Chapter 10

Fast and Accurate Accessible Surface Area Prediction Without a Sequence Profile

Eshel Faraggi, Maksim Kouza, Yaoqi Zhou, and Andrzej Kloczkowski

Abstract

A fast accessible surface area (ASA) predictor is presented. In this new approach no residue mutation profiles generated by multiple sequence alignments are used as inputs. Instead, we use only single sequence information and global features such as single-residue and two-residue compositions of the chain. The resulting predictor is both highly more efficient than sequence alignment based predictors and of comparable accuracy to them. Introduction of the global inputs significantly helps achieve this comparable accuracy. The predictor, termed ASAquick, is found to perform similarly well for so-called easy and hard cases indicating generalizability and possible usability for de-novo protein structure prediction. The source code and a Linux executables for ASAquick are available from Research and Information Systems at <http://mamiris.com> and from the Battelle Center for Mathematical Medicine at <http://mathmed.org>.

Key words Accessible surface area, Protein structure, Sequence only prediction

1 Introduction

Accessible surface area (ASA) has played a crucial role in understanding biological function and genome relationships. The reason is that on the molecular level interactions are mostly limited to regions in the protein that are accessible to solvent or other molecules. Hence, ASA serves as a good indicator whether a given genetic or protein sequence region can potentially interact with other molecules, i.e., be involved in protein interactions and hence participate in biological function. Specific targeting of such genetic regions promises great hope for the advancement of medicine and biological understanding. Predicting ASA for a given protein sequence helps in finding these regions.

The ASA plays an important role in the stability, aggregation, enzyme activity, and binding affinity of proteins. With the help of the ASA, protein stability can be investigated by analyzing the unfolding transition as monitored by difference in the ASA between the denatured and folded state of a protein [1]. Exposure of

hydrophobic residues has been shown to facilitate the protein aggregation and fibril formation linked to different aggregation-related diseases such as Alzheimer's and Parkinson's diseases [2, 3], Menkes disease [4], and amyotrophic lateral sclerosis [5]. Based on experimental data [6] it was suggested that the ASA of a protein can be used to compute its solvation free energy [7]. The free energy of protein solvation has been shown to be strongly correlated with the ASA in a continuum approach [8, 9]. Simple linear relationship between solvation free energy and ASA has been used in developing fast implicit solvent models [10, 11].

When protein structure is available, all-atom models [12–15] and analytical methods [16, 17] have been used to compute ASA. However, the number of protein structures solved experimentally by X-ray crystallography and NMR and deposited in the Protein Data Bank [18] is about 114,000 (as of January of 2016), while the number of known protein sequences reached 80 millions in 2014 [19]. The number of protein sequences continues to grow with remarkably faster rate than the number of protein structures deposited in PDB. The widening gap between protein sequences and structures solved experimentally motivates the prediction of ASA from the protein sequence.

Marsh and Teichmann [20] studied the role of the relative solvent ASA in the monomeric state in predicting the magnitude of binding-induced conformational changes. The study indicated that relative solvent ASA of monomer protein is strongly correlated with its flexibility and conformational changes upon binding. It has been found that the larger the relative solvent ASA the larger its conformational changes upon binding and flexibility. On the other hand, peptide flexibility [21] and the population of fibril-prone conformation in the monomeric state [22, 23] have been shown to be one of the main factors governing fibril formation times. These results not only suggest that ASA plays key role in docking of proteins, but also might open new routes to understand the docking and aggregation of proteins just by looking at the solvent ASA of monomeric forms.

De novo prediction of the ASA of a protein is a challenging problem. Understanding the process by which the protein surface is predicted from its sequence has attracted attention of researchers for many years. Wolfgang Pauli, the 1945 Nobel Prize winner in Physics, claimed that “God made the bulk; the surface was invented by the devil” emphasizing that the behavior of atoms in the surface is very difficult to understand.

The ASA [7, 24–37] is defined as the surface area of a protein or residue that is accessible to a solvent and is given here in units of square Angstroms. First described by Lee and Richards [7], ASA is typically calculated using the ‘rolling ball’ algorithm [38, 39]. In this approach a computational ball is rolled on the surface of the protein coordinates and probes for which and how much of the

residues are accessible, i.e., in contact with the ball. Recently Klenin et al. introduced a fast analytical method to calculate the ASA using power diagrams [40]. Other efforts in characterizing the surface of proteins were also carried out [41–47]. Since the ASA describes the amount of surface a given residue has that is accessible to the solvent or for other intermolecular interactions, it is easy to understand why the ASA is important for recognizing functional sites along the chain [48]. Accessibility is a prerequisite for a residue to be involved in external interactions. Hence knowledge of the solvent exposed residues along the residue chain of the protein can facilitate various approaches associated with function prediction and targeted mutations.

The topic of ASA prediction has been well documented in many publications [24, 27, 28, 31, 37, 49–59]. Current, actively used methods for predicting the ASA use multiple sequence alignments. A computationally costly technique that grows slower by ever growing datasets of resolved genomes. The advent of modern computational power and automated processes enables significant amounts of patient or specimen specific genetic sequencing. Moreover, researchers and clinicians may be interested in the effect of varying the genetic sequence. Both of these interests mean that in many cases one would be interested in obtaining ASA for a large number of genetic sequences. Hence, the amount of time each prediction of ASA takes may be a crucial consideration. In addition, since most methods rely heavily on mutation profiles generated by multiple sequence alignments, and since some mutated sequences would generate identical profiles, algorithms using only single sequence information may have an advantage in some cases.

2 Materials and Methods

ASAquick [37] is such an algorithm. It was programmed in FORTRAN 90 and is constructed out of several subroutines that process the data, initialize the model, and train it. It is also capable of producing predictions from existing single models or producing ensemble predictions with expected deviations. Its execution is terminal based. It was built on Ubuntu Linux, under the BASH environment. It is a window-based predictor. We predict consecutively the ASA for all residues. We aim for a fast ASA predictor, and use sequence information alone. We represent the sequence in several ways. For a given residue we include a set number of its neighbors (window) and represent the residue types with a constant length vector according to the BLOSUM62 substitution matrix representation of residues and with physicochemical parameters [28, 31]. The first representation allows for some information on the mutations between residues while the second captures some of the chemical properties of the molecules involved.

Since we do not use multiple sequence alignment profiles there is less information sharing between sequentially similar chains. Hence, we have less of a problem of over-training. Potentially one may consider then that a more inclusive representation of the entire PDB [18] will facilitate learning from as many separate instances and will improve the overall prediction accuracy. This is indeed strongly suspected and is the topic of future work. For our case here, to maintain a reasonable comparison with previous methods we use a PISCES list [60, 61] of non-homologous protein chains with resolution better than 3 Å and sequence identity lower than 40%. Structure files were downloaded for these chains from the PDB. This non-redundant set was 14,361 proteins in size. A concern may arise here that we are using a higher sequence identity for clustering than is done in the training of most profile based predictors. One should realize that the common value of 25% sequence identity comes about in an attempt to eliminate redundancy in the training sets. Since proteins with sequence identity greater than 25% produce similar PSSMs. Hence, using more than one such sequence would mean a degenerate training example. However, ASAquick is trained at the sequence level. At this level, training examples are unique at 40% and even higher sequence identity.

For each residue in these chains we calculate the ASA using the DSSP program [62]. We then find the residue-type-dependent minimum and maximum values and use these to linearly normalize the ASA to get the relative solvent accessible area (RASA) between -1 (completely buried) and 1 (completely exposed). This scale choice was motivated by the inherent bipolarity of the neural network we use; a decision that is based on our previous work in this field [31, 63]. When searching for the maximum ASA we arbitrarily ignore the largest 1% of values. This is done to allow for exceptional behavior and various mishaps. These RASA values are the target output for GENN, and can be transformed back to their corresponding ASA values. The final predictor reports back the ASA in Angstrom squared.

As mentioned previously, inputs for ASAquick were chosen with speed in mind. Hence, no alignment profiles were used. Instead, each residue type is represented by seven parameters characterizing their physicochemical properties, and by 20 parameters related to residue mutation probabilities taken from the BLOSUM62 matrix [64–66]. For a given residue we use a window of neighboring residues as inputs to capture the local sequential environment of the residue.

To make up for some of the information lost because of abandoning multiple alignments we also found the following global parameters to be useful: the length of the chain divided by 1000, the residue type composition of the whole chain (25 values), and the directional two-residue composition (625 values). We have 25 residue types because we allow for the various characters reported

by DSSP, these include atypical residues, unknowns and chain gaps in the structure file. The output and each of the local inputs is stored in a separate file in a directory named for that chain. The global inputs are all stored in a single file in the same directory. Refer to examples provided with the distribution of GENN [67] for further information.

The parameters of the neural network were optimized using the following approach. First we tested various values for the hidden layer size and settled on a size of 31 nodes per hidden layer as a balance between speed, generalizability, and over-training avoidance. We then tested the dependence of the accuracy on the input window size. We also used this data to analyze the benefit of each of the input parameters. Refer to the original ASAquick publication for justification of parameter choice.

The accuracy of ASAquick was tested on two different datasets. In the first test we randomly partition our non-redundant set with 14,361 proteins in two parts. The first contains 5000 proteins and is used to train different neural network instances with different inputs as described below. We also use a subset of this first set for setting aside an over-protection set. The second partition contains 9361 and we use this set only to test the accuracy of prediction. We have found that inclusions of more proteins in the training set was beneficial. We use the Pearson's correlation coefficient to measure the accuracy. It is important to note that these correlations are between the ASA and not RASA. Correlation between the RASA can be significantly different.

In general we found that the prediction accuracy peaks when using an average of about five neural networks, if global input features are used. We also ranked the weights according to their accuracy on the over fit protection set and progressively added the best neural networks to the ensemble. We found that in this case too an average of around five neural networks reaches a similar maximum accuracy. Note that using global features not only improves the prediction accuracy but also allows the accuracy to reach a peak after a relatively small number of networks (around five). In contrast, the predictor without global features seems to not peak even after 24 networks. Refer to the original publication of ASAquick [37] for a more complete discussion of these results.

Further testing was conducted on the CASP10 [68] set. We used the ASAquick server and the SPINE-X [35, 69] server to predict the ASA. SPINE-X has been repeatedly found to be among the top predictors of ASA available [33]. We find that ASAquick with global features achieves a correlation of 0.66, while SPINE-X achieves a correlation coefficient of 0.71. Among the proteins in the CASP competition one subset of proteins are known as "hard targets." For the hard targets in CASP10 we find that the ASAquick correlation for ASA prediction is 0.66, while for SPINE-X the correlation coefficient is 0.68. The difference between sequence-based

and profile-based predictions is almost completely erased when considering proteins with no experimentally solved structures of homologs.

To further test the accuracy of ASAquick and its usefulness for harder targets we randomly selected 500 protein chains from PDB chains clustered at 25 % sequence identity. This set of 500 proteins has approximately 112,000 residues. For each one of these proteins we calculated a prediction using ASAquick and using SPINE-X. For each protein we also ran a BLAST search against the PDB and calculated the product of the top five e-values excluding the query protein. This measure, the e-value product, allows us to quantify the amount of similar sequences with a structure deposited in the PDB. Those query proteins with many similar sequences in the PDB will have an e-value product much smaller than one, while those with few homologs will have a value of order one or greater. We found that for proteins with similar sequence deposited in the PDB, a consistent advantage for using SPINE-X and multiple alignments approaches over ASAquick is evident. However, as we move towards harder and harder targets, we found many cases where using a pure sequence approach is advantageous and on average both perform similarly [37]. It is interesting to note that the pure sequence approach seems more useful for improving the MAE than the correlations.

As we described at the onset, one reason for designing ASAquick was to drastically increase the speed of predictions. Indeed, analysis on the time it takes to generate a prediction was carried out. It was found that ASAquick takes less than a second to produce a prediction on a single Intel Xeon E5410 at 2.33 GHz processor. On the other hand, it was found that SPINE-X takes considerably more time. Since the major bottleneck for producing prediction in SPINE-X is the generation of the profile, by default SPINE-X attempts to use four processors to carry out this job. It was found that for a protein for which ASAquick needs less than a second and a single processor, SPINE-X needs over 2 min and four processors to complete the job. Hence, ASAquick reduces the time and resources necessary to get a prediction by almost three orders of magnitude.

3 Notes

1. Source and Availability ASAquick was implemented using GENN [67], a general neural network designed to train on ad-hoc data. GENN can take any numerical input/output problem and prepare a corresponding, non-memorizing, model representation. Data can be organized in files containing individual instances or a collection of ordered instances where each line is an individual input/output target. Running ASAquick requires supplying a

sequence file in the FASTA format. A few Perl wrappers then take this sequence and quantify them to use as input features for the set of optimized weights. Because of the modularity of GENN, modifications and additions to this approach can be easily implemented. GENN and ASAquick with helpful documentation and examples are available from Research and Information Systems at <http://mamiris.com>, from the Battelle Center for Mathematical Medicine at <http://mathmed.org>, and by contacting the authors.

2. License The license for both GENN and ASAquick is available from a file called “LICENSE” in their respective directory. It allows for academic users the opportunity to use and modify it freely with proper citations while retaining some rights for commercial use.

3. Installation To install ASAquick for full functionality one should install GENN by: (1) Download the tar gzipped archive file, (2) Uncompress it: `tar -xzf GENN+ASAquick.tgz`, (3) Go to the GENN directory: `cd GENN+ASAquick`, (4) And install: `./install`. If one wishes to only use ASAquick for several predictions, one may simply type `./ASAquick your-fasta-file`, however, you will need to copy the contents of the ASAquick bin directory to your main bin directory so these scripts are found during the ASAquick run. To install GENN you must have gfortran, sed, and bc. Type these at a BASH terminal prompt to see if you have them or how to get them. Further information is available in the “README” files that are packed with the distribution.

4. Usage To run ASAquick use `ASAquick your-file` where *your-file* is either a fasta file or a dsspget [67] file. dsspget is a personal system of recording the crystallographer index, amino acid type, 8-state-dssp-ss, 3-stat-dssp-ss, ASA, phi, psi from a given PDB file. You can make such a file from a PDB file using `pdb2dsspget.pl`, or from a fasta file using `fasta2dsspget.pl`. In the second case arbitrary values are used for the unknown structural values. Both programs are provided with the distribution and should be in your `/bin/` and ready to use if the installation was completed successfully. For the usage of GENN, refer to the original publication [67].

Acknowledgements

We gratefully acknowledge the financial support provided by the National Institutes of Health (NIH) through Grants R01GM072014 and R01GM073095 to Andrzej Kloczkowski, Grant R01GM085003 to Yaoqi Zhou, and the National Science Foundation through Grant NSF MCB 1071785. This work is also supported in part by the National Health and Medical Research Council (1059775) of Australia to Yaoqi Zhou. Maksim Kouza would like to acknowledge support from the Polish Ministry of

Science and Higher Education Grant No. IP2012 016872 and “Mobilnosc Plus” No. DN/MOB/069/IV/2015; National Science Center grant [MAESTRO 2014/14/A/ST6/00088].

References

1. Myers JK, Pace CN, Scholtz JM (1995) Denaturant m values and heat capacity changes: relation to changes in accessible surface areas of protein unfolding. *Protein Sci* 4(10):2138–2148
2. Dobson CM (1999) Protein misfolding, evolution and disease. *Trends Biochem Sci* 24(9):329–332
3. Soto C (2001) Protein misfolding and disease; protein refolding and therapy. *FEBS Lett* 498(2–3):204–207
4. Kouza M, Gowtham S, Seel M, Hansmann UHE (2010) A numerical investigation into possible mechanisms by that the A629P mutant of ATP7A causes Menkes disease. *Phys Chem Chem Phys* 12(37):11390–11397
5. Munch C, Bertolotti A (2010) Exposure of hydrophobic surfaces initiates aggregation of diverse ALS-causing superoxide dismutase-1 mutants. *J Mol Biol* 399(3):512–525
6. Nozaki Y, Tanford C (1971) The solubility of amino acids and two glycine peptides in aqueous ethanol and dioxane solutions: establishment of a hydrophobicity scale. *J Biol Chem* 246(7):2211–2217
7. Lee B, Richards FM (1971) The interpretation of protein structures: estimation of static accessibility. *J Mol Biol* 55(3):379–IN4
8. Richards FM (1977) Areas, volumes, packing, and protein structure. *Annu Rev Biophys Bioeng* 6(1):151–176
9. Eisenberg D, McLachlan AD (1986) Solvation energy in protein folding and binding. *Nature* 319:199–203
10. Lazaridis T, Karplus M (1999) Effective energy function for proteins in solution. *Prot Struct Funct Genet* 35:133–152
11. Ooi T, Oobatake M, Nemethy G, Scheraga HA (1987) Accessible surface areas as a measure of the thermodynamic parameters of hydration of peptides. *Proc Natl Acad Sci USA* 84(10):3086–3090
12. Humphrey W, Dalke A, Schulten K (1996) VMD - visual molecular dynamics. *J Mol Graph* 14:33–38
13. Van Der Spoel D, Lindahl E, Hess B, Groenhof G, Mark AE, Berendsen HJC (2005) GROMACS: fast, flexible, and free. *J Comput Chem* 26(16):1701–1718
14. Eisenmenger F, Hansmann UHE, Hayryan S, Hu CK (2005) An enhanced version of SMMP—open-source software package for simulation of proteins. *Comput Phys Commun* 174(5):422–429
15. Fraczekiewicz R, Braun W (1998) Exact and efficient analytical calculation of the accessible surface areas and their gradients for macromolecules. *J Comput Chem* 19:319–333.
16. Busa J Jr, Hayryan S, Wu MC, Busa J, Hu CK (2005) An enhanced version of SMMP - open-source software package for simulation of proteins. *Comput Phys Commun* 174:422
17. Wu MC, Li MS, Ma WJ, Kouza M, Hu CK (2011) Universal geometrical factor of protein conformations as a consequence of energy minimization. *Europhys Lett* 96:68005
18. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat T, Weissig H, Shindyalov IN, Bourne PE (2000) The protein data bank. *Nucleic Acids Res* 28(1):235–242
19. The UniProt Consortium (2015) UniProt: a hub for protein information. *Nucleic Acids Res* 43(D1):D204–D212
20. Marsh JA, Teichmann SA (2011) Relative solvent accessible surface area predicts protein conformational changes upon binding. *Structure* 19(6):859–867
21. Kouza M, Co NT, Nguyen PH, Kolinski A, Li MS (2015) Preformed template fluctuations promote fibril formation: insights from lattice and all-atom models. *J Chem Phys* 142(14):145104
22. Nam HB, Kouza M, Hoang Z, Li MS (2010) Relationship between population of the fibril-prone conformation in the monomeric state and oligomer formation times of peptides: insights from all-atom simulations. *J Chem Phys* 132(16):165104
23. Chiti F, Stefani M, Taddei N, Ramponi G, Dobson CM (2003) Rationalization of the effects of mutations on peptide and protein aggregation rates. *Nature* 424:805–808
24. Chothia C (1974) Hydrophobic bonding and accessible surface area in proteins. *Nature* 248(5446):338–339
25. Holbrook SR, Muskalk SM, Kim SH (1990) Predicting surface exposure of amino acids from protein sequence. *Protein Eng* 3(8):659–665

26. Rost B, Sander C (1994) Conservation and prediction of solvent accessibility in protein families. *Proteins Struct Funct Genet* 20(3):216–226
27. Moret M, Zebende G (2007) Amino acid hydrophobicity and accessible surface area. *Phys Rev E* 75(1):011920
28. Dor O, Zhou Y (2007) Real-spine: an integrated system of neural networks for real-value prediction of protein structural properties. *Proteins Struct Funct Bioinf* 68(1):76–81
29. Durham E, Dorr B, Woetzel N, Staritzbichler R, Meiler J (2009) Solvent accessible surface area approximations for rapid and accurate protein structure prediction. *J Mol Model* 15(9):1093–1108
30. Zhang H, Zhang T, Chen K, Shen S, Ruan J, Kurgan L (2009) On the relation between residue flexibility and local solvent accessibility in proteins. *Proteins Struct Funct Bioinf* 76(3):617–636
31. Faraggi E, Xue B, Zhou Y (2009) Improving the prediction accuracy of residue solvent accessibility and real-value backbone torsion angles of proteins by guided-learning through a two-layer neural network. *Proteins Struct Funct Bioinf* 74(4):847–856
32. Zhang T, Zhang H, Chen K, Ruan J, Shen S, Kurgan L (2010) Analysis and prediction of rna-binding residues using sequence, evolutionary conservation, and predicted secondary structure and solvent accessibility. *Curr Protein Pept Sci* 11(7):609–628
33. Gao J, Zhang T, Zhang H, Shen S, Ruan J, Kurgan L (2010) Accurate prediction of protein folding rates from sequence and sequence-derived residue flexibility and solvent accessibility. *Proteins Struct Funct Bioinf* 78(9):2114–2130
34. Nunez S, Venhorst J, Kruse CG (2010) Assessment of a novel scoring method based on solvent accessible surface area descriptors. *J Chem Inf Model* 50(4):480–486
35. Faraggi E, Zhang T, Yang Y, Kurgan L, Zhou Y (2012) Spine x: improving protein secondary structure prediction by multistep learning coupled with prediction of solvent accessible surface area and backbone torsion angles. *J Comput Chem* 33:259–267
36. Wang C, Xi L, Li S, Liu H, Yao X (2012) A sequence-based computational model for the prediction of the solvent accessible surface area for α -helix and β -barrel transmembrane residues. *J Comput Chem* 33(1):11–17
37. Faraggi E, Zhou Y, Kloczkowski A (2014) Accurate single-sequence prediction of solvent accessible surface area using local and global features. *Proteins Struct Funct Bioinf* 82(11):3170–3176
38. Shrake A, Rupley J (1973) Environment and exposure to solvent of protein atoms. Lysozyme and insulin. *J Mol Biol* 79(2):351–371
39. Hasel W, Hendrickson TF, Still WC (1988) A rapid approximation to the solvent accessible surface areas of atoms. *Tetrahedron Comput Methodol* 1(2):103–116
40. Klenin KV, Tristram F, Strunk T, Wenzel W (2011) Derivatives of molecular surface area and volume: simple and exact analytical formulas. *J Comput Chem* 32(12):2647–2653
41. Liang J, Edelsbrunner H, Fu P, Sudhakar PV, Subramaniam S (1998) Analytical shape computation of macromolecules: I. Molecular area and volume through alpha shape. *Proteins Struct Funct Genet* 33(1):1–17
42. Yan C, Dobbs D, Honavar V (2003) Identification of surface residues involved in protein-protein interaction—a support vector machine approach. In: *Intelligent systems design and applications*. Springer, Berlin, pp 53–62
43. Yan C, Dobbs D, Honavar V (2004) A two-stage classifier for identification of protein-protein interface residues. *Bioinformatics* 20(Suppl 1):i371–i378
44. Binkowski TA, Joachimiak A, Liang J (2005) Protein surface analysis for function annotation in high-throughput structural genomics pipeline. *Protein Sci* 14(12):2972–2981
45. Yan C, Wu F, Jernigan RL, Dobbs D, Honavar V (2008) Characterization of protein-protein interfaces. *Protein J* 27(1):59–70
46. Li B, Turuvekere S, Agrawal M, La D, Ramani K, Kihara D (2008) Characterization of local geometry of protein surfaces with the visibility criterion. *Proteins Struct Funct Bioinf* 71(2):670–683
47. Venkatraman V, Sael L, Kihara D (2009) Potential for protein surface shape analysis using spherical harmonics and 3d Zernike descriptors. *Cell Biochem Biophys* 54(1–3):23–32
48. Liang J, Woodward C, Edelsbrunner H (1998) Anatomy of protein pockets and cavities: measurement of binding site geometry and implications for ligand design. *Protein Sci* 7(9):1884–1897
49. Naderi-Manesh H, Sadeghi M, Arab S, Movahedi AAM (2001) Prediction of protein surface accessibility with information theory. *Proteins Struct Funct Bioinf* 42(4):452–459
50. Pollastri G, Baldi P, Fariselli P, Casadio R (2002) Prediction of coordination number and relative solvent accessibility in proteins. *Proteins Struct Funct Bioinf* 47(2):142–153

51. Ahmad S, Gromiha MM, Sarai A (2003) Real value prediction of solvent accessibility from amino acid sequence. *Proteins Struct Funct Bioinf* 50:629–635
52. Yuan Z, Huang B (2004) Prediction of protein accessible surface areas by support vector regression. *Proteins Struct Funct Bioinf* 57:558–564
53. Adamczak R, Porollo A, Meller J (2004) Accurate prediction of solvent accessibility using neural networks-based regression. *Proteins Struct Funct Bioinf* 56:753–767
54. Adamczak R, Porollo A, Meller J (2005) Combining prediction of secondary structure and solvent accessibility in proteins. *Proteins Struct Funct Bioinf* 59(3):467–475
55. Garg A, Kaur H, Raghava G (2005) Real value prediction of solvent accessibility in proteins using multiple sequence alignment and secondary structure. *Proteins* 61:318–324
56. Xu Z, Zhang C, Liu S, Zhou Y (2006) QBES: predicting real values of solvent accessibility from sequences by efficient, constrained energy optimization. *Proteins Struct Funct Bioinf* 63:961–966
57. Wang J, Lee H, Ahmad S (2005) Prediction and evolutionary information analysis of protein solvent accessibility using multiple linear regression. *Proteins Struct Funct Bioinf* 61:481–491
58. Pollastri G, Martin AJ, Mooney C, Vullo A (2007) Accurate prediction of protein secondary structure and solvent accessibility by consensus combiners of sequence and structure information. *BMC Bioinf* 8(1):201
59. Rost B (2009) Prediction of protein structure in 1d—secondary structure, membrane regions, and solvent accessibility. In: *Structural bioinformatics*, pp 679–714
60. Wang G, Dunbrack RL (2003) PISCES: a protein sequence culling server. *Bioinformatics* 19(12):1589–1591
61. Wang G, Dunbrack RL (2005) PISCES: recent improvements to a PDB sequence culling server. *Nucleic Acids Res* 33(Suppl 2):W94–W98
62. Kabsch W, Sander C (1983) Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers* 22:2577–2637
63. Xue B, Dor O, Faraggi E, Zhou Y (2008) Real value prediction of backbone torsion angles. *Proteins Struct Funct Bioinf* 72:427–433
64. Henikoff S, Henikoff JG (1992) Amino acid substitution matrices from protein blocks. *Proc Natl Acad Sci* 89(22):10915–10919
65. Eddy SR et al (2004) Where did the blosum62 alignment score matrix come from? *Nat Biotechnol* 22(8):1035–1036
66. Styczynski MP, Jensen KL, Rigoutsos I, Stephanopoulos G (2008) Blosum62 miscalculations improve search performance. *Nat Biotechnol* 26(3):274–275
67. Faraggi E, Kloczkowski A (2014) GENN: a GEneral Neural Network for learning tabulated data with examples from protein structure prediction. In: *Artificial neural networks: methods and applications*. *Methods in molecular biology*
68. Moult J, Fidelis K, Kryshchuk A, Tramontano A (2011) Critical assessment of methods of protein structure prediction (CASP) round IX. *Proteins Struct Funct Bioinf* 79(S10):1–5
69. Faraggi E, Yang Y, Zhang S, Zhou Y (2009) Predicting continuous local structure and the effect of its substitution for secondary structure in fragment-free protein structure prediction. *Structure* 17:1515–1527