Software

ProtSA: a web application for calculating sequence specific protein solvent accessibilities in the unfolded ensemble

Jorge Estrada1,2, Pau Bernadó*3, Martin Blackledge4 and Javier Sancho*1,2

Address: 1Departamento de Bioquímica y Biología Molecular y Celular, Facultad de Ciencias, Universidad de Zaragoza, 50009 Zaragoza, Spain, 2Biocomputation and Physics of Complex Systems Institute (BIFI), Universidad de Zaragoza, 50009 Zaragoza, Spain, 3Biomolecular NMR, Institute for Research in Biomedicine, Parc Científic de Barcelona, 08028 Barcelona, Spain and 4Institut de Biologie Structurale Jean-Pierre Ebel, 38027 Grenoble, France

Email: Jorge Estrada - jorge.estrada@unizar.es; Pau Bernadó* - pau.bernado@irbbarcelona.org; Martin Blackledge - martin.blackledge@ibs.fr; Javier Sancho* - jsancho@unizar.es

* Corresponding authors

Abstract

Background: The stability of proteins is governed by the heat capacity, enthalpy and entropy changes of folding, which are strongly correlated to the change in solvent accessible surface area experienced by the polypeptide. While the surface exposed in the folded state can be easily determined, accessibilities for the unfolded state at the atomic level cannot be obtained experimentally and are typically estimated using simplistic models of the unfolded ensemble. A web application providing realistic accessibilities of the unfolded ensemble of a given protein at the atomic level will prove useful.

Results: ProtSA, a web application that calculates sequence-specific solvent accessibilities of the unfolded state ensembles of proteins has been developed and made freely available to the scientific community. The input is the amino acid sequence of the protein of interest. ProtSA follows a previously published calculation protocol which uses the Flexible-Meccano algorithm to generate unfolded conformations representative of the unfolded ensemble of the protein, and uses the exact analytical software ALPHASURF to calculate atom solvent accessibilities, which are averaged on the ensemble.

Conclusion: ProtSA is a novel tool for the researcher investigating protein folding energetics. The sequence specific atom accessibilities provided by ProtSA will allow obtaining better estimates of the contribution of the hydrophobic effect to the free energy of folding, will help to refine existing parameterizations of protein folding energetics, and will be useful to understand the influence of point mutations on protein stability.

Background

A detailed understanding of protein folding energetics is fundamental for ab initio prediction of protein 3-D structures from sequences, for the rational engineering of new proteins, and for understanding diseases related to protein misfolding or aggregation [1,2]. The unfolded state of proteins is central in developing the theoretical framework of folding processes because it represents the starting point from which proteins evolve to the native state. The hydrophobic effect operating on apolar side chains is an important factor driving protein folding [1], and the change in solvent accessible surface area (SASA) of a pro-
tein upon folding can be used to estimate the contribution of the hydrophobic effect to the free energy of folding [1,3]. Empirical models relate changes in SASA (total, polar or apolar) upon folding to the heat capacity, enthalpy, or entropy of folding, and to equilibrium m-values in chemical unfolding [4,5].

The SASA of a protein was defined [6] as the surface described around the protein by the centre of a solvent sphere in contact with the van der Waals surface of the molecule. Experimental determination of accurate SASAs of folded proteins at the atom level is not yet possible. Fortunately, computation of SASA values in the native state is straightforward when 3-D structures are available. Accurate SASA values for the unfolded state are not only difficult to determine but also difficult to calculate. In attempts to calculate the changes in SASA associated to the protein folding reaction, a variety of models of the unfolded state have been proposed. They include tripeptides [7-9], peptide-fragment collections in both native and extended conformations, extracted from a set of native structures [10,11], ensembles of Ac-(Ala)_{3-}X-(Ala)_{3}-Nme peptides [12], and ensembles of polypeptide conformations of a specific selected protein [13]. A common characteristic of all these models, but the last one, is that they provide mean solvent accessibilities for the 20 residue types, but they do not take into account the possibility that these accessibilities are modulated by the specific sequence context of the residue of interest.

We have recently developed a way to estimate SASA at atomic resolution in the unfolded ensemble. The method provides individual SASAs for each atom of each residue in a given protein sequence [14]. The structural model chosen to describe the unfolded state consists of hundreds to thousands of unfolded conformations generated by Flexible-Meccano, an algorithm that performs conformational sampling using a coil-library and a simple volume exclusion term [15]. The ensembles generated in this way successfully describe backbone fluctuations of several intrinsically unfolded proteins probed by Nuclear Magnetic Resonance (NMR) and small-angle X-ray scattering (SAXS) [15-19]. Our analysis of solvent exposures in unfolded ensembles of proteins generated with this method clearly indicates that the SASA of any residue is strongly influenced by its sequence neighbours [14] and, therefore, using generic residue-type values is not justified. A detailed benchmarking of the method has been described [14].

Here, we present a web application that calculates SASA of protein unfolded-state ensembles, detailed per residue and atom, using the methodology described [14]. As far as we know, only two related servers exist. BPPred [20] calculates, from the number of residues, an overall protein change of SASA upon folding. Unfolded implements the approach by [11], which is based in generic residue-type values. None of these two servers calculates SASA values on a sequence specific representation of the unfolded-state ensemble of the protein of interest. In this sense, ProtSA is an innovative web application that will provide researchers with more accurate accessibility data for the parameterization and interpretation of protein folding thermodynamics.

Implementation
ProtSA architecture consists of three parts: the user web browser, a middle tier Common Gateway Interface (CGI) application running on a web server, and the server part that calculates SASA of the protein unfolded-state ensemble. The server part uses three external software programs to perform the calculations: Flexible-Meccano for backbone-conformation generation, SCCOMP [21] for side chain building, and ALPHASURF [22] for SASA calculations of each conformation of the unfolded ensemble of the requested protein. The interaction between the ProtSA parts is as follows: the user fills in the input form using the web browser; the browser sends the input data to the CGI application, which checks its completeness and validity, and redirects complete and valid requests to the server part; the server part is a multithreaded program, with one network thread for receiving requests, and several worker threads for processing requests (one request per thread); the network thread receives the request from the CGI application, checks for resource availability and replies to the CGI application with an acceptance or refusal message; the CGI application informs the user whether the request is accepted or not, with the reason for refusal in the latter case; if resources are available, the network thread in the server part queues the request and, when a worker thread becomes available and no earlier requests are queued, that worker thread processes the request, calculating SASA of the protein unfolded-state ensemble; finally, the worker thread emails the results to the user. Both the CGI application and the server part were programmed using C++.

ProtSA basically follows the method shown in [14] for calculating SASA of a protein unfolded-state ensemble, though ProtSA uses ALPHASURF instead of NACCESS [23] for the calculations of each unfolded-state protein conformation. ALPHASURF was chosen because it uses an exact analytical method (based on the alpha shape theory) and is free software; the results section shows that ALPHASURF and NACCESS give very similar results. The steps of the ProtSA method are:

1. Check that all residues in the protein sequence belong to the set of 20 standard types. If the user provides a 3-D structure, check it for gaps or missing atoms.

2. Generate, from the protein sequence, a set of unfolded-state backbone-only conformations using Flexible-Meccano.
Flexible-Meccano’s Monte Carlo algorithm for generating the backbone of the unfolded-state conformations uses a subset of the database of amino-acid-specific Φ- and Ψ-torsion angles described in [24]; the subset is obtained by exclusion of all residues in α-helices and β-sheets. The database includes symmetric values for glycine Φ- and Ψ-torsion angles, and has special cases for residues preceding a proline. For each protein unfolded-state conformation the algorithm constructs the backbone starting at the C-terminal, although it has been shown that building directionality does not influence SASA results [14]. Residue i is connected to residue i+1 by selecting a random pair of Φ- and Ψ-angles, for the type of residue i, from the torsional subset database. If residue i presents clashes with other residues (where residues are represented as spheres centred at the Cβ atom -the Ca atom for glycine residues-using radii derived from Levitt’s force-field [25]), the Φ- and Ψ-torsion-angle pair is rejected, and another one is randomly selected. If, after 500 tries, the algorithm does not find a non-clashing Φ- and Ψ-torsion-angle pair, the partially-built conformation is rejected and the algorithm starts again at the C-terminal residue.

A key factor to the sequence-specificity of SASAs calculated for unfolded ensembles is the decoration of each polypeptide backbone with energetically realistic conformers of the sequence residues. This is performed using the iterative method implemented in S RCMP. Using the rotamers of a backbone-dependent library, and a backbone independent one for special locations in the protein chain (such as the first and last residues), S RCMP assigns rotamers, residue by residue, optimizing a scoring function with terms accounting for atom-atom contacts, steric overlaps, torsion energy, and the hydrophobic effect. S RCMP repeats the complete assignment of rotamers to the protein residues until either there is no change in structure in two consecutive iterations or the limit of allowed iterations is reached.

**Results**

The original method described to calculate solvent accessibilities in unfolded ensembles [14] used NACCESS, while the ProtSA application relies in ALPHASURF. We have compared the performance of these two methods by recalculating unfolded solvent accessibilities for the set of 19 proteins used in the original implementation. Another popular program to calculate exposures, DSSP [26], yields values 5% higher than those of NACCESS and ALPHASURF (not shown).

The results of the new calculations performed with ALPHASURF are shown in Table 1 compared with those obtained with NACCESS and previously reported [14]. The two algorithms provide very similar exposures for the same protein with overall SASA values differing less than 0.36%. The average, minimum and maximum SASA accessibilities found for each residue type within the unfolded ensembles of the 19 proteins are shown in Table 2. Differences with the original data reported [14] are also minimal. For average residue SASAs, the biggest difference (0.53%) is for methionine (Table 2 and data in [14]), and the mean of the differences observed for all the residues using the two methods is 0.19%. Similarly, for the minimum value of SASA found for each residue type within the 19 ensembles, the biggest difference is at 2.04% for one specific threonine residue, with a mean of 0.77% for the twenty residue types. For the maximum SASA values for residue types, the biggest difference is 3.15% for one specific glutamic acid residue (mean difference of all maximally exposed residues being 0.81%). The main utility of ProtSA calculations is that they can highlight strong divergences in the exposure of specific residues from the average values exhibited by their corresponding residue types in the unfolded ensemble (Table 2). These divergences are sequence context dependent and can only be revealed with sequence specific calculations.

ProtSA is available at [27]. The input web form in ProtSA is very simple (Fig. 1). The user can supply a protein sequence, a PDB-formatted file, or a PDB id. The user must also specify the number of protein conformations to generate, and the radius of the solvent probe. Specification of probe radius may be used for calculating surface accessibility to different ions, not just water molecules. The user also specifies the email address where ProtSA will mail the results. When the user supplies a protein sequence (which must be a single-chain one), ProtSA calculates only the SASA for the unfolded ensemble. When the user supplies a PDB file or a PDB id (which ProtSA uses to fetch the corresponding PDB file from the Protein Data Bank [28]), ProtSA also calculates the SASA for the 3-D structure, which is assumed to represent the folded state. ProtSA emails the user the calculated results. For each atom and residue, the results include the average sequence-specific SASA in the unfolded ensemble and, if it was calculated, the SASA in the folded state, and the difference (SASA_{folded} - SASA_{unfolded}).

To highlight those residues with unusually high or low exposures in the unfolded ensemble relative to typical values (calculated as the average exposures in the 19 test pro-
Table 1: ProtSA solvent accessibilities of unfolded ensembles of test proteins

| Protein PDB code | Number of residues | Accessibilities by ProtSA | Accessibilities by NACCESS |
|------------------|--------------------|---------------------------|----------------------------|
| PDB:1LN4        | 98                 | 10520                     | 10497                      |
| PDB:1T1D        | 100                | 11170                     | 11136                      |
| PDB:1BKR        | 109                | 11683                     | 11666                      |
| PDB:1BGF        | 124                | 13843                     | 13808                      |
| PDB:1B3T        | 131                | 14300                     | 14261                      |
| PDB:2L1S        | 136                | 15320                     | 15302                      |
| PDB:1QGY        | 142                | 15648                     | 15632                      |
| PDB:1EY4        | 149                | 16437                     | 16430                      |
| PDB:1EPD        | 185                | 20568                     | 20529                      |
| PDB:1L3K        | 196                | 21442                     | 21436                      |
| PDB:1BY1        | 224                | 23476                     | 23393                      |
| PDB:1E59        | 232                | 25251                     | 25250                      |
| PDB:1IIS        | 233                | 25330                     | 25258                      |
| PDB:1WER        | 334                | 37937                     | 37845                      |
| PDB:1FCQ        | 348                | 39539                     | 39492                      |
| PDB:1FQ9        | 350                | 39941                     | 39877                      |
| PDB:1ES5        | 416                | 43551                     | 43429                      |
| PDB:1G3SO       | 431                | 45765                     | 45628                      |
| PDB:2BCE        | 579                | 62744                     | 62631                      |

1Solvent accessibilities (in Å²) calculated with ProtSA, using ALPHASURF. They are averaged over 2000 unfolded structures of each protein. The first and the last five residues of each sequence are not taken into account.
2Solvent accessibilities (in Å²) calculated with NACCESS in otherwise identical conditions (taken from Table 3 in reference [14]).

Table 2: Solvent accessibilities (Å²) of amino acid residues in protein unfolded ensembles calculated with ProtSA

| Residue | Number of Residues ¹ | Average ² | Minimum ³ | Maximum ⁴ | % Difference ⁵ |
|---------|-----------------------|-----------|-----------|-----------|---------------|
| Ala     | 349                   | 73.2 (73.1)| 58.2      | 84.2      | 31            |
| Arg     | 233                   | 178.9 (178.6)| 155.3    | 192.8      | 19            |
| Asn     | 198                   | 109.2 (109.1)| 91.1     | 121.5      | 25            |
| Asp     | 255                   | 102.2 (102.0)| 83.8     | 117.2      | 29            |
| Cys     | 51                    | 88.7 (88.3)| 76.7      | 98.2       | 22            |
| Glu     | 287                   | 126.0 (125.9)| 108.9    | 140.9      | 23            |
| Gin     | 171                   | 125.9 (125.6)| 108.6    | 141.6      | 23            |
| Gly     | 312                   | 54.3 (54.2)| 36.6      | 65.6       | 44            |
| His     | 115                   | 129.5 (129.3)| 109.0    | 140.0      | 22            |
| Ile     | 229                   | 122.5 (122.2)| 107.2    | 135.5      | 21            |
| Leu     | 407                   | 131.9 (131.5)| 110.3    | 147.7      | 25            |
| Lys     | 247                   | 149.9 (149.8)| 131.2    | 167.2      | 22            |
| Met     | 102                   | 134.3 (133.6)| 122.0    | 149.1      | 18            |
| Phe     | 174                   | 146.1 (146.1)| 130.7    | 163.2      | 20            |
| Pro     | 217                   | 100.3 (100.0)| 81.8     | 123.4      | 34            |
| Ser     | 198                   | 76.0 (75.8)| 59.3      | 90.5       | 34            |
| Thr     | 245                   | 93.3 (93.2)| 79.7      | 107.6      | 26            |
| Trp     | 70                    | 173.2 (173.0)| 161.8    | 185.1      | 13            |
| Tyr     | 148                   | 156.9 (156.8)| 140.1    | 173.2      | 19            |
| Val     | 319                   | 102.2 (102.0)| 84.8     | 115.3      | 26            |

1Total number of residues of that kind found in the 19 protein sequences simulated.
2Residue-specific solvent exposure averages. The numbers in parenthesis were calculated with NACCESS (see [14])
3Minimum solvent exposure found in one of the 19 denatured ensembles.
4Maximum solvent exposure found in one of the 19 denatured ensembles.
5Percentage difference between maximum and minimum solvent exposures found for one residue type: 100(max-min)/max
Input form

Please, fill in all fields. Only ASCII printable characters are allowed.

Title for this job
Send file:
PDB id
Protein sequence (1 chain only)
Solvent radius (Å) (greater than 0.0, 1.4 is a usual value)
Unfolded conformations to generate (between 1-2000; 2000 suggested, see below)
E-Mail address to return results

Send the request | Clear form

Figure 1
Input form of the ProtSA web application. Through this simple input form the user can identify the request with a title, and submit the protein information in three different ways (as a PDB file, as a PDB id, or as a chain of residues in text form). The user sets the solvent radius and the number of unfolded conformations to generate for each chain in the protein. ProtSA sends the results of the calculations to the email address of the user.

Figure 2
Colour-coded sequence for the unfolded SASA of [PDB:5cro], calculated with ProtSA. Almost all residues are in the range 90%–110% of the average unfolded SASA for its residue type (from a set of 19 proteins). As expected, residues in the extremes of the chains are more exposed than the average. The data shown is for 1.4 Å of solvent radius and 2000 conformations. Tyr 26 is labelled.
Figure 3
Colour-coded sequence for the folded SASA of [PDB:5cro], calculated with ProtSA. The specific degree of burial in the folded state of a given residue type varies widely. Tyr 26 is clearly more exposed than the average tyrosine residues present in the test set of 11 proteins (see text), which could be suggestive of a possible 'reverse hydrophobic effect'. This effect can be discarded by inspection of figure 4a. The data shown is for 1.4 Å of solvent radius and 2000 conformations. Tyr 26 is labelled.

Figure 4
Colour-coded sequences of the ratio of folded and unfolded SASA for [PDB:5cro] and [PDB:1ucj]. Upon folding, most residues tend to be less exposed to solvent than in the unfolded state, as shown by the majority of reddish residues in the figure. However, some residues may be more exposed in the folded state (bluish residues). a) Tyr 26 (labelled) of [PDB:5cro] has a ratio of folded SASA over unfolded SASA in the range 0.9–1.1, suggesting that no reverse hydrophobic effect is taking place (see text). b) Thr 36 (labelled) of [PDB:1ucj], as most residues, is also in a reddish shade. Since this residue bears a polar side chain, its burial is suggestive of a destabilizing contribution to the native conformation. The data shown in both cases is for 1.4 Å of solvent radius and 2000 conformations.
(figure 3) reveals that the exposure of this tyrosine residue in the folded state is well above average, which could have influenced the initial interpretation. The second example refers to Thr 36 in the V36T mutant of RNase Sa. Compared to the wild type protein, the presence of a threonine at position 36 destabilizes the folded state [31]. ProtSA depicts Thr 36 in the ratio sequence on an orange background (Fig. 4b), which visually indicates this residue loses a large percentage of its solvent exposure upon folding, and is expected to destabilize the folded conformation. Incidentally, the ratio sequence for this protein shows that additional polar residues appear more buried in the folded conformation than in the unfolded ensemble. To assess whether they are likely to destabilize the native conformation their local environments should be analysed. They may establish appropriate compensating polar interactions or else they should be considered as potentially destabilising residues.

From the execution times of the 19 test proteins in a computer running CentOS 5, with 2 GB of RAM and a Core 2 Duo-2.4 GHz CPU (data not shown) we can deduce a linear dependence on the size of the input sequence. For the more demanding requests corresponding to ensembles of 2000 protein molecules there is a fix cost of about 80 minutes and a variable cost of about 0.5 minutes per residue in the sequence.

**Conclusion**

ProtSA, the freely-available web application presented in this work, represents a novel tool for the researcher interested in protein folding energetics. The sequence-specific protein solvent accessibilities in the unfolded state ensemble calculated by ProtSA will provide researchers a more precise view of unfolded state ensembles, and will help to understand the influence of mutations on protein stability.

**Availability and requirements**

Project name: ProtSA

Project home page: http://webapps.bifi.es/protsa/

Operating system(s): Platform independent

Programming language: C/C++

Other requirements: JavaScript-enabled web browser

Any restrictions to use by non-academics: None

**Authors’ contributions**

JE developed ProtSA. PB and JS directed the design. MB created software parts. All authors wrote the manuscript, and read and approved the final version.
Acknowledgements

We thank Sara Ayuso (Univ. Zaragoza, Spain), for help in testing; Patrice Koehl (Univ. California Davis, USA), for help with ALPHASURF; Guillermo Losilla (BIFI, Univ. Zaragoza, Spain), for technical help and José Ramón Peréríga (Univ. Zaragoza, Spain), for comments and discussions. The molecular graphics image in Fig. 5 was produced using POV-Ray [32] and the UCSF Chimera package [33] from the Resource for Biocomputing, Visualization, and Informatics at the University of California, San Francisco (supported by NIH P41 RR-01081). We acknowledge financial support from MEC (Spain): grant BFU2007-61476/BMC and BIO2007-63458, and from DGA (Spain): grant PI07/08. PB holds a Ramón y Cajal contract financed by MEC (Spain) and the Institute for Research in Biomedicine. JE was a recipient of an FPU doctoral fellowship from MEC (Spain).

References

1. Baldwin RL: Energetics of protein folding. J Mol Biol 2007, 371:283-301.
2. Chen Y, Ding F, Nie H, Serohijos AW, Sharma S, Wilcox KC, Yin S, Dokholyan NV: Protein folding: then and now. Arch Biochem Biophys 2008, 469:4-19.
3. Wesson L, Eisenberg D: Atomic solvent parameters applied to molecular dynamics of proteins in solution. Protein Sci 1992, 1:227-235.
4. Robertson AD, Murphy KP: Protein structure and the energetics of protein stability. Chem Rev 1997, 97:1251-1268.
5. Myers JK, Pace CN, Scholtz JM: Denaturant m values and heat capacity changes: relation to changes in accessible surface areas of protein unfolding. Protein Sci 1995, 4:2138-2148.
6. Lee B, Richards FM: The interpretation of protein structures: estimation of static accessibility. J Mol Biol 1971, 55:379-400.
7. Miller S, Janin J, Lask AM, Chothia C: Interior and surface of monomeric proteins. J Mol Biol 1987, 196:421-456.
8. Shrike A, Rupley JA: Environment and exposure to solvent of protein atoms. Lysozyme and insulin. J Mol Biol 1973, 79:351-371.
9. Rose GD, Geselowitz AR, Lesser GJ, Lee RH, Zehfus MH: Hydrophobicity of amino acid residues in globular proteins. Science 1985, 229:834-838.
10. Cramer TP, Srivivasan R, Rose GD: Modeling unfolded states of peptides and proteins. Biochemistry 1995, 34:16245-16250.
11. Cramer TP, Srivivasan R, Rose GD: Modeling unfolded states of proteins and peptides. II. Backbone solvent accessibility. Biochemistry 1997, 36:2832-2835.
12. Gong H, Rose GD: Assessing the solvent-dependent surface area of unfolded proteins using an ensemble model. Proc Nat Acad Sci USA 2008, 105:3321-3326.
13. Goldenberg DP: Computational simulation of the statistical properties of unfolded proteins. J Mol Biol 2003, 326:1615-1633.
14. Bernardó P, Blackledge M, Sancho J: Sequence-specific solvent accessibilities of protein residues in unfolded protein ensembles. Biophys J 2006, 91:4536-4543.
15. Bernardó P, Blanchard L, Timmins P, Marion D, Ruigrok RW, Blackledge M: A structural model for unfolded proteins from residual dipolar couplings and small-angle x-ray scattering. Proc Nat Acad Sci USA 2005, 102:17002-17007.
16. Bernardó P, Bertocnici CW, Griesinger C, Zweckstetter M, Blackledge M: Defining long-range order and local disorder in native α- synuclein using residual dipolar couplings. J Am Chem Soc 2005, 127:17968-17969.
17. Mukrasch MD, Markwick P, Biernat J, Göke R, Bernardó P, Griesinger C, Handelkow E, Zweckstetter M, Blackledge M: Highly populated turn conformations in native unfolded tau protein identified from residual dipolar couplings and molecular simulation. J Am Chem Soc 2007, 129:5235-5243.
18. Wells M, Tidow H, Rutherford TJ, Markwick P, Jensen MR, Mylonas E, Svergun DI, Blackledge M, Forsht AR: Structure of tumor suppressor p53 and its intrinsically disordered N-terminal transactivation domain. Proc Nat Acad Sci USA 2008, 105:5762-5767.
19. Jensen MR, Houben K, Lescop E, Blanchard L, Ruigrok RW, Blackledge M: Quantitative conformational analysis of partially folded proteins from residual dipolar couplings: application to the molecular recognition element of Sendai virus nucleoprotein. J Am Chem Soc 2008, 130:8055-8061.