifs by similar subgraph mining.
3) Score the pockets.

Figure 1 shows an outline of these steps.

The first step is the extraction of the pockets of the target and referential proteins. The CASTp algorithm [12] is utilized to create the pocket data. The CASTp server can provide the information of atoms that correspond to the pockets that are extracted from the protein structure data. Moreover, graph representation corresponding to each pocket can be generated by enumerating the vertices of the polygons located near the atoms provided by the CASTp server. The second step is searching for the surface motifs in the pockets of the target protein by referring to the pockets of the referential proteins. The motif search is achieved by comparing the pockets of the target protein and the referential proteins. To compare the pockets, they are represented as graphs and similar subgraph mining is applied. Finally, each pocket of the target protein is scored based on the frequency of the subgraphs that appear in positively referential proteins (the proteins binding to the same ligand as the target protein) or negatively referential proteins (other proteins).

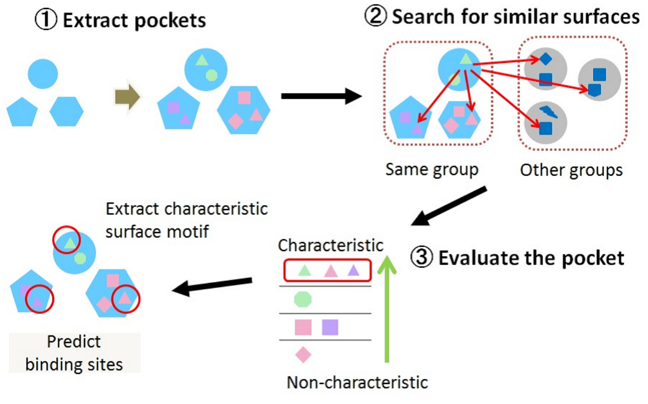

Fig. 1. General flow of binding site extraction

C. Extraction of surface motifs

To discover the common structures in the pocket of the target protein to the pockets of several referential proteins, the pockets of the target and referential proteins are compared and their similarity is evaluated. However it is difficult for all the vertices included in the pockets to have a one-on-one correspondence by global structural alignment. Therefore, in our method, we compare pockets by finding common subgraphs extracted from the pockets of several proteins.

To compare two graphs of pockets, the attribute values (structural data and physical properties) attached to the vertices are compared and a pair of vertices whose attribute values are similar is regarded as similar vertices. The vertices connected to the similar vertices are also compared to determine whether they are similar. Thus, similar subgraphs are extracted by extending similar regions, i.e., the set of similar vertices connected by edges.

We utilize gApprox, which is one popular graph mining algorithm, to extract the pattern of similar regions [13]. The pocket graphs of the target and referential proteins are given and the subgraph of the pocket of the target protein that resembles the subgraph of the referential proteins are searched for by extending the pattern.

The following steps show the detailed algorithm for extracting similar subgraphs using gApprox:

1) Let $P_T = \{t_1, t_2, \ldots, t_N\}$ be a set of the pockets of target protein T, let $P_R^i = \{r_1^i, r_2^i, \ldots, r_N^i\}$ be a set of the pockets of referential protein R_i extracted from the set of referential proteins $R = \{R_1, R_2, \ldots, R_M\}$, let G^t be a graph representing pocket $t \in P_T$, and let $V^t = \{v_1^t, v_2^t, \ldots, v_n^t\}$ be a set of vertices constituting G^t. Each vertex v_j^t is assigned an ID number $I(v_j^t)$.

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

2) A vertex with a minimum number is denoted by

$$v^t_{min} = \{v^t_j | \min_j I(v^t_j), v^t_j \in V^t\}. \quad (1)$$

A subgraph of pocket t consisting of only one vertex v^t_{min} is denoted by g_t. Let V^r be a set of vertices constituting graph G^r of pocket $r \in P^k_R$ of referential protein R_k. $VR(v^t_{min})$, which is a set of vertices $vr_j(v^t_{min})$ in the pockets of the referential proteins similar to v^t_{min} denoted as follows, is searched for:

$$\begin{aligned} VR(v^t_{min}, r) = \{&vr_1(v^t_{min}), vr_2(v^t_{min}), \\ &\ldots, vr_l(v^t_{min}) | vr_j(v^t_{min}) \in V^r, \\ &j = 1, \ldots, l\}. \end{aligned} \quad (2)$$

The similarities between two vertices are calculated by comparing their attribute values. Let e_v, h_v, K_v, and H_v be the electrostatic potential, hydrophobicity, Gaussian curvature, and mean curvature of vertex v, respectively, and $\sigma_e, \sigma_h, \sigma_K$, and σ_H be the thresholds of the differences between each attribute value. $vr(v^t_{min})$ is calculated by the following equation:

$$\begin{aligned} vr(v^t_{min}) \;=\; \{&v_k \in V^r || e_{v^t_{min}} - e_{v_k}| < \sigma_e, \\ &|h_{v^t_{min}} - h_{v_k}| < \sigma_h, \\ &|K_{v^t_{min}} - K_{v_k}| < \sigma_K, \\ &|H_{v^t_{min}} - H_{v_k}| < \sigma_H\}. \end{aligned} \quad (3)$$

All vertices satisfying this equation are enumerated. Let $rg^z_w(g_t)$ be a similar graph of g_t that exists in pocket r^z_w of referential protein R_z and $RG_z(g_t) = \{rg^z_1(g_t), rg^z_2(g_t), \ldots, rg^z_N(g_t)\}$ is a set of the graphs of the pockets in referential protein R_z that are similar to g_t. Let $RG(g_t) = \{RG_1(g_t), RG_2(g_t), \ldots, RG_M(g_t)\}$ be a set of RG_z for all referential proteins, $vr_k(v^t_{min})$ are searched for all pockets P_R of all referential proteins R, and then $RG(g_t)$ is found.

g_t is discarded if $RG(g_t) = \emptyset$, otherwise it is regarded as a similar subgraph corresponding to the surface motif. Let $Gp = \{Gp_1, Gp_2, \ldots, Gp_L\}$ be a set of the groups of referential proteins and, the number of referential proteins $N_k(g_t)$ such that $RG_w(g_t) \neq \emptyset$ for the referential protein R_w that belongs to $Gp_k \in Gp$ is counted with a similar subgraph as the frequency of similar subgraphs.

3) The similarities between v_x that connects to g_t and v_y that connects to $RG_z(g_t)$ are examined. If v_x and v_y are similar, v_x is added to g_t, and v_y is added to $RG_z(g_t)$. This step is done for every $RG_z(g_t) \in RG(g_t)$.
4) g_t is extended until similar graph $RG_z(g_t)$ is not found and g_t is enumerated every time it is extended.
5) Steps 2-4 are iterated for all vertices $V^{t\prime} = \{\{v^t_1, v^t_2, \ldots, v^t_m\} | v^t_{min} \notin V^{t\prime}\}$ that constitute $G^{t\prime}$, which is obtained by removing v^t_{min} from G^t.
6) Output all similar subgraph patterns and their frequencies for every group $N_k(g)$ for each similar graph pattern g by repeating step 5 until V^t has no vertex.

All similar subgraph patterns in the pockets of the target protein can be extracted as surface motifs in the target protein by these steps.

D. Scoring the pockets

All the surface motifs, which are the subgraphs of the pockets, are extracted using the above algorithm, but not all the extracted surface motifs are important.

Even if the extracted surface motifs can be observed in referential proteins that bind the same ligand as the target pockets, some may not be binding sites if they are also observed in the other referential proteins that bind different ligands.

We introduce a method of evaluating each pocket p_h $(h = 1, 2, \ldots, M)$ that is extracted from target protein T by subgraph mining. Let $G = \{g_1, g_2, \ldots, g_N\}$ be a set of groups, let $S = \{S_1, S_2, \ldots, S_L\}$ be a set of extracted surface motifs, and let n_k be the size of each surface motif S_k that is defined as the number of the vertices of S_k. The following observation must be considered to identify the surface motifs as binding sites.

- Binding sites tend to have surface motifs that are commonly observed in group g_i which T is classified into but are rarely observed in other groups $g_j (j \neq i, j = 1 \ldots N)$.
- The large size of surface motifs is more appropriate for a binding site than one that happens frequently in a very small region.

The scoring function is defined as:

$$\psi(S_k, g_i) = \frac{F_{g_i}(S_k)}{F_{g_j \neq g_i}(S_k) + b} \times n_k. \quad (4)$$

$F_{g_i}(S_k)$ means the frequency of the proteins that have surface motif S_k out of the proteins in group g_i where T belongs, $F_{g_j \neq g_i}(S_k)$ means the frequency of the proteins that have S_k out of the proteins in group g_j other than g_i, and b is a constant value for numerical stability.

This equation can give high scores to the large surface motifs that are frequently in g_i and infrequently in g_j. The following equation calculates $\gamma(p_h, T)$; the score of pocket $p_h (h = 1, 2, \ldots, M)$ of T is based on the surface motif with the highest score:

$$\gamma(p_h, T) = \max_{S_k} \psi(S_k, g_i). \quad (5)$$

Pockets p_h are enumerated in descending order of score $\gamma(p_h, T)$ as binding sites.

E. Group integration

Our method predicts binding sites by searching for the surface motifs that are frequently in the same group and infrequently in other groups, assuming no other binding site structures are common in several groups. Generally, however, different ligands are sometimes partially similar and then their binding sites are often similar. Although these structures are binding site structures, they may not be extracted because they are not regarded as a specific structure for the particular group by the scoring function.

Therefore, common binding site structures in several groups are extracted by integrating several groups into a single group.

1) Binding site prediction using group integration: In predicting binding sites using group integration, the evaluation method of the surface motifs is extended. Let $IG = \{ig_1, ig_2, \ldots, ig_N\}$ be a set of groups that are targets of the integration to group g_i, where IG is a subset of the set of all groups $G = \{g_1, g_2, \ldots, g_N\}$. Then the surface motifs are scored:

$$\psi(S_k, g_i, IG) = \frac{F_{g_i}(S_k) + F_{IG}(S_k)}{F_{g_j \neq g_i \notin IG}(S_k) + b} \times n_k. \tag{6}$$

$F_{IG}(S_k)$ indicates the frequency of the proteins that have surface motif S_k out of the proteins in group IG.

The score of pocket p_h of target protein T is given by the following equation:

$$\gamma(p_h, T, F_{g_i}) = \max_{S_k} \psi(S_k, F_{g_i}, F_{IG}). \tag{7}$$

2) Criterion of group integration: Some patterns of the combination of groups to be integrated may improve predictions and other patterns may degrade predictions. Therefore, we introduce a criterion that appropriately selects the combination of groups for integration.

The variance of the scores of all pockets of the target protein is used as the integration criterion. In integrating several groups, if the variance is high, a large difference exists between high and low score pockets, suggesting the existence of surface motifs that are specific to the integrated groups. So the variance value may be a clue to determine the appropriate combination of groups for integration.

We define $V(g)$, which is the variance of the scores of all pockets from a target protein that belongs to group g, as follows:

$$V(g) = \frac{1}{N}\sum_{k=1}^{N}(\gamma_k - \gamma_\mu)^2, \tag{8}$$

where N means the number of pockets of the target protein in g, γ_k means the score of the k-th pocket, and γ_μ means the average scores of all pockets.

If $V(g_i \cup IG) > V(g_i)$, group IG is integrated into group g_i.

III. Results and Discussion

A. Evaluation experiment

To verify the effectiveness of our proposed method, we experimentally extracted binding sites for the protein structural data where their binding sites or binding ligands are known. All experiments were conducted on a PC with a 3.40 GHz CPU and 16 GB main memory.

B. Dataset

The dataset for the experiments was constructed by referring to 48 protein-ligand complexes, which were used in the reference [14] as benchmark data. 60 types of ligands from 48 complexes are listed in [14], but we only selected five types of relatively large ligands. Proteins that bind to these five ligands were retrieved from PDB and filtered out based on $< 30\%$ sequence homology. Each ligand and the selected proteins binding to it compose a protein group summarized in Table I.

TABLE I
Ligands and protein groups

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

[1] 2'-DEOXYURIDINE 5'-MONOPHOSPHATE

[2] 4-(4-METHYL-PIPERAZIN-1-YLMETHYL)-N-[4-METHYL-3-(4-PYRIDIN-3-YL-PYRIMIDIN-2-YLAMINO)-PHENYL]-BENZAMIDE

[3] 2-DEOXY-2,3-DEHYDRO-N-ACETYL-NEURAMINIC ACID

C. Binding site prediction without group integration

Binding site prediction was conducted for binding site known proteins under the assumption that their binding sites are unknown. The scores of the pockets of the target protein were calculated and the prediction success was evaluated to determine whether the top ranked pocket or the top three pockets are true binding sites. Fig. 2 shows the accuracy of the binding site prediction. The horizontal axis indicates the name of five ligands, each of which corresponds to a protein group. For each group, one protein was selected and regarded as the target protein. The rest of the proteins in the same group

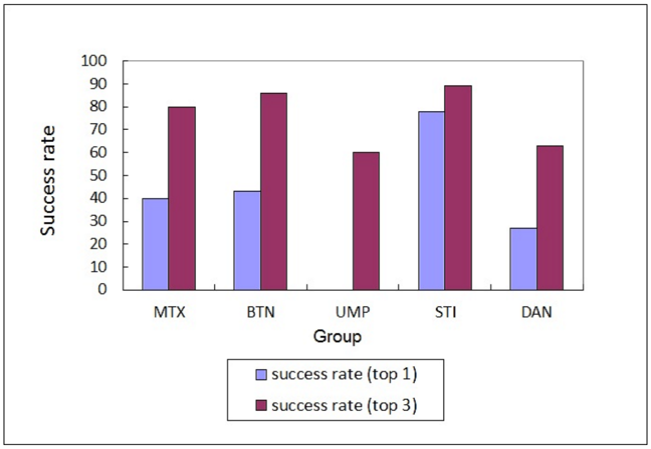

Fig. 2. Success rate of prediction

and the proteins in other groups are regarded as referential proteins. The success rate means the rate of successful proteins for binding site prediction out of the proteins in the group.

Fig. 2 shows that the binding sites were predicted with the highest score for about 40% of proteins in groups MTX and BTN. For group STI, almost all binding sites were successfully extracted. However, the prediction completely failed for group UMP. On the other hand, the success rates in the top three predictions exceeded 60% for all the groups and 80% for more than half.

D. Binding site prediction using group integration

To predict binding sites, the group to which the target protein belongs is integrated with other groups, and the target protein is assumed to belong to that group.

The results of binding site prediction with group integration were performed based on the criterion mentioned in Section II.E. A comparison of the results of the binding site prediction with and without group integration is shown in Fig. 3.

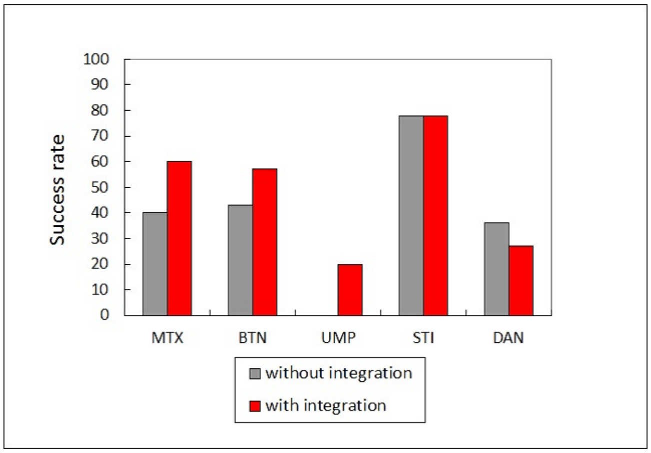

Fig. 3. Comparison of success rates of prediction with and without group integration

The gray bar on the vertical axis indicates the success rate of the binding site prediction without group integration, and the red bar indicates the success rate of the binding site prediction using group integration. For the top prediction, the success rate for DAN slightly worsened by integration, but the success rate for the other groups improved or remained high. This may indicate that the automatic integration of groups improves prediction accuracy.

IV. Conclusion

We proposed a method of extracting the binding sites of proteins using 3D structural and classification information. To extract binding sites, the inputs are pockets that are compared to find the surface motifs by representing them as graphs and applying similar subgraph mining.

The score function, which can extract the surface motifs that are frequently observed among several proteins in the same group and are rarely observed among the proteins in the other groups as binding sites, achieves favorable prediction results. In addition, we confirmed that the framework of the group integration improves prediction accuracy.

While our method can remove universally observed similar local surfaces from the candidates of binding site, true binding sites may also be discarded if the size of data sets is small, which is one of drawbacks of our method.

Therefore, a future challenge is to enlarge our dataset. This paper only shows the prediction results for five protein groups. Increasing the number of groups also increases the number of combinations of group integration, which may improve prediction. In addition, we will compare the proposed method with other methods through experiments with large scale data sets to confirm the effectiveness of our method in near future.

Acknowledgment

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

References

[1] S. Goto, T. Nishioka, and M. Kanehisa, "Ligand: Chemical database for enzyme reactions," *Bioinformatics*, vol. 14, no. 7, pp. 591–599, 1998.

[2] C. Chothia and J. Janin, "Principles of protein-protein recognition," *Nature*, vol. 256, pp. 705–708, Aug. 1975.

[3] S. Jones and J. M. Thornton, "Principles of protein-protein interactions," *Proceedings of the National Academy of Sciences*, vol. 93, pp. 13–20, Jan. 1996.

[4] I. M. Nooren and J. M. Thornton, "Diversity of protein-protein interactions," *The EMBO Journal*, vol. 22, no. 14, pp. 3486–3492, 2003.

[5] N. L. Shrestha, Y. Kawaguchi, T. Nakagawa, and T. Ohkawa, "A method of filtering protein surface motifs based on similarity among local surfaces," *Intelligent Data Engineering and Automated Learning*, vol. 3177, pp. 39–45, Aug. 2004.

[6] Y. Shimizu, N. L. Shrestha, and T. Ohkawa, "Parallel processing method of protein surface motifs extraction," in *Proc. of International Conference on Advances in Computer Science and Technology (ACST 2004)*, St. Thomas, US Virgin islands, Nov. 2004, pp. 265–270.

[7] S. Jones and J. M. Thornton, "Prediction of protein-protein interaction sites using patch analysis," *Journal of Molecular Biology*, vol. 272, pp. 133–143, Sep. 1997.

[8] 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, Oct. 2008.

[9] T. Dai *et al.*, "A new protein-ligand binding sites prediction method based on the integration of protein sequence conservation information," *BMC Bioinformatics*, vol. 12, pp. 87–97, Dec. 2011.

[10] J. G. P. Brady and P. F. W. Stouten, "Fast prediction and visualization of protein binding pockets with pass," *Journal of Computer-Aided Molecular Design*, vol. 14, no. 4, pp. 383–401, 2000.

[11] M. Weisel, E. Proschak, and G. Schneider, "Pocketpicker: analysis of ligand binding-sites with shape descriptors," *Chemistry Central Journal*, vol. 1, Mar. 2007. [Online]. Available: http://journal.chemistrycentral.com/content/pdf/1752-153X-1-7.pdf

[12] T. A. Binkowski, S. Naghibzadeg, and J. Liang, "Castp: Computed atlas of surface topography of proteins," *Nucleic Acid Resarch*, vol. 31, no. 13, pp. 3352–3355, 2003.

[13] C. Chen, X. Yan, F. Zhu, and J. Han, "gapprox: Mining frequent approximate patterns from a massive network," in *Proc. of International Conference on Data Mining (ICDM 2007)*, Ohama NE, USA, Oct. 2007, pp. 445–450.

[14] B. Huang and M. Schroeder, "$LIGSITE^{CSC}$:predicting ligand binding sites using the connolly surface and degree of conservation," *BMC Structural Biology*, vol. 6, Sep. 2006. [Online]. Available: http://www.biomedcentral.com/content/pdf/1472-6807-6-19.pdf