

A New Program to Compute the Surface Properties of Biomolecules

Dan V. Nicolau Jr.^{*}, Florin Fulga[†], Dan V. Nicolau Sr.[†]

Department of Mathematics and Statistics, Faculty of Science
University of Melbourne, Parkville 3010, Victoria

sarmisegetusa@yahoo.com

Industrial Research Institute
Swinburne University of Technology, Hawthorn 3122, Victoria

dnicolau@swin.edu.au, ffulqa@swin.edu.au

Abstract

The interactions of large molecules with surfaces and with each other are strongly dependent upon their surface, rather than their bulk properties. In addition, the local properties of biomolecular surfaces are very important in their own right in biomedicine and other areas, for example for locating binding sites. Following to previous work, we have developed a program to compute to compute amino acid and atom-based surface descriptors, and used it to generate a small database of charge and hydrophobicity-related surface properties for a set of proteins. The program requires the user to input two text files: one assigning a real number to each atom of each amino acid, and one assigning a real number to each amino acid. Although we have so far only computed surface charge (atom-based) and surface hydrophobicity (amino acid-based), we note that this program could be used to compute any surface parameter whatsoever, since the user can assign arbitrary atom-by-atom and amino acid properties. We discuss possible applications of this program and describe one current application, the Biomolecular Adsorption Database.

Keywords: molecular surface

1 Introduction

The geometric properties of molecular surfaces have received constant attention during the past two decades due to their importance to inter-molecular, and in particular, biochemical interactions. These properties include solvent-accessible surface area and surface roughness. In addition, a few other parameters, e.g. electrostatic surface potential can also currently be computed, for example using the program GRASP.

Information about the solvent-accessible surface can be used, for example, to make inferences about the interactions of molecules with each other and with surfaces, or to help in locating binding sites (Pettit & Bowie, 1999) and so on. However, despite their importance, non-geometric biomolecular surface properties have received very little attention; and the studies that do exist have remained for the most part qualitative.

On the other hand, breakthroughs in this area could shed light on other important problems. As a prime example, despite considerable attention over the last twenty years, the phenomenon of protein adsorption has resisted all attempts at predictive modelling. This is due to several factors, but arguably foremost among these is that molecular detail is usually ignored or oversimplified. Since most proteins are at the smaller end of the colloid spectrum, the bulk of the interaction between a protein and a surface can be attributed to the interaction of the protein surface with the solid. For example, both electrostatic and hydrophobic interactions occur largely between the two surfaces, and hence the systematic and quantitative investigation of the magnitude and distribution of such surface parameters should prove fruitful.

To this end, we have developed a set of algorithms to compute arbitrary local properties of molecular surfaces. The corresponding global parameters can be obtained by approximating the integral of the local parameter over the molecular surface. We divide the computational foundations of these algorithms into two categories. The first is "atom-based", where we assume that we can assign constant and homogeneously distributed properties to individual atoms of the same element; the second assumes that properties can be assigned to individual amino acids rather than atoms.

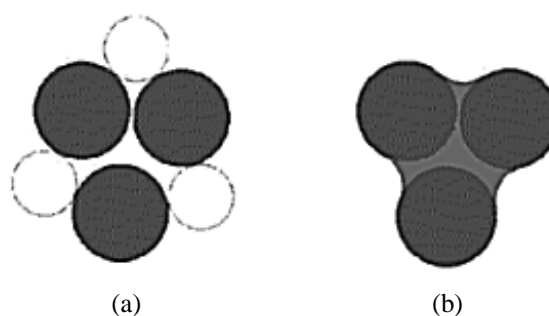


Figure 1. Illustration of Connolly's algorithm for generating the solvent-accessible surface; (a) the probe sphere is "rolled" over the cloud of balls representing the atoms, being placed tangentially to all pairs of atoms in the process (b) the solvent-accessible surface is the resulting locus of the probe sphere's contact point with the atomic spheres.

2 Computational Methods

We have already briefly described some of the ideas behind these algorithms with respect to the computation of the special cases of atom-by-atom charge and amino-acid based hydrophobicity in recent work (Nicolau and Nicolau, 2002).

In this section, we present general algorithms for the computation of arbitrary surface properties of molecules and biomolecules. These form the computational basis for our program.

2.1 Overview of Connolly's algorithm

Connolly's algorithm has been described in great detail elsewhere, for example in Connolly (1983) and we will only briefly describe it here. The rationale behind the computation of the "accessible" surface of a molecule is that of the part of the molecule that is exposed to the solvent, only those features that are at least the same size as a solvent molecule actually have interactions with the solution: smaller features are not "accessible" to the solvent.

Based on this idea, Connolly's algorithm rolls an imaginary sphere of a radius close to the effective size of a solvent molecule (in the case of water, this is 1.4Å) over the 3-dimensional structure of the molecule, which consists of a cloud of spheres (the constituent atoms) of appropriate radii (usually the van der Waals radius of each atom is used). The "pivot points" of the ball are recorded as it is rolled over the molecule, and together, these constitute an approximation to the solvent-accessible surface.

From this set of points, one can compute the surface area of the molecule, surface curvature and other features. The "ball-rolling" procedure is illustrated in Figure 1.

2.2 Computing atom-based surface properties

Here we concern ourselves with computing an arbitrary surface property Π_S (such as surface charge or surface hydrophobicity), constant values of which can be assigned, under reasonable assumptions, to each atom. Moreover, we assume that the property is concentrated entirely at the surface of the atomic sphere and distributed evenly over that surface.

Then let the solvent-accessible surface of a molecule be $\mathbf{S}(r)$ (where r is the radius of the probe sphere used). Clearly \mathbf{S} will be a three-dimensional vector, intuitively, an "empty shell". Consider a function $\pi: \mathbf{S} \rightarrow \mathfrak{R}$, which assigns to each point (x_S, y_S, z_S) of the surface \mathbf{S} a real number $\pi(x_S, y_S, z_S)$. Then we will compute the global surface property Π_S as

$$\Pi_S = \int_{\mathbf{S}} \pi(x_S, y_S, z_S) dS \quad (1)$$

where the integral will be taken over the surface \mathbf{S} .

In practice, we will approximate Π_S by assigning to each point generated by our algorithm a small area element, and summing the contributions of all the elements. In order to do this, we assign to each atom i a real number π_i , and scale it by the surface area of the atom in question. We have done this by computing the van der Waals area of the atom using $A_i = 4\pi r_{\text{vdW}}^2$, where r_{vdW} is the van der Waals radius of each atom. We obtained these radii from the paper by Gavezotti (1983).

The resulting approximation can be stated as

$$\pi(x_S, y_S, z_S) \approx \pi_i / A_i \quad (2)$$

where the relation holds only if the point (x_S, y_S, z_S) "belongs" to the surface of atom i . More precisely, a point "belongs" to a certain atom's surface if the distance between the point and that atom's centre is the smallest such distance.

This leads us to the approximation

$$\Pi_S \approx \sum_{\mathbf{S}} \frac{\pi_i}{A_i} \Delta A(x_S, y_S, z_S) \quad (3)$$

where $\Delta A(x_S, y_S, z_S)$ is the area of each surface element and the dependence of the coordinates indicates that there is a small variation in these areas due to the local surface curvature.

Therefore, our algorithm for determining Π_S can be stated as follows:

for each surface area element

determine to what atom it belongs

look up the value of π_i/A_i for this atom in a user-supplied table

record the value of π_i/A_i at this point in a file

increment Π_S by this value

The property that we have computed in this way is surface charge. First, we calculated formal charges for each atom of each isolated amino acid using the program HyperChem from Hypercube™, disregarding the effect of connecting amino acids together. Then, we calculated the surface area of all atom types using their van der Waals radii. Finally, we divided the two values and stored them in a table.

Our program was then used to compute several surface charge descriptors such as surface charge at each point, total positive charge at the surface, total negative charge at the surface, average charge per square Angstrom at the surface, and standard deviation of surface charge at the surface. This work is described in Nicolau and Nicolau (2002). We have also developed a scheme to assign hydrophobicities to each atom in this way (to be published).

We note that any property which can be assigned atom-by-atom values under reasonable circumstances could be computed in this way, as long as the assumptions that it is concentrated at the surface of the atom and is homogeneously distributed over that surface are also reasonable. In fact, it would even be possible to assign to each atom a complex (instead of real) or vectorial value, in which case the summation in (2) would have to be carried out two or more times, once for each independent component of the vector.

2.3 Computing amino acid-based surface properties

Some properties of proteins cannot be assigned to individual atoms, but can be assigned to the amino acids. For example, hydrophobicity scales usually attribute a hydrophobicity or hydrophilicity value to each amino acid (e.g. Kyte-Doolittle hydrophobicities). Therefore, we have added a variation of the above algorithm to our program, capable of computing an arbitrary amino-acid bases surface property of a molecule.

The algorithm is very similar to the atom-based one, but the approximation to π this time takes the form

$$\pi(x_S, y_S, z_S) \approx \pi_{aa} / A_{aa} \quad (4)$$

and holds true for all points (x_S, y_S, z_S) that belong to the surface of the amino acid denoted by the subscript "aa". In this case, a point "belongs" to the surface of an amino acid if it belongs (in the previous sense) to the surface of an atom that is part of that amino acid. A_{aa} , the area of the amino acid in question, was computed by us as the solvent-accessible surface area of that amino acid using a probe radius of 1.4 Å. This was done for all the amino acids, and must be used to scale the assigned value π_{aa} for the same reasons as above. Therefore, (3) now becomes

$$\Pi_S \approx \sum_S \frac{\pi_{aa}}{A_{aa}} \Delta A(x_S, y_S, z_S) \quad (5)$$

The modified algorithm is then

for each surface area element

determine to what amino acid it belongs

look up the value of π_{aa}/A_{aa} for this amino acid in a user-supplied table

record the value of π_{aa}/A_{aa} at this point in a file

increment Π_S by this value

So far, the amino-acid based property feature has only been used by us to compute hydrophobicities, following the Kyte-Doolittle scale (a default file is supplied with the program). However, any property could again be assigned to the amino acids (for example, adsorption on a chromatographic column).

3 Description of program

We have given our program the acronym PSPC (Protein Surface Properties Calculator), although of course its use is not restricted to proteins or even to biological molecules. It was written in Visual Fortran and runs under Windows™ environments.

We briefly describe the main features of the program in terms of input and output characteristics and a few remarks on running time and space requirements. A screenshot of the program window is shown in Figure 2.

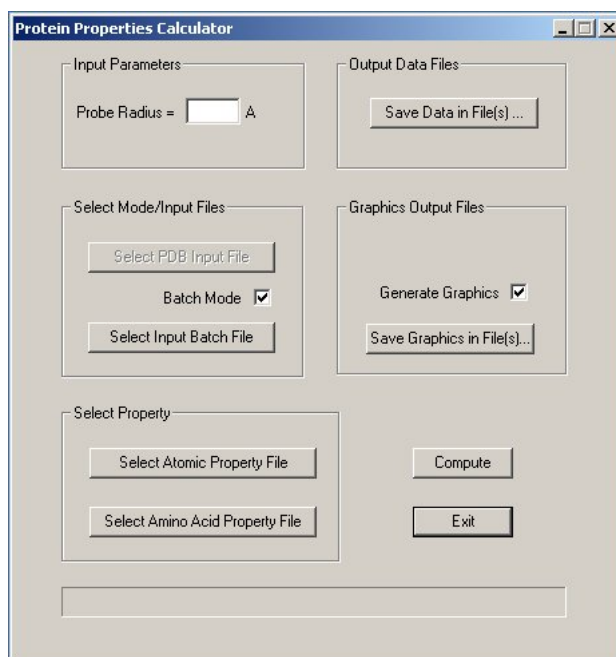


Figure 2. Screenshot of the program window.

3.1 Description of input

3.1.1 Probe Radius

This is the radius, in Angstroms, of the sphere used to probe the molecular surface. This parameter must be entered in the Probe Radius input text box at the top of the window. Note that a sphere representing a water molecule has a radius of about 1.4 Å, and this value is usually used to compute the solvent-accessible surface of molecules in aqueous solution.

3.1.2 Selecting Mode/Input files

Here the user can enter a file, or set of files to perform the calculations on. If Batch Mode is not checked, then one must select the file of interest. This file must be in either PDB/ENT (.pdb/.ent) format or in Alchemy (.hin) format. If Batch Mode is checked, then one must select a text file containing the names of the PDB files of interest (one name per line). For example, to perform calculations on the three files I1GT.ENT, I1LYZ.ENT and I1MYO.ENT, the batch file would contain

IIGT
ILYZ
IMYO

3.1.3 Selecting a Property

Here the user can specify the files containing the atom-based and amino-acid based properties to be used in calculations. For atom-based properties, this can either be an atomic charge file (*.chg), an atomic hydrophobicities file (.hph), or in principle any file in which a real number is assigned to each atom of each amino acid, in the format:

```
Amino-Acid 3-letter code, Atom Code,  
Property/van der Waals area of atom  
  
Amino-Acid 3-letter code, Atom Code,  
Property/van der Waals area of atom  
  
...
```

A file named aminochg.chg, included with the program, is the default atomic charge file and is an example of this format. It assigns to each atom of each amino acid, a value equal to the charge on that atom divided by the atom's van der Waals surface area. Atom-based hydrophobicities follow an identical format, with each atom being assigned a value equal to its predicted hydrophobicity/hydrophilicity contribution divided its van der Waals area. A default file is supplied with the program in this case also. In general, the property assigned to the atom must be always divided by the atom's surface area, since the program computes surface properties by summing the contributions of small area elements on the molecular surface, as described above.

In similar fashion, the user can select the amino-acid based property file, in which a real number is assigned to each amino acid instead of each atom. The format is:

```
Amino-Acid 3-letter code, Property/Surface  
Accessible Area of Amino Acid  
  
Amino-Acid 3-letter code, Property/Surface  
Accessible Area of Amino Acid  
  
...
```

For the same reasons as above, the property must again be divided by the area of the amino-acid (we have used the solvent-accessible areas of the amino acids, computed using a probe radius of 1.4 Å -- this file will be supplied on request).

3.2 Description of Output

3.2.1 Output Data Files

The primary output is a text file, containing the numerical results of the run, including both atom-based and amino-acid based property results, together with some information about the input file. If the program is running in Batch Mode, then more than one primary output file will be generated.

3.2.2 Generating Graphics.

Three sets of files will be generated if this feature is used. The first is a point-by-point table of the atomic-based property; the second is a point-by-point table of the amino-acid based property. Finally, two MATLAB™ files will be generated, allowing the user to view the graphics generated. Running these in MATLAB™ will produce a 3-D, 2-color, rotatable surface showing red for a positive property and blue for a negative property. This can be used to qualitatively answer questions such as "is this molecular surface like a zebra or like a leopard, with respect to this property?"

Finally, a progress bar at the bottom of the window provides a rough indication of the current status of the computation. Running times vary greatly with input file size, molecular shape (globular molecules take less time than elongated ones) and probe radius. To a first approximation, the time and space complexities are quadratic in probe radius and linear in the number of atoms, as expected.

An average run time for a small molecule, such as crambin, using a small probe radius, such as 1.4 Å, is no more than 30 seconds on an average desktop machine running Windows NT, for example. On the same machine, running a computation on a large molecule such as IgG, using a large probe radius, say 20 Å, can take as long as 30 minutes, if graphics are also generated. In general, the program is quite demanding in terms of both CPU and memory usage, and takes a large share of the system resources, even on a fast machine.

Although initial space requirements are very small (the program itself, together with a few auxiliary files occupies less than 200 KB of hard disk space), the temporary file generated during the run (and deleted afterwards) can reach a few MB in size, and if graphics are generated, each of the six graphics files can also be as large as a few MB. Consequently, we recommend that it be run on a machine with at least 10-20 MB of free space, and if Batch Mode is used, then this should be scaled accordingly.

4 Discussion

We have purposefully built a large degree of flexibility into this software: the user can and is encouraged to not only supply custom atomic and amino acid property data files, but he or she can actually

determine what property is to be computed in each case. The limits of this flexibility are only determined by the availability of per-atom or per-amino acid data related to the property of interest, and of course the validity of assigning numerical properties in the manner described here. In keeping with this spirit, the eventual applications of this software are also left to the technical interests and area of expertise (and the imagination) of the user.

This is one of several programs that can be used to investigate molecular surfaces; GRASP and STING are among the better-known of these — the former in particular is used extensively for this task. Several important differences exist between our program and these. GRASP, STING and similar programs are entire molecular visualisation and analysis packages, with functionality split more or less evenly between these very general tasks. The focus on visualisation and/or the broad scope of these packages mean that the molecular surface analysis functionality is fixed and limited. Even with GRASP, only a small number of properties can be calculated, and there is very little flexibility in computing these.

The program we present here is dedicated entirely to the analysis of molecular surfaces, and while it is possible to produce images, the intended purpose of these is to aid the user in interpreting the data produced, rather than on visualisation itself. On the other hand, every effort has been made to give the user as much latitude as possible in surface analysis. Not only is it possible to vary the probe radius and the input property files, but both “atom-based” and “amino-acid-based” can be computed simultaneously. Another difference from general-purpose software is that SPPC returns a large number of parameters for each property computed. Possibly the greatest degree of flexibility comes from the ability to compute *any* surface property whatsoever (for which the user can supply the needed input files). Although the immediately obvious use of the program is the calculation of surface charge and hydrophobicity and input files are supplied for these tasks, the true power of the software would come from the computation of arbitrary surface properties. Finally, computations on several or many molecules can be run “in batch mode”.

To exemplify the above comparisons, we will briefly describe three possible applications of the program. The first is the calculation of the part of an exposed molecular surface that is due to a certain amino acid or group of amino acids. All the user needs to do to achieve this is to assign a 1 to the amino acid(s) of interest and a 0 to all others. An example input file might be:

```
ALA, 1
LYS, 0
MET, 0
... (all others 0)
```

The primary output file would contain the exposed area, in \AA^2 , which can be attributed to Alanine.

The same can be done to discover what part of a molecular surface can be attributed to a particular element or set of elements; in this case, it is the atomic property file that would contain the ones and zeroes. In fact, it would even be possible to find out if and to what extent one particular atom or some other part of one particular amino acid, e.g. the nitrogen or COO⁻ group in Alanine is exposed, in the same way.

A second possible application is the determination of the fractal dimension or surface roughness of molecular surfaces. The fractal dimension of a two-dimensional molecular surface can be defined by

$$D = 2 - \frac{\partial \log A}{\partial \log r} \quad (5)$$

where D is the fractal dimension, A is the surface area of the molecule and r is the radius of the probe sphere. Since the derivative is generally smaller than 0, the fractal dimension is larger than 2 — an indication of the fractal nature of protein surfaces.

Although it was previously possible to compute the fractal dimension (which is an indication of the surface roughness of a molecule) of the molecular surface geometry, using our program one can study the variation (and possibly fractal nature) of any surface property with probe radius. For example, preliminary investigations suggest that the properties of protein surfaces exposed to water-sized probe spheres are vastly different quantitatively, but also qualitatively, to those exposed to larger probe spheres, representing perhaps other molecules or other solvents. Investigations of this nature should prove of some service in molecular biology and bioinformatics.

Finally, we will mention a more pragmatic and bioinformatics-related application. We have used these algorithms to compute the following surface properties for a set of some tens of proteins of interest in protein adsorption to solid surfaces (Nicolau and Nicolau, 2002): surface area, positive/negative surface areas, total positive/negative surface charges, average surface charge, standard deviation of surface charge, total hydrophobic/hydrophilic surface areas, total surface hydrophobicity/hydrophilicity and average surface hydrophobicity.

This data was compiled into a small database (the Biomolecular Descriptors Database). We used this, together with our Biomolecular Adsorption Database (B.A.D.) — which contains experimental data describing protein adsorption to solid surfaces from solution, to build semi-empirical models of protein adsorption from solution (Nicolau and Nicolau, 2002). This database is available at www.bionanoeng.com/bad/.

Finally, we note that although the intended application areas for this software are bioinformatics and molecular biology, in principle these algorithms can be applied to any structures which can be represented in the

same way as molecules are in the basic PDB format — as a set of spheres in three dimensions, the coordinates of whose centres and whose radii are given. Discussion of this possibility is beyond the scope of the present work.

5 Conclusion

We have developed general algorithms for the computation of arbitrary atomic and amino-acid based properties of molecular and especially biomolecular surfaces. These have been implemented in a Windows™ program, which we call the Protein Surface Properties Calculator (PSPC). The algorithms in question are described, as are the functionalities of the program. We also give some examples of possible applications. The program can be downloaded from the website www.bionanoeng.com/.

6 References

- NICOLAU D.V. Jr and NICOLAU D.V. (2002) A Database Comprising Biomolecular Descriptors Relevant to Protein Adsorption on Microarray Surfaces, *SPIE Proceedings*, San Jose, USA, **4626**: 109-116
- NICOLAU D.V. Jr and NICOLAU D.V. (2002) A Model of Protein Adsorption to Solid Surfaces from Solution, *SPIE Proceedings*, San Jose, USA, **4626**: 109-116
- PETTIT, F. and BOWIE, J.U. (1999): Protein Surface Roughness and Small Molecular Binding Sites, *J. Mol. Biol.* **285**: 1377-1382.
- CONNOLLY, M.L. (1983): Solvent-accessible Surfaces of Proteins and Nucleic Acids, *Science*, **221**: 709-713
- CONNOLLY, M.L. (1983): Analytical Molecular Surface Calculation, *Journal of Applied Crystallography*, **16**: 548-558.
- CONNOLLY, M.L. (1986): Measurement of Protein Surface Shape by Solid Angles, *Journal of Molecular Graphics*, **4**: 3-6.
- NICOLAU, D. Jr and NICOLAU, D. Biomolecule adsorption database. <http://www.bionanoeng.com/bad/>
- PAKULA, A.A. and SAUER, R.T. (1990): Reverse hydrophobic effects relieved by amino-acid substitutions at a protein surface, *Nature* **344**: 363-364.
- KRAULIS, P.J. (1991) MOLSCRIPT: a program to produce both detailed and schematic plots of protein structures, *J. Appl. Crystallogr.* **24**: 946-950.
- ZAMYATNIN, A.A. (1984) Amino acid, peptide, and protein volume in solution. *Annu. Rev. Biophys. Bioeng.* **13**, 145-165.