Protein Surface Atoms Extraction: Voxels as an Investigative Tool

Ling Wei Lee, and Andrzej Bargiela, *Member, IEEE*

***Abstract*—Formed of amino acids combinations, proteins are essential components of living beings which participate in the regulation of bodily functions. Each protein is assigned specific function(s) and reacts to external agents of the best fits. Most binding activities take place on surfaces specifically in regions of high complementarity. The structure and chemical composition of each site are defined by the arrangement of residues and their corresponding atoms. As such the extraction of protein surface atoms provides a good listing for investigation of surface properties as well as aids in reducing the amount of processing required in computer-aided drug design (CADD) programs. Many algorithms are available for the analysis of protein surfaces including but not limited to methods of probe or geometrical nature, calculation of hot spots and energy functions, triangulation and Voronoi tessellations etc. Grid units have been used for locating potential cavities in early programs such as POCKET and LIGSITE but the role of voxels in the extraction of surface atoms has not been thoroughly investigated. A method is presented here which enlists voxels as its main experimental tool with constraints applied. These constraints come in the form of voxel occupancy as well as the degree of belonging of an atom to the voxel. Application of these rules to the voxels lead to considerable improvements in the extracts, with further enhancements made through the implementation of a 'peeling' method for removing internal atoms found in the output. The study was carried out on sets of proteins with the results visualised and compared to output from the MSMS and Surface Racer programs.**

***Index Terms*—constraints-filtering, protein-surface-atoms, space-tesellation, voxel-based-studies.**

I. Introduction

PROTEINS are essential components of all living beings and they actively participate in the regulation of bodily functions, digestion as well as a host of other responsibilities. According to [1] there are some 60,000 varieties existing within human cells wherein each protein is tasked with its own specific set of functions. There are 20 different types of amino acids which may be categorised as hydrophilic, hydrophobic or charged based on their properties [2]. These amino acids combine to form stable polypeptides resulting in the many different species and families of proteins available today.

There are 4 different structural levels associated with proteins. Sequences of proteins are classified as primary level entries, and the folding of groups of primary chains resulting in conformations such as α-helices and β-sheets lead to the formation of secondary structures. Grouping of these secondary structures contribute to the protein as a whole while quaternary structures are obtained from the bonding of several proteins into specific arrangements. In view of the various structural levels, research works have been carried out in the investigation of properties associated to these levels for better understanding of proteins.

Early research initiatives place focus on the comparison of protein sequences. String encodings were aligned in the search for the longest common sub-sequence or for repetitive patterns between the compared entries. Examples of such algorithms include the Smith-Waterman [3] and the Needleman-Wunsch [4]. Systems have been developed as well for large scale sequence comparisons such as PSI-BLAST [5] and CLUSTALW [6]. Although computationally efficient, with attempts taken to predict protein folding based on strings, however sequences are not capable of suggesting the actual structures proteins fold into. A new research field emerged thereafter – an area wherein emphasis is placed on the study of proteins based on crystallographically-determined structures. Common sub-structures become easily identified using the given spatial coordinates and atomic details of each protein. The increasing number of structures lead to the development of databases such as SCOP (Structural Classification of Proteins) [7] and CATH (Class, Architecture, Topology, Homologous superfamily) [8] which classify structures based on unique hierarchies.

Both the sequence and structure-based studies are capable of grouping proteins into meaningful families based on detected similarities. Such classifications provide meaningful suggestions on possible functions associated with newly discovered proteins and are especially helpful in drug discovery procedures. Each protein is selective by nature and binds to specific agents or ligands. This behavior may be interpreted from the arrangement of residues and atoms on the surfaces of proteins. According to Via et al [9], "protein surface comparison is a hard computational challenge and evaluated methods allowing the comparison of protein surfaces are difficult to find". Many implementations are available as of current and these include approaches such as triangulation, Voronoi tessellations, lattice modeling, multi-resolution modeling, geometric hashing etc [10]-[13]. One of the more commonly used geometric-based methods is the convex hull which is a subset of the Delauney triangulation and is defined as the smallest convex polyhedron enclosing all atom centers. Most methods

Manuscript received August 21, 2012; revised August 21, 2012.

Ling Wei Lee is with the School of Computer Science, University of Nottingham Malaysia Campus, Jalan Broga, 43500 Semenyih, Selangor Darul Ehsan, Malaysia. (email: leelingwei@yahoo.co.uk).

Andrzej Bargiela is with the School of Computer Science, University of Nottingham Jubilee Campus, Wollaton Road, Nottingham NG8 1BB, UK and Institute of Informatics, Cracow Technical University, Poland (tel: 01158467279 email: Andrzej.Bargiela@nottingham.ac.uk).

developed based on polyhedrons return estimations of protein surfaces, and in the process sacrificed small details which may be vital to the definition of the subjects. However the approach of generalising surfaces is usually computationally inexpensive and typically retains larger (and significant) regions.

An early method used for investigating protein surfaces is the Connolly method [14] in which a water-molecule-sized probe is used to roll over surfaces for inspection of the exterior. The concepts of Solvent Excluded Surface (SES) [15] and Solvent Accessible Surface (SAS) [16] are manifestations of the probe technique. In a program by Sanner termed MSMS [17] the author implemented reduced surface versions of both SES and SAS for the fast examination of proteins. Options are available for changing the probe size as well as commands for the list of output files to be generated. The AREA files from the program are used in this study for comparison of identified surface atoms against the output obtained from the voxel-based method. On the other hand the Surface Racer program [18] implements calculations for the exact accessible surface area, molecular surface area as well as the average curvature of molecular surface. Similar to the MSMS, users are allowed to change the size of the probe and to specify the algorithms they would like executed.

Grid spaces or voxels have been used in the study of proteins as displayed in early programs like POCKET [19] and LIGSITE [20]. The POCKET program uses grids and a test sphere to identify potential cavities but suffers from grid-related orientation issues. LIGSITE identifies the weaknesses of POCKET and includes solutions to overcome the issues by introducing additional rigorous scanning. From a 3-directional scan in the former the checking has been increased to 7 in the latter, with the additional scans targeted at the 4 diagonals. LIGSITE is claimed to be fast and is capable of locating potential binding sites to high precisions. In a study by [21] cubes are used for the soft docking of proteins. The authors showed that pure geometric docking with conformational changes is sufficient to determine the matching entities in the test cases. However not all proteins form rigid docking bonds and as such there are limitations to the use of only geometrical criteria for docking studies.

Other existing methods investigate chemical complementarity as well as energy functions associated with active sites. The surface of each protein is divided into patches and each patch includes the parameters of solvation potential, hydrophobicity, planarity, accessible surface area etc with rankings carried out based on the values of all potential patches [22][23]. It was found that energy does not distribute evenly across the surfaces of proteins but is concentrated on dock areas. GlamDock [24] was developed for flexible ligand docking and is based on a simple Monte Carlo approach which includes combinations of energy functions and search space definitions. Molecular dynamics were used as well in the design of docking algorithms [25]. The method was shown to be fast and accurate in rigid protein-flexible ligand complexes. It was reported that the average structure of a protein experiences changes when sufficient favourable energy is available for an induced-fit bind [26].

Grid spaces have been used in a number of programs as experimental environments [27]-[29] therefore suggesting their robustness as an investigative tool. A grid space constitutes of units or voxels of various resolutions as its subsets, and objects may be represented and defined in the number of units occupied, the estimated total surface area and volume, and the overall shape based on descriptions of the voxels cluster. Partitioning can be carried out to subdivide a space into units of desired sizes. Although the orientation problem associated with the rigidity of grid spaces is inherent in most implementations, however measures can be taken to reduce its impact. The same object of two different orientations often produces two varied sets of voxels in which some units may be the same while differing in the rest. One way of overcoming the orientation problem is to pre-rotate the protein based on fixed rules such that it is always aligned to a particular axis. Another way of ensuring data consistency is through the use of statistical studies. Both these approaches were employed in the research, the procedures of which are presented in the next section.

A surface atoms extraction algorithm returns full atoms listings which may be used for the study of binding regions on protein surfaces. The exclusion of internal atoms also reduces processing time in computer-assisted drug design (CADD) programs by eliminating internal entries which do not actively participate in any activity. A recent study was carried out by Kim et al [30] on a real-time method for locating boundary surface atoms using a GPU. In a similar experiment [31] surface entries were identified based on the computed atomic contribution to the SAS area. To investigate if voxels are capable of extracting surface atoms with results of competitive nature, a method is proposed here which introduces experimentally-determined constraints to voxels for the selection and filtering of surface atoms. Using only spatial coordinates, atom type information and van der Waals radii this study proceeds to show that a combination of voxel properties and constraints are capable of delivering promising results. All methods and procedures are described in the next section, beginning with the definition of surface atoms followed by steps taken for pre-processing and concluding with the algorithm for extraction of surface atoms.

II. Surface Atoms and Pre-Processing

A. Definition

What defines an atom as being a ‘surface’ entry? In general, a surface atom can be described as one that is located on the exterior or outermost layer of a molecule. It should be exposed to the external environment, and should be viewable to the investigator. A surface atom may or may not be completely/partially occluded by neighbouring atoms, and the probability of the atom in participating in interactions with other entities depends on its exposed area. A fully exposed atom is considered a definite surface entry; an occluded atom is categorised as internal with extremely low or almost no chances of contributing to binding activity. However an atom that is partially occluded has equal chances of being accepted or rejected. The condition for such an atom to be accepted is that the externally exposed area must be

sufficiently large for interaction with at least an external atom or in the event of probe experiments – large enough for the probe to produce contact with. According to [31] a surface entry "must not only be exposed at the van der Waals surface…but must also be exposed at the so-called SAS of the macromolecule".

B. *Pre-Processing*

Three sets of data were selected for this study. The first set consists of three FK506-bound proteins [PDB: 1YAT, 1BKF, 1FKF]. The second set contains experimental entries from [31] which are [PDB: 6CHA, 1RA2, 3FXN, 7TLN, 1TIM, 3RTA] and the third set are proteins from [30] namely [PDB: 2PLT, 1A19, 1Q3Y, 1QBS, 1EA1]. All the protein files are downloaded from the RCSB Protein Data Bank (http://www.rcsb.org) in PDB format. Each of the file contains crystallographically-determined information of the atoms including details such as spatial coordinates, atom type, residue chain, residue type etc. Fig. 1 shows a print screen of an excerpt from a PDB protein file.

7 data items are required in the implementation namely the x, y, z coordinates, atom element type, atom number, residue name, and the van der Waals radius for each atom, which is introduced separately. The VDW radius gives estimations of the sizes of the atoms, thus corresponding well to the generalised approach of voxels. Each protein file first undergoes pre-processing for extraction and compilation of the required information. The spatial coordinates of the atoms are translated to the all-positive domain and the values scaled up to facilitate image visualisations of the atoms.

III. Method

A. *Defining the Environment*

The protein is first projected into a 3D grid environment using the pre-processed information. The size of the environment is dependent on the protein but it must be in multiples of 4.0 as this value defines the smallest unit in the test space. The van der Waals radii for atoms range from about 1.0 Å to 2.0 Å (diameter ~2.0 Å to 4.0 Å) – with the exception of several elements – therefore a value of 4.0 Å was chosen as the value of the smallest unit voxel in the grid space. This ensures that each unit is sufficiently large to encapsulate most atoms.

To reduce the amount of processing time required, only the space enclosing the protein is targeted. This area is determined by locating the maximum and minimum coordinates of the protein for each of the axes and then identifying the upper and lower bound values which are numbers in multiples of 4.0. Take for example a protein with the following maximum/minimum values for (x, y, z) – (27.2, 45.8, 68.6)/(5.6, 8.9, 11.2). The smallest enclosing space about the protein is then determined by casting the maximum coordinates to the upper bound values and vice versa, therefore the coordinates for the bounding space are (28.0, 48.0, 72.0)/(4.0, 8.0, 8.0).

Prior to the selection of 4.0 Å as the voxel size, experiments have been carried out to determine the optimum value. Test proteins were analysed at voxel units of different sizes beginning with a unit size of 8.0 Å. It was found that proteins were coarsely represented at this level with some significant features left out. At a unit size of 4.0 Å, the structures became better defined with a good number of voxels capturing the shape of the proteins. Vital areas such as cavities and protruding regions were clearly shown compared to the use of a larger voxel size. To determine if a voxel size smaller than the van der Waals radii of some atoms is capable of returning results of higher precisions, a size of 2.0 Å was tested as well. Proteins were shown to be very well described but at the same time another problem was encountered. Small empty regions of low significance were picked up in the process. As most atoms are >2.0 Å in diameter, henceforth it is unlikely for small spaces of ~2.0 Å to contribute significantly to binding activity. A reasonable assumption can be made that that a voxel of 4.0 Å is sufficient to accommodate most atoms with respect to the radii range.

B. *Proteins and the Grid Space*

After the environmental parameters have been defined, the next stage is to identify units occupied by the protein in the experimental grid space. A 3-dimensional environment induces higher complexities and processing load compared to a 2-dimensional one, as such an approach was taken to reduce the dimensionality of the protein in the test space. Conceptually similar to the Z-buffer algorithm, one of the three axes was selected for a 'slicing' process, a method in which the 3-dimensional protein (described in terms of voxels) is converted into a series of 2-dimensional images. The number of images obtained is equivalent to the number of segments occupied by the protein in the selected axis. For example if the protein has atoms occupying units 24 Å to 52 Å in the selected axis, with the smallest unit space being 4 Å, the 'slicing' process then proceeds to generate an image for every 4 Å internal, thus giving a total of 8 images. Spatial data for the other two dimensions together with atomic information were used for visualisation for each of the output. More specifics of this method are documented in [32]. Fig. 2 shows sample output of different layers after the 'slicing' process has been applied.

ATOM	63	CB	LYS	A	3	66.750	29.725	6.408	1.00	20.42	C
ATOM	64	CG	LYS	A	3	67.656	29.230	7.510	1.00	21.60	C
ATOM	65	CD	LYS	A	3	68.223	27.854	7.207	1.00	22.32	C
ATOM	66	CE	LYS	A	3	69.119	27.347	8.319	0.01	22.13	C
ATOM	67	NZ	LYS	A	3	69.699	26.028	7.962	0.01	22.19	N
ATOM	68	N	ILE	A	4	65.984	32.679	5.545	1.00	20.69	N
ATOM	69	CA	ILE	A	4	65.858	33.533	4.385	1.00	21.50	C
ATOM	70	C	ILE	A	4	67.244	33.620	3.722	1.00	22.43	C
ATOM	71	O	ILE	A	4	68.209	33.831	4.452	1.00	22.94	O
ATOM	72	CB	ILE	A	4	65.330	34.969	4.663	1.00	20.49	C
ATOM	73	CG1	ILE	A	4	63.969	34.898	5.367	1.00	20.36	C
ATOM	74	CG2	ILE	A	4	65.223	35.784	3.360	1.00	19.99	C
ATOM	75	CD1	ILE	A	4	63.362	36.305	5.694	1.00	20.97	C
ATOM	76	N	ASP	A	5	67.251	33.383	2.445	1.00	23.57	N

Fig. 1. Print screen of an excerpt from a PDB file. The sample protein used is 1YAT. The columns are defined as (from left to right) the record name, atom number, atom name, residue name, residue chain, residue sequence number, x coordinate, y coordinate, z coordinate, atom occupancy, temperature factor and atom element.

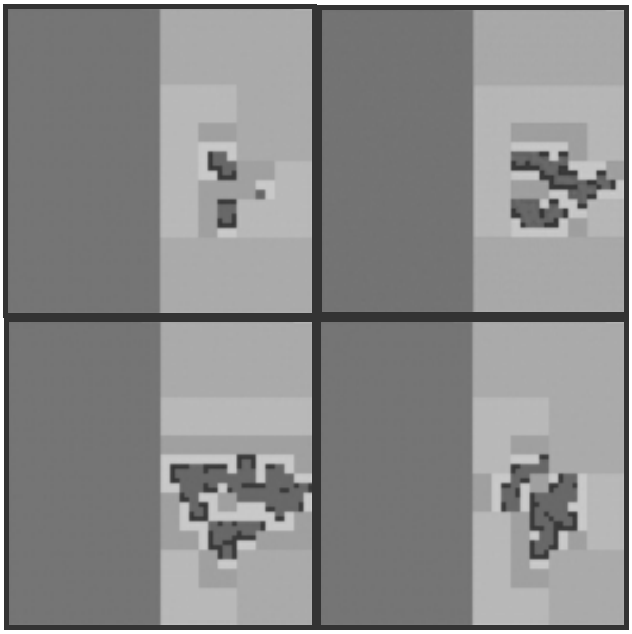

Fig. 2. Slicing of protein PDB300D with partitioning applied. Each image is a different layer of the protein. Re-merging these layers gives a reconstructed voxels representation of the protein.

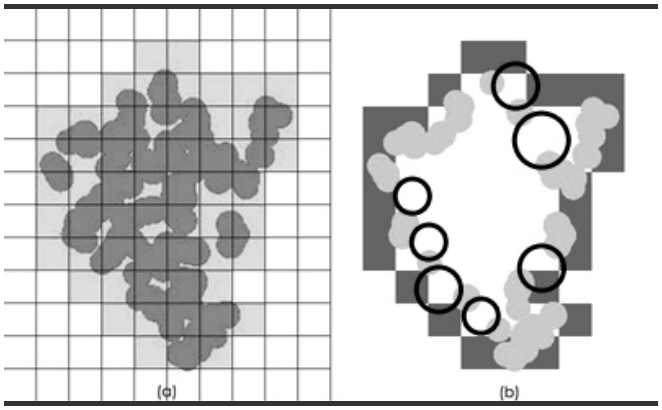

Fig. 3. Application of a >0% voxel occupancy to the protein. Voxels are selected even if a single pixel belonging to the protein is detected. (a) Layer 400 of protein 1FKF, with all atoms visible. (b) Layer 400 of protein 1FKF, with only surface voxels selected and surface atoms visible. The circled areas show the regions where surface atoms have not been selected. This lead to experiments carried out to determine the optimum value for the best coverage of atoms within the voxels.

TABLE I
PSEUDOCODE FOR THE BI-PARTITIONING ALGORITHM

```
PartitionSize = The Test Space Dimension

FOR 0 to Test Space Dimension
    FOR each PartitionSize

        Check if partition contains atom
            IF YES
                Leave partition unshaded
            ELSE
                Shade partition based on current partition color

        END OF CHECK
    END FOR

    Decrease PartitionSize by half

END FOR
```

The images in Fig. 2 depict as well the outcome after a bi-partitioning algorithm has been applied. Pseudocode for the algorithm is given in Table 1. The first few runs executed at coarser resolutions limit the experimental space to the region where the protein resides in. Further runs of the algorithm (until the smallest voxel has been attained) begin to indicate to the user the potential regions of the protein as well as providing clearer definitions of the shape.

Simple image processing techniques were applied to the images to determine if the voxels at each level of partitioning contain parts of the protein. An imprint of the protein is first created on a temporary image in a solid color. Then the pixels are checked on this temporary image according to the partitioned regions in the original image. If the unit is found to be void of any atom pixels then a colour or shade is assigned to the unit. By default a voxel is selected as long as a single pixel belonging to the protein is detected. This is not feasible as it leads to the selection of only a limited number of surface atoms (Fig. 3). The circled areas show regions wherein surface atoms have not been extracted. A condition for the selection of voxels containing the optimum number of atoms is required and is introduced in the form of a constraint termed the 'voxel occupancy'. This constraint is defined as a ratio of protein pixels to the total number of pixels within a voxel. A series of statistical studies were carried out to determine the optimum value and a percentage of 40%-100% was found to produce the highest coverage of atoms in a reasonable number of voxel units [33]. Fig. 4 shows images for different voxel occupancies applied to the units.

C. Selection of Surface Atoms and their Corresponding Atoms

Referring to Fig. 3 it can be observed that surface atoms are contained within the outermost voxels. The next step is to implement rules which automatically list out all surface voxels. Extracting these voxels which lie on the exterior of the cluster consequently leads to the selection of atoms on or close to the surfaces of proteins. The outline of the algorithm can be summarised as follows:

1. Pre-process the protein and compile the required information.
2. Project the protein into the experimental cubic grid space using the compiled data.
3. Select all voxels containing the presence of any of the atoms within the protein. This stage also includes checking for the occupancy percentage of each voxel (i.e. 40%-100%)
4. Filter all voxels having 1 or more faces fully exposed. These are categorised as the surface voxels.

Step 4 is the key process in the selection of surface voxels. In a cluster, internal voxels have all 6 faces fully connected to other voxels. However voxels on the surface or exposed to the external environment often have 1 or more non-connected faces. Cases wherein empty units occur within the protein with the neighbouring (internal) voxels labeled as surface entries suggest the possibility of an internal

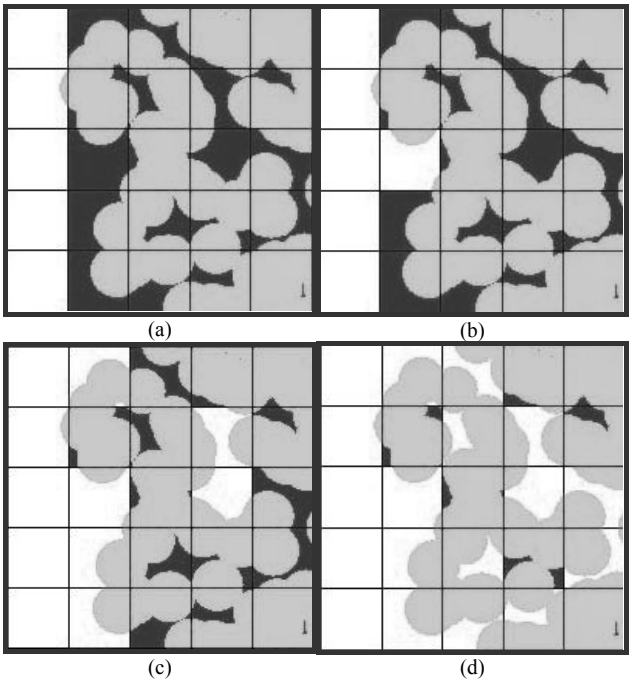

Fig. 4. The images show screenshots of the voxels applied with different voxel occupancies. (a) 5%-100% occupancy. (b) 25%-100% occupancy. (c) 60%-100% occupancy. (d) 85%-100% occupancy. At 5% occupancy the shape of the protein is not clearly defined while at 85% occupancy the protein has lost its structure. Therefore a conclusion can be made that the optimum voxel occupancy value has to be a value between 5% and 85%.

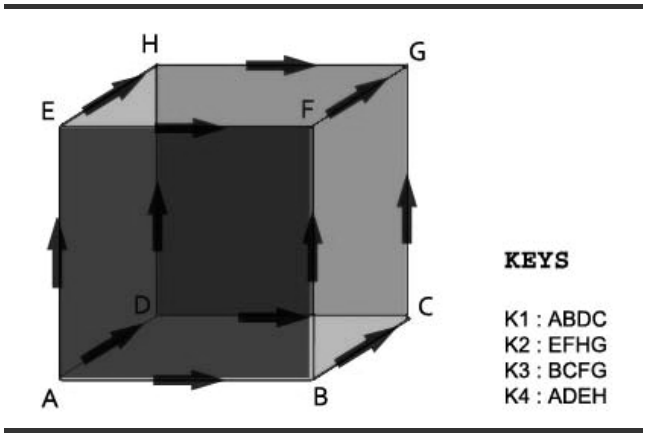

Fig. 5. A voxel with labeled points and parallel edges together with their corresponding keys.

binding area or part of a cavity extended inwards. Based on this observation, the points of each voxel are used to create vector sets which describe each of the 6 faces (Fig. 5). These vectors provide fixed definitions with emphasis placed on the order of the points as edges in one voxel are parallel to those of another. The four accepted faces depicted in Fig. 5 are ABDC, EFHG, BCFG and ADEH. Both the front and back faces are not used due to the 'slicing' process described earlier as the inclusion of these 2 faces invariably leads to the selection of all voxels both internal and on the surface.

The following step is to extract all the surface atoms from the protein. As the atoms are directly correlated with the voxels, the algorithm needs only iterate through the list of atoms for each layer and calculate the atoms contained within all associated surface atoms based on the coordinate points and the van der Waals radius of each atom. If an overlap is found with any of the identified surface voxels, the atom is selected. However it is also possible for internal atoms to occupy part of the surface voxels. Visual inspections may aid in the elimination and removal of these internal entries but it is infeasible to carry out such checking on a large set of test proteins. The alternative is to employ a fast and approximated method targeted at improving the extractions by removing unnecessary atoms. This procedure is described in the following sub-section.

D. Refinement of the Extracts

To obtain a refined list of surface atoms with minimal interference from internal entries, the extracts were subjected to a series of procedures based on well-defined rules. In a notion similar to voxel occupancy, the degree-of-belonging of an atom to a voxel is introduced. Defined as the percentage of an atom belonging to a voxel, a value of 5% and greater was found to be most effective in filtering out internal atoms. Again this value was selected based on experiments carried out – larger degree-of-belonging values were observed to remove surface entries as well. The degree-of-belonging value needs to be reasonably small to exclude atoms of low significance within the surface voxels.

However there still exist cases in which internal atoms possess large degree-of-belonging values resulting in these atoms bypassing the filtering stage. As such another method is devised to determine the acceptability of each atom within the surface voxels. Internal entries with less exposure to the external environment are consequently 'peeled' off. Secondary information obtained from the surface atoms are used in the 'peeling' process. Firstly the algorithm checks for the furthest atom within each voxel from the averaged center of the protein. All furthest atoms are marked in their respective voxel domains. Consecutive iterations then determine the exposure of each atom (within each voxel but excluding the marked furthest atom) to the external environment. The atoms are shortlisted if external exposure was found to be higher than internal exposure. The details of the implementation are given in [34].

IV. Results

A. Experiment Set I

Results from the implemented voxel-based method were visualised and compared against the output from the MSMS program [17] . Fig. 6 shows the differences in the output for various stages of the algorithm as well as the images obtained from the benchmark program for proteins 1YAT, 1BKF and 1FKF. The MSMS program employs a reduced surface method for both the probe implementations of SES and SAS. For comparison purposes and to ensure complete coverage of surface atoms obtained from both the probe methods, the results of both SES and SAS were merged together. Two layers were presented for each protein. The circled areas in the original extracts show regions where internal atoms are present. These atoms were shown to be excluded/removed in the output for the compared program as well as the implemented constraints-based filtering.

The other benchmark program used is Surface Racer [18] which returns output for SAS and molecular surface (MS)

areas. Similarly, both the results for SAS and MS were merged into a single list for each of the proteins. In Fig. 6 the first column shows cross-sectional images of two layers of proteins with a 40%-100% voxel occupancy applied. Only the surface voxels were selected alongside with the atoms contained or associated with these voxels. The images clearly depict a good coverage of surface atoms but also include a number of internal atoms in certain regions. In the second column the merged output from the MSMS program is visualised. A probe with radius of 2.0 Å was used to ensure consistency with the size of the voxels (diameter 4.0 Å) used in the experiment. Due to the reduced surface method some surface atoms have been excluded during investigation using the probe. Images in the third column show slight improvements compared to the output in the first column after application of the degree-of-belonging of 5%-100% to all atoms. Finally the last column shows further improvements to the voxel-based extracts through the execution of a 'peeling' algorithm which removes atoms with low exposure to the external environment.

Statistics for the extracts compared against MSMS and Surface Racer are given in Tables II and III, with manual extractions used as the benchmark results.

B. Experiment Set II

The second experiment consists of the 6 proteins from the study by Deanda and Pearlman [31] given as [PDB: 6CHA, 1RA2, 3FXN, 7TLN, 1TIM, 3RTA]. The results have not been replicated in this study due to notable differences detected in the total atoms count for 4 out of 6 proteins based on files obtained from the RCSB Protein Data Bank. The exact reason for the discrepancies in the data is unknown. For the experiment comparisons are only carried out between the implemented voxel method, the MSMS and Surface Racer programs as well as visual-based manual extractions. A probe radius of 1.4 Å (diameter 2.8 Å) was used for the programs, while the unit size for the voxel method was retained at 4.0 Å. The compared results are presented in Tables IV and V respectively.

C. Experiment Set III

The third experiment consists of 5 proteins collected from the study by Kim et al [30] namely [PDB: 2PLT, 1A19, 1O3Y, 1QB5, 1EAI]. All surface extracts are first compared between the implemented method, the chosen programs as well as manual extractions followed by a comparison to the results reported by the authors. Probe radius was maintained at 1.4 Å (diameter 2.8 Å) for the programs whereas the voxel size was kept at 4.0 Å. The results are presented in Tables VI, VII and VIII. Analysis of all tabulated data in Experiments I, II and III show that the voxel-based method returns higher numbers of potential surface atoms. Manual extractions were used as benchmarking datasets as the selections have been visually validated.

TABLE II
COMPARISON OF EXTRACTIONS FROM THE VOXEL-BASED METHOD AND THE MERGED OUTPUT OF MSMS PROGRAM AGAINST RESULTS OBTAINED FROM MANUAL EXTRACTIONS FOR EXPERIMENT I

Protein	T_A	Manual Extracts	Identified Surface Atoms		Matching Number of Atoms to ME*	
			Voxel	MSMS	Voxel	MSMS
1YAT	849	602	569	475	479	373
1BKF	827	587	514	489	448	366
1FKF	832	566	529	495	453	435

* **ME** – manual extracts
† T_A – total number of atoms in the protein. All following tables are similarly defined.
†† The identified surface atoms for voxel are based on the final count after the filtering and 'peeling' processes were applied. All following tables are similarly defined.

TABLE III
COMPARISON OF EXTRACTIONS FROM THE VOXEL-BASED METHOD AND THE MERGED OUTPUT OF SURFACE RACER (SR) PROGRAM AGAINST RESULTS OBTAINED FROM MANUAL EXTRACTIONS FOR EXPERIMENT I.

Protein	T_A	Manual Extracts	Identified Surface Atoms		Matching Number of Atoms to ME*	
			Voxel	SR	Voxel	SR
1YAT	849	602	569	485	479	434
1BKF	827	587	514	495	448	429
1FKF	832	566	529	498	453	433

TABLE IV
COMPARISON OF EXTRACTIONS FROM THE VOXEL-BASED METHOD AND THE MERGED OUTPUT OF MSMS PROGRAM AGAINST RESULTS OBTAINED FROM MANUAL EXTRACTIONS FOR EXPERIMENT II.

Protein	T_A	Manual Extracts	Identified Surface Atoms		Matching Number of Atoms to ME*	
			Voxel	MSMS	Voxel	MSMS
6CHA	3472	1892	1814	1519	1418	1062
1RA2	1268	925	881	677	766	499
3FXN	1073	644	623	584	502	495
7TLN	2432	1332	1302	1003	1026	768
1TIM	3740	2090	2015	1853	1549	1409
3TRA	1362	1206	1145	1116	1042	826

TABLE V
COMPARISON OF EXTRACTIONS FROM THE VOXEL-BASED METHOD AND THE MERGED OUTPUT OF SURFACE RACER (SR) PROGRAM AGAINST RESULTS OBTAINED FROM MANUAL EXTRACTIONS FOR EXPERIMENT II.

Protein	T_A	Manual Extracts	Identified Surface Atoms		Matching Number of Atoms to ME*	
			Voxel	SR	Voxel	SR
6CHA	3472	1892	1814	1842	1418	1379
1RA2	1268	925	881	730	766	673
3FXN	1073	644	623	584	502	498
7TLN	2432	1332	1302	1117	1026	953
1TIM	3740	2090	2015	1841	1549	1469
3TRA	1362	1206	1145	1052	1042	962

TABLE VI
COMPARISON OF EXTRACTIONS FROM THE VOXEL-BASED METHOD AND THE MERGED OUTPUT OF MSMS PROGRAM AGAINST RESULTS OBTAINED FROM MANUAL EXTRACTIONS FOR EXPERIMENT III.

Protein	T_A	Manual Extracts	Identified Surface Atoms		Matching Number of Atoms to ME*	
			Voxel	MSMS	Voxel	MSMS
2PLT	727	472	406	387	324	293
1A19	1438	978	911	774	759	660
1O3Y	2662	1637	1603	1277	1269	912
1QB5	3750	2038	2009	1595	1533	1143
1EAI	4540	2642	2854	2190	2175	1703

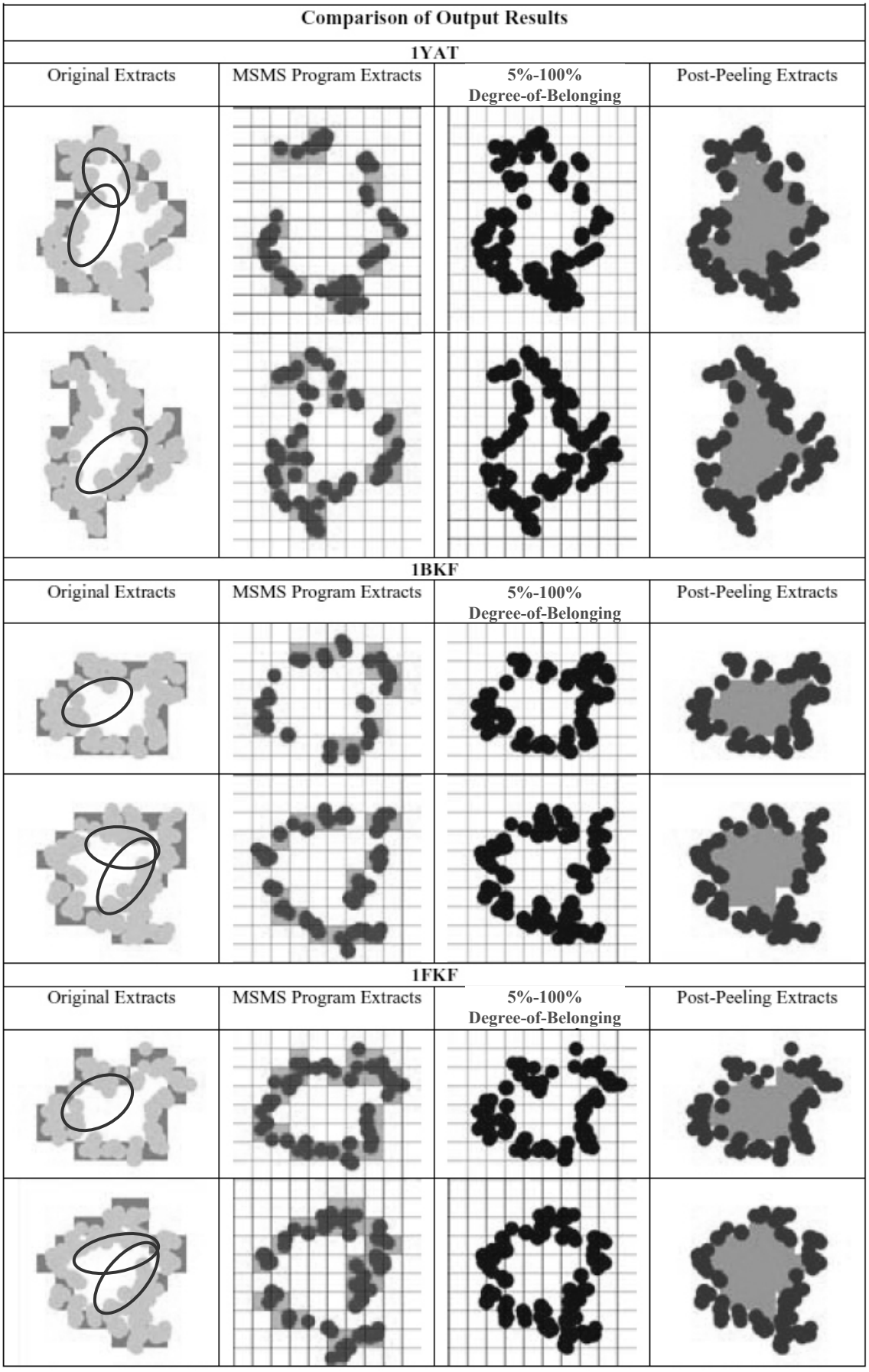

Fig. 6. The figure shows comparisons between the original extracts with voxel occupancy of 40%-100% against the results obtained from the MSMS program. Also presented are the results with constraints and the 'peeling' method applied. The probe radius for the MSMS program was set to 2.0 Å (diameter of 4.0 Å to ensure consistency in experimental conditions as the size of the voxels used was 4.0 Å). Based on the images, the original extracts showed the presence of a number of internal atoms as shown in the circled areas. Application of the degree-of-belonging constraint resulted in the exclusion of a good number of internal entries. Finally the internal 'peeling' algorithm eliminates entries of which were not filtered by the degree-of-belonging constraint.

TABLE VII
COMPARISON OF EXTRACTIONS FROM THE VOXEL-BASED METHOD AND THE MERGED OUTPUT OF SURFACE RACER (SR) PROGRAM AGAINST RESULTS OBTAINED FROM MANUAL EXTRACTIONS FOR EXPERIMENT III.

Protein	T_A	Manual Extracts	Identified Surface Atoms		Matching Number of Atoms to ME*	
			Voxel	MSMS	Voxel	MSMS
2PLT	727	472	406	373	324	338
1A19	1438	978	911	758	759	670
1O3Y	2662	1637	1603	1398	1269	1233
1QB5	3750	2038	2009	-	1533	-
1EAI	4540	2642	2854	2296	2175	1903

TABLE VIII
COMPARISON OF EXTRACTIONS FROM THE VOXEL-BASED METHOD AGAINST THE REPORTED OUTPUT BY THE AUTHORS FOR EXPERIMENT III.

Protein	T_A	Identified Surface Atoms	
		Reported	Voxel
2PLT	727	338	406
1A19	1438	681	911
1O3Y	2662	1261	1603
1QB5	3750	1482	2009
1EAI	4540	2151	2854

V. DISCUSSION

Fig. 6 shows how the constraint-based filtering of degree-of-belonging of an atom to a voxel was shown to improve the extractions by removing unnecessary internal atoms. There was a small number of internal entries remaining of which were eliminated through the 'peeling' algorithm. The remaining atoms were then classified as surface entries and shown to be competitive against the results from the MSMS program. To determine the efficiency of the method the post-peeling extracts were compared against a series of probe-based output with different probe radii applied from the MSMS program. Visualisations of the cross-sectional images are presented in Fig. 7 and Fig. 8. The program produces good results from a radius of 1.6 Å and below. However some internal atoms have been included as well as depicted in the circled regions. A comparison of the images showed that the voxel-based method performed considerably well at extracting boundary atoms.

Based on the tables the voxel implementation classified a higher number of atoms as surface entries in all cases compared to the MSMS and Surface Racer programs. It has to be noted that the higher count does not necessarily indicate all surface atoms have been identified correctly. In comparison to the manual extracts, the voxel-based method was shown to produce a higher number of matches in all cases except for protein 2PLT of Experiment III. Each method contains a number of unique entries not identified by the compared approach. A series of images was generated to show the common atoms generated between the implementation and the compared programs. The visualisations also illustrate the atoms uniquely identified by each of the methods and are shown as patterned atoms in Fig. 9. The entry 1QB5 in Table VII has been highlighted due to the failure of Surface Racer in processing the protein.

In Fig. 9, a total of 3 sample proteins from the experiments were chosen to show the differences in the extracted atoms for each of the methods. Two layers are picked for each of the proteins. The first and third columns show the cross-sectional visualisations of the voxel-based extracts. In the second column images for the MSMS program are presented whereas outputs from the Surface Racer program are given in the last column. All atoms shaded in dark gray are identified as the common atoms identified between the compared methods. The patterned atoms are entries unique to each of the approach. A conclusion can be made that the voxel-based method successfully identified a higher number of unique entries classified as surface atoms. The extracts were found to display high similarities to the output of Surface Racer program as opposed to the MSMS program.

The size of the voxel used leads to a larger number of atoms selected as each unit encapsulates a higher number of entries. Although the constraints-based filtering and 'peeling' method eliminated many of the internal entries, however there still remain a very minimal number of internal atoms. Removal of these entries may require the use of filters of higher complexities. Nevertheless the availability of these atoms could aid in the study of binding sites which consider both surface atoms and atoms located near to the boundary. Alternatively, these atoms are capable of contributing to the prediction of dock sites and have shown success in the identification of active regions [35].

Different parameters settings were used in the experiments to determine the optimum settings for this study. Voxels of both larger and smaller sizes were tested with the findings that smaller voxel sizes often lead to the inclusion of higher number of internal atoms and increases in execution time. On the other hand larger unit sizes do not return much information due to generalisation of the structures of the proteins. Usage of a higher or lower voxel occupancy value lead to varied sets of surface voxels being selected – the quality of extracted atoms was found to follow a Gaussian distribution curve wherein the best extractions (defined as the maximum number of atoms covered in the minimum number of voxels) corresponds to a 40%-100% voxel occupancy value [33].

VI. CONCLUSION

A voxel-based approach is employed in this study for the extraction of protein surface atoms. Constraints were introduced in the form of voxel occupancy and degree-of-belonging of atoms to voxels with the results of the experiments displaying improvements in the quality of the extractions – internal entries were eliminated through application of these rules. The output from the implemented method was compared against two other programs, namely MSMS and Surface Racer. The investigation revealed that the use of constraints-applied voxels is capable of achieving competitive results. In comparisons of output images between the method and the programs, the voxel-based method was shown to produce higher count of surface atoms and showed more similarities to those of Surface Racer. Common atoms of all methods were identified with the atoms unique to each method distinctly marked for comparisons. Again the voxel-based method displayed a higher number of unique entries. This study proved that with proper constraints applied voxels can be used as a tool for proteins analysis.

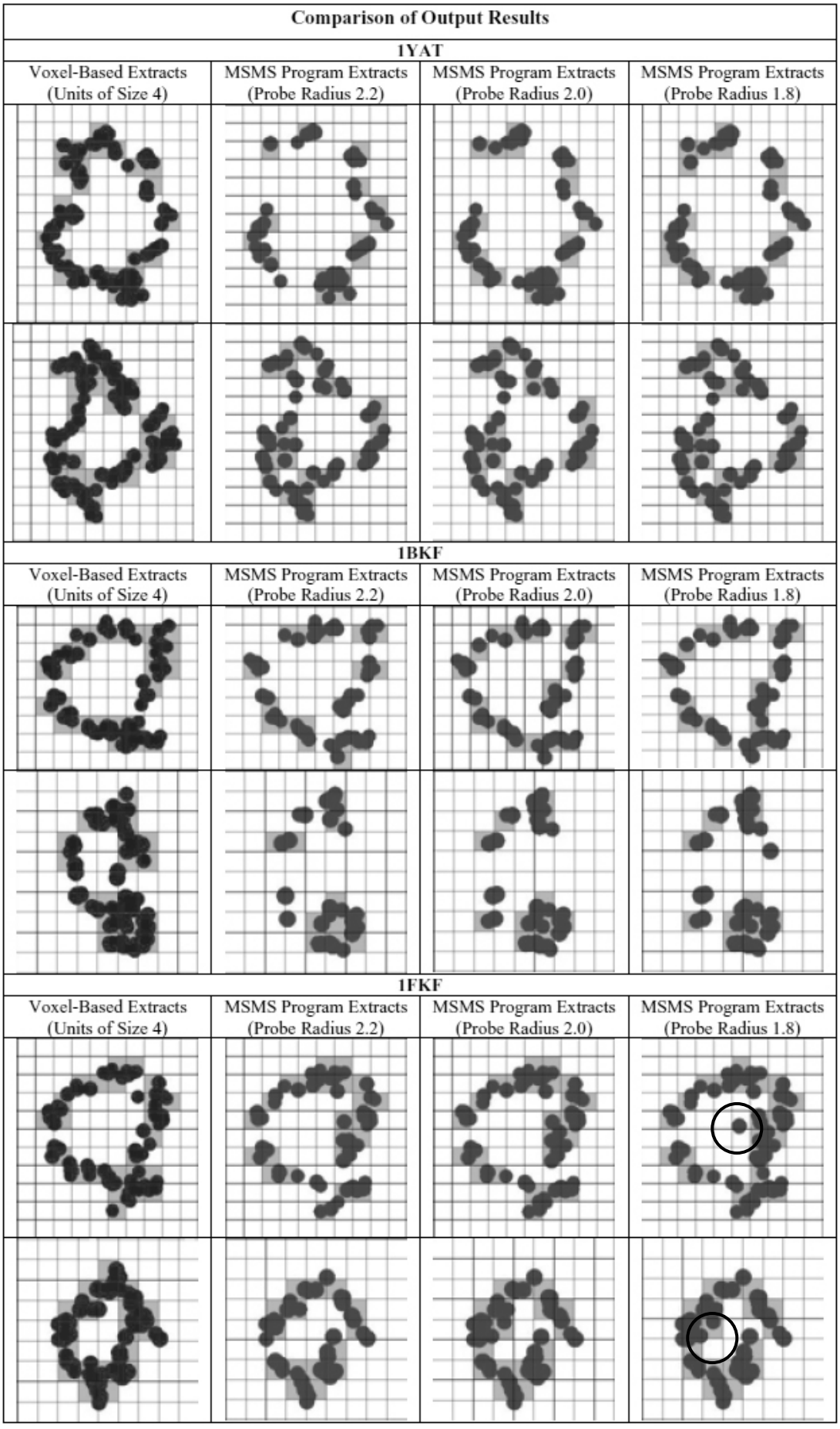

Fig. 7. Comparison between post-peeling extracts from the voxel-based method to the obtained range of output from application of different probe sizes in the MSMS program. The probe sizes in this image range from 1.8 Å to 2.2 Å. A larger probe radius results in comparably generalised outlines of the proteins. As the probe size becomes smaller, more surface atoms are detected. However the presence of internal atoms consequently increases as well (circled areas).

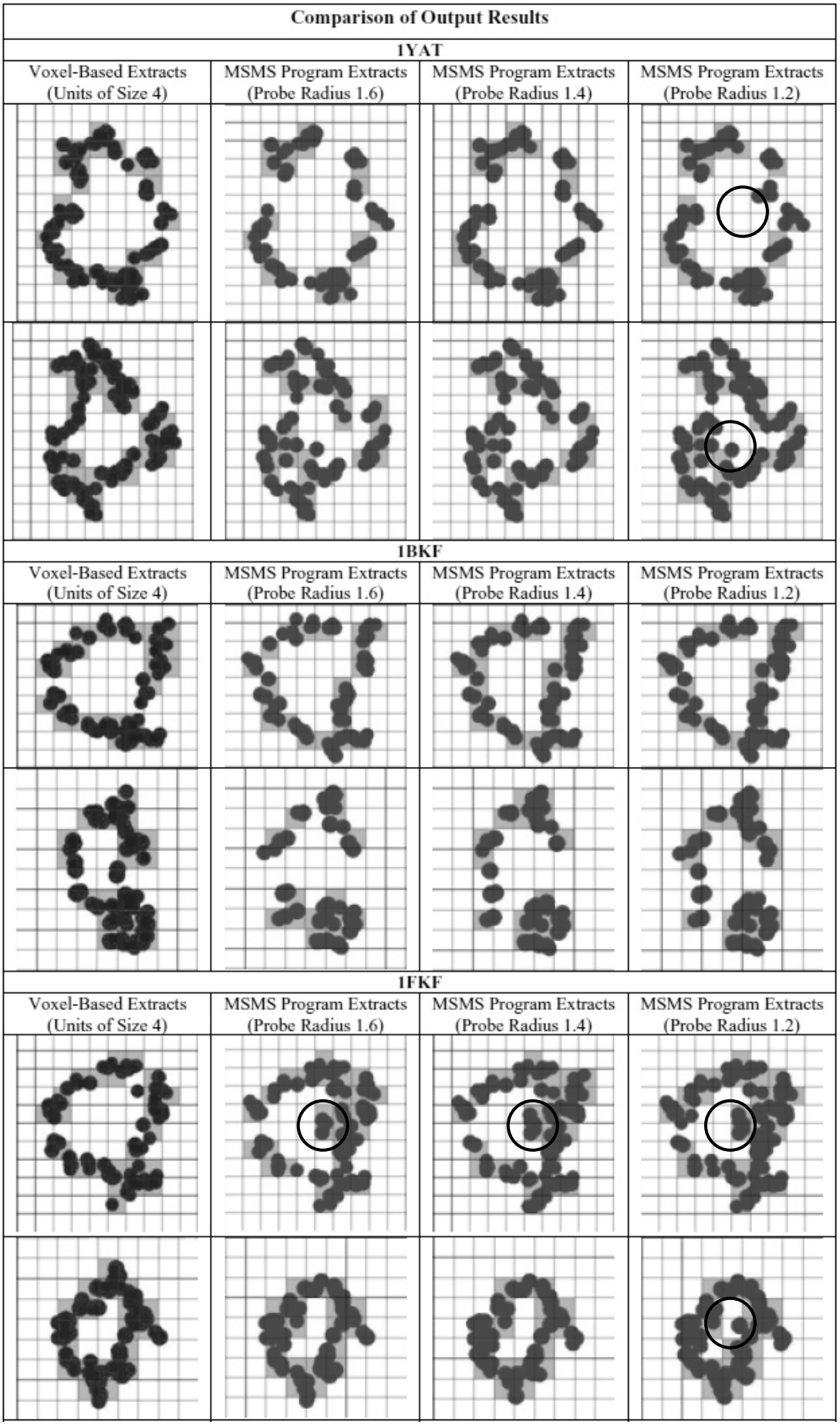

Fig. 8. Comparison between post-peeling extracts from the voxel-based method to the obtained range of output from application of different probe sizes in the MSMS program. The probe sizes in this image range from 1.2 Å to 1.6 Å. The usage of smaller probe radii leads to the presence of more internal atoms in the results (circled areas).

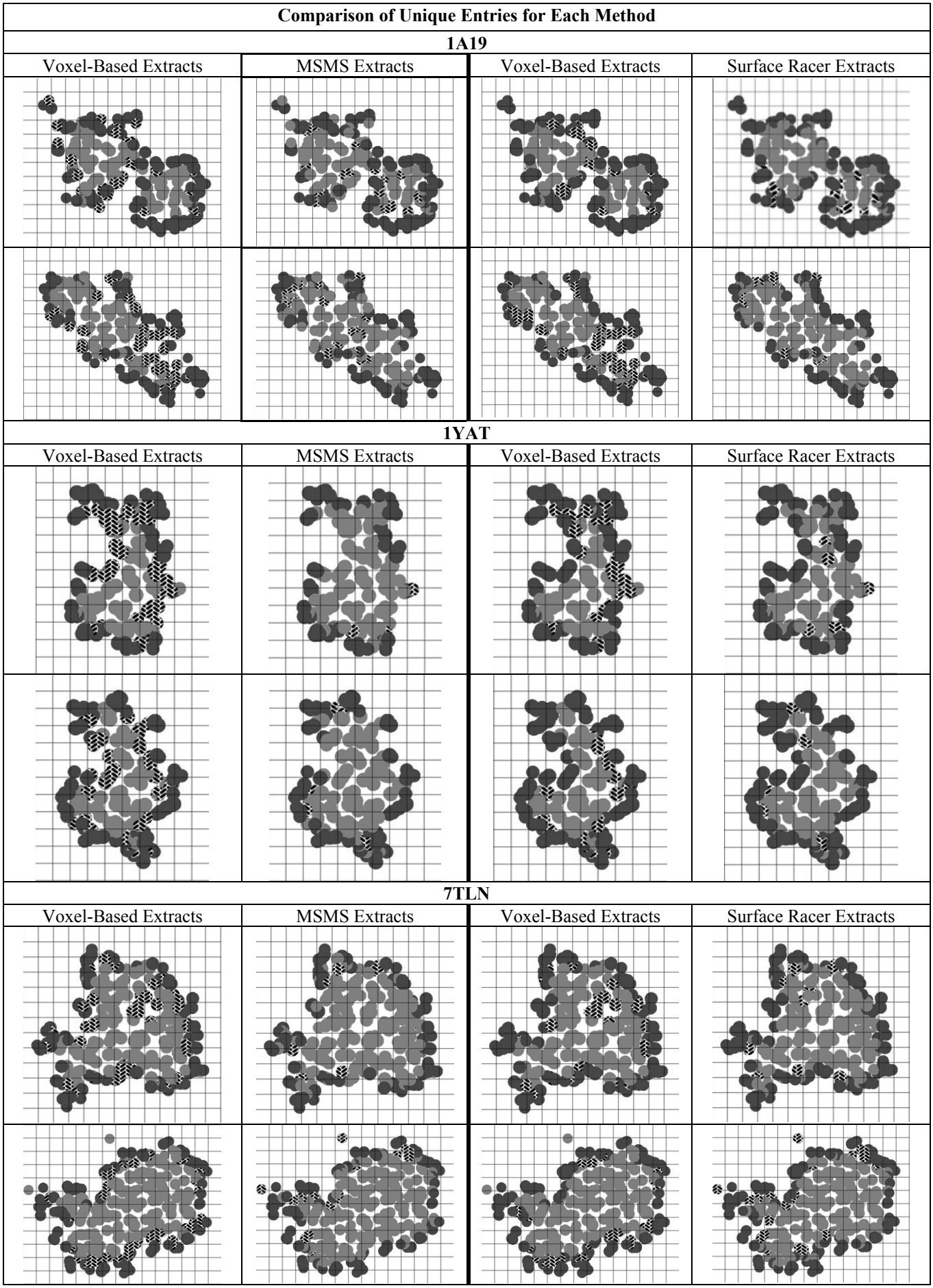

Fig. 9. Visualisations of the unique atoms found for each method for protein 1A19, 1YAT and 7TLN. The first two columns show comparisons between the voxel-based method and the MSMS program whereas the latter two columns show comparisons between the voxel-based method and the Surface Racer program. All identified common atoms are coloured in dark gray. Patterned atoms depict entries which are unique to the particular method and not found in the compared counterpart..

REFERENCES

[1] Ball, P. "Molecular Biology", *Abridged Version from Molecules: A Very Short Introduction*, Oxford University Press.

[2] Aftabuddin, M., Kundu, S. "Hydrophobic, Hydrophilic and Charged Amino Acid Networks within Protein", *Biophysical Journal*, Vol. 93, pp. 225 – 231, 2007.

[3] Smith, T.F., Waterman, M.S. "Identification of Common Molecular Subsequences", *J. Mol. Biol*. Vol. 147, pp. 195 – 197, 1981.

[4] Needleman, S.B., Wunsch, C.D. "A General Method Applicable to the Search for Similarities in the Amino Acid Sequence of Two Proteins", *J. Mol. Biol*. Vol. 48, pp. 443 – 453, 1970.

[5] Cates, S. "PSI-BLAST", *Connexions*, Retrieved from the Connexions Web Site: http://cnx.org/content/m11040/2.13/.

[6] Thompson, J.D., Higgins, D.G., Gibson, T.J. "CLUSTAL W: Improving the Sensitivity of Progressive Multiple Sequence Alignment through Sequence Weighting, Position-Specific Gap Penalties and Weight Matrix Choice", *Nucleic Acids Research*, Vol. 22, pp. 4673 – 4680, 1994.

[7] Andreeva, A., Howorth, D., Chandonia, J.M., Brenner, S.E., Hubbard, T.J.P., Chothia, C., Murzin, A.G. "Data Growth and its Impact on the SCOP Database: New Developments", *Nucleic Acids Research*, 36 (Database Issue), D419 – D425, 2007.

[8] Orengo, C.A., Michie, A.D., Jones, S., Jones, D.T., Swindells, M.B., Thornton, J.M. "CATH – A Hierarchic Classification of Protein Domain Structures", *Structure*, Vol. 5, pp. 1093 – 1108, 1997.

[9] Via A., Ferre F., Brannetti B., Helmer-Citterich M. "Protein Surface Similarities: A Survey of Methods to Describe and Compare Protein Surfaces." *Cell. Mol. Life Sci.* Vol. 57, pp. 1970 – 1977, 2000.

[10] Shoichet, B.K., Kuntz, I.D. "Protein Docking and Complementarity." *J. Mol. Biol*. Vol. 221, pp. 327 – 346, 1991.

[11] Brakoulias, A., Jackson, R.M. "Towards a Structural Classification of Phosphate Binding Sites in Protein-Nucleotide Complexes: An Automated All-Against-All Structural Comparison using Geometric Matching" *Proteins* Vol. 56, pp. 250 – 260, 2004.

[12] Shulman-Peleg, A., Nussinov, R., Wolfson, H.J. "Recognition of Functional Sites in Protein Structures." *J. Mol. Biol.* Vol. 339, pp. 607 – 633, 2004.

[13] Wang, H. "Grid-Search Molecular Accessible Surface Algorithm for Solving the Protein Docking Problem." *J. Comp. Chem*. Vol. 12, pp. 746 – 750, 2004.

[14] Connolly, M.L. "Solvent-Accessible Surfaces of Proteins and Nucleic Acids." *Science,* Vol. 221, pp. 709 – 713, 1983a.

[15] Connolly, M.L. "Analytical Molecular Surface Calculation." *J. Appl. Cryst*. Vol. 16, pp. 548 – 558, 1983.

[16] Lee, B., Richards, F.M. "The Interpretation of Protein Structures: Estimation of Static Accessibility." *J. Mol. Biol*. Vol. 55, pp. 379 – 400, 1971.

[17] Sanner, M. F., Olson, A. J. "REDUCED SURFACE: an Efficient Way to Compute Molecular Surfaces." *Biopolymers* Vol. 38, pp. 305 – 320, 1996.

[18] Tsodikov, O.V., Record, M.T.Jr, Sergeev, Y.V. "A Novel Computer Program for Fast Exact Calculation of Accessible and Molecular Surface Areas and Average Surface Curvature." *J. Comput. Chem*. Vol. 23, pp. 600 – 609, 2002.

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

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

[21] Jiang, F., Kim, S.H. "'Soft Docking': Matching of Molecular Surface Cubes", *J. Mol. Biol.*, Vol. 219, pp. 79 – 102, 1991.

[22] Jones, S., Thornton, J.M. "Analysis of Protein-Protein Interaction Sites using Surface Patches", *J. Mol. Biol.*, Vol. 272, pp. 121 – 132, 1997a.

[23] Jones, S., Thornton, J.M. "Prediction of Protein-Protein Interaction Sites using Patch Analysis", *J. Mol. Biol.*, Vol. 272, pp. 133 – 143, 1997b.

[24] Tietze, S., Apostolakis, J. "GlamDock: Development and Validation of a New Docking Tool on Several Thousand Protein-Ligand Complexes", *J. Chem, Inf. Model*, Vol. 47, pp. 1657 – 1672, 2007.

[25] Taufer, M., Crowley, M., Price, D., Chien, A.A., Brooks III, C.L. "Study of a Highly Accurate and Fast Protein-Ligand Docking Algorithm based on Molecular Dynamics", *Proceedings of the 18th International Parallel and Distributed Processing Symposium*, pp. 188 – 195, 2004.

[26] Carlson, H.A., McCammon, J.A. "Accommodating Protein Flexibility in Computational Drug Design", *Molecular Pharmacology*, Vol. 57, pp. 213 – 218, 2000.

[27] De Jonge, M.R., Vinkers, H.M., Van Lenthe, J.H., Daeyaert, F., Bush, I.J., Van Dam, H.J.J., Sherwood, P., Guest, M.F. "Ab Initio Potential Grid Based Docking: From High Performance Computing to In Silico Screening", *COMPLIFE*, Vol. 940, pp. 168 – 178, 2007.

[28] Fidanova, S. "An Improvement of the Grid-Based Hydrophobic-Hydrophilic Model", *International Journal of Bioautomation*, Vol. 14, pp. 147 – 156, 2010.

[29] Bajaj, C., Siddavanahalli, V. "An Adaptive Grid Based Method for Computing Molecular Surfaces and Properties", *ICES Technical Report*, TR-06-57, 2006.

[30] Kim, B., Kim, K-J., Choi, J-H., Baek, N., Seong, J-K., Choi, Y-J. "Finding Surface Atoms of a Protein Molecule on a GPU." *Proceedings of the SIGGRAPH Asia 2011 Posters (SA' 11)*, 2011.

[31] Deanda, F., Pearlman, R.S. "A Novel Approach for Identifying the Surface Atoms of Macro-molecules." *J. Mol. Graph. Model*. Vol. 20, pp. 415 – 425, 2002.

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

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

[34] Lee, L.W., Bargiela, A. "An Investigation on the Use of Constraints-Applied Voxels in the Search for Protein Surface Atoms", *Lecture Notes in Engineering and Computer Science: Proceedings of the World Congress on Engineering 2012*, WCE 2012, 4-6 July, 2012, London, UK, pp. 655 – 661.

[35] Lee L.W., Bargiela A. "Prediction and Modelling of Ligand Binding Sites using An Integrated Voxel Method." *Proceedings of the 26th European Conference on Modelling and Simulation (ECMS 2012)*, pp. 96 – 102, 2012.