**Structural bioinformatics**

**Three-dimensional computation of atom depth in complex molecular structures**

Daniele Varrazzo\(^1\), Andrea Bernini\(^1,2\), Ottavia Spiga\(^1,2\), Arianna Ciutti\(^2\), Stefano Chiellini\(^2\), Vincenzo Venditti\(^1\), Luisa Bracci\(^1\) and Neri Niccolai\(^1,2,\)*

\(^1\)Biomolecular Structure Research Center and Department of Molecular Biology, Università di Siena, I-53100 Siena, Italy and \(^2\)SienaBioGrafix Srl, I-53100 Siena, Italy

Received on February 9, 2005; revised on March 30, 2005; accepted on April 7, 2005

Advance Access publication April 12, 2005

**ABSTRACT**

**Motivation:** For a complex molecular system the delineation of atom–atom contacts, exposed surface and binding sites represents a fundamental step to predict its interaction with solvent, ligands and other molecules. Recently, atom depth has been also considered as an additional structural descriptor to correlate protein structure with folding and functional properties. The distance between an atom and the nearest water molecule or the closest surface dot has been proposed as a measure of the atom depth; but, in both cases, the 3D character of depth is largely lost. In the present study, a new approach is proposed to calculate atom depths in a way that the molecular shape can be taken into account.

**Results:** An algorithm has been developed to calculate intersections between the molecular volume and spheres centered on the atoms whose depth has to be quantified. Many proteins with different size and shape have been chosen to compare the results obtained from distance-based and volume-based depth calculations. From the wealth of experimental data available for hen egg white lysozyme, H/D exchange rates and TEMPOL induced paramagnetic perturbations have been analyzed both in terms of depth indexes and of atom distances to the solvent accessible surface. The algorithm here proposed yields better correlations between experimental data and atom depth, particularly for those atoms which are located near to the protein surface.

**Availability:** Instructions to obtain source code and the executable program are available either from http://sienabiografix.com or http://sadic.sourceforge.net

**Contact:** niccola@unisi.it

**Supplementary information:** http://www.Sienabiogzefix.com/publication

**INTRODUCTION**

Structural biology is nowadays rapidly growing, due to a synergistic post-genomic effect of the large developments of X-ray crystallography, nuclear magnetic resonance (NMR) and bioinformatics. In the Protein Data Bank (PDB) (Berman et al., 2000), a wealth of resolved and predicted protein structures are available and, on this basis, structural descriptors have been developed to correlate accessible molecular surface (Lee and Richards, 1971; Richmond, 1984; Quillin and Matthews, 2000; Totrov and Abagyan, 1996; Gerstein et al., 1995), molecular volumes (Richmond, 1984) and potential binding sites (Lo Conte et al., 1999; Shulman-Peleg et al., 2004; Tsuchiya et al., 2004; Innis et al., 2004; Gutteridge et al., 2003) with functional features, protein folding and structural stability (Serrano et al., 1992).

Recently, the calculation of atom depth from the protein surface has been proposed as an additional criterion to define protein structures more accurately (Pintar et al., 2003b; Chakravarty and Varadarajan, 1999) by exploring the interior of the molecule, as the strength of van der Waals (VdW) and electrostatic interactions might be dependent on the distance from the molecular surface (Chakravarty and Varadarajan, 1999; Richards, 1977). Moreover, once the deepest molecular moieties can be defined, a systematic analysis of their properties can be carried out to gain information on molecular structure and stability.

Atom depth can be defined as the distance between an atom and the nearest surface water molecule, either experimentally defined or hypothetically present. However, by exploring the closest distance between an atom and a dot of the solvent accessible surface (Chakravarty and Varadarajan, 1999) or the distance between an atom and its closest solvent accessible neighbor (Pintar et al., 2003a), contributions from the 3D molecular shape to the actual atom depth are largely lost.

Hence, to estimate atom depth a new approach reflecting the molecular shape is proposed here by measuring intersections between the molecular volume and a sphere of a suitable radius, centered on the atom whose depth has to be quantified. It is apparent, indeed, that smaller the exposed volume, deeper is the 3D insertion of the investigated atom inside the molecular structure. Since, in general, depth is a very relative quantity which depends on the overall size and shape of the object under discussion, an atom depth index is suggested as a more appropriate parameter to discuss atom insertions within each investigated molecular systems.

These depth indexes, calculated by using SADIC (Simple Atom Depth Index Calculator) algorithm, for instance, could constitute a new rational basis to reanalyze inner and outer amino acid compositions of proteins or to improve the analysis of depth-related physical phenomena. Among the latter molecular processes, H/D isotopic exchange of protein amide hydrogens is particularly relevant being commonly referred to as molecular surface exposures. Exchange rates are very frequently determined from NMR (Roder et al., 1985) or mass spectrometry (Miranker et al., 1996) studies to explore...
protein conformations and dynamics. It has also been shown that NMR studies of through-the-space interactions, occurring between paramagnetic probes and protein nuclei, can be interpreted in terms of protein surface exposures (Niccolai et al., 2001, 2003; Pintacuda and Otting, 2002).

In the present report, volume-based and distance-based atom depth have been evaluated and compared for proteins of different size and shape. Calculated depths have been also correlated with H/D exchange and paramagnetic perturbation data available for hen egg white lysozyme (HEWL).

**ALGORITHM**

SADIC algorithm is based on the simple idea of sampling the space around each atom of a given molecule by evaluating, for selected distances from the atom center, the portion of volume that is external to any protein atom. In other words, such volume, henceforth called the exposed volume, represents the space external to the molecular surface comprised at a distance \( r \) in all directions around the atom. Therefore, the size of the exposed volume is a direct measure of atom depth with respect to the molecular surface, as smaller the exposed volume, deeper is the atom within the molecular structure. When dealing with exposed volumes instead of linear distances, as previously proposed for depth calculations (Chakravarty and Varadarajan, 1999; Pintar et al., 2003b), the information on surface shape is considered. SADIC algorithm can yield an accurate indication of atom depth, since distances from the atom center to the solvent exposed surface are simultaneously evaluated in all directions. It follows that atoms located in protruding loops have exposed volumes greater than those exhibited by atoms which are equally close to the surface but located at the bottom of a pocket.

In principle, this algorithm can be used to analyze local depth for objects having any size and shape, provided that only an ‘inside’ and an ‘outside’ can be unambiguously assigned. The 3D model of a molecule, as an assembly of sphere shaped atoms, satisfies this requirement, since all the points located inside one of these spheres, i.e. closer to an atom center than the VdW atom radius, can be considered ‘inside’ the molecule.

In order to approximate the volume of the intersection between the molecule and a sphere with a given radius \( r \) and center \( C \), the sphere interior is split into units whose volume is known. For each volume unit a representative point is taken: we approximate the exposed volume by testing the sampling points against the molecule and summing the volume relative to all the points outside the molecular model.

The choice of \( r \) is of critical importance, since too small or too large \( r \) values would yield null or large exposed volumes, respectively at a very similar extent for all the atoms of the investigated molecule. It should be noticed that the values obtained by sampling inside a sphere of radius \( r \) can be effectively used to calculate exposition for each sphere of radius \( r' \leq r \). To exploit this possibility better, sampling points are chosen over concentric spheres with growing radii \( r_0 \ldots r_n \).

A simplistic pattern to sample a sphere interior consists of a regular grid in spherical coordinates. This method has the drawback to produce the same number of samples at each radius, thus yielding more packed points toward the poles, and the center and coarser points toward the equator and the outside. In order to overcome the problem, a different sampling pattern is used by SADIC, as described in detail in the Supplementary information section.

**IMPLEMENTATION**

The current SADIC implementation is written in Python and C programming languages (see Supplementary information). The program consists of an object-oriented library providing classes responsible to generate sampling patterns, to model solid objects (they can be subclassed to add new capabilities to the framework) and to parse PDB. (Berman et al., 2000) entity files. An executable with a command line interface is provided: the program can read PDB entity files either from a local file system or from an external database through its URL (http, ftp and file protocols are supported) or its pdb ID code. The user can perform a molecule sampling on a list of given points in the space or on a selection of entity atoms, either absolutely referred by serial number or selected by atom name (e.g. using ‘CA’ to refer to protein backbone α-carbon), residue number, chain identifier. If the file entity contains more than one structure, as is the case of NMR determined structures, the sampling can be separately performed on a selection of structures. In this case, average and SD of the results may be automatically calculated.

**RESULTS AND DISCUSSION**

SADIC program has been developed to obtain a new tool for a structural characterization of complex molecules, such as proteins and nucleic acids, by considering the atom depth. Since depth is a characteristic which can be conveniently discussed only in relation to the size of each investigated system, SADIC outputs are more conveniently analyzed as atom depth indexes, \( D \), rather than absolute exposed volumes.

Thus, for an atom \( i \) of a given molecule and a sampling radius \( r \), a depth index \( D_i, r \) may be defined as

\[
D_i, r = \frac{2V_{i, r}}{V_{0, r}}, \quad (1)
\]

where \( V_{i, r} \) is the exposed volume of a sphere of radius \( r \) centered on atom \( i \) and \( V_{0, r} \) is the exposed volume of the same sphere when centered on an isolated atom.

As already pointed out in the Algorithm section, to avoid flattening of the algorithm outputs towards similar \( D_i, r \) values, the selection of the \( r \) value represents a very critical step. In Figure 1 the evolution of \( D_i, r \) of a representative selection of HEWL Ca carbons is shown: Thr47 and Ile58 Ca atoms are both equally close to the solvent exposed surface, but in the convex and concave molecular regions, respectively. Each of the Trp28 and Ser50 Ca atoms are, instead, deeply inserted in one of the two HEWL domains. For small and large \( r \) values, all \( D_{i, r} \) converge to 0 and 2, respectively and the atom depth index calculated by SADIC loses its structural information. Conversely, in an intermediate region of \( r \) values, centered for HEWL at \( \sim 9 \) Å, a large dispersion of \( D_{i, r} \) can be observed. Then, to analyze conveniently atom depths it seems appropriate to choose the biggest sphere radius which determines the condition \( D_{i, r} = 0 \) only for one nth inner most atom. In the case of HEWL this condition is met by Trp28 Ca carbon, thus resulting as the most internal atom, at a \( r \) value of 9 Å.

Thus, once a suitable sphere radius has been chosen, calculated \( D_{i, r} \) readily describe the topology of each atom, as values close to 0 or 1 defines the inner or outer atoms (Fig. 2). Furthermore, the
D. Varrazzo et al.

Fig. 1. The evolution of $D_{ij}$ values with the sphere radius $r$ calculated for selected backbone Cα carbons of HEWL (Trp28: crosses, Thr47: squares, Ser50: filled circles and Ile58: triangles).

Fig. 2. Space fill representation of lysozyme (pdb file ID code 4lzt); the enzyme is halved into two complementary moieties to show some of the inner heavy atoms colored according to their $D_{ij}$ values. $D_{ij} > 1$ condition defines atoms which are very close to a convex molecular surface, as in the case of HEWL α carbons of Thr47, Asp48 and Gly117, whose $D_{ij}$ are 1.22, 1.10 and 1.15, respectively.

The validity of the proposed algorithm has been tested on many proteins by comparing SADIC outputs with distance-based atom depths. Thus, as shown in Figure 3, $D_{ij}$s of a small spherical protein and of a large oblate one are compared with atom distances calculated from the closest exposed neighbor, $dpx_i$ (Pintar et al., 2003b) and from the nearest surface water molecule (Chakravarty and Varadarajan, 1999), $dnw_i$. Among the different sets of data, a good agreement exists, as the Cα carbons exhibiting the highest $D_{ij}$ values correspond to the shortest $dpx_i$s and $dnw_i$s. Conversely, for the Cα carbons having the longest $dpx_i$ and $dnw_i$ values, $D_{ij}$s close to 0 are found. It is also evident that a higher detail in describing atom depth, particularly for those atoms which are located near to the protein surface, is reached by SADIC. This feature directly derives from the fact that only $D_{ij}$s depend both on surface distances and molecular shape and that equally distant atoms from the surface, but close to concave or convex surface regions, exhibit very different depth indexes.

To check how $D_{ij}$s can be useful in the structural interpretation of experimental data, reported H/D exchange rates of HEWL amide protons (Pedersen et al., 1993) and paramagnetic perturbations of NMR signals (Niccolai et al., 2003) have been analyzed in terms of atom depth. As shown in Figure 4, $D_{ij}$s, $dpx_i$s and $dnw_i$s correlate similarly with H/D exchange rates, $K_{ex}$, and paramagnetic signal attenuations, $A_i$, only in the case of the innermost HEWL atoms. For the outer ones, $K_{ex}$s and $A_i$s are, indeed, all grouped in a very narrow range of both $dpx_i$ and $dnw_i$ values. By simple inspection of Figure 4, it is apparent that $D_{ij}$s generally exhibit a higher correlation than distance-based depths with the experimental data, as all the slowest exchange rates of HEWL amide hydrogens have been calculated for atoms with $D_{ij} < 0.6$. It should be noted that for the latter amide hydrogens a large variety of distance-based atom depths is derived. Furthermore, any overlapping of the experimentally derived parameters observed at the closest surface distances is largely resolved, while the slow exchange rates measured for Ala10, Phe34 and Leu83 amide groups are more consistent with the corresponding $D_{ij}$s.
3D Computation of atom depth

Correlations of TEMPOL induced paramagnetic attenuations, $A_i$, of HEWL $^{13}$C-$^1$H HSQC NMR signals of $\alpha$ carbons with (a) $D_{ij}$, (b) $dp_{ij}$ and (c) $dn_{ij}$. Correlations between $K_{ex_i}$ of HEWL backbone amide protons and (d) $D_{ij}$, (e) $dp_{ij}$ and (f) $dn_{ij}$. Identical $dp_{ij}$ and $dn_{ij}$ values are exhibited by the most exposed atoms: the latter atoms are highlighted with triangles in the $D_{ij}$ plot. Labels refer to amino acid residue positions in the protein sequence.

A linear or higher level dependence of $D_{ij}$s on HEWL $K_{ex_i}$s and $A_i$'s cannot be delineated, as fast H/D exchange rates were measured for the deeply inserted Thr40 and Cys94 amide hydrogens, in spite of their close proximity to a concave surface. Moreover, a small $A_i$ value is exhibited by the surface exposed Asp48 $\alpha$ carbon, while for the buried Ile98 a strong paramagnetic attenuation is observed. These four cases represent the most evident discrepancies, but many other anomalous behaviors of $K_{ex_i}$s and $A_i$'s versus both volume-based and distance-based depth can be seen in the data shown in Figure 4. In this respect, it should be stressed that both experimental parameters depend on atom depth only at a first approximation and that a more detailed discussion of exchange rates and paramagnetic perturbations in terms of atom depth would be needed. The H/D exchange process, determined by the dynamics of the hydrogen bond network within the protein and its hydration shell, is commonly related to solvent accessibility. The fact that SADIC outputs are more consistent with $K_{ex_i}$s than atom depths obtained from distance-based calculations, suggests that a step forward in the structural interpretation of the latter experimental parameter might be achieved. On the other hand, the weaker correlation observed between paramagnetic perturbations induced by TEMPOL and atom depths confirms that complex dynamics control the interaction of protein surfaces with paramagnetic probes (Nicolai et al., 2003). On the basis of the $B$ factors reported in the crystal structure of HEWL with PDB (Berman et al., 2000) ID code 4lzt, it is apparent that local structural flexibility is not responsible for the limited correlation between atom depth and accessibility dependent experimental data.

It can be concluded that the use of SADIC algorithm might favor improved depth-oriented discussions on experimental data, possibly enhancing our understanding of structure stability and dynamics of complex molecules.

ACKNOWLEDGEMENTS

Thanks are due to grants from the Italian Ministry of University PRIN03-059395 and from the University of Siena (PAR 2002). Special thanks are also due to Francesco Niccolai for technical assistance.

REFERENCES

