ELSEVIER

Journal of Molecular Graphics and Modelling 20 (2002) 415–425

Journal of Molecular Graphics and Modelling

www.elsevier.com/locate/JMGM

A novel approach for identifying the surface atoms of macromolecules

Felix Deanda*, Robert S. Pearlman

Laboratory for the Development of Computer-Assisted Drug Discovery Software, College of Pharmacy, University of Texas, Austin, TX 78712, USA

Accepted 10 December 2001

Abstract

A significant number of atoms lie buried beneath the "molecular surface" of proteins and other biologic macromolecules. Interactions between ligands and these macromolecules are dominated by interactions with the "surface atoms". Although interactions with the "buried" or interior atoms of the macromolecule certainly contribute to the total intermolecular interaction energy, many computer-assisted drug design (CADD) strategies can benefit from the identification of those atoms "on the surface" of proteins and other macromolecules.

We have developed a simple, yet novel method to distinguish the surface atoms of macromolecules from the interior atoms which is based on computing the atomic contributions to the solvent-accessible surface (SAS) area. This report describes that method and demonstrates that it compares very favorably with four alternative methods.

Keywords: Macromolecular surface; Surface atoms; Interior atoms; Solvent-accessible surface area

1. Introduction

Various computer-assisted drug design (CADD) tasks require or are facilitated by distinctions between "interior atoms" and "surface atoms" of proteins or other biologic macromolecules. Perhaps the most obvious example is the computer-graphic visualization and real-time manipulation of macromolecular surfaces. Knowing which atoms are on the surface of a macromolecule eliminates the need to update positions or consider rendering surfaces of atoms that do not meaningfully contribute to the macromolecular surface [1,2].

Since molecules interact at their surfaces, an understanding of molecular surface characteristics can be extremely useful for studying their interactions. This is obviously true for the "lock-and-key" scenario of receptor–ligand binding. However, before one can analyze receptor–ligand interactions, the binding site must first be located. Computational methods have been developed which explore macromolecular surfaces in search of potential ligand binding sites. Typically, these algorithms focus on "surface atoms" but the identification of those surface atoms is usually ill-defined [3–10].

Investigators have used molecular dot surfaces to represent the surfaces of macromolecules in efforts to locate potential ligand binding sites. In order to keep the number of surface dots at a manageable level, the dot density is typically reduced to just a single dot per atom [4–10,21]. In theory, such low-density dot surfaces would enable distinction between surface atoms and interior atoms. In practice, however, we will illustrate that such algorithms often yield incorrect results.

After locating putative binding sites, the next step in addressing the "docking problem" is to estimate binding constants by estimating various components of the free energy of interaction. Estimates of the free energy required to desolvate the binding site regions of macromolecules are greatly facilitated by distinctions between "surface atoms" and "interior atoms". In addition, although interior atoms certainly contribute to the total electronic interaction energy between ligands and macromolecules, estimates of interaction energies computed by considering only the surface atoms can be extremely useful for high-throughput strategies in computer-assisted molecular design [11–20].

In this paper, we present a simple yet novel approach which distinguishes the surface atoms of macromolecules from the interior atoms. Our approach for identifying surface atoms is based on atomic contributions to the solvent-accessible surface (SAS) area of macromolecules and, hence, is referred to as the SAS approach. In addition to describing this approach, four alternative methods are

* Corresponding author. Present address: Computational, Analytical and Structural Sciences, GlaxoSmithKline, Five Moore Drive, Research Triangle Park, NC 27709, USA. Tel.: +1-919-483-9482; fax: +1-919-315-0430.
E-mail address: fd69145@glaxowellcome.com (F. Deanda).

1093-3263/02/$ – see front matter
PII: S1093-3263(01)00142-5

also described and compared (both in terms of speed and accuracy) with the SAS approach.

2. Definitions of molecular surfaces

To determine which atoms lie on the surface of macromolecules, we must first choose a suitable representation of the molecular surface from which to develop our definition of "surface atom". Despite the fact that the electron density surrounding a molecule has no well-defined boundary surface, molecular surface models have found widespread use in the field of molecular modeling. Not surprisingly, one of their most common applications has been in modeling receptor–ligand interactions [1–4,22–27]. Three types of molecular surface models have been described in the literature. They include the van der Waals surface, Hermann's SAS [28], and Richards' contact/re-entrant surface [29].

Illustrated in Fig. 1A is a cross-section of the familiar van der Waals surface. In constructing this molecular surface model (as well as the other two molecular surface models mentioned), each atom in the molecule is represented as a hard sphere, whose radius is equal to the van der Waals radius of the atom, r_{atm}. Given this spherical representation of the atoms, the van der Waals surface can be defined as the union of all portions of all atomic sphere surfaces not occluded by neighboring atomic spheres. Although the van der Waals surface is a reasonable approximation of the molecular surface, it does not address the issue of whether or not an atom is accessible to the solvent environment. For this, a different representation of the molecular surface was needed.

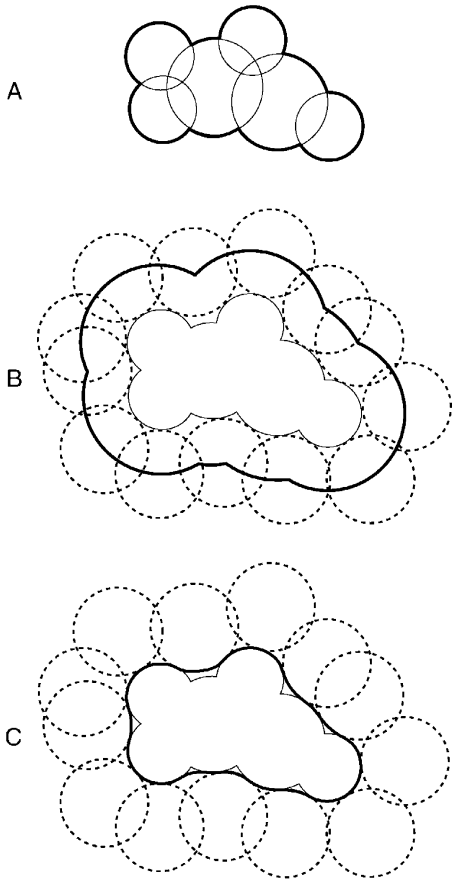

Fig. 1. Cross-sections of the (A) van der Waals surface, (B) SAS and (C) contact/re-entrant surface.

Hermann [28] was the first to point out that, depending on the size of the solvent molecule, atoms located in narrow crevices may not come into van der Waals contact with the solvent. In light of this observation, he proposed a new description of the molecular surface, which is now commonly referred to as the solvent accessible surface. A cross-section of the SAS is illustrated in Fig. 1B. The SAS can be described as the molecular surface created by the center of a probe (or solvent) sphere, when the probe is rolled over the entire van der Waals surface of a molecule. This is equivalent to a van der Waals surface in which the atomic radii have been extended by the probe radius. Generally, the probe sphere is assigned a radius, r_{probe}, that best approximates the dimensions of either the entire solvent or the atom or group of atoms on the solvent most likely to make van der Waals contact with the molecular surface. As an example, a radius of 1.50 Å is typically used to represent the effective radius of a water molecule [30,31].

Richards [29] has proposed a third description of the molecular surface, illustrated in Fig. 1C, which also takes into account the solvent accessibility of atoms in a molecule. This molecular surface is defined as that surface traced out by the inward-facing surface of a probe sphere. When the probe is in contact with a single atom, this is equivalent to the outward-facing van der Waals surface of that atom. Richards referred to this portion of the molecular surface as the "contact surface". The other part of the molecular surface corresponds to the inward-facing surface of the probe sphere when the probe is simultaneously in contact with two or more solute atoms. Richards referred to this portion of the molecular surface as the "re-entrant surface". The molecular surface, thus, resembles the van der Waals surface of the atoms, except that crevices too small to accommodate the probe sphere are eliminated and clefts between atoms are smoothed over [1,21,29–31,34].

3. Definition of surface atom

Based on the definition of the van der Waals surface, we can argue that, if an atom is not completely occluded by its neighbor atoms, that atom must be partially exposed at the surface of a macromolecule and, hence, define that atom as a surface atom. However, the question arises as to whether or not that atom would be accessible for interaction with a probe sphere representing some other atom in either a solvent molecule or ligand molecule. Depending on the size of the probe sphere, atoms located in narrow crevices may be inaccessible to solvent or ligands. Thus, a "true

surface atom" must not only be exposed at the van der Waals surface of the macromolecule but must also be exposed at the so-called SAS of the macromolecule.

These true surface atoms could also be identified as those which contribute to the contact portion of Richards' contact/re-entrant surface. However, computing the contact portion of that surface offers no advantage over computing the SAS for which much faster, equally accurate algorithms have been developed [28–40]. Amongst these, the algorithm developed by Pearlman [30] as implemented within the SAVOL3 (surface area and volume) package [41] is particularly advantageous for use in macromolecular contexts as will be indicated in Section 4.

An atom will be classified as a "true surface atom" if its SAS area is greater than zero ($SA_{atm}^{acc} > 0\,Å^2$). However, the following important question arises: should we classify atoms with extremely small SAS areas as surface atoms? The answer to this question depends to some extent upon the precision with which the atomic contributions to the SAS area are computed and upon the intended application of the surface/interior distinction. Accordingly, an atom will be classified as an "effective surface atom" if its SAS area is greater than a user-specified minimum threshold value for the atomic SAS area, SA_{min}^{acc}.

4. Computational methods for identifying surface atoms

4.1. Solvent-accessible surface (SAS) approach

SAVOL3 [41] is a widely distributed program for calculating molecular surface area, molecular volume, and appropriately partitioned atomic contributions thereto. SAVOL3 can perform these calculations using either the analytic SAVOL2 algorithm [42] or the SAVOL1 algorithm [30] based on numerical integration. For calculations on small molecules, the analytic algorithm is much faster and exact. For macromolecular structures, the SAVOL1 algorithm is advantageous due, primarily, to more efficient handling of multiple overlaps (which occur frequently in large structures but less frequently in small structures). Briefly, the SAVOL1 algorithm computes the molecular (or solvent-accessible) surface area by summing the non-occluded surface area of each atom in the molecule. A slicing-plane is rotated incrementally about an axis of each atomic sphere, thereby cutting the sphere into many double-lunar segments (imagine, for example, two slicing-planes intersecting at the center of a sphere at an angle of one degree from each other; the resulting small, pair-wise spherical segments are what is referred to as a double-lunar segment). Clearly, the precision of the surface area calculation is determined by the angle increment used for rotating the slicing-plane. The non-occluded surface area of the atom is obtained by summing the non-occluded surface area of all double-lunar segments. The non-occluded surface area of each double-lunar segment is computed by analytically determining which portions of the segment are not contained within the van der Waals sphere of a neighboring atom [30–33].

The author-recommended angle increment of 9° was used in our SAVOL1 calculations. This provides atomic surface area values that are accurate to approximately $\pm 0.003\,Å^2$, which is quite sufficient for our purposes. SAVOL3 offers several choices of sets of atomic (van der Waals) radii. The radii proposed by Bondi [43] were used in this study. The probe radius (solvent radius) was assigned the value of 1.50 Å, which is the value commonly used to represent the effective radius of a water molecule.

Recall our definition of an effective surface atom as one for which the SAS exceeds the threshold value, SA_{min}^{acc}. Indication of a value for SA_{min}^{acc} would complete our specification of the SAS approach. Given that a slicing-plane angle increment of 9° yields atomic surface areas accurate to approximately $\pm 0.003\,Å^2$, the value for SA_{min}^{acc} should be at least as large. However, as will be indicated in Section 5, somewhat larger values provide greater efficiency with no loss of accuracy.

4.2. Number of intersecting neighbors (NIN) approach

One alternative approach considered for the identification of surface atoms is based on the number of intersecting neighbors (N_{int}) for each atom in the macromolecule. Atoms i and j are intersecting neighbors if and only if $r_{ij} < R_i + R_j$, where r_{ij} is the distance between the centers of atomic spheres i and j, and R_i and R_j are the atomic radii (including the probe radius).

The basic concept behind this simple approach is as follows. Atoms which lie in the interior of a macromolecule are surrounded on all sides by other atoms. Obviously, this is not the case for atoms which lie on the surface of the macromolecule. Thus, one might expect that the N_{int} of interior atoms would be greater than the number of neighbors of surface atoms and, thus, one might expect that N_{int} could be used as the basis for a classification algorithm.

4.3. Sum of vectors (SOV) approach

This second alternative approach might be regarded as an extension of the NIN approach described above. Rather than simply considering the number of neighbors of a given atom, this approach considers the sum of vectors (SOV) extending from the given atom to all of its neighbor atoms.

The basic concept behind this second approach is that if an atom lies within the interior of a macromolecule, then, by symmetry, one might expect that the norm of the sum of vectors to its neighbors would be small, i.e. $||\boldsymbol{V}_{sum}|| \cong 0$. In contrast, since surface atoms have neighbors towards the interior but none in the opposite direction, one might expect that the norm of the sum of vectors to its neighbors should be substantially greater than zero ($||\boldsymbol{V}_{sum}|| \gg 0$) and

that $||\boldsymbol{V}_{sum}||$ could be used as the basis for a classification algorithm.

4.4. UCSF approach

The third alternative approach considered for the identification of surface atoms is based on an algorithm developed at the University of California at San Francisco by Bash et al. [2]. The UCSF approach was developed to enable removal of interior atoms from computer-graphic displays of macromolecular structures in order to present a simplified (improved) image and to reduce the CPU-time required to display and manipulate that image.

The UCSF algorithm can be summarized as follows. First, a 3D lattice is constructed to contain the macromolecule. Within this lattice, the size of an atom is represented by a cube of 27 ($3 \times 3 \times 3$) lattice points with the center of the cube corresponding to (or nearest to) the center of the atom. For each lattice point, atom numbers (indices) of all atoms represented by the given lattice point are recorded in an atom-list. Once this is done for all lattice points, "surface points" are then identified as those with at least one adjacent lattice point having a null (empty) atom-list. Finally, the surface atoms of the macromolecule are identified as those atoms contained in the atom-list of at least one surface point.

The lattice spacing used by Bash et al. was 1.60 Å. Details as to how they chose this value were not provided in the journal article. The authors simply stated that it was the value "typically chosen" in their work [2]. To determine whether this value was appropriate for our objective, we considered a reasonable range of lattice spacings.

4.5. Molecular dot surface (MDS) approach

In Section 1, we discussed the use of molecular dot surfaces for locating potential ligand binding sites. As a fourth alternative approach, we used the molecular dot service to identify atoms that are on the surface of macromolecules. Each dot on the molecular surface approximately corresponds to a specific amount of surface area equal to the inverse of the user-specified dot density (expressed as dots/Å^2). If each dot on the molecular surface is associated with a particular atom, the number of dots per atom provides a rough approximation of the atom's contribution to the SAS area and, hence, might be used in the same way that SAVOL1-derived surface areas were used in the SAS approach to identifying surface atoms. Alternatively, by specifying a dot density consistent with the value which would have been chosen for SA_{min}^{acc}, any atom associated with one or more dots could be classified as an effective surface atom.

The accuracy of this MDS approach is determined not only by the dot density but also by the uniformity with which dots are distributed on the molecular surface. Most dot surface programs were developed for the sole purpose of computer-graphic display. Consequently, those dot surface algorithms suffer two flaws which have little or no effect on the quality of computer-graphic displays but which render the algorithms completely unsatisfactory for quantitative applications such as the MDS approach proposed above.

1. The dots on a given atomic sphere are distributed in a non-uniform fashion.
2. The dot density varies (sometimes quite substantially) between atomic spheres of different radii.

In contrast with all other dot surface algorithms, those developed by Brusniak and co-workers [44,45] were specifically developed to enable quantitative applications. More specifically, their Quick Dot Surface (QDS) algorithm not only avoids both deficiencies mentioned above but also is much faster than other methods (e.g. [21,22]). For these reasons, the QDS algorithm was used for this work.

5. Results and discussion

As will be indicated below, the SAS approach is clearly the best of the five methods described in Section 4. Therefore, in this section, we first present results from the SAS approach and then assess the performance of the other methods relative to the performance of the SAS method. We compiled a small but representative set of biologic macromolecules to serve as the basis for these comparisons. The 3D structures were obtained from the Brookhaven Protein Data Bank (PDB) [46,47] and the PDB codes are listed in Table 2.

All methods were developed and tested on an SGI Indigo workstation with an R4400 processor. All methods would run substantially faster on the faster processors now available, but the relationships between the speeds of the various methods would remain essentially the same.

5.1. Solvent-accessible surface (SAS) approach

The first step in assessing the SAS approach was to investigate how the choice of SA_{min}^{acc} affects the results and to determine a suitable choice of SA_{min}^{acc} for routine applications. For this purpose, we performed a series of calculations on the serine protease, α-chymotrypsin (6CHA). The atomic contributions to the SAS area (SA_{atm}^{acc}) of 6CHA were calculated for all atoms using the SAVOL1 algorithm of the SAVOL3 program. Fig. 2 provides a summary analysis of the frequency distribution for the set of SA_{atm}^{acc} values.

From Fig. 2, the following simple observations can be made. First, since any atom for which $SA_{atm}^{acc} > 0\,Å^2$ is a true surface atom, all atoms in the second and higher columns of the histogram certainly lie on the surface of the macromolecule. Secondly, since atoms for which $SA_{atm}^{acc} = 0\,Å^2$ lie in the interior of the macromolecule, all true interior atoms are contained in the first column of the histogram. However, the classification of atoms whose SA_{atm}^{acc} values fall within the interval $0\,Å^2 < SA_{atm}^{acc} \leq 2.50\,Å^2$ remains to be addressed. In other words, where within this interval should we

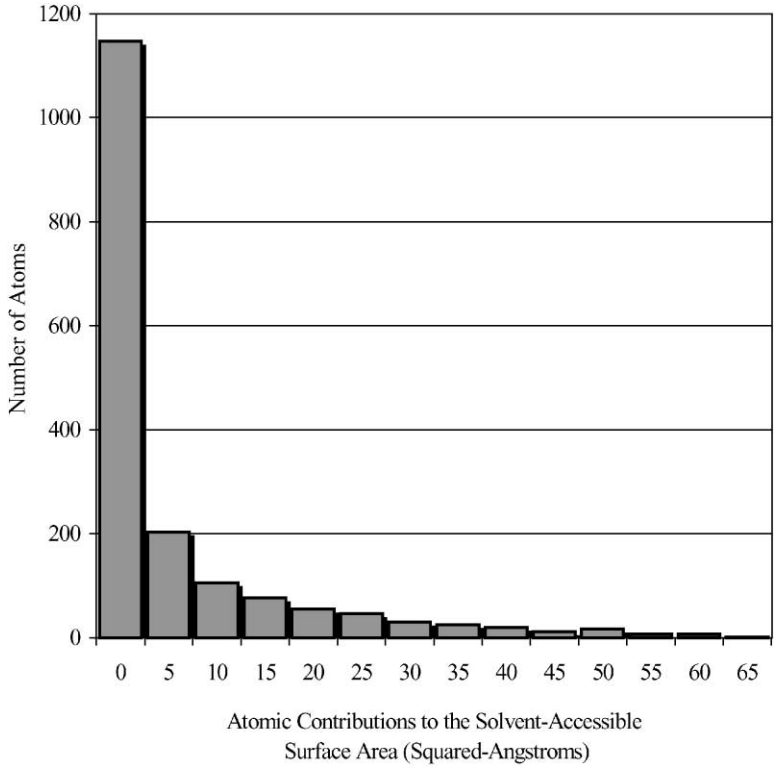

Fig. 2. The frequency distribution of atomic contributions (SA^{acc}_{atm}, Å^2) to the SAS area of 6CHA. Each column represents a range of SA^{acc}_{atm} values. Second column is a typical example. It spans the range $2.50\,Å^2 < SA^{acc}_{atm} \leq 7.50\,Å^2$ of which the column label, $5\,Å^2$, is the midpoint. The only exception is the first column, which spans the range $0\,Å^2 \leq SA^{acc}_{atm} \leq 2.50\,Å^2$.

define the threshold between effective surface atoms and effective interior atoms? Clearly, the histogram in Fig. 2 does not enable specification of the exact value for SA^{acc}_{min}.

To address the question above, we considered a number of potential values for SA^{acc}_{min} spanning the range between 0 and 2.50 Å^2. Table 1 indicates these values as well as the corresponding cumulative percent of the total SAS area (TSA) and number of atoms which would be classified as surface atoms using that value for SA^{acc}_{min}. Note that as we decrease the value of SA^{acc}_{min}, we account for a larger portion of the total molecular surface area but, naturally, we also classify a larger number of atoms as surface atoms. If we choose 1 Å^2 as the value for SA^{acc}_{min}, we can account for approximately 99.3% of the total molecular surface area with only 41.4% of the total number of atoms in 6CHA. If we want to account for 100% of the molecular surface area, we can choose either 0.005 or 0.01 Å^2 as the value for SA^{acc}_{min}. Note that both values are well within the level of precision of our computations (recall that the SAVOL1 program provides atomic values accurate to approximately ± 0.003 Å^2 when using the author-recommended angle increment of 9°). For purposes demanding 100% accuracy, we would choose 0.01 Å^2 as the value for SA^{acc}_{min}. However, since accounting for 99.3% of the surface should be quite sufficient for most purposes and since accounting for the last 0.7% of surface required classifying an additional 11.8% of atoms as surface atoms, we can argue that the value of SA^{acc}_{min} should be set equal to 1 Å^2.

Table 1
Potential values for SA^{acc}_{min} [a]

SA^{acc}_{min} (Å^2)	Cumulative % TSA	Cumulative % atoms
2.500	97.430	34.644
1.000	99.266	41.368
0.500	99.715	45.242
0.250	99.902	48.148
0.100	99.985	51.111
0.050	99.993	51.795
0.025	99.998	52.650
0.010	100.00	53.162
0.005	100.00	53.219

[a] Results are for 6CHA.

The SAS approach was also applied to several other biologic macromolecules as indicated in Table 2. Note that a value of 1 Å^2 for SA^{acc}_{min} is sufficient to account for at least 98.8% of the total molecular surface area. Interestingly, for most cases, less than 50% of the total number of atoms was needed to account for ~99% of the total molecular surface. The only exception was tRNA (3TRA) for which (due to its relatively smaller size and non-globular shape) almost 58% of its total number of atoms account for approximately 99.3% of its total molecular surface. Nevertheless, the results in Table 2 certainly provide further support for our

Table 2
Summary of SAS results as applied to several example macromolecules using $SA_{min}^{acc} = 1\,Å^2$

Compound[a]	Total number of atoms	Number of surface atoms	Cumulative % atoms	Cumulative % TSA	CPU-time[b] (s)
6CHA	1755	726	41.368	99.266	10.76
1RA2	1268	604	47.634	99.439	7.57
3FXN	2117	645	30.468	98.906	19.70
7TLN	2436	891	36.576	99.166	15.73
1TIM	3740	1565	41.845	98.834	23.45
3TRA	1603	929	57.954	99.314	9.74

[a] PDB codes: 6CHA, α-chymotrypsin; 1RA2, dihydrofolate reductase; 3FXN, flavodoxin (oxidized form); 7TLN, thermolysin; 1TIM, triose phosphate isomerase; 3TRA, tRNA.
[b] CPU-time on an SGI Indigo with an R4400 processor.

recommendation that the value of SA_{min}^{acc} be set equal to $1\,Å^2$.

Table 3 shows the same sort of results as Table 2, but with SA_{min}^{acc} set to $0.01\,Å^2$ rather than $1\,Å^2$. Note that, as found with 6CHA, this smaller value of SA_{min}^{acc} identifies atoms which account for 100% of the surface area but, as expected, the numbers of atoms classified as surface atoms are roughly 12% greater than in Table 2. Significantly, the CPU-times are exactly the same in Tables 2 and 3 because the surface areas are being calculated to exactly the same precision. Only the threshold between atoms classified as effective surface atoms or effective interior atoms was changed. Note, however, that the larger number of atoms classified as effective surface atoms would increase the CPU-time required by whatever software subsequently makes use of the list of surface atoms.

Although the CPU-times indicated in Tables 2 and 3 are certainly not excessive (and would be substantially smaller using the faster processors found in typical PCs as well as faster SGI machines), we were prompted to consider alternative methods for the identification of surface atoms of macromolecules. Several factors should be considered when making such comparisons: (1) CPU-time, (2) percentage of surface area represented by the surface atoms identified, (3) number of surface atoms correctly identified and (4) numbers of surface and interior atoms incorrectly classified. Since the results in Table 3 provide an unambiguous distinction between surface and interior atoms and since the SAS approach is based on atomic surface areas which are fundamental to the distinction between surface atoms and interior atoms, the results from the alternative methods are contrasted with the results of the SAS method shown in Table 3.

5.2. Number of intersecting neighbors (NIN) approach

The first alternative method considered for the identification of surface atoms, the NIN approach, was applied to each macromolecule in our test set. As a typical example, Fig. 3 illustrates the frequency distribution of the NIN results for 6CHA. The CPU-time required to generate these results was only 1.30 s. Compared to the SAS approach, the NIN approach was roughly eight times faster. However, the histogram in Fig. 3 clearly demonstrates that the NIN approach does not provide a clear distinction between the surface atoms and interior atoms. Essentially identical results were seen for all the other macromolecules in our test set.

The NIN approach was based on the intuitive notion that the number of intersecting neighbors would be far greater for interior atoms than for surface atoms, given that interior atoms are surrounded by neighbors on all sides but surface atoms are not. However, it is important to recall that the atomic radii used to identify neighbors must, necessarily, include the probe radius. Thus, as can be confirmed by molecular graphic display, depending on the curvature of the macromolecular surface, many surface atoms are intersected by macromolecule atoms on the "solvent-side" of the macromolecular surface.

Table 3
Summary of SAS results as applied to several example macromolecules using $SA_{min}^{acc} = 0.01\,Å^2$

Compound[a]	Total number of atoms	Number of surface atoms	Cumulative % atoms	Cumulative % TSA	CPU-time[b] (s)
6CHA	1755	933	53.368	100.00	10.76
1RA2	1268	730	57.634	100.00	7.57
3FXN	2117	872	41.190	100.00	19.70
7TLN	2436	1181	48.576	99.999	15.73
1TIM	3740	2216	59.845	100.00	23.45
3TRA	1603	1120	69.954	100.00	9.74

[a] For PDB codes see Table 2.
[b] CPU-time on an SGI Indigo with an R4400 processor.

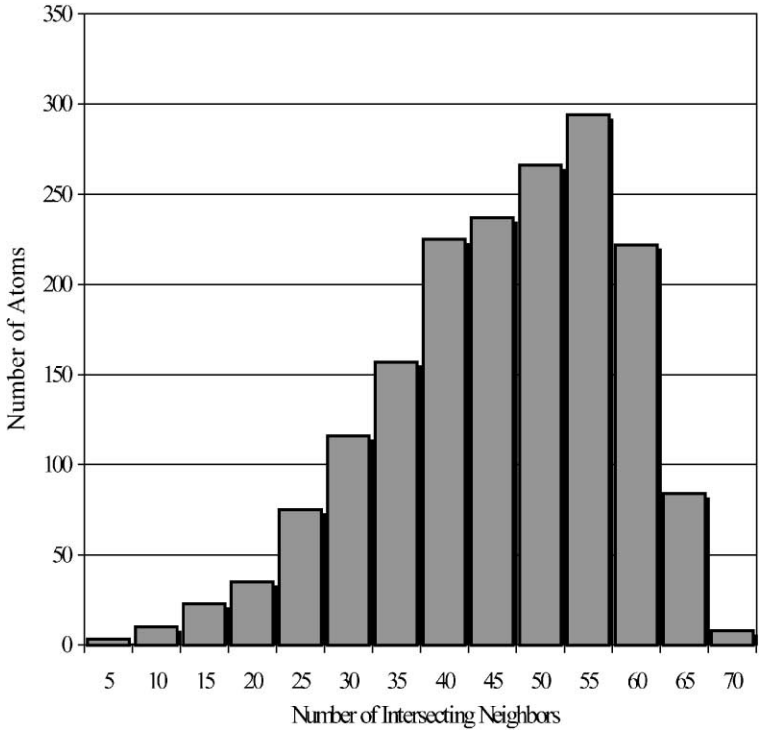

Fig. 3. The frequency distribution of the NIN method applied to 6CHA. Each column represents a range of N_{int} values. For example, the first column includes values in the interval $3 \leq N_{int} \leq 7$. Obviously, 5 intersecting neighbors is the midpoint of this interval.

5.3. Sum of vectors (SOV) approach

The second alternative method considered for the identification of surface atoms was the SOV approach. Once again, the results for 6CHA (illustrated in Fig. 4) were typical of the results obtained for all members of our test set. Like the NIN approach, 1.38 CPU s were required to generate these results but, as with the NIN approach, the results in Fig. 4 clearly indicate that $||\boldsymbol{V}_{sum}||$ does not enable a distinction between surface atoms and interior atoms. Comparable results were also seen for all other macromolecules in our test set. The SOV approach fails for the same reason as indicated for the failure of the NIN approach.

5.4. UCSF approach

As one would expect, the results obtained using the UCSF method depend strongly upon the lattice spacing. As a first step, we used the author-recommended spacing of 1.60 Å and the results are summarized in Table 4. One sees almost immediately that the UCSF approach classified many more atoms as surface atoms compared to the results using the SAS approach displayed in Table 3. For example, the UCSF approach classified 1282 atoms of 6CHA as surface atoms. This is approximately 73% of the total number of atoms in this molecule and many of those atoms actually had SAS areas that were equal to zero. In contrast, the SAS approach classified 933 atoms as surface atoms—roughly 53% of the total number of atoms. According to our definition of an effective surface atom (i.e. $SA_{atm}^{acc} \geq SA_{min}^{acc} = 0.01\,Å^2$), only 869 of the 1282 UCSF surface atoms are actually surface atoms. The remaining 413 atoms are interior atoms. Moreover, of the 473 atoms which the USCF method classified as interior atoms, 64 (with surface areas of 0.02 $Å^2$ to as much as 2.55 $Å^2$) should have been classified as a surface atom.

The UCSF method yielded similarly unsatisfactory results for the other macromolecules in our test set. Typically, it classified over 70% of the atoms in a macromolecule as surface atoms even though a large fraction of these atoms had near-zero surface areas and should have been classified as interior atoms. In addition, many of the atoms classified as interior atoms by the UCSF method should have been classified as surface atoms. Some of the atoms incorrectly classified as interior atoms had surface areas as large as 43 $Å^2$. The UCSF results for flavodoxin (3FXN) were slightly better than for the other five macromolecules in that only 56.4% of the total number of atoms were classified as surface atoms. However, this still included 399 incorrectly classified interior atoms.

Although the results obtained using the UCSF method with the recommended lattice spacing were unsatisfactory, its speed—roughly 50 times faster than the SAS method—prompted us to attempt to improve its performance by repeating the analysis of Table 4 using other values for the lattice spacing. In this study, lattice spacings of 1.20–2.10 Å in increments of 0.10 Å were investigated. Since the lattice

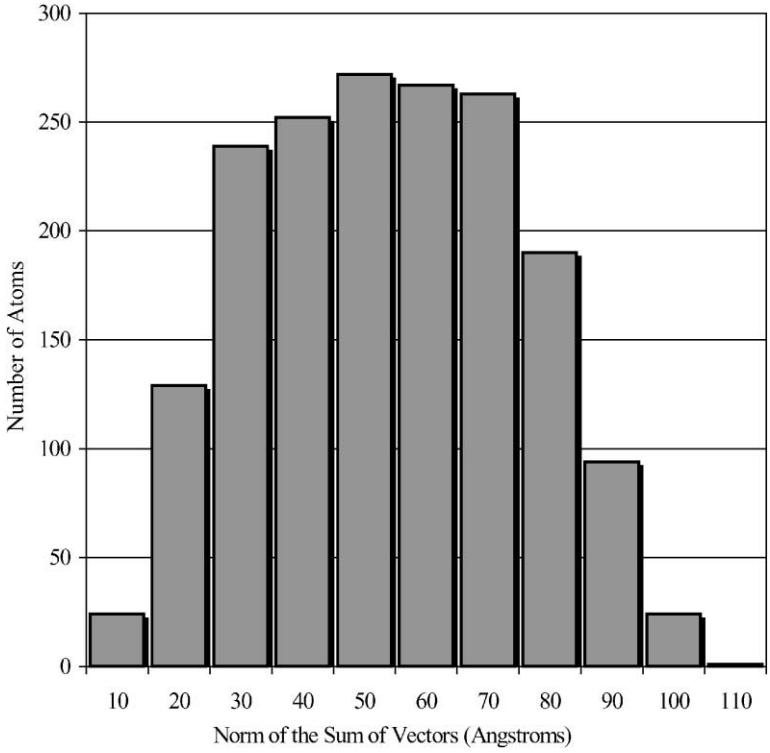

Fig. 4. The frequency distribution of the SOV for 6CHA. Each column represents a range of $\|\boldsymbol{V}_{\text{sum}}\|$ values. For example, the first column includes values in the interval $5 < \|\boldsymbol{V}_{\text{sum}}\| \leq 15$ Å. Obviously, 10 Å is the midpoint of this interval.

spacing is somewhat analogous to the radius of an atom, this range of values was at least consistent with the range of values seen with the atomic radii used in the SAS approach.

For lattice spacings less than 1.60 Å, the USCF results were even more unsatisfactory (data not shown). As the lattice spacing decreased, the total number of atoms classified as surface atoms significantly increased. Although there was a small increase in the number of atoms correctly classified as surface atoms, there was a very substantial increase in the number of interior atoms incorrectly classified as surface atoms.

For lattice spacings greater than 1.60 Å, the USCF results were also unsatisfactory. Table 5 summarizes the UCSF results obtained using lattice spacings of 1.70, 1.80 and 1.90 Å. The observed trend for these values (as well as for 2 and 2.10 Å) was that, as the lattice spacing increased, the total number of atoms classified as surface atoms decreased and the number of interior atoms incorrectly classified as surface atoms decreased. However, we also observed that the number of surface atoms correctly classified as surface atoms decreased and that, therefore, the number of surface atoms not classified as surface atoms increased.

Although the UCSF approach is much faster than the SAS approach, it yields unreliable results. The inability of the UCSF approach to reliably identify surface atoms is not surprising given that it is based on a very crude representation of the atoms of a macromolecule. Recall that the size of an atom is represented by a lattice cube and that the coordinates

Table 4
Summary of UCSF results using a 1.60 Å lattice spacing

Compound[a]	Total number of atoms	UCSF surface atoms	Classification of UCSF surface atoms[b]		Classification of UCSF interior atoms[b]		CPU-time[c] (s)
			Surface atoms	Interior atoms	Surface atoms	Interior atoms	
6CHA	1755	1282	869	413	64	409	0.24
1RA2	1268	1013	722	291	8	247	0.22
3FXN	2117	1193	794	399	78	846	0.20
7TLN	2436	1748	1131	617	50	638	0.38
1TIM	3740	2894	2046	848	170	676	0.56
3TRA	1603	1523	1108	415	12	68	0.37

[a] For PDB codes see Table 2.
[b] Classification of UCSF surface and interior atoms as compared to the SAS results using $SA_{\text{min}}^{\text{acc}} = 0.01$ Å^2.
[c] CPU-time on an SGI Indigo with an R4400 processor.

Table 5
Summary of UCSF results using lattice spacings of 1.70, 1.80 and 1.90 Å

Compound[a]	Total number of atoms	Lattice-point spacing (Å)	UCSF surface atoms	Classification of UCSF surface atoms[b]		Classification of UCSF interior atoms[b]		CPU-time[c] (s)
				Surface atoms	Interior atoms	Surface atoms	Interior atoms	
6CHA	1755	1.70	1213	839	374	94	448	0.22
		1.80	1137	817	320	116	502	0.19
		1.90	1122	805	317	128	505	0.17
1RA2	1268	1.70	997	715	282	15	256	0.19
		1.80	965	705	260	25	278	0.17
		1.90	935	698	237	32	301	0.15
3FXN	2117	1.70	1159	784	375	88	870	0.18
		1.80	1151	769	382	103	863	0.16
		1.90	1133	760	373	112	872	0.14
7TLN	2436	1.70	1611	1099	512	82	743	0.34
		1.80	1552	1076	476	105	779	0.29
		1.90	1542	1068	474	113	781	0.26
1TIM	3740	1.70	2594	1910	684	306	840	0.50
		1.80	2341	1777	564	439	960	0.44
		1.90	2309	1748	561	468	963	0.39
3TRA	1603	1.70	1520	1105	415	15	68	0.33
		1.80	1485	1090	395	30	88	0.29
		1.90	1483	1085	398	35	85	0.25

[a] For PDB codes see Table 2.
[b] Classification of UCSF surface and interior atoms as compared to the SAS results using $SA_{min}^{acc} = 0.01\,Å^2$.
[c] CPU-time on an SGI Indigo with an R4400 processor.

of the center of the atom are approximately represented by the center of that cube. Also, the size representation of an atom is exactly identical to all others despite the fact that different atom types have different atomic radii. As a result, the USCF approach can only provide crude information regarding the macromolecular surface (which clearly depends on both the 3D coordinates and the van der Waals radii of the atoms).

5.5. Molecular dot surface (MDS) approach

Given that the value of SA_{min}^{acc} was set equal to $0.01\,Å^2$ for comparative purposes, a dot density of 100 dots/Å^2 would be needed to identify all atoms with SAS areas greater than or equal to $0.01\,Å^2$ as surface atoms. However, a dot density of 100 dots/Å^2 is too large and impractical. The CPU-time required to generate a molecular dot surface using such a high dot density would far exceed the CPU-time needed by all other methods considered. The dot densities typically reported in the literature range between 1 and 10 dots/Å^2. Hence, we experimented within this range.

Table 6 summarizes the MDS results for our test set of macromolecules. The dot densities used to generate these results were 5 and 10 dots/Å^2. In the limit of infinite dot density, the molecular dot surface and the solvent accessible surface are identical. Therefore, it is reasonable to expect that, when using practical dot densities, the MDS approach will yield results similar to those from the SAS approach. However, because we are using dot densities significantly less than 100 dots/Å^2, the MDS approach should not be expected to yield results as good as those in Table 3.

As indicated in Table 6, the MDS approach identified several atoms as surface atoms which the SAS approach classified as interior atoms. This was observed for all macromolecules in our test set at the indicated dot densities. Further investigation revealed that, for approximately 20% of these atoms, their SAS areas were within the interval $0.005\,Å^2 \leq SA_{atm}^{acc} < 0.01\,Å^2$. Recall that, given the angle increment of 9° chosen for use in the SAVOL1 program, the smallest value we can reasonably assign to SA_{min}^{acc} is $0.005\,Å^2$. If we simply used a smaller angle increment and set the value of SA_{min}^{acc} equal to $0.005\,Å^2$, the SAS approach would also classify these particular atoms as surface atoms. However, Table 6 also clearly demonstrates that the MDS approach failed to identify many surface atoms which the SAS approach (even when using the 9° angle increment) classified correctly.

Some of the results obtained using the MDS approach were somewhat puzzling. For example, at 5 dots/Å^2, the MDS approach classified one 6CHA atom as a surface atom even though its SA_{atm}^{acc} was equal to zero. Visual inspection of the molecular surface using the Sybyl Molecular Modeling Software [48] also confirmed that this atom was completely occluded by neighboring atoms. This atom was correctly classified when the dot density was increased to 10 dots/Å^2, suggesting that better results could be obtained by modest increases in dot density. However, at 10 dots/Å^2, three other 6CHA interior atoms were incorrectly classified as surface

Table 6
Summary of MDS results using dot densities of 5 and 10 dots/Å^2

Compound[a]	Total number of atoms	Dot density (dots/Å^2)	MSD surface atoms	Classification of MDS surface atoms[b]		Classification of MDS interior atoms[b]		CPU-time[c] (s)
				Surface atoms	Interior atoms	Surface atoms	Interior atoms	
6CHA	1755	5	887	886	1	47	821	7.65
		10	911	908	3	25	819	18.11
1RA2	1268	5	711	702	9	28	529	5.20
		10	724	718	6	12	532	10.42
3FXN	2117	5	796	793	3	79	1242	10.34
		10	805	805	4	67	1241	21.39
7TLN	2436	5	1114	1110	4	71	1251	10.74
		10	1147	1134	13	47	1242	21.40
1TIM	3740	5	2067	2058	9	158	1515	15.82
		10	2144	2128	16	88	1508	32.00
3TRA	1603	5	1091	1089	2	31	481	6.32
		10	1104	1101	3	19	480	12.69

[a] For PDB codes see Table 2.
[b] Classification of MDS surface and interior atoms as compared to the SAS results using $SA_{min}^{acc} = 0.01$ Å^2.
[c] CPU-time on an SGI Indigo with an R4400 processor.

atoms that had been correctly classified at 5 dots/Å^2. Similar results were seen for all other macromolecules in our test set. Interestingly, each of the interior atoms that were incorrectly classified as surface atoms had only one dot "exposed" on their surface.

There are two probable causes for the slightly inconsistent results seen with the MDS approach: (1) the fact that dots necessarily appear at different positions when different dot densities are specified and (2) the fact that some degree of numerical round-off error is unavoidable within the QDS program used to generate the MDS. The CPU-time required by the MDS approach using 5 dots/Å^2 is slightly less than the time needed by the SAS approach but the reverse is true at 10 dots/Å^2. Moreover, the MDS results at these dot densities are not as accurate as the results seen with the SAS approach. Increasing the dot densities in order to achieve more accuracy would substantially increase the CPU-time.

6. Summary

There are numerous applications which either rely upon or could benefit from the reliable identification of the effective surface atoms within a macromolecular structure. We have proposed a simple yet reliable method which, significantly, is based upon the fundamental definition of a "surface atom". We have validated this SAS approach and we have compared it to four alternative methods. While some of the alternatives are somewhat faster, none of them are as reliable. Given that the CPU-time required by the SAS approach is merely a matter of seconds, its accuracy certainly justifies its use as the preferred method for the identification of surface atoms of macromolecular structures.

References

[1] R. Langridge, T.E. Ferrin, I.D. Kuntz, M.L. Connolly, Real-time color graphics in the studies of molecular interactions, Science 211 (1983) 661–666.
[2] P.A. Bash, N. Pattabiraman, C. Huang, T.E. Ferrin, R. Langridge, van der Waals surfaces in molecular modeling: implementation with real-time computer-graphics, Science 222 (1983) 1325–1327.
[3] K.P. Peters, J. Fauck, C. Frömmel, The automatic search for ligand sites in proteins of known 3D structures using only geometric criteria, J. Mol. Biol. 256 (1997) 201–213.
[4] C.A. Del Carpio, Y. Takahashi, S. Sasaki, A new approach to the automatic identification of candidates for ligand receptor sites in proteins. I. Search for pocket regions, J. Mol. Graphics 11 (1993) 23–29.
[5] C.M.W. Ho, G.R. Marshall, Cavity search: an algorithm for the isolation and display of cavity-like binding regions, J. Comput. Aided Mol. Des. 4 (1990) 337–354.
[6] R. Norel, D. Fischer, H.J. Wolfson, R. Nussinov, Molecular surface recognition by a computer vision-based technique, Protein Eng. 7 (1994) 39–46.
[7] I.D. Kuntz, J.M. Blaney, S.J. Oatley, R. Langridge, T.E. Ferrin, A geometric approach to macromolecular–ligand interactions, J. Mol. Biol. 161 (1982) 269–288.
[8] B.K. Shoichet, I.D. Kuntz, Matching chemistry and shape in molecular docking, Protein Eng. 6 (1993) 723–732.
[9] D. Fischer, R. Norel, H. Wolfson, R. Nussinov, Surface motifs by a computer vision technique: searches, detection, and implications for protein–ligand recognition, Proteins Struct. Funct. Genet. 16 (1993) 278–292.
[10] M.L. Connolly, Shape complementarity at the hemoglobin $\alpha_1\beta_1$ subunit interface, Biopolymers 25 (1986) 1229–1247.
[11] G. Jones, P. Willett, R.C. Glen, Molecular recognition of receptor sites using a genetic algorithm with a description of desolvation, J. Mol. Biol. 24 (1995) 43–53.
[12] P.K. Weiner, R. Langridge, J.M. Blaney, R. Schaefer, P.A. Kollman, Electrostatic potential molecular surfaces, Proc. Natl. Acad. Sci. U.S.A. 79 (1982) 3754–3758.

[13] H.A. Gabb, R.M. Jackson, M.J.E. Sternberg, Modeling protein docking using shape complementarity, electrostatics and biochemical information, J. Mol. Biol. 272 (1997) 106–120.
[14] M. Meyer, P. Wilson, D. Schomburg, Hydrogen bonding and molecular surface shape complementarity as a basis for protein docking, J. Mol. Biol. 264 (1996) 199–210.
[15] A.R. Leach, Ligand docking to proteins with discrete side-chain flexibility, J. Mol. Biol. 235 (1994) 345–356.
[16] F. Jiang, S. Kim, Soft docking: matching of molecular surface cubes, J. Mol. Biol. 219 (1991) 79–102.
[17] B.K. Shoichet, I.D. Kuntz, Protein docking and complementarity, J. Mol. Biol. 221 (1991) 327–346.
[18] D.J. Bacon, J. Moult, Docking by least-squares fitting of molecular surface patterns, J. Mol. Biol. 225 (1992) 849–858.
[19] P. Andrews, Functional groups, drug–receptor interactions and drug design, TIPS 7 (1986) 148–151.
[20] Ajay, M.A. Murcko, Computational methods for predicting binding free energy in ligand–receptor complexes. J. Med. Chem. 38 (1995) 38 4954–4967.
[21] M.L. Connolly, Solvent-accessible surfaces of proteins and nucleic acids, Science 221 (1983) 709–713.
[22] M.L. Connolly, The molecular surface package, J. Mol. Graphics 11 (1993) 139–141.
[23] N.C. Cohen, J.M. Blaney, C. Humblet, P. Gund, D.C. Barry, Molecular modeling software and methods for medicinal chemistry, J. Med. Chem. 33 (1990) 883–894.
[24] J.B. Moon, W.J. Howe, A fast algorithm for generating smooth molecular dot surface representations, J. Mol. Graphics. 7 (1989) 109–112.
[25] N.L. Max, Computer representation of molecular surfaces, J. Mol. Graphics 2 (1984) 8–13.
[26] T.J. Richmond, Solvent-accessible surface area and excluded volume in proteins: analytical equations for overlapping spheres and implications for the hydrophobic effect, J. Mol. Biol. 178 (1984) 63–89.
[27] E.L. Plummer, The application of quantitative design strategies in pesticides design, in: K.B. Lipkowitz, D.B. Boyd (Eds.), Reviews in Computational Chemistry, Vol. 7, VCH Publishers, New York, 1994, pp. 119–168.
[28] R.B. Hermann, Theory of hydrophobic bonding. II. The correlation of hydrocarbon solubility in water with solvent cavity surface area, J. Phys. Chem. 76 (1972) 2754–2759.
[29] F.M. Richards, Areas, volumes, packing, and protein structure, Ann. Rev. Biophys. Bioeng. 6 (1977) 151–176.
[30] R.S. Pearlman, Molecular surface area and volume: their calculation and use in predicting solubilities and free energies of desolvation, in: W.J. Dunn III, J.H. Block, R.S. Pearlman (Eds.), Partition Coefficient: Determination and Estimation, Pergamon Press, New York, 1986, pp. 3–20.
[31] R.S. Pearlman, Molecular surface areas and volumes and their use in structure/activity relationships, in: S.H. Yalkowsky, A.A. Sinkula, S.C. Valvani (Eds.), Physical Chemical Properties of Drugs, Marcel Dekker, New York, 1980, pp. 321–347.
[32] R.S. Pearlman, SAREA: calculation of van der Waals (or accessible) surface areas of molecules, QCPE Bull. 1 (1980) 15.
[33] M.L. Connolly, Computation of molecular volume, J. Am. Chem. Soc. 107 (1985) 1118–1124.
[34] F.M. Richards, Calculation of molecular volumes and areas for structures of known geometry, Meth. Enzymol. 115 (1985) 440–464.
[35] L.R. Dodd, D.N. Theodorou, Analytical treatment of the volume and surface area of molecules formed by an arbitrary collection of unequal spheres intersected by planes, Mol. Phys. 72 (1991) 1313–1345.
[36] B. Lee, F.M. Richards, Interpretation of protein structures: estimation of static accessibility, J. Mol. Biol. 55 (1971) 379–400.
[37] G. Perrot, B. Cheng, K.D. Gibson, J. Vila, K.A. Palmer, A. Nayeem, B. Maigret, H.A. Scheraga, MSEED: a program for the rapid analytical determination of accessible surface areas and their derivatives, J. Comput. Chem. 13 (1992) 1–11.
[38] M. Petitjean, On the analytical calculation of van der Waals surfaces and volumes: some numerical aspects, J. Comput. Chem. 15 (1994) 507–573.
[39] M. Totrov, R. Abagyan, The contour build-up algorithm to calculate the analytical molecular surface, J. Struct. Biol. 116 (1996) 138–143.
[40] R. Fraczkiewicz, W. Braun, Exact and efficient analytical calculation of the accessible surface areas and their gradients for macromolecules, J. Comput. Chem. 19 (1998) 319–333.
[41] SAVOL3, Version 5.0, University of Texas, Austin, TX.
[42] J.M. Skell, Software tools for computer-assisted molecular design, Ph.D. Dissertation, The University of Texas, Austin, 1993 (Chapter 3).
[43] A. Bondi, van der Waals volumes and radii, J. Phys. Chem. 68 (1964) 441–451.
[44] M.-Y.K. Brusniak, Development and application of software for CADD, Ph.D. Dissertation, The University of Texas, Austin, 1996 (Chapter 2).
[45] M.-Y.K. Brusniak, R.B. Balducci, R.S. Pearlman, Novel algorithms for accurate and rapid generation of molecular dot surfaces, in press.
[46] E.E. Abola, F.C. Bernstein, S.H. Bryant, T.F. Koetzle, J. Weng, Protein Data Bank, in: F.H. Allen, G. Bergerhoff, R. Sievers (Eds.), Crystallographic Databases-Information Content, Software Systems, Scientific Applications, Data Commission of the International Union of Crystallography, Bonn, 1987, pp. 107–132.
[47] F.C. Bernstein, T.F. Koetzle, G.J.B. Williams, E.F. Meyer Jr., M.D. Brice, J.R. Rodgers, O. Kennard, T. Shimavouchi, M. Tasumi, Protein Data Bank: a computer-based archival file for macromolecular structures, J. Mol. Biol. 112 (1977) 535–542.
[48] SYBYL Molecular Modeling Software, Version 6.4.2, Tripos Inc., St. Louis, MO, 1998.