PREDICTION AND MODELLING OF LIGAND-BINDING SITES USING AN INTEGRATED VOXEL METHOD

Ling Wei Lee
School of Computer Science
The University of Nottingham Malaysia Campus
Jalan Broga, 43500 Semenyih
Selangor Darul Ehsan, Malaysia

Andrzej Bargiela
School of Computer Science
The University of Nottingham Jubilee Campus
Wollaton Road,
Nottingham NG8 1BB, UK.

KEYWORDS
Ligand-Binding Sites, Voxelisation, Voxel-Based Analysis, Grid-Based Method, Binding Site Residues

ABSTRACT

The interaction between proteins and their binding agents take place on surfaces and involve factors such as chemical and shape complementarity. It was shown in past studies that protein-protein interactions involve flatter regions whereas protein-ligand bindings are associated with crevices. Many approaches have been implemented which focus on the identification of such sites using various measures. Here we present an integrated method based on the use of voxels and computer-vision in the search for ligand-binding areas. Each identified site is modeled and analysed in 2D with the corresponding residues listed out. We carried out our experiment on a set of 3 FK506-bound proteins and 2 heme-bound proteins and showed that the integrated method is capable of identifying correctly the sites of interest.

INTRODUCTION

Proteins are made up of combinations of amino acids and they carry out binding to external agents usually on their surfaces. A host of factors contribute to the reactivity of a site of interest including hydrophobicity, electronegativity, chemical composition, shape complementarity etc (Fisher et al, 1993; Cheng and Weng, 2003; Venkatachalam, 2003; Kellenberger et al, 2004; Nayal and Honig, 2006; Weisel et al, 2009). Bind site characteristics can be attributed to the arrangement of atoms therefore leading to the activation and deactivation of certain atoms and thereby giving the protein its unique set of functions. Protein surface analysis is capable of returning better exterior information compared to sequential or structural studies. Via et al (2000) stated that "protein surface comparison is a hard computational challenge and evaluated methods allowing the comparison of protein surfaces are difficult to find". One of the properties which allow a group of proteins to bind to the same ligand is the probable conservation of features within the dock sites indicating the proteins may be descended from the same family. However in the event of mutations non-related proteins may carry similar features as well (Kinnings and Jackson, 2009).

Many attempts have been undertaken in the past to study protein surfaces and identify potential dock sites. Some of the earliest programs available include POCKET (Levitt and Banaszak, 1992) and LIGSITE (Hendlich, Rippmann and Barnickel, 1997). The former uses an experimental sphere of a specified radius to examine protein surfaces for pockets in a 3-dimensional grid space, although the algorithm is still exposed to orientation-related problems. The latter identified this issue and introduced rigorous scannings to reduce the severity of the problem. From a 3-directional check the scan is increased to 7, the additional directions being the 4 diagonals.

Jones and Thornton (1997) proposed the use of surface patches for the detection of interaction sites on a protein. A series of parameters are calculated for each patch including the solvation potential, hydrophobicity, planarity, accessible surface area etc with the patch rankings determined based on these parameters. Bogan and Thorn (1998) presented the concept of 'hot spots' which correlates to active sites. The authors found that binding energy does not distribute evenly across the surface of a protein but tend to be highly concentrated on dock areas. In a work by Fernandez-Recio et al (2005) the Optimal Docking Area (ODA) method was presented focusing on hot spots. ODA identifies patches through experimentations of different atomic solvation parameters. It was also found that larger interfaces generally consist of multiple patches with at least a pair of patches equivalent in size to a single patch interface (Chakrabati and Janin, 2002).

Understanding of the factors contributing to an active site is vital for successful detection of such areas. Most grid-based approaches prioritise shape complementarity in the search for potential sites. A crevice has to be sufficiently large to accommodate a ligand for interaction to take place. In our approach we present an integrated method using a combination of computer-vision techniques and voxel-based environment for the identification and modeling of potential binding sites. All associated atoms and corresponding residues are extracted as well.

Proceedings 26th European Conference on Modelling and Simulation ©ECMS Klaus G. Troitzsch, Michael Möhring, Ulf Lotzmann (Editors)
ISBN: 978-0-9564944-4-3 / ISBN: 978-0-9564944-5-0 (CD)

BACKGROUND

The surface of a protein is an interesting landscape of concave and convex areas. Each protein has its own set of defined functionalities. The use of grid spaces or voxels in protein studies is no longer a new paradigm as demonstrated in programs like POCKET and LIGSITE. The grid space offers a fast and robust solution to many applications. In our implementation we introduce a cubic grid-space for the identification of potential dock sites based on computer vision-inspired techniques.

A cubic grid-space is first constructed large enough to contain within it the entire protein. The experimental space is then tessellated into smaller units, with each unit having a size of 4 Å (the size of a voxel which fully encapsulates most atoms). All data sources for the test proteins are obtained from the RCSB Protein Data Bank in PDB format. We then extract all required information from the files including the spatial coordinates of all the atoms, the atom types and included the van der Waals radii for the atoms. These information are compiled into a new file that will be used as input for the algorithm.

We have chosen well-tested proteins for the study. The first set consists of FK506-bound proteins [PDB: 1FKF, 1BKF, 1YAT] which are molecules having a single active binding site each for one substrate. The protein 1FKF has been experimentally determined through a wet lab approach (Van Duyne et al, 1993) and attempted as well using a geometrically-based search coupled with geometric hashing (Peters, Fauck and Frommel, 1996). With proven results this protein serves as a good test subject. The second set consists of heme-bound proteins [PDB: 4HHB, 4MBN]. All input proteins are first projected into the 3D grid environment and tessellation of the space is carried out. As the 3D space induces a higher complexity compared to 2D processing, as such a 'slicing' process is carried out which converts the 3D environment into a series of 2D images by selection of a chosen dimension for conversion (Lee and Bargiela, 2009). This is conceptually similar to the Z-buffer algorithm in 3D graphics.

Each obtained image is processed using simple image processing techniques to identify voxels related to the protein. Surface voxels are then identified such that a list of surface atoms can be obtained (Lee and Bargiela, 2010). As each protein is encapsulated in a grid space, one is only able to obtain 6 views of the protein based on the characteristics of the cube. In a visual sense, crevices on the protein can be discerned through a sense of depth and clarity. We attempt to identify potential dock sites based on the depth attribute. A cuboid is first grown within the protein until it hits a plateau in each of the 6 faces. This defines the starting plane of visual projections executed from within to the surface. A depth count is then carried out and any area in which the count is smaller than the surface average or a user-specified

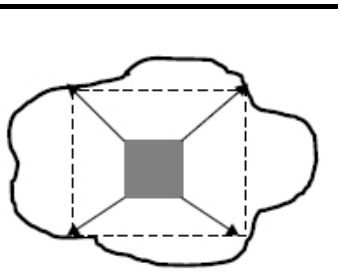


Figure 1. The construction of a cuboid beginning from the center of the protein. The algorithm terminates when the largest fully-filled cuboid has been obtained.

threshold is defined as a potential dock site. Finally all associated atoms and residues are projected for the sites.

THE ALGORITHM

The human vision is capable of perceiving areas protruding from the protein and 'valleys' of which binding agents of matching shape and chemical configurations may bind to. However to translate this into a simulated system is computationally challenging. As such most algorithms seek to minimise complexities and represent the problem in simpler domains. In our implementation, the protein is first enclosed in a grid space. By translating the human view to the 6 faces of the grid space (since the grid space is made up of voxels and each voxel has only 6 faces) one can then obtain 6 views of the protein.

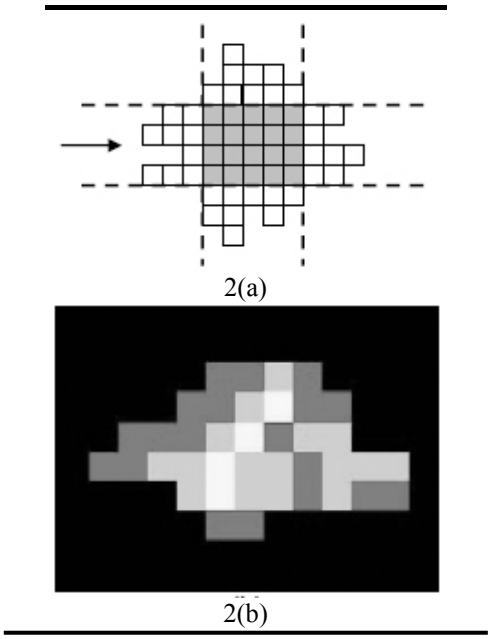
For each of these 6 faces, we proceed to locate the crevices within the viewing boundary of each face. The 'starting plane' for each face is first defined by identifying the largest cuboid beginning from the averaged center of the protein. The algorithm terminates when a perfect plane (one in which the plane is fully occupied by voxels) in any axis is no longer encountered. A 2D visualisation of this approach is given in Figure 1. The surfaces of this inner cuboid become the 'starting planes' for all analyses working outwards beginning from the voxels situated on the plane. Each higher level of the voxels builds on the previous level, therefore rendering some voxels on the lower levels hidden to the external environment. Such a move is capable of determining potential sites if deeper clefts are found and are externally exposed. Due to the inside-out numbering of the levels binding sites have smaller depth-level values and outermost regions have larger numbers.

A breakdown of the cuboid-growing process is given below.

1. All identified surface voxels and surface atoms are first loaded and stored into memory.
2. The full list of voxels defining the protein – including both internal and surface voxels – is accessed.

3. Identify the largest possible cuboid constructed from voxels from the center of the protein. the process begins with the initiation of an 'infant cube' in the from of an equilateral cuboid. The rule is such that the cuboid must not contain any parts devoid of voxels.
4. Once the 'infant cube' has been created, the method then proceeds to stretch all sides of the cube until the largest possible cuboid is attained that is completely filled with voxels.

Following that a color grid map is created with different levels of voxels distinctly color-marked based on the depth level. A matrix is instantiated alongside this grid map for the checking of potential sites. A rule-based 3x3 window is designed for use in the scans. Figure 2(a) shows how the internal cuboid provides a 'starting plane' for internal-to-external voxel layering with the color map for the arrow-designated face presented in Figure 2(b). The matrix for the color map is given in Figure 3 – all 0s represent blank spaces, 1 for the deepest levels, 2 for a level higher and so on.



2(a)

2(b)

Figure 2. (a) The cuboid acting as a starting plane for internal-to-external visualisation of the voxels for all faces. (b) The color-map projection of the tessellated protein from the view specified by the arrow in (a). Different colors indicate different depth levels. The image has been converted to grayscale.

0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	1	1	2	1	0	0	0	0
0	0	0	0	1	1	2	3	1	1	0	0	0
0	0	1	1	1	2	3	1	2	2	0	0	0
0	1	1	2	2	3	2	2	1	2	2	2	0
0	0	0	0	2	3	2	2	1	2	1	1	0
0	0	0	0	0	1	1	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0

Figure 3. The resulting matrix for the color-map presented in Fig. 2(b).

The 3x3 window is then used to filter the matrix for the sites. The purpose of this matrix is to reduce the computation time of processing a face from the voxel. Scanning of the color-map takes longer time as each voxel has 1600 pixels (40x40) and there are many voxels in a face. With the use of a matrix the computation time is effectively reduced – only an integer array of NxN dimensions is involved. This is many times faster than processing of the color-map. The filter-window is dependent on the threshold value defined by the user. In most cases a value equivalent to the average of all depth levels suffices. However should the user wished to obtain a larger or smaller model of the dock site, the threshold value may be adjusted accordingly.

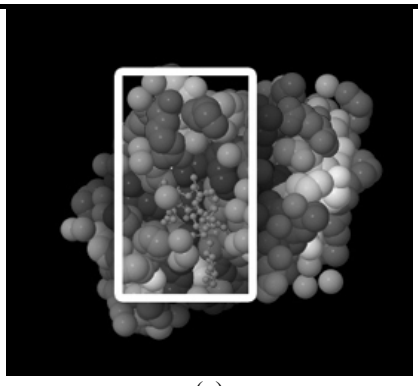
The final step in the process is to check all surface atoms against the selected range of dock site voxels. The atoms are shortlisted if they are found to be partially or wholly contained within the voxels. A list of residues associated with the identified atoms is compiled and compared against visualisations from the RCSB PDB and for the case of protein 1FKF, comparisons are made against both the documented wet lab and geometric hashing results.

RESULTS

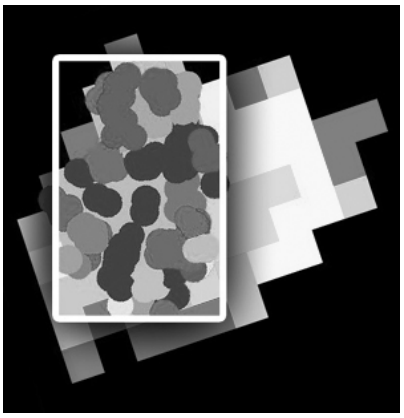
The results obtained for this study are divided into two sections. The first section presents the output for the FK506-bound proteins, with extra comparisons carried out for the protein 1FKF to published results, and the second section gives the output for the heme-bound proteins. The identified sites of each protein is presented in image-form and compared to screenshots from the RCSB PDB. Note that all screen shots from the PDB include solvent molecules whereas these molecules have been omitted in the implementation.

(A) FK506-Bound Proteins

Protein 1FKF



(a)



(b)

Figure 4. Protein 1FKF. (a) Visualisation from RCSB PDB. (b) Dock site identification from voxel-based integrated approach.

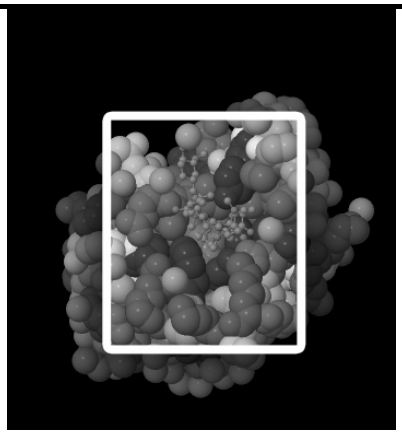
Table 1. Comparison of extracted residues from a wet lab experimentation, a geometric hashing-based approach and the implemented method for protein 1FKF.

	Residue	WL*	GH**	Voxel
1	TYR26	Y	Y	Y
2	PHE36	Y	Y	Y
3	PHE46	Y	Y	Y
4	VAL55	Y	Y	Y
5	ILE56	Y	Y	Y
6	ARG57	-	Y	Y
7	TRP59	Y	Y	Y
8	ALA81	Y	Y	Y
9	TYR82	Y	Y	Y
10	PHE99	Y	Y	Y
11	ASP37	Y	-	Y
12	ARG42	Y	-	Y
13	GLU54	Y	-	Y
14	HIS87	Y	-	Y
15	ILE91	Y	-	Y

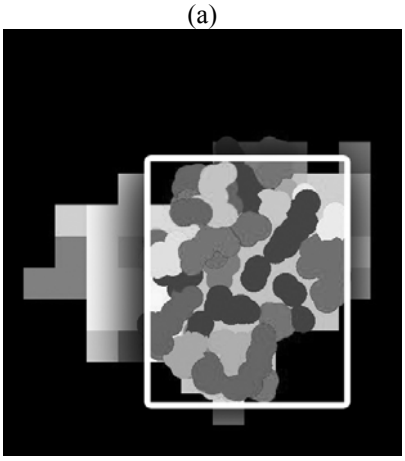
* WH – Wet Lab
** GH – Geometric Hashing

Based on the results from the table it can be seen that the integrated voxel-based method correctly extracts all the residues involved in the binding site. However the number of excess residues obtained is high as well – although this can be considered a small problem as there are bound to be atoms from unconcerned residues contained within the identified dock site voxels. This is a compromise that has to be made due to the use of a lower level representation of the protein in voxel units. the excess residues are listed as ILE90, TYR80, PRO45, ASP79, ARG40, GLN53, PRO78, ASP41, PHE48, ASN43, GLY58, SER39, GLY83, LYS44, PRO88, HIS25 and LYS47.

Protein 1BKF



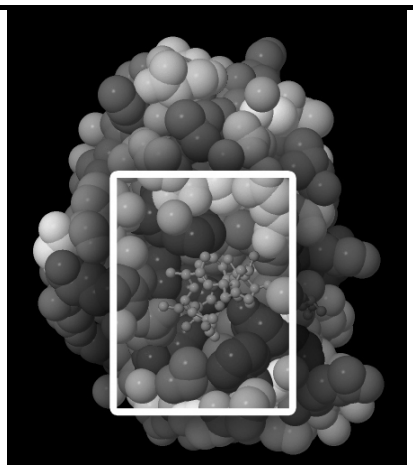
(a)



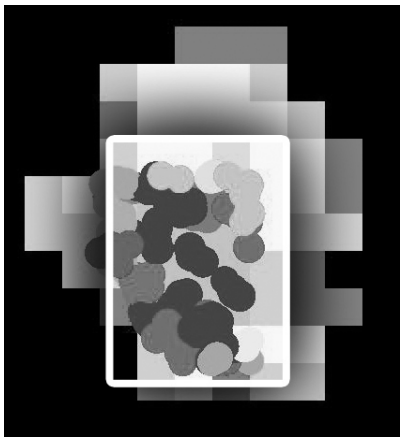
(b)

Figure 5. Protein 1BKF. (a) Visualisation from RCSB PDB. (b) Dock site identification from voxel-based integrated approach.

Protein 1YAT



(a)

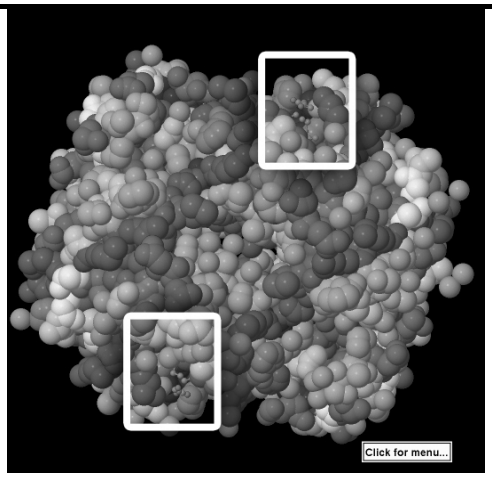


(b)

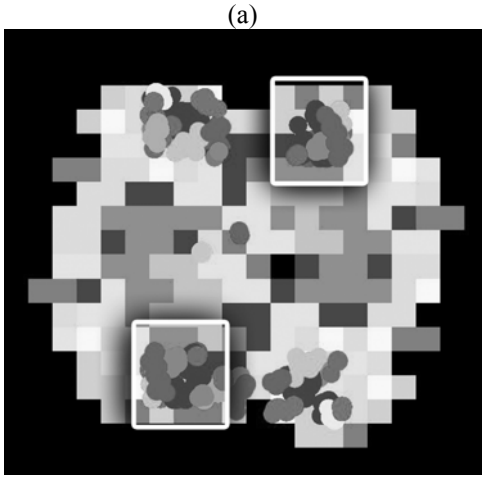
Figure 6. Protein 1YAT. (a) Visualisation from RCSB PDB. (b) Dock site identification from voxel-based integrated approach.

(B) Heme-Bound Proteins

Protein 4HHB



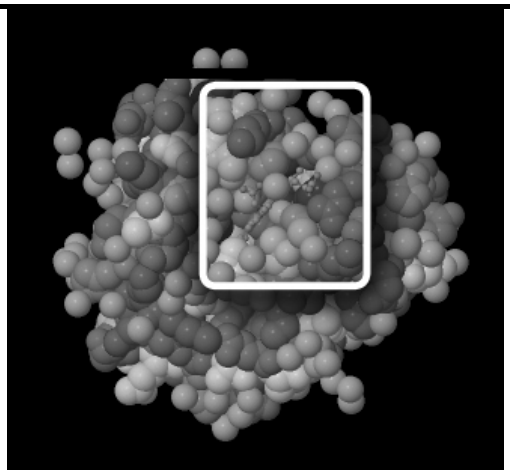
(a)



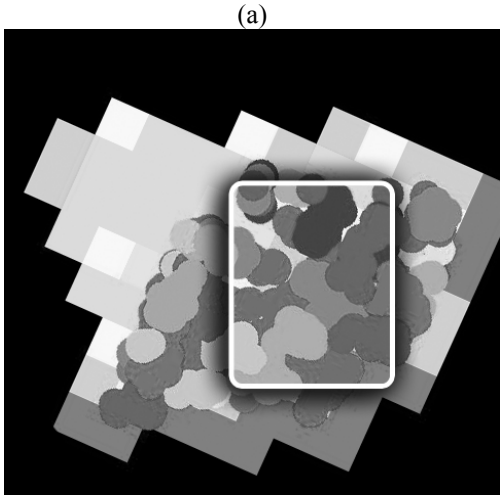
(b)

Figure 7. Protein 4HHB. (a) Visualisation from RCSB PDB. (b) Dock site identification from voxel-based integrated approach.

Protein 4MBN



(a)



(b)

Figure 8. Protein 4MBN. (a) Visualisation from RCSB PDB. (b) Dock site identification from voxel-based integrated approach.

DISCUSSION

The results reported above show that a voxel-based approach integrated with some visuals understanding is capable of identifying potential dock sites on the surfaces of proteins. The algorithm correctly identifies the residues contributing to the dock site, and is reasonably efficient computationally. A protein with ready datasets of surface voxels and atoms requires approximately 10 seconds of processing time on a standard PC to identify and output potential dock sites from all 6 viewing platforms, although the time is largely dependent on the size of the protein as well. The areas and depth of all 6 faces are contributing factors to the computational complexity of the method. A larger and deeper area results in a higher number of voxels being processed starting with the identification and arrangement of the voxels in generating the levels and color-map to the filtering of the corresponding matrix in the search for potential regions. As the matrix is composed of a set of numbers therefore the filtering process is fast with a complexity of O(n).

However as with most algorithms the method has its limitations as well. The algorithm works well when potential binding sites are located parallel to any of the 6 faces of the voxel. The contradiction comes in the form of a site located on any edge and is split between any two (or more) faces. This calls for further enhancements for tackling of such issues.

The proposed method is simple to implement and is effective in identifying potential dock sites. Due to the representations of the protein in voxels terms which effectively reduces the resolution, the probability of each voxel containing atoms belonging to several residues is high. This explains the excess residues obtained from the identified dock sites for the proteins, and is listed in comparison for protein 1FKF. Despite that the excess residues are mostly neighbouring entries which help indicate the location of the site of interest although they do not contribute directly to the site.

CONCLUSION

An integrated approach based on a voxelisation method equipped with understanding of visuals is presented here for the identification and modeling of potential dock sites on proteins. The experiment was carried out on 2 sets of proteins (FK506-bound proteins and heme-bound proteins) with the dock sites correctly identified. Compilation of the list of residues for the site of protein 1FKF showed successful extractions comparable to both the wet-lab and geometric hashing approaches. Dock sites were identified using a 'depth-level' scanning with potential regions targeted at areas having lowest numbers. Once all site-related voxels have been identified, the algorithm lists all atoms and residues associated with the site. Although excess extractions were obtained, the method remains a promising solution based on the quality of the results obtained.

REFERENCES

Bogan, A.A., Thorn, K.S. (1998) "Anatomy of Hot Spots in Protein Surfaces." J. Mol. Biol. 280, 1 – 9.

Chakrabati, P., Janin, J. (2002) "Dissecting Protein-Protein Recognition Sites." Proteins 47, 334 – 343.

Cheng, R., Weng, Z. (2003) "A Novel Shape Complementarity Scoring Function for Protein-Protein Docking" Proteins 51, 397 – 408.

Fernandez-Recio, J., Totrov, M., Skorodumov, C., Abagyan, R. (2005) "Optimal Docking Area: A New Method for Predicting Protein-Protein Interaction Sites." Proteins 58, 134 – 143.

Fisher, D., Norel, R., Wolfson, H., Nussinov, R. (1993) "Surface Motifs by A Computer Vision Technique: Searches, Detections, and Implications for Protein-Ligand Recogniton." Proteins 16, 278 – 292.

Hendlich, M., Rippmann, F., Barnickel, G. (1997) "LIGSITE: Automatic and Efficient Detection of Potential Small Molecule-Binding Sites in Proteins." J. Mol. Graph. Model. 15, 359 – 363.

Jones, J., Thornton, J.M. (1997) "Analysis of Protein-Protein Interaction Sites using Surface Patches." J. Mol. Biol. 272, 121 – 132.

Kellenberger, E., Rodrigo, J., Muller, P., Rognan, D. (2004) "Comparative Evaluation of Eight Docking Tools for Docking and Virtual Screening Accuracy" Proteins 57, 225 – 242.

Kinnings, S.L., Jackson, R.M. (2009) "Binding Site Similarity Analysis for the Functional Classification of the Protein Kinase Family." J. Chem. Inf. Model., 49, 318 – 329.

Lee L.W., Bargiela A. (2009) "Space-Partition Based Identification of Protein Docksites." Proceedings of the 23rd European Conference on Modelling and Simulation (ECMS 2009), 848 – 854.

Lee L.W., Bargiela A. (2010) "Statistical Extraction of Protein Surface Atoms based on A Voxelisation Method." Proceedings of the 24th European Conference on Modelling and Simulation (ECMS 2010), 344 – 349.

Levitt, D.G., Banaszak, L.J. (1992) "POCKET: A Computer Graphics Method for Identifying and Displaying Protein Cavities and Their Surrounding Amino Acids." J. Mol. Graph., 10, 229 – 234.

Nayal, M., Honig, B. (2006) "On the Nature of Cavities on Protein Surfaces: Application to the Identification of Drug-Binding Sites." Proteins 63, 892 – 906.

Peters, K.P., Fauck, J., Frommel, C. (1996) "The Automatic Search for Ligand Binding Sites in Proteins of Known Three-Dimensional Structure Using only Geometric Criteria." J. Mol. Bio. 256, 201 – 213.

Weisel, M., Proschak, E., Kriegl, J.M., Schneider, G. (2009) "Form follows Function: Shape Analysis of Protein Cavities for Receptor-based Drug Design" Proteomics 9, 451 – 459.

Van Duyne, G.D., Standaert, R.F., Karplus, P.A., Schreiber, S.L., Clardy, J. (1993) "Atomic Structures of the Human Immunophilin FKBP-12 Complexes with FK506 and Rapamycin." J. Mol. Biol. 229, 105 – 124.

Venkatachalam, C.M., Jiang X., Oldfield, T., Waldman, M. (2003) "LigandFit: A Novel Method for the Shape-Directed Rapid Docking of Ligands to Protein Active Sites" J. Mol. Graph. Model. 21, 289 – 307.

Via A., Ferre F., Brannetti B., Helmer-Citterich M. (2000) "Protein Surface Similarities: A Survey of Methods to Describe and Compare Protein Surfaces." Cell. Mol. Life Sci. 57, 1970-1977.

AUTHOR BIOGRAPHIES

LING WEI LEE was born in Kuala Lumpur and studied at the University of Nottingham Malaysia Campus where she took up Computer Science and obtained her honours degree in 2007. She worked for about a year as an analyst programmer with a local company before deciding to pursue her postgraduate studies. Her current research focuses on the use of multiresolution and computational methods in the area of proteins. She is currently in her final year of completion. She can be reached at leelingwei@yahoo.co.uk.

ANDRZEJ BARGIELA is Professor in the School of Computer Science at the University of Nottingham, UK. He served as Director of the School of Computer Science at the Malaysia Campus of the University of Nottingham from 2007 to 2010. He is currently President of the European Council for Modelling and Simulation and serves as Associate Editor of the IEEE Transactions on Systems Man and Cybernetics and Associate Editor of the Information Sciences. His research involves investigation into Granular Computing, human-centred information processing as a methodological approach to solving large-scale data mining and system complexity problems. He can be reached at Andrzej.Bargiela@nottingham.ac.uk.