Protein-ligand Binding Region Prediction based on Geometric Features and CUDA Acceleration

Ying-Tsang Lo[1], Hsin-Wei Wang[1], Tun-Wen Pai[1,3*], Wen-Shoung Tzou[2,3], Hui-Huang Hsu[4], Hao-Teng Chang[5,6]

[1]Department of Computer Science and Engineering, National Taiwan Ocean University, Keelung, Taiwan, R.O.C.

[2]Department of Life Sciences, National Taiwan Ocean University, Keelung, Taiwan, R.O.C.

[3]Center of Excellence for Marine Bioenvironment and Biotechnology, National Taiwan Ocean University, Keelung, Taiwan, R.O.C.

[4]Department of Computer Science and Information Engineering, Tamkang University, Taipei, Taiwan, R.O.C.

[5]Graduate Institute of Molecular Systems Biomedicine, China Medical University, Taichung, Taiwan, R.O.C.

[6]China Medical University Hospital, Taichung, Taiwan, R.O.C..

YTL: yt.lo@mail.ntou.edu.tw

HWW: wanghsinwei@msn.com

TWP: twp@mail.ntou.edu.tw

WST: wstzou@ntou.edu.tw

HHH: h_hsu@mail.tku.edu.tw

HTC: htchang@mail.cmu.edu.tw

*Corresponding author: Dr. Tun-Wen Pai, Department of Computer Science and

Engineering & Center of Excellence for Marine Bioenvironment and Biotechnology, National Taiwan Ocean University, No. 2, Peining Road, Keelung, 20224, Taiwan, R.O.C.. TEL: +886-2-24622192 ext. 6618, FAX: +886-2-24623249, E-mail: twp@mail.ntou.edu.tw

Abstract

Background

Protein-ligand interactions are key processes of triggering and controlling biological functions within cells. Prediction of protein binding regions on surface assists in understanding the mechanisms and principles of molecular recognition. *In silico* geometrical shape analysis plays a primary step to analyze spatial characteristics of protein binding regions and facilitates applications of bioinformatics in drug discovery and designing.

Methods

An efficient way based on CUDA parallel technologies was designed to extract solid angle feature of each surface atom. Among all surface residues, representative anchors were assigned according to ranking of solid angles. In addition, cavity depths and volumes were obtained through scanning multiple directional vectors within each selected cavity. Both depth and volume features were combined with various weighting coefficients for ranking predicted potential binding regions.

Results

Two testing datasets from LigASite, each containing 388 of bound and unbound structures were applied to predict binding regions, and the results were compared to two well-known prediction systems, SITEHOUND and MetaPocket2.0. It has shown that our proposed system outperformed the other systems with accuracy rates of 94.3% for unbound proteins and 95.5% for bound proteins through a ten-fold cross validation mechanism. Additionally, the CUDA parallel computing architecture was designed to enhance the computational efficiencies, and an average of 11-fold faster was obtained for computing geometric features on the testing datasets.

Conclusions

In silico binding region prediction has been considered as one of the initial procedures for structure-based drug design. To improve the efficacy of biological experiments for drug development, a system employing geometrical features only can achieve a surprisingly good overall performance on protein-ligand binding region prediction. Based on the same approach and rationale, these features also can be applied to predict carbohydrate-antibody interaction for further design and development of carbohydrate-based vaccines. The prediction website is freely available at http://SAVE.cs.ntou.edu.tw/ .

Keywords: solid angle, cavity depth, cavity volume, CUDA, geometrical feature, ligand binding

1. INTRODUCTION

The study of protein binding site prediction assists in understanding the mechanisms and principles of molecular recognition, and provides determining information for protein function annotation, construction of protein-protein interaction networks, drug design and discovery, and vaccine design and development [1, 2]. In recent years, various *in silico* methods for prediction of protein-protein and protein-ligand binding sites have been extensively developed [3]. However, the number of protein structures and protein complex structures grows exponentially in last decade, and it causes that a fast and effective algorithm to identify binding regions on a protein is still urgently required. Especially, an important application of carbohydrate vaccine development has gained much attention in last few years. A bioinformatics predictor could assist to predict binding pockets between a glycan and antibody since the carbohydrate-based vaccine is one of new strategies against pathogen infection and cancers [2]. The binding affinity of carbohydrate-based antibody is normally weaker than that of protein-based antibody. Therefore, a prediction tool for revealing characteristics of carbohydrate binding sites

could provide sufficient information for the development of carbohydrate-based vaccines. In the past, different approaches based on geometric characteristics, physicochemical properties, or their combinations were frequently employed to predict the protein interaction regions. For example, in terms of geometric characteristics, a sparse global surface description of proteins obtained from Connolly surface and geometric characteristics of proteins in a two-dimensional projection space were discussed [4]. Among them, several physical shape characteristics were frequently employed to analyze and identify surface interfaces, such as accessible surface areas [5, 6], sequence conservation [7, 8], and amino acid compositions [9]. In addition, a number of different approaches adopted the Fourier-based concepts, transforming a three-dimensional grid onto a set of orthogonal basis functions and calculating overlapped areas by employing Fast Fourier Transform (FFT) techniques [10-12]. On the other side, the contents of interface residues and their corresponding physicochemical properties were also significantly considered and statistically analyzed for predicting binding sites. For example, the aliphatic and aromatic residues were usually enriched in the interface regions compared to the charged residues and several specific composition of amino acid residues indeed appeared with higher frequencies at binding interfaces than non-binding surface regions [13-15]. Although most of previous approaches predicting protein binding regions

adopted similar ideas for analyzing protein-protein interfaces and protein-ligand binding regions, different characteristics existed between these two major types of binding mechanisms, such as different binding architectures and different sizes of binding regions [16]. In this study, we aimed to design an improved prediction system for the type of protein-ligand binding mechanisms. The query proteins were assumed as rigid components for a straightforward approach by considering the geometric characteristics including solid angle, cavity depth and volume. Similar with most existing algorithms, this study also applied the concept of shape complementary as the primary filter to rank all potential binding regions. In addition, this paper also focused on grid-based structure construction techniques for surface residue identification and parallel processing mechanisms for efficient computations on geometric features. Accordingly, identified cavities and pockets with irregular shapes on protein surfaces can be efficiently determined and ranked as protein-ligand binding regions.

In this study, the solid angle and associated features were considered as the main geometric attributes for protein-ligand binding region analysis. Connolly proposed the first solid angle approach according to protein surface binding characteristics that if two points could be fitted into each other, then the sum of two compactly matched solid angles would be 4π in a three-dimensional space [17]. There are mainly two major methods for

computing solid angles. The first approach employs the Gauss-Bonnet formula to find solid angles on surface points while the second adopts a virtual sphere concept on the protein surface by calculating the steradian formed by the protein surface and the virtual sphere, and dividing by the square of the radius of the virtual sphere. Both methods provide the solid angle of a specified surface point. Subsequently, several papers utilized the superior characteristics of solid angles and significant results were published in the fields of protein docking [18, 19] and structure alignment [20]. Nevertheless, due to huge amount of surface atoms on protein surfaces and which required tremendous computational power and time for solid angle calculations, this paper utilized the NVIDIA's Compute Unified Device Architecture (CUDA) technology to enhance execution speed of the proposed algorithms. CUDA is a parallel computing architecture that utilizes graphics processing units (GPUs) for general purpose computing. GPU was originally employed to speed up graphics display and it could quickly and easily generate a lot of threads. Moreover, the floating point operation and memory bandwidth performance are much faster than CPUs [21]. The multi-core architecture allows each thread to perform an identical computing task simultaneously. Since the introduction of CUDA in 2007, harnessing the power of the GPU becomes easier. Recently, numerous GPU based algorithms in bioinformatics have been proposed, including sequence alignment [21-24], protein docking [25], surface area

[26, 27], molecular dynamic simulation [28] and systems biology [29]. In this paper, we also adopted the CUDA architecture to reduce the required computational time and developed an effective prediction system to identify binding regions through evaluating geometric features of solid angles, depths and volumes of cavity on protein surfaces. From the experimental validations and compared performance to other existing systems, the proposed system can be considered as a good selection for detecting protein-ligand binding regions, and further applications on drug and vaccine development.

2. MATERIALS AND METHODS

The proposed prediction system was composed of five major steps (see **Error! Reference source not found.**), and a PDB identifier or a PDB file could be imported into the system for analysis. Multiple chains or specific range of a protein structure could be assigned and evaluated simultaneously according to users' requests. The CUDA mechanism was developed to parallelly compute the spatial features of huge amount of atoms on the query protein surfaces. Each step is briefly introduced as follows.

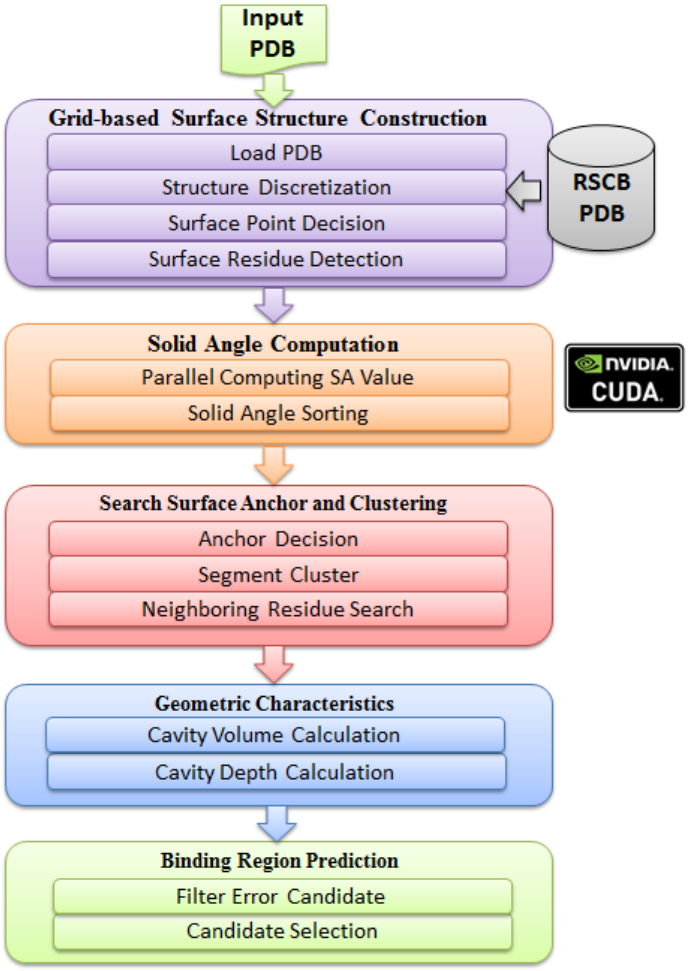


Figure 1 Flowchart of PLB-SAVE.

Grid-based Protein Structure Construction

The retrieved file from PDB database based on a PDB identifier or the input structural data file is stored in the PDB format [30] which contains complete spatial coordinate information of molecules through X-ray crystallography, NMR spectroscopy, Cryo-electron microscopy, or *in silico* prediction approaches. In this proposed system, the coordinates of atoms and corresponding van der Waals radii were transformed into

corresponding volume pixels (voxels) within a grid structure, which facilitated rapid identification of protein surfaces and efficient calculation of solid angle for each atom. After discretization processes, the query protein was represented as a set of discrete voxels which were categorized to inside, outside and surface portions of the query protein respectively.

Solid Angle Computation

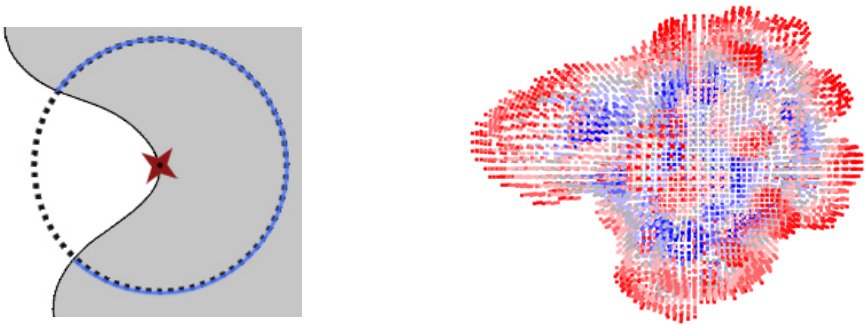
For each surface voxel within a protein, the proposed system computedits corresponding solid angle by using the following formula:

$$SA = (V_{in} / V_{sphere}) * 4\pi \quad (1)$$

where SA is the value of solid angle, V_{in} denotes the number of voxels located on both surface of sphere and inside the protein, and V_{sphere} represents the total number of voxels located within the sphere surface. In this step, the recommended radius of the sphere by Connolly was defined as 6 Å for all surface voxels[17]. The proposed system employed CUDA coding modules to compute solid angles on all surface voxels parallelly to enhance the computational performance.

Figure 2(a) illustrates the idea of how to efficiently calculate a solid angle from equation (1) and an example of calculated solid angles for all surface voxels of the query

protein. In Figure 2(b), the red spheres represented the solid angles with small values, and those surface voxels were also generally expressed as the spheres located on convex regions. Reversely, the blue colored spheres represented relatively large values of solid angles on the protein surface, and these voxels reflected concave areas. Flat surface areas were represented by spheres with white or lighter-shaded grey colors.



(a) (b)

Figure 2 (a) A 2D representation for solid angle calculation, V_{in} is the volume of virtual sphere located within the interior regions of the query protein (blue circle line), and V_{sphere} represent the volume of the total sphere (black dotted circle line). (b) Calculated solid angles on surface areas of the query protein (PDB ID: 1TPA). Red colored spheres are recognized as protruding regions; white or lighter-shaded spheres represent flat regions; blue colored spheres represent concave regions on protein surfaces.

Surface Anchor Residues and Clustering

Since we were looking for binding cavities from a query protein, in this study, only surface voxels with solid angles ranked in top 20% were clustered into representative groups. Two surface voxels would be clustered into an identical group when they were neighboring voxels located within a threshold distance (8 Å) and both voxels possessed solid angles at a similar level. A surface voxel with the largest solid angle within the selected clustered

group was considered as the representative anchor of the group.

is an example of surface voxels after clustering processes. Different colors represented different representative groups and red dots denote the anchors for various groups. These identified groups generally possessed larger average solid angles (concave regions) and were stored separately to facilitate the future applications on binding region identification.

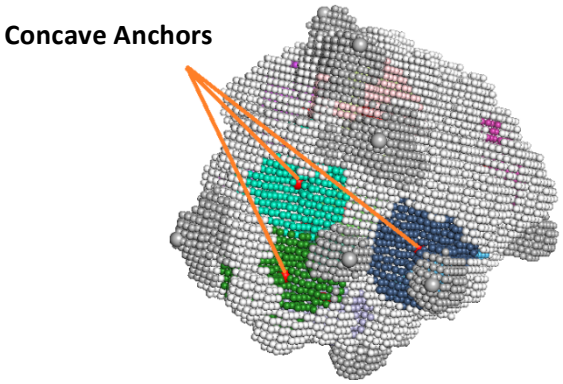


Figure 3 The clustered surface residues of the query protein structure (PDB ID: 1TPA) according to the ranking of solid angles. Red dots denoted the representative anchor of the clustered group.

Geometric Feature Calculation

After the assignment of clustered groups and representative anchors, the system calculated additional geometric characteristics for each group, including average depth and volume of identified anchor regions. These selected features were required to possess

rotation- and translation-invariant characteristics since a resolved protein structure by different experimental methods or from different laboratories was annotated in dramatically different locations. The following sections describe the geometric features in details:

Average Depth of Cavity

According to our observations, a defined surface anchor might possess a large solid angle, but not necessary conditions for all its neighboring surface residues. A cluster of surface residues containing distinct levels of solid angles sometimes caused wrong binding region prediction in this study. Hence, to avoid such high variations of neighboring surface residues within a group, an enhanced feature of average depth of a potential cavity was calculated and verified. The average depth was heuristically defined and evaluated according to the following formula:

$$Depth = \begin{cases} 5 & if\ SA > 0.9 * 4\pi \\ 4 & if\ 0.8 * 4\pi < SA \le 0.9 * 4\pi \\ 3 & if\ 0.7 * 4\pi < SA \le 0.8 * 4\pi \\ 2 & if\ 0.6 * 4\pi < SA \le 0.7 * 4\pi \\ 1 & if\ 0.5 * 4\pi < SA \le 0.6 * 4\pi \\ 0 & else \end{cases} \quad (2)$$

where the SA represented the solid angle for each clustered neighboring surface residues in a group. The average depth indicator was obtained by taking an average of transformed depths in the cluster. One example of 6 surface residues within a cluster was illustrated in

Figure 4, and the corresponding depth indicator was obtained by averaging the transformed values between solid angles and mapped depth values.

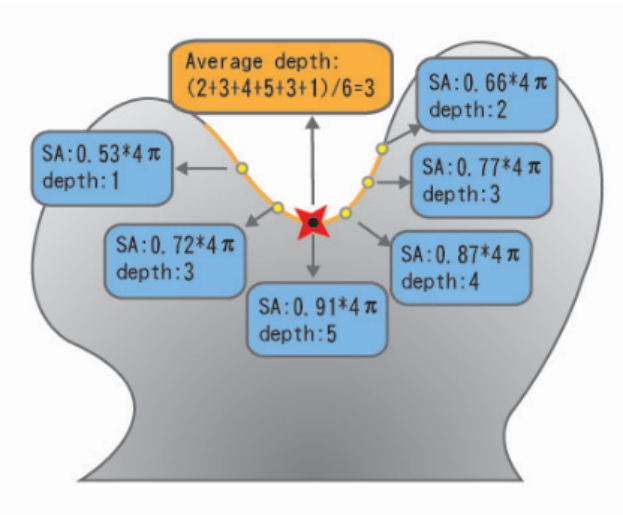
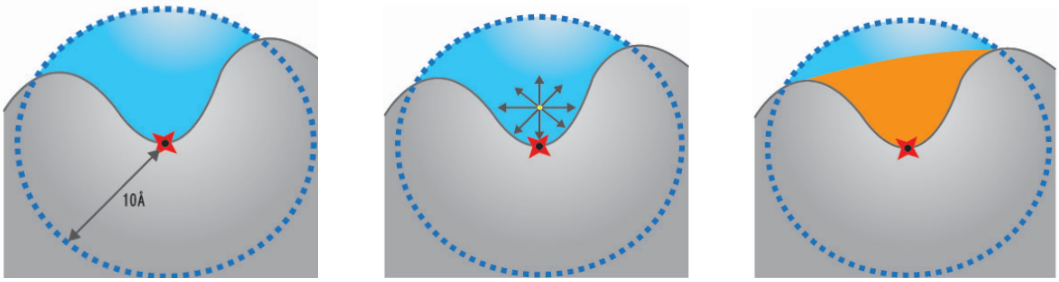


Figure 4 A simplified example of an average depth indicator for an anchor cluster with 6 neighboring surface residues.

Volume of a potential cavity

The volume of selected cavities provides identifiable discrimination between binding and non-binding regions. In this study, the volume indicator of a cluster was obtained by taking the anchor surface residue as a center and formulating a virtual sphere with a radius of 10 Å. Those voxels located within the virtual sphere but not inside the query protein were then evaluated by taking 7 directional vectors including the edge and diagonal vectors of a cube. If extending both directions of one of the directional vectors could intersect with the query protein simultaneously, then this directional vector was assigned as the interior directional vector. For each voxel under investigation, if it possesses more

than or equal to 4 verified interior vectors, this voxel is defined as part of the volume within the cavity. After examining all voxels in the virtual sphere, total interior voxel counts could provide as the volume value for the cluster. An example is shown in Figure 5. In this figure, each interior voxel was verified and evaluated through 7 extended directional vectors to see if the voxel belongs to the volume indicator.



(a) (b) (c)

Figure 5 An example for volume indicator: (a) a virtual sphere of 10 Å located at the center of anchor residue was constructed to evaluate the total number of potential volume voxels; (b) 7 extended directional vectors of a candidate volume voxel; (c) If a voxel possesses more than or equal to 4 extened directional vectors intersecting with the protein, the voxel was defined as one of the volume voxels and represented in orange, on the other words, the voxels within the virtual sphere but not belonged to volume contents, these voxels were depicted in blue.

When both geometric features of average depth and volume were obtained, a measuring score combining with linear weighting coefficients was performed for ranking all identified potential binding regions. The formula is written as the following equation:

$$RV(p) = \frac{CD(p)_{avg}}{CD_{max}} \times w_1 + \frac{CV(p)}{CV_{max}} \times w_2$$

Where the RV(p) is the ranking value for anchor residue p; $CD(p)_{avg}$ is the average depth

value of p; CD_{max} is the maximum depth ofthe query protein; CV(p) is the volume ofp; CV_{max} is the maximum volumn of the query protein; the sum of both weighting coefficients of w_1 and w_2 is equal to 1.

3. Experimental Results and Discussion

Experimental Datasets and Measurements

The testing protein datasets include two types of bound and unbound proteins which were collected from LigASite version 9.5 [31] (http://www.bigre.ulb.ac.be/Users/benoit/LigASite/index.php). Each dataset contains 388 representative and non-redundant protein structures. The binding sites of each protein were also provided for verification. Five evaluation parameters were calculated to compare the performance with other prediction systems including sensitivity, specificity, accuracy, positive predictive value (PPV) and Matthew's correlation coefficient (MCC). These measurements were obtained by the following formulae:

$$\text{Sensitivity} = \frac{TP}{TP + FN}$$

$$\text{Specificity} = \frac{TN}{TN + FP}$$

$$\text{Accuracy} = \frac{TP + TN}{TP + FP + TN + FN}$$

$$PPV = \frac{TP}{TP + FP}$$

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}}$$

System comparison

The proposed system was named as PLB-SAVE and freely available at http://SAVE.cs.ntou.edu.tw/. The prediction performance of PLB-SAVE was evaluated under a ten-fold cross-validation scheme. Both bound (HOLO) and unbound (APO) protein sets, each containing 388 representative proteins, were randomly partitioned into 10 subsets, respectively. Each partitioned subset was retained as the validation proteins for evaluating the prediction model, and the remaining 9 subsets were then applied as training data for setting best default parameters. The cross-validation process is repeated for ten times and each of the ten subsets was applied exactly once as the validation subset. The final measurements were obtained by taking average from individual ten prediction results. The final prediction results were shown in Table . Both prediction performances achieved superior and stable performance compared to most of previously published systems. It can be noticed that the performance for bound dataset performed better in all measurements than the unbound dataset in general. This is mainly due to that the trained parameters were obtained from verified binding regions and which were in the bound conditions. Hence, it can be expected that the bound dataset possessed better performance.

Table 1 Prediction system evaluated under a ten-fold cross-validation mechanism.

PLB-SAVE Cross-verification	APO-388 Proteins	HOLO-388 Proteins
Sensitivity	**0.579043**	**0.642564**
Specificity	**0.972336**	**0.976363**
Accuracy	**0.942588**	**0.955269**
PPV	**0.634765**	**0.651935**
MCC	**0.566041**	**0.613089**

To demonstrate the superior performance of the PLB-SAVE system, we compared the prediction results with two existed systems: SiteHound and MetaPocket2.0 (MPK2). The first SiteHound prediction system was published on Nucleic Acid Research, 2009 [32], and which identified ligand binding sites by computing interactions between a chemical probe and a protein structure, and adopted the profiles of affinity map and total interaction energy to rank the predicted binding sites. The second MPK2 system was published on Bioinformatics, 2011 [33], which integrated 8 approaches including LIGSITE^{CSC}, PASS, QsiteFinder [34], SURFNET, Fpocket [35], GHECOM, ConCavity [36], and POCASA. The prediction results were obtained by voting mechanisms to decide the predicted protein-ligand binding sites. Previously mentioned bound and unbound proteins in two testing datasets were uploaded one-by-one to these two prediction systems and their performances were listed in Table and Table 3. However, only partial proteins could be successfully predicted by both systems through on-line analysis. Though the proposed PBL-SAVE could successfully predicted all 388 protein structures, for fairly comparison, we

only selected identical structures which were able to be individually processed by these two existing systems. In Table 2, since 373 proteins from APO (uncound structures) could be analyzed by SiteHound and 181 proteins for MPK2, the comparisons of prediction measurements were listed respectively. It can be observed that except the sensitivity for the 181 proteins was worse than MPK2, all other measures were higher than these two system. However, the overall accuracy rate of PLB-SAVE was much higher than the MPK2. In addition, all 388 unbound structures could be successfully identified by the proposed system. Similarly, for bound proteins in HOLO dataset, our proposed system could be successfully predicted all 388 entries, but only 374 proteins for SiteHound and 148 proteins for MPK2. Neglecting the unpredictable proteins for both SiteHound and MPK2 systems, we evaluated the system performance according to the successfully predicted cases from these two systems respectively. Accordingly, from the Table 3 (a), the PLB-SAVE provided superior performance than SiteHound in terms of sensitivity, specificity, accuracy, PPV, and MCC indicators for bound proteins. In Table 3(b), the average prediction results were also better than MPK2 in most aspects except the sensitivity and MCC were slightly lower for these 148 protein structures. Nevertheless, we believe that if all proteins were required to be evaluated for the performance, the proposed PLB-SAVE system should outperform these two existing systems in all aspects. In addition, it could be observed as

previously mentioned phenomenon that the performance of three prediction systems for bound proteins were generally better than the unbound ones, which is due to the bound proteins possessing less flexibility on protein surface conformation and perhaps lower static energies as well. Interestingly, we found that the performance of the proposed system is more stable than the other two systems regarding the bound and unbound protein structures. For example, the prediction results from unbound to bound proteins, in terms of sensitivity, the increased performances were 11.1%, 42.0%, and 21.3% for the PLB-SAVE, SiteHound, and MPK2 respectively. The stable performance of a prediction system is important since the practical applications for unknown protein binding site prediction would mainly be unbound structures. Therefore, from the performance of our proposed system and compared results, it could reveal that simple and reliable features could provide a quite stable performance for protein binding region analysis.

Table 2 Prediction results of PLB-SAVE system on APO unbound dataset were compared to two existing systems respectively. All measurements were obtained according to successfully predicted proteins by SiteHound and MPK2. Bold faced data represent the best performance between two prediction systems. (a) PLB-SAVE compared with SiteHound with respect to 373 proteins. (b) PLB-SAVE compared with MPK2 with respect to 181 proteins.

APO (unbound structures)	PLB-SAVE (373 proteins)	SiteHound (373 proteins)
Sensitivity	**0.527**	0.379
Specificity	**0.968**	0.955
Accuracy	**0.934**	0.912

PPV	**0.583**	0.399
MCC	**0.509**	0.332

(a)

APO (unbound structures)	PLB-SAVE (181 proteins)	MPK2 (181 proteins)
Sensitivity	0.567	**0.710**
Specificity	**0.953**	0.904
Accuracy	**0.905**	0.878
PPV	**0.609**	0.478
MCC	**0.524**	0.500

(b)

Table 3 Prediction results of PLB-SAVE system on HOLO bound dataset were compared to two existing systems respectively. All measurements were obtained according to successfully predicted proteins by SiteHound and MPK2. Bold faced data represent the best performance between two prediction systems. (a) PLB-SAVE compared with SiteHound with respect to 373 proteins. (b) PLB-SAVE compared with MPK2 with respect to 181 proteins.

HOLO (bound structures)	PLB-SAVE (374 proteins)	SiteHound (374 proteins)
Sensitivity	**0.623**	0.538
Specificity	**0.975**	**0.975**
Accuracy	**0.953**	0.952
PPV	**0.629**	0.625
MCC	**0.589**	0.585

(a)

HOLO (bound structures)	PLB-SAVE (148 proteins)	MPK2 (148 proteins)
Sensitivity	0.673	**0.861**
Specificity	**0.959**	0.912
Accuracy	**0.927**	0.905
PPV	**0.654**	0.556

MCC	0.615	**0.634**

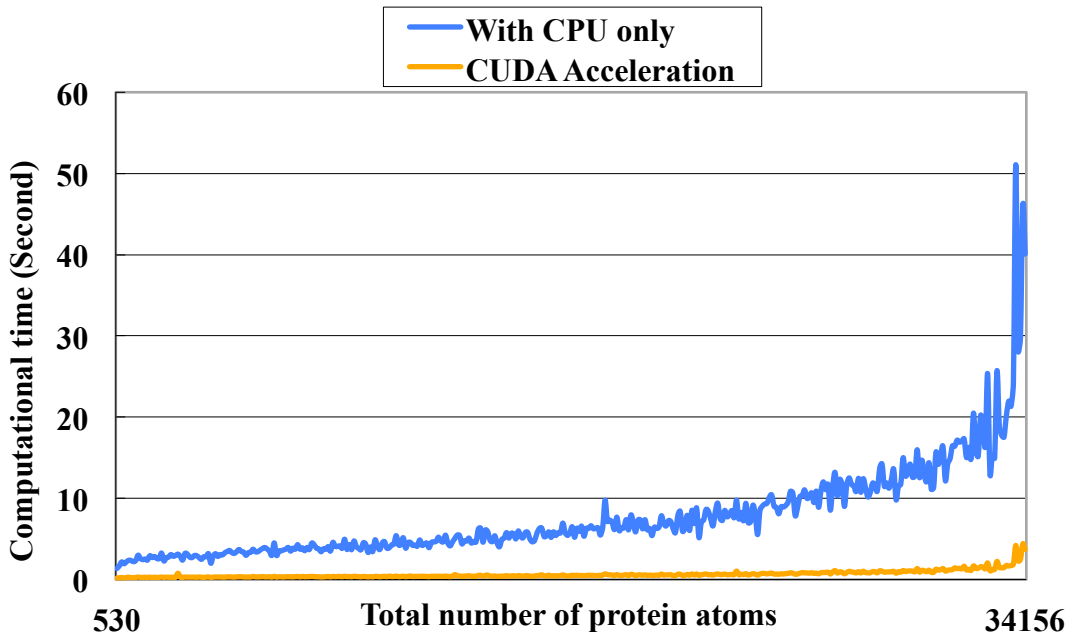
(b)

Computational Performance by CUDA

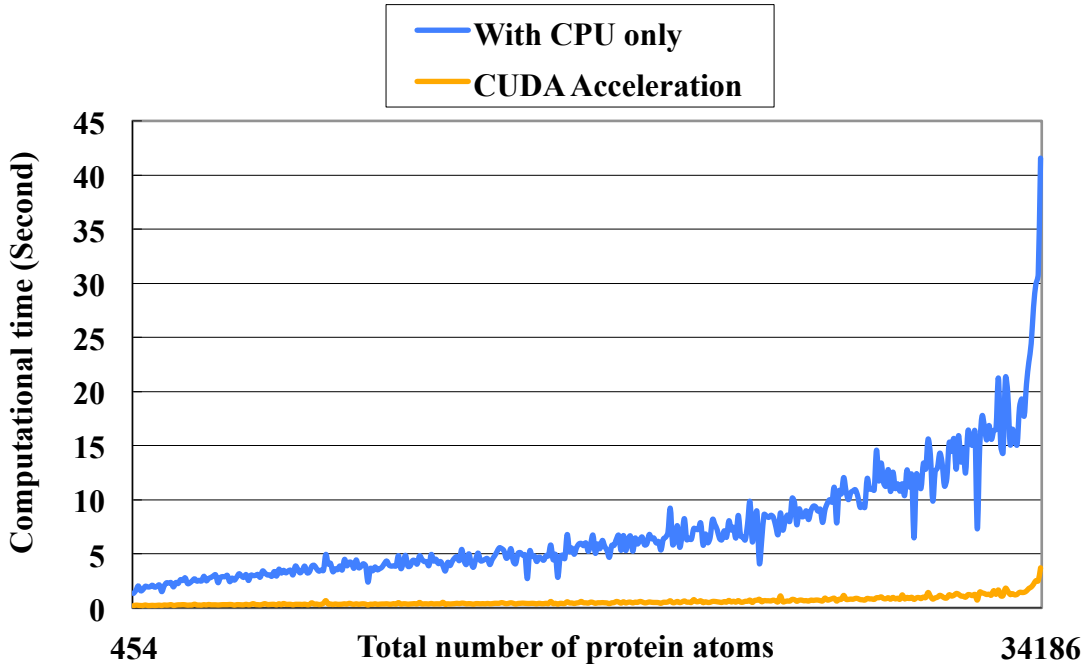
We employed the CUDA toolkit version 3.2 and Visual Studio 2008 to implement the proposed algorithms. For comparison, both adopting CPU architecture alone and utilizing CUDA computing architecture were implemented respectively. The experimental results were carried out on an Intel Dual-Core CPU E8400 2.6 GHz with 4 GB DDR2memory and a GeForce GTS 450 graphics card using the Microsoft Windows XP operating system.

The sizes of unbound protein structures in the APO dataset range from 58 to 4521 amino acids, atom number from 530 to 34,156, and surface points from 4,513 to 162,159 voxels. Required average computational time for computing solid angles through CPU and CUDA acceleration could be reduced from 7.03 seconds to 0.64 seconds. Similarly, the sizes of bound protein structures within complexes in the HOLO dataset range from 58 to 4520 amino acids, atom number from 454 to 34,186, and surface points from 4,510 to 141,201 voxels. Required average computational time for computing solid angles through CPU and CUDA acceleration could be reduced from 6.51 seconds to 0.59 seconds. The relationship between required computational time and the total number of atoms in both datasets was depicted in Figure 6. It can be observed that the performance of utilizing CUDA architecture can significantly reduce required computational time and gain more

improvement as the protein sizes increased. The CUDA implementation could obtain tremendously improved performance and near 11 fold faster in average for both bound and unbound testing datasets.



(a)



(b)

Figure 6 Required running time for geometric feature computation from both CPU alone and CPU incorporating with GPU on (a) unbound structure (APO) and (b) bound structure (HOLO) datasets.

4. CONCLUSION

The feature of solid angle in bioinformatics was originally proposed as early as 1986 by Connolly, and it is powerful and frequently applied to verify the uneven nature of surface bindings in a three-dimensional space. In this paper, based on the selected anchor surface residues from ranked solid angles, two extra important geometric features including the depth and volume of the selected potential cavities were employed. We developed an efficient and effective identification system for predicting the protein-ligand binding regions. The novel and important combinatorial features based on CUDA parallel processing technologies were proposed in this study. Accompany with the feature extraction algorithms for solid angles, clustering processes, anchor determination, and two extra geometric features were designed to facilitate the binding region selection. These identified protein-ligand binding regions on protein surface usually belong to concave structure from previous observations. Hence, all possible interactively combined anchors from the query proteins can be efficiently identified for the applications of drug and vaccine design strategies. In addition, not only protein-based vaccine but

carbohydrate-based vaccine is used in clinical prevention. The binding sites between the antibody and antigen are crucial for the efficacy of the protective effects. Recently, the carbohydrate-based vaccine has gained more and more attention due to the serotypes of various bacterial or viral strains. As well as the glycans exposed on the surface of cancer cells, carbohydrate has been developed as a target to be neutralized by an antibody for inducing the antibody-dependent cell-mediated cytotoxicity for cancer therapy [37]. Therefore to develop carbohydrate-based vaccine could be expected specifically to protect hosts against the infection and eliminate the cancer cells by immunotherapy. Thus, prediction of the ligand, such as carbohydrate or glycan, binding sites would contribute a lot of contributions to the field of vaccine development. The main contribution of this paper does not only emphasize on identifying the accurate protein-ligand binding regions, but also try to provide a practical way under the CUDA parallel computing architecture. Two testing datasets include 388 unbound and bound proteins were evaluated and compared to two existing well-known systems. The results have shown that the proposed parallel algorithms achieved an average accuracy rate of 94.9% for correctly identifying protein-ligand binding regions on two unbound and bound proteins, and an average of 11-fold faster for computing geometric features by employing CUDA architecture on these testing datasets. This PLB-SAVE can be applied as one of the first prediction tools for

protein surface analysis and protein-ligand binding region detection for practical applications on drug and vaccine development.

5. Competing interests

No competing interests.

6. Authors' contributions

YTL and HWW designed the algorithms and performed the experimental data analysis. TWP and HTC conceived of the study, participated in its design and coordination and helped to draft the manuscript. WST and HHH participated in the design and helped to review the manuscript. All authors read and approved the final manuscript.

7. Acknowledgements

This work is supported by the Center of Excellence for Marine Bioenvironment and Biotechnology of National Ocean University and National Science Council, Taiwan, R.O.C. (NSC 101-2321-B-019-001 and NSC 100-2627-B-019-006 to T.-W. Pai) and an award from the Taiwan Department of Health Clinical Trial and Research Center of Excellence (DOH101-TD-B-111-004).

REFERENCES

1. Jones S, Thornton JM: **Prediction of protein-protein interaction sites using patch analysis**. *J Mol Biol* 1997, **272**(1):133-143.
2. Astronomo RD, Burton DR: **Carbohydrate vaccines: developing sweet solutions to**

sticky situations? *Nat Rev Drug Discov* 2010, **9**(4):308-324.

3. Leis S, Schneider S, Zacharias M: **In silico prediction of binding sites on proteins**. *Curr Med Chem* 2010, **17**(15):1550-1562.
4. Walls PH, Sternberg MJ: **New algorithm to model protein-protein recognition based on surface complementarity. Applications to antibody-antigen docking**. *J Mol Biol* 1992, **228**(1):277-297.
5. Chung JL, Wang W, Bourne PE: **Exploiting sequence and structure homologs to identify protein-protein binding sites**. *Proteins* 2006, **62**(3):630-640.
6. Bradford JR, Westhead DR: **Improved prediction of protein-protein binding sites using a support vector machines approach**. *Bioinformatics* 2005, **21**(8):1487-1494.
7. Neuvirth H, Raz R, Schreiber G: **ProMate: a structure based prediction program to identify the location of protein-protein binding sites**. *J Mol Biol* 2004, **338**(1):181-199.
8. Liang SD, Zhang C, Liu S, Zhou YQ: **Protein binding site prediction using an empirical scoring function**. *Nucleic Acids Res* 2006, **34**(13):3698-3707.
9. Soga S, Shirai H, Kobori M, Hirayama N: **Use of amino acid composition to predict ligand-binding sites**. *J Chem Inf Model* 2007, **47**(2):400-406.
10. Chen R, Li L, Weng Z: **ZDOCK: an initial-stage protein-docking algorithm**. *Proteins* 2003, **52**(1):80-87.
11. Katchalski-Katzir E, Shariv I, Eisenstein M, Friesem AA, Aflalo C, Vakser IA: **Molecular surface recognition: determination of geometric fit between proteins and their ligands by correlation techniques**. *Proc Natl Acad Sci U S A* 1992, **89**(6):2195-2199.
12. Kozakov D, Brenke R, Comeau SR, Vajda S: **PIPER: an FFT-based protein docking program with pairwise potentials**. *Proteins* 2006, **65**(2):392-406.
13. Janin J, Seraphin B: **Genome-wide studies of protein-protein interaction**. *Current opinion in structural biology* 2003, **13**(3):383-388.
14. Bahadur RP, Chakrabarti P, Rodier F, Janin J: **A dissection of specific and non-specific protein-protein interfaces**. *J Mol Biol* 2004, **336**(4):943-955.
15. Bahadur RP, Zacharias M: **The interface of protein-protein complexes: analysis of contacts and prediction of interactions**. *Cellular and molecular life sciences : CMLS* 2008, **65**(7-8):1059-1072.
16. Burgoyne NJ, Jackson RM: **Predicting protein interaction sites: binding hot-spots in protein-protein and protein-ligand interfaces**. *Bioinformatics* 2006, **22**(11):1335-1342.
17. Connolly ML: **Measurement of Protein Surface Shape by Solid Angles**. *J Mol Graphics* 1986, **4**(1):3-&.

18. Hendrix DK, Kuntz ID: **Surface solid angle-based site points for molecular docking**. *Pac Symp Biocomput* 1998:317-326.
19. Shentu Z, Al Hasan M, Bystroff C, Zaki MJ: **Context shapes: Efficient complementary shape matching for protein-protein docking**. *Proteins* 2008, **70**(3):1056-1073.
20. Lanzavecchia S, Cantele F, Bellon PL: **Alignment of 3D structures of macromolecular assemblies**. *Bioinformatics* 2001, **17**(1):58-62.
21. Vouzis PD, Sahinidis NV: **GPU-BLAST: Using graphics processors to accelerate protein sequence alignment**. *Bioinformatics* 2010.
22. Manavski SA, Valle G: **CUDA compatible GPU cards as efficient hardware accelerators for Smith-Waterman sequence alignment**. *BMC Bioinformatics* 2008, **9 Suppl 2**:S10.
23. Schatz MC, Trapnell C, Delcher AL, Varshney A: **High-throughput sequence alignment using Graphics Processing Units**. *BMC Bioinformatics* 2007, **8**:474.
24. Trapnell C, Schatz MC: **Optimizing Data Intensive GPGPU Computations for DNA Sequence Alignment**. *Parallel Comput* 2009, **35**(8):429-440.
25. Ritchie DW, Venkatraman V: **Ultra-fast FFT protein docking on graphics processors**. *Bioinformatics* 2010, **26**(19):2398-2405.
26. Dynerman D, Butzlaff E, Mitchell JC: **CUSA and CUDE: GPU-accelerated methods for estimating solvent accessible surface area and desolvation**. *J Comput Biol* 2009, **16**(4):523-537.
27. Juba D, Varshney A: **Parallel, stochastic measurement of molecular surface area**. *J Mol Graph Model* 2008, **27**(1):82-87.
28. Friedrichs MS, Eastman P, Vaidyanathan V, Houston M, Legrand S, Beberg AL, Ensign DL, Bruns CM, Pande VS: **Accelerating molecular dynamic simulation on graphics processing units**. *J Comput Chem* 2009, **30**(6):864-872.
29. Dematte L, Prandi D: **GPU computing for systems biology**. *Brief Bioinform* 2010, **11**(3):323-333.
30. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE: **The Protein Data Bank**. *Nucleic Acids Res* 2000, **28**(1):235-242.
31. Dessailly BH, Lensink MF, Orengo CA, Wodak SJ: **LigASite--a database of biologically relevant binding sites in proteins with known apo-structures**. *Nucleic Acids Res* 2008, **36**(Database issue):D667-673.
32. Hernandez M, Ghersi D, Sanchez R: **SITEHOUND-web: a server for ligand binding site identification in protein structures**. *Nucleic Acids Res* 2009, **37**(Web Server issue):W413-416.
33. Zhang Z, Li Y, Lin B, Schroeder M, Huang B: **Identification of cavities on protein**

surface using multiple computational approaches for drug binding site prediction. *Bioinformatics* 2011, **27**(15):2083-2088.

34. Laurie AT, Jackson RM: **Q-SiteFinder: an energy-based method for the prediction of protein-ligand binding sites**. *Bioinformatics* 2005, **21**(9):1908-1916.

35. Le Guilloux V, Schmidtke P, Tuffery P: **Fpocket: an open source platform for ligand pocket detection**. *BMC Bioinformatics* 2009, **10**:168.

36. Capra JA, Laskowski RA, Thornton JM, Singh M, Funkhouser TA: **Predicting protein ligand binding sites by combining evolutionary sequence conservation and 3D structure**. *PLoS Comput Biol* 2009, **5**(12):e1000585.

37. Matthay KK, George RE, Yu AL: **Promising therapeutic targets in neuroblastoma**. *Clin Cancer Res* 2012, **18**(10):2740-2753.