Term Project

Solvent Accessible Surface Area — spring 2024

IIT KHARAGPUR

Akash Das — 20CS10006
Rudrak Patra — 20CS30044

April 15, 2024

1 Introduction

The concept of solvent accessibility of residues in globular proteins was first introduced by Lee and Richards in 1971. They defined the term "solvent accessible surface area" (SASA) as the extent to which atoms on the surface of a protein can form contacts with solvent. The solvent in biological systems is taken as water with a radius of 1.4 Å. It represents the locus of the center of the solvent molecule as it rolls along the protein, making the maximum permitted van der Waals contacts without penetrating any other atom. It was also described as the "static accessible surface area" because potential flexibility was not included.

For calculation purposes, a sphere is focused at each atomic position in the co-ordinate list as shown in Figure 1. It is assigned a radius equal to the sum of the radius of the atom and that of the solvent molecule. According to the first algorithm proposed by Lee and Richards, SASA can be calculated using the following relations:

$$SASA = \left[\frac{R}{(R^2 - Z\_i^2)}\right] L\_i \cdot D, D = \frac{\Delta Z}{2} + \Delta' Z \tag{1}$$

where R is the radius of the sphere, L_i is the length of the arc drawn on a given section i, Z_i is the perpendicular distance from the centre of the sphere to the section i, ΔZ is the spacing between the sections, and Δ' is $\Delta Z/2$ or $R - Z_i$, whichever is smaller. Summation is all over the arcs drawn for the given atom.

The accessibility to the solvent can be calculated as follows:

$$\text{Accessibility} = \frac{100 \cdot \text{SASA}}{4\pi R^2} \tag{2}$$

The relative ASA (RSA) is the percentage of real ASA with respect to ASA of amino acids in the Ala-X-Ala extended state . SASA is generally measured in $Å^2$. The early computer programs were developed to compute the SASA of a number of model compounds prior to their application to proteins. The first model compounds were tripeptides of the form Gly-X-Gly and Ala-X-Ala, where 'X' is the residue whose SASA was to be computed. From his studies on shape and surface area of folded proteins, Gates arrived at several important conclusions:

- The SASA of a protein has been stated to be proportional to two-thirds of the real volume.
- It was also concluded that the shape is not important in estimating the solvent accessibility, and spheres are good models for protein subunits.
- When the volume of the ordered figure used to calculate the SASA is calculated geometrically, the accessible area becomes proportional to the 0.77 power of that volume.

SASA is a quantity of particular interest in protein folding and functional studies. It plays an important role in understanding the structure-function relationship of proteins and their residues. It is a well-known fact that burial of hydrophobic residues is a key factor in protein folding. Naturally, the exposure of these residues to the solvent and the hydrophobic core is directly related to the stability of the protein. Moret and Zebende analyzed the disparity of the SASA of amino acids in small fragments of proteins to investigate the hydrophobic effect of the residues. The estimation of the loss of SASA proved to be a measure of the probability of an amino acid to have a non-polar or polar side chair. Likewise, the environment free energy of amino acids depends on an accurate and rapid estimation of solvent accessible surface area . The stability of proteins is also governed by the heat capacity, enthalpy, and entropy changes, which are strongly associated with the change in SASA experienced by the polypeptide chain.

A precise prediction of tertiary structure of proteins has been one of the most challenging problems for biologists. The accuracy of predictions depends on a number of factors. It has been proposed that the relative solvent accessibility of residues might be an effective factor for increasing the accuracy of protein secondary and tertiary structure prediction. investigated the effect of the alteration of the solvent accessibility threshold on the accuracy of protein structure prediction. Kurt and Cavagnero investigated the influence of chain truncation on the disclosure of nonpolar solvent-accessible surface area (NSASA) for (a) unfolded state, represented as a fully extended chain, and (b) native-like folded state. Changes in protein stability upon point mutations have been analyzed in terms of Relative Solvent Accessibility (RSA) as classifiers for potentials. RSA can also be used to describe the physical and evolutionary properties of a protein. The concept of SASA has also been used to understand the protein-protein and protein-nucleic acid interactions. Protein-DNA recognition is significant for different cellular processes and the binding free energy of protein-DNA complex is related to the change in interface accessible surface area upon binding. Protein-protein interactions play a very important role nearly in all biological processes. analyzed the SASA of protein-protein interactions and established a method for predicting the interface surface area. carried out kinetic experiments to estimate the relation between SASA and bindings specificity of thrombin-thrombomodulin interface.

Analyzing the role of SASA to understand the structure-function relationship of protein-protein networks, geometric representations of proteins and ligands including SASA values can be applied to characterize interactions between and within proteins, ligands, and the solvent. These studies have revealed the potential use of SASA in the field of drug design and discovery. The SASA analysis can also be used as a probe to study the protein-DNA binding interface and protein mobility. used information on predicted SASA to analyze co-evolution based protein interactions that are quantified in the form of phylogenetic trees.

In view of the immense significance of SASA in the field of protein structure, function, and stability, our aim is to review various available methods for the calculation of SASA. In this review, we have focused on various types of

SASA and its application in the field of structural biology and protein biochemistry. Furthermore, we have described computational algorithms and prediction methods for calculating/predicting SASA from the crystal structure of a protein. Since proteins cannot crystallize in the unfolded conformation, such an estimation of SASA in this conformation is not possible experimentally. Therefore, various theoretical models for predicting SASA in the unfolded state have also been emphasized. We have also compared different values of SASA for a particular residue of the protein using different SASA prediction methods.

SASA can be broadly classified as (i) unfolded state SASA, which is represented as XASA, and (ii) folded state SASA, that is represented as FASA. Both types of SASA can be applied to the individual residues, domains, subdomains, and even to the whole protein molecule.

The absolute or numerical SASA for a residue/protein is estimated using atomic coordinates of the crystal structure (for that residue/protein) in Protein Data Bank (PDB) by different computational algorithms. The relative solvent accessibility for the ith amino acid (RSAi) is defined as the ratio of the absolute SASA of that residue observed in a given structure, denoted as SAi, and the maximum attainable value of the solvent-exposed surface area for this residue, denoted as MSAi. Thus, RSAi can have a value between 0 and 100, with 0 corresponding to a fully buried and 100 to a fully accessible residue, respectively.

$$RSAi(\%) = 100 \times \left(\frac{SAi}{MSAi} \right) \tag{3}$$

2 Algorithms to Calculate SASA in the Folded State

Various algorithms have been developed to calculate the Solvent Accessible Surface Area (SASA) in the folded state of proteins. Two commonly used methods are the Z-layer Integration Method and the Intersection Method.

2.1 Z-layer Integration Method

The Z-layer Integration Method calculates SASA by dividing the protein into layers along the Z-axis. Each layer is then integrated to compute the surface area accessible to the solvent. The method involves discretizing the protein surface into a grid of points and integrating the area of each grid cell that lies within a specified distance from the protein surface. By summing the areas of all accessible grid cells, the total SASA is obtained.

2.2 Intersection Method

The Intersection Method determines SASA by simulating the penetration of spheres representing water molecules into the protein structure. It involves

0.45
(a)
b

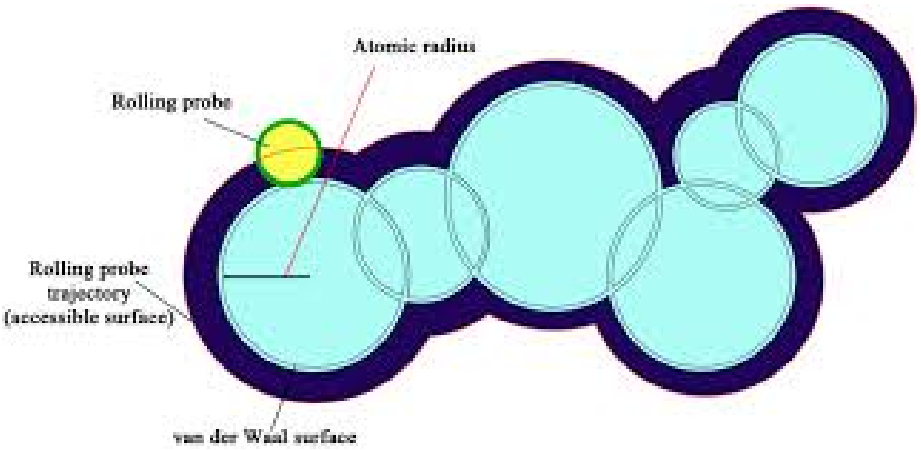

Figure 1: A cross-section of a part of a macromolecule in space rolling probe, van der Waals and accessible surface areas are indicated in the figure.

0.45
(a)
b

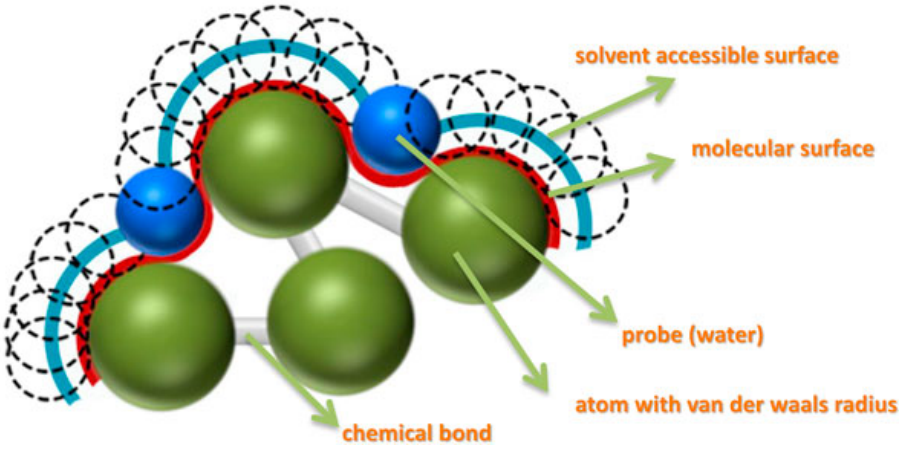

Figure 2: Three-dimensional representation of SASA of a protein on which water molecule is rolling (Green). The relative size of water molecule can be compared with nitrogen (blue) and oxygen (red).

generating a set of spheres around each atom in the protein, representing the maximum allowable solvent contact. The SASA is then calculated by determining the surface area of the protein that intersects with these spheres. This method provides a more accurate representation of the actual solvent-accessible surface of the protein.

Both of these methods have advantages and limitations depending on the complexity of the protein structure and the computational resources available.

2.3 Shrake and Rupley Algorithm

The Shrake and Rupley Algorithm is a widely used method for calculating SASA in the folded state of proteins. It utilizes a surface definition based on the rolling probe sphere model. The algorithm works by placing a probe sphere of a specific radius (usually corresponding to the size of a water molecule) on each atom of the protein and rolling it along the protein surface. The surface area accessible to the probe sphere without intersecting with other atoms is then calculated by discretizing the surface into small elements and summing their areas. By repeating this process for each atom in the protein, the total SASA is obtained.

The Shrake and Rupley Algorithm is known for its efficiency and accuracy in estimating SASA, making it a popular choice for computational studies of protein structure and function.

2.4 Linear Combinations of Pairwise Overlaps (LCPO) Method

The Linear Combinations of Pairwise Overlaps (LCPO) Method is a computational approach for calculating Solvent Accessible Surface Area (SASA) in the folded state of proteins. This method operates by decomposing the protein surface into a set of overlapping atomic spheres. The SASA is then computed by considering the pairwise overlaps between these spheres.

The LCPO Method employs a linear combination of pairwise overlap functions to approximate the solvent-accessible surface of the protein. By optimizing the coefficients of these functions, the method aims to accurately represent the SASA while minimizing computational complexity.

The LCPO Method offers advantages such as improved accuracy and efficiency compared to traditional grid-based methods. It has been successfully applied in various computational studies of protein structure and function, making it a valuable tool in structural biology research.

2.5 Power Diagram Method

The Power Diagram Method is a computational technique used for calculating Solvent Accessible Surface Area (SASA) in the folded state of proteins. This method operates by partitioning space into convex cells, each representing the influence region of an atom in the protein.

The Power Diagram Method utilizes the concept of power diagrams, which are extensions of Voronoi diagrams. In a power diagram, each cell is defined by a point (the atom center) and a weight (related to the atom's radius). By computing the intersections of these cells with the protein surface, the SASA is determined.

This method offers advantages such as accuracy and efficiency in representing the protein's surface. It has been employed in various computational studies to analyze protein structures and dynamics, contributing to our understanding of protein-ligand interactions and molecular recognition processes.

3 SASA Calculation for the Unfolded State

In the unfolded state of proteins, the calculation of Solvent Accessible Surface Area (SASA) poses unique challenges due to the lack of a well-defined tertiary structure. However, various computational methods have been developed to estimate SASA in the unfolded state.

3.1 Upper-Lower Bound Model

The Upper-Lower Bound Model is a theoretical approach used to estimate Solvent Accessible Surface Area (SASA) in the unfolded state of proteins. This model employs upper and lower bounds on the SASA values based on geometric considerations and statistical principles. By defining geometric constraints on the protein chain and incorporating statistical distributions of side-chain conformations, the Upper-Lower Bound Model provides estimates of SASA that are consistent with the unfolded state ensemble.

3.2 Statistical Model

The Statistical Model for SASA calculation in the unfolded state utilizes principles of statistical mechanics to predict the accessible surface area of unfolded protein chains. This model considers the conformational ensemble of unfolded states and employs statistical distributions to describe the distribution of solvent-accessible regions. By integrating parameters such as chain flexibility and excluded volume effects, the Statistical Model offers insights into the SASA distribution and dynamics in the unfolded state.

3.3 Monte Carlo Method

The Monte Carlo Method is a computational technique commonly used to estimate Solvent Accessible Surface Area (SASA) in the unfolded state of proteins. This method involves sampling the conformational space of unfolded protein chains using stochastic simulations. By randomly generating conformations of the protein chain and evaluating their surface accessibility, the Monte Carlo Method provides ensemble averages of SASA values. This approach allows for

the exploration of a wide range of protein conformations and provides statistical estimates of SASA in the unfolded state.

Despite the challenges associated with characterizing the unfolded state, computational techniques continue to advance our understanding of SASA in this context. By integrating experimental data with computational predictions, researchers can gain valuable insights into the structural and dynamic properties of unfolded proteins.

4 Tools for SASA Prediction

Several computational programs and tools have been developed for the prediction of Solvent Accessible Surface Area (SASA) in proteins. Among these, Artificial Neural Network (ANN), Support Vector Machine (SVM), and Markov Chain Model are commonly used methods.

4.1 Artificial Neural Network (ANN)

Artificial Neural Network (ANN) is a computational model inspired by the structure and function of biological neural networks. ANN-based approaches have been employed for SASA prediction by training neural networks on large datasets of protein structures with known SASA values. By learning the complex relationships between amino acid sequences and SASA, ANN models can accurately predict SASA for unseen protein structures.

4.2 Support Vector Machine (SVM)

Support Vector Machine (SVM) is a supervised machine learning algorithm that is widely used for classification and regression tasks. SVM-based methods for SASA prediction utilize kernel functions to map protein features into a higher-dimensional space, where a hyperplane is constructed to separate different classes of SASA values. SVM models have shown promising results in predicting SASA for proteins with diverse structural characteristics.

4.3 Markov Chain Model

Markov Chain Model is a probabilistic model that describes a sequence of events where the probability of each event depends only on the state attained in the previous event. In the context of SASA prediction, Markov Chain Models have been applied to analyze the structural dynamics of proteins and infer SASA values based on the transitions between different conformational states. These models capture the inherent stochastic nature of protein folding and provide insights into the fluctuations of SASA over time.

These programs and tools provide valuable resources for researchers to predict SASA in proteins, facilitating studies on protein structure, function, and dynamics.

S.No.	Program	Specification	URL	Developer
1	PDBePISA	Gives residue SASA for folded state of protein	http://www.ebi.ac.uk/msdsrv/prot_int/pistart.html	EBI (EMBL), U.K.
2	CCP4	Complete structure analysis suite	http://www.ccp4.ac.uk/	RcaH, STFC Rutherford A. Labs, U.K.
3	ProtSA	Sequence specific average SASA of unfolded ensemble	http://oldwebapps.bifi.es/protSA/	BIFI, Spain
4	GETAREA	Analytical calculation of SASA based on MC simulation	http://curie.utmb.edu/getarea.html	SCSB, University of Texas, U.S.A.
5	DSSP	Database of secondary structure assignments	http://swift.cmbi.ru.nl/gv/dssp/	CMBI, Nijmegen, Netherlands
6	NACCESS	Calculates atomic and residue ASA for proteins and nucleic acids	http://wolf.bms.umist.ac.uk/naccess/	University of Manchester, U.K.
7	ACCESS	Gives SASA of backbone and side chain atom of each residue	http://www.csb.yale.edu/	Yale University, U.S.A
8	POPS-R	Fast and analytical method, residue based approach for large assemblies like ribosomes	http://mathbio.nimr.mrc.ac.uk/~ffranca/POPS	NIMR, London, U.K.
9	SERF	Facilitates the use of SASA in structure analysis like changes during binding and complexation	http://guitar.rockefeller.edu/pub/jpo/serf.tar	DPMS, U.K.
10	ASAP	SVM based tool for calculating SASA of transmembrane residues	http://ccb.imb.uq.edu.au/ASAP/	University of Queensland, Australia
11	SABLE	Linear regression based method for RSA prediction	http://sable.cchmc.org	CHRF, Cincinnati, U.S.A

Table 1: List of computational programs available for SASA calculation.

5 Programs for SASA Calculation

5.1 PDBePISA

PDBePISA is an interactive tool developed by Krissinel and Henrick for exploring interfaces, structures, and assemblies of proteins. It is based on chemical thermodynamics and provides a single-button analysis of X-ray structures, including the assessment of multimeric state, symmetry number, protein-protein interfaces, accessible and buried surface areas of the residues and proteins . PDBePISA is available on the European Bioinformatics Institute (EBI) web server as a part of the European Molecular Biology Lab (EMBL) in the United Kingdom.

5.2 CCP4[1]

The CCP4 software suite is a collection of about 200 autonomous programs and software libraries used for estimating various physico-chemical and structural parameters based on X-ray crystallographic data of proteins. It is a community-based resource that supports researchers and academicians. The software is regularly updated, and the current release (v.6.3.0) can be downloaded freely . To calculate SASA, a separate program called AreaMol is available, which takes a .pdb file as input and provides SASA values for individual residues and proteins.

5.3 ProtSA

ProtSA is a web application that calculates sequence-specific SASA in the unfolded state . It can estimate the change in the accessible surface after protein folding. The server utilizes external software tools such as Flexible-Meccano for backbone conformation generation, SCCOMP for side chain building, and ALPHASURF for SASA calculations of each conformation of the unfolded protein ensemble. The output includes chain-wise description of average residue SASA, main chain SASA, side chain SASA, polar and non-polar SASA in both folded and unfolded states.

5.4 GETAREA

GETAREA is an analytical method developed by Fraczkiewicz and Braun for calculating SASA and its gradient for proteins. It finds solvent-exposed vertices of intersecting atoms, avoiding buried vertices that are not required to estimate the SASA. The CPU time for accurate determination of SASA has been significantly reduced compared to previous approaches. The input is the Cartesian coordinates of the protein stored in PDB format, and the output is SASA in different formats.

[1] Collaborative Computational Project, Number 4

5.5 Define Secondary Structure of Protein (DSSP)

DSSP is a widely accepted tool for estimating the SASA of individual residues of proteins . Originally developed by Kabsch and Sander in standard Pascal, it standardizes the secondary structure of proteins and sets up a database of SASA for secondary structure assignments. Recently, Hekkelman developed a new software, also called DSSP, which provides the same output as the original DSSP but is faster and easier to maintain.

6 Conclusions

Solvent Accessible Surface Area (SASA) has been considered as one of the most important physical parameters in protein science for the prediction of various physico-chemical and thermodynamic properties. In this review, we have defined the various types of SASA and their significance. Early and recent models to compute SASA of protein groups in the folded state have been explained in detail. It has been observed that the prediction of values of SASA of folded protein groups using various prediction methods is very accurate and model independent.

We have also described the various algorithms used to estimate the SASA of residues in unfolded states of proteins. Contrary to the folded state SASA, there does not exist any exact, accurate, and error-free method for SASA calculations of the unfolded proteins. Various online resources available for estimating SASA have also been listed.

The applications, uses, and correlations of SASA with various physicochemical and thermodynamic properties have been thoroughly reviewed. Some of these properties include molecular weight, hydrophobicity, radius of gyration, free energy changes during protein folding, transfer-free energies, intermolecular hydrogen bonding, partition coefficients, etc. Furthermore, the concept of solvent accessibility is very essential for finding protein-protein interfaces, analyzing the effects of mutations on protein stability, thermodynamic studies of proteins, discovery of de novo drugs, in silico molecular modeling, and protein engineering.