ELSEVIER

Biophysical Chemistry 124 (2006) 145–154

Biophysical Chemistry

http://www.elsevier.com/locate/biophyschem

Atom-wise statistics and prediction of solvent accessibility in proteins

Y. Hemajit Singh [a], M. Michael Gromiha [b], Akinori Sarai [c], Shandar Ahmad [a,*]

[a] *Department of Biosciences, Jamia Millia Islamia University, New Delhi-110025, India*

[b] *Computational Biology Research Center (CBRC), National Institute of Advanced Industrial Science and Technology (AIST), 2-42 Aomi, Koto-ku, Tokyo 135-0064, Japan*

[c] *Department of Bioscience and Bioinformatics, Kyushu Institute of Technology, Iizuka, Fukuoka 820 8502, Japan*

Received 20 May 2006; received in revised form 21 June 2006; accepted 22 June 2006
Available online 24 July 2006

Abstract

In this work, we explore a novel method to broaden the scope of sequence-based predictions of solvent accessibility or accessible surface area (ASA) to the atomic level. All 167 heavy atoms from the 20 types of amino acid residues in proteins have been studied. An analysis of ASA distribution of these atomic groups in different proteins has been performed and rotamer-style libraries have been developed. We observe that the ASA of some atomic groups (e.g., backbone C and N atoms) can be estimated from the sequence environment within a mean absolute error of 2–3 Å^2. However, some side chain atoms such as CG in Pro, NH1 in Arg and NE2 in Gln show a strong variability making it more difficult to estimate their ASA from sequence environment. In general, the prediction of ASA becomes more difficult for atomic positions at the side chain extremities of long amino acid residues (aromatic side chain terminals being the exception). Several atomic groups are frequently exposed to solvent. Some of them have a bimodal distribution, suggesting two stable conformations in terms of their solvent exposure. More detailed understanding and prediction of solvent accessibility, i.e., at an atomic level is expected to help in bioinformatics approaches to structure prediction, functional relevance of atomic solvent accessibilities and other interaction analyses.

Keywords: Solvent accessibility; Structure prediction; Solvation energy; Neural network

1. Introduction

Prediction of solvent accessibility of amino acid residues has been actively pursued by several researchers in the recent past [1–14]. We have previously developed a novel method to predict real-value accessible surface area of amino acid residues instead of the conventional predictions based on arbitrarily defined exposure states [15,16]. Real-value prediction method was quickly adopted in other works [17,18]. All reported prediction methods, to date including our previous method, make a prediction of (suitably normalized) total surface area or solvent accessibility of residues in the protein, use the information about the identity of target residue and its neighbors, represented in various ways.

It has been reported that the contribution of hydrophobic free energy computed with "atomic solvent accessibilities" is more reliable than that from "residue solvent accessibilities" for understanding the folding and stability of proteins [20,21]. In this work, we present another novelty in the direction of predicting solvent accessibility in proteins by analyzing and then making prediction models for accessible surface area of each atom of amino acid types in proteins. We started with the largest of the data sets used in our previous studies and found that the atomic solvent accessibility is more sensitive to the quality and completeness of coordinate data, and hence, a large number of proteins from those data sets could not satisfactorily be used for the calculation of atomic solvent accessibility. We therefore developed a new non-redundant data set of 2277 protein domains (nearly 1.4 million atoms). General statistics of ASA in all 167 atomic groups have been collected. Neural networks have been developed and trained to predict atomic ASA of these atomic groups.

2. Materials and methods

2.1. Protein data sets (ASTRAL-NR data sets)

The domain information about proteins has been taken from SCOP [22]. PDB style coordinates for isolated domains have

* Corresponding author.

0301-4622/$ - see front matter © 2006 Elsevier B.V. All rights reserved.
doi:10.1016/j.bpc.2006.06.013

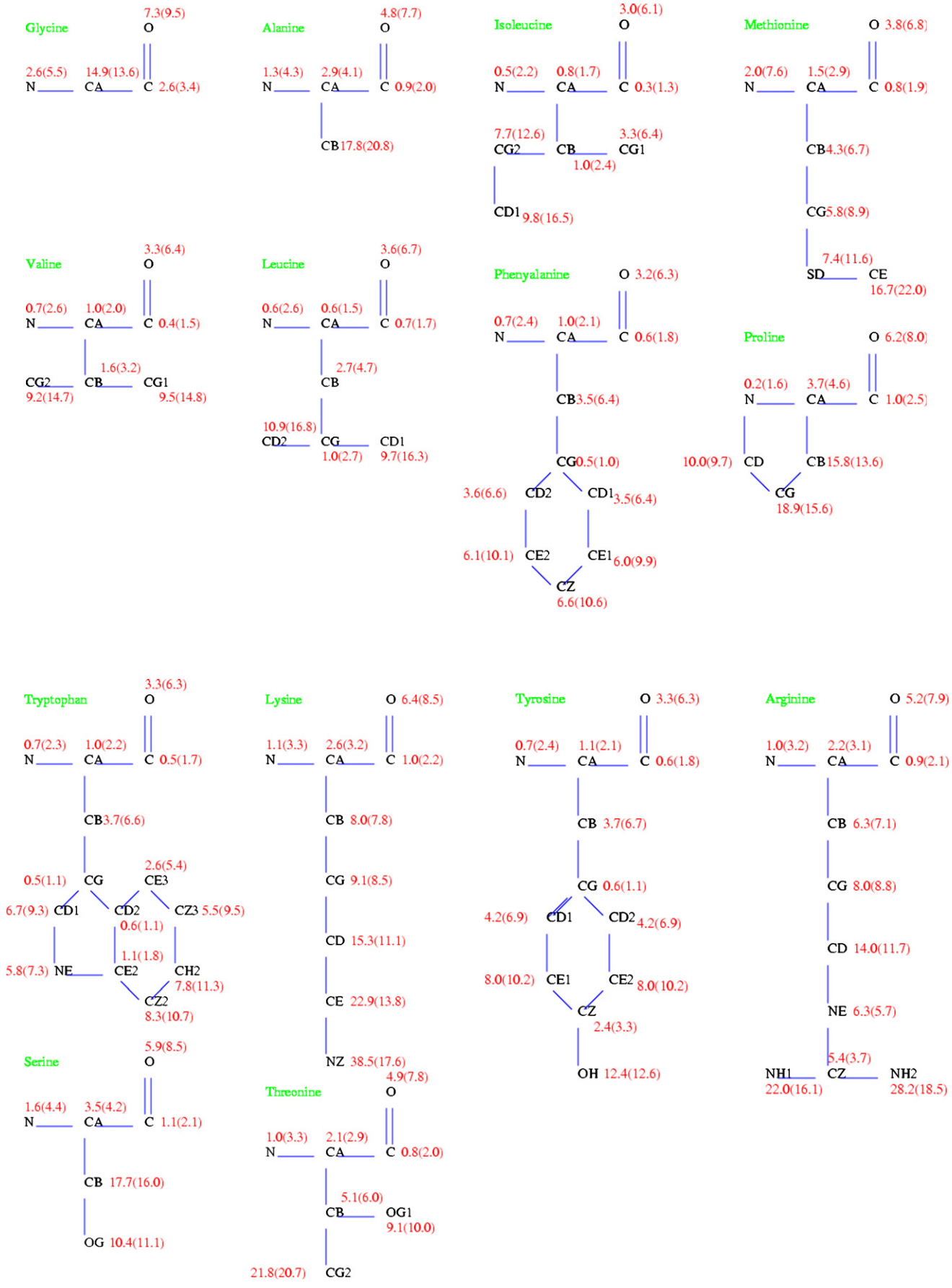

Fig. 1. Mean and standard deviation values (in brackets) of ASA in different atomic positions of 20 amino acid residues. All values are in $Å^2$.

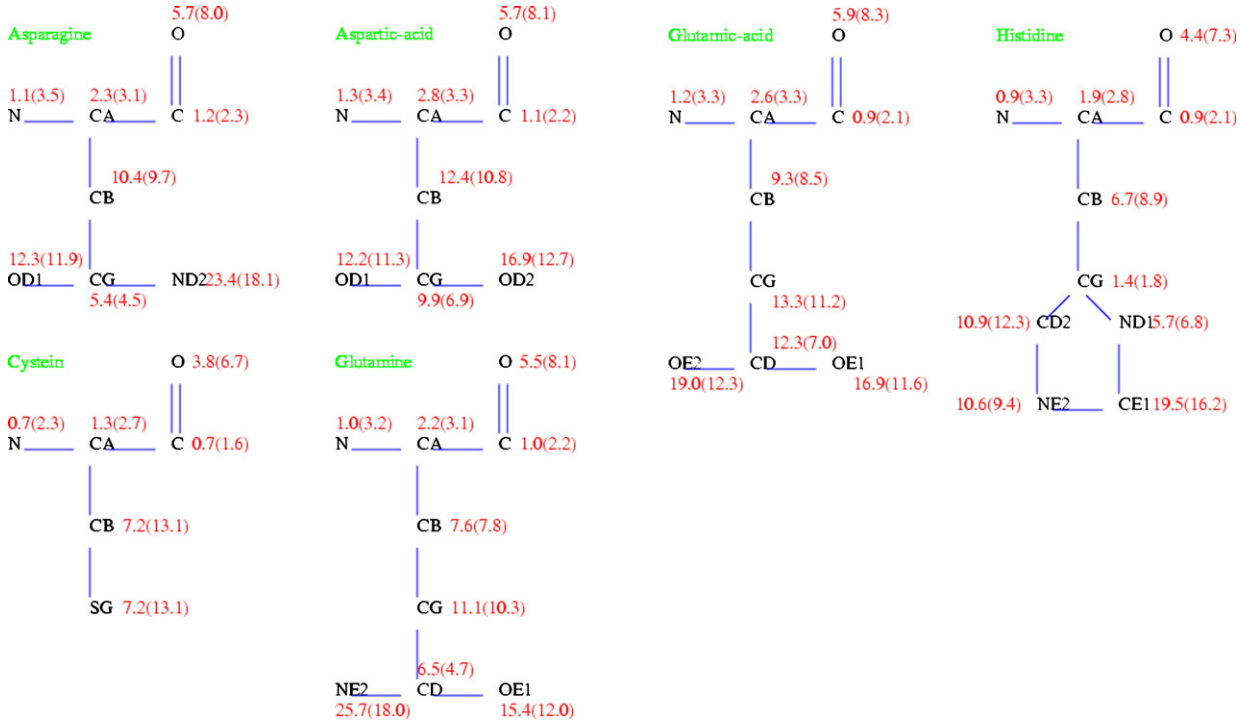

Fig. 1 (*continued*).

been taken from ASTRAL database [23]. Version 1.63 of SCOP lists consists of 5226 domains from 3332 collected proteins such that no two domain sequences have more than 40% sequence identity [22]. Redundancy was further removed by clustering this sequence data at 25% identity threshold using blastclust [24]. This resulted in 4536 domains. ASC (see next section) could successfully produce atomic ASAs of only 2612 of them (due to insufficient quality of available PDB coordinates in other proteins). From this data, domains whose structures were obtained by NMR or at a resolution poorer than 3.0 Å have been excluded from analysis, leaving behind 2277 protein domains. No attempt was made to find a functional or structure classification of these domains, although some improvements may be expected by treating membrane and unstructured proteins as well as monomeric and complexed proteins separately. Coordinate files used here contain information about the domain residues only and the presence of non-domain residues from the same or other chains, ligand atoms and other heterogeneous atoms is ignored.

2.2. Calculation of solvent accessibility

We previously used the Dictionary of Secondary Structure of Proteins (DSSP) program of Kabsch and Sander [25] for the calculation of residue solvent accessibility or accessible surface area (ASA). However, DSSP does not provide information about atomic solvent accessibility, and hence, we used another readily available program ASC, which analytically calculates solvent accessibility of each atom in the protein [26]. This method has been widely used, e.g., to assign the location of mutant site residues for understanding the stability of protein [27–29].

2.3. Grouping atoms

For a quick analysis, all 167 groups of carbon, nitrogen and sulfur atoms were separately sorted in terms of their mean ASA values. They were then classified into highly accessible to mostly buried atomic groups. The data belonging to several atomic groups with similar mean ASA values were combined and corresponding normalized histograms were plotted. Details of such groups are presented in Results and discussion.

2.4. Neural network design and training

We have previously used a neural network approach to model total real values of solvent accessibility and also for category prediction of this property [10,15]. We used similar neural networks for the prediction of ASA. However, instead of extensively evaluating predictions based on window size, we restricted ourselves to the development of 167 neural networks—one for each atom type with the identity of one and two residue neighbors being fed into the neural network. These results are compared with the mean deviation in the overall statistics of the corresponding atomic ASA values. Evolutionary information has been previously used by researchers to improve prediction [17–19], but we have restricted ourselves to the sequence information only, trying to analyze the effect of neighbours and sequence dependence of ASA. The difference in the mean deviation of ASA in the data and the mean absolute error of prediction using one and two neighboring residue information has been attributed to the effect of neighbors in constraining the accessible surface area of protein atoms.

2.5. Measurement of accuracy

Mean absolute error (MAE) has been used to measure the quality of predictions. MAE is defined as the absolute error of prediction per residue or per atom.

3. Results and discussion

The complete set of plots and data sets are provided in the supplementary online material (www.netasa.org/atomic-asa/). Key features of the statistics shown in supplementary tables are captured in the form of the following discussion and graphs. Fig. 1 shows rotamer style diagrams for the mean and standard deviation values of ASAs of all 167 atomic groups belonging to 20 amino acid residues. Supplementary Table 1 and Supplementary Figure 1 show detailed statistics of the distribution of atomic groups in different ranges of ASA. Supplementary Table 2 lists the summary of the results of predictions using neural network.

The most obvious property of ASA shown in these graphs is that far from being a Gaussian distribution, the data are highly skewed towards lower ASA values (see Supplementary Figure 1 histograms). In most cases, the frequency of occurrence of an atom in a given range of ASA falls abruptly very close to zero value. Obvious meaning is that there are a very large majority of totally buried atoms of all types. Fewer and fewer atoms have higher values of exposed surface area. So much so that there are almost no atoms with an exposed ASA of more than 50 Å^2. NH1 and NH2 in Arg, ND2 in Asn, NE2 in Gln and CG2 in Thr (all with their mean ASA close to 30 Å^2) have the highest mean ASA values.

A schematic representation of all amino acids with all atoms labeled by their mean ASA and the standard deviation is shown in Fig. 1. Supplementary Table 1 contains the values of mean, maximum, standard deviation and mean deviation values of all 167 heavy atomic types from these proteins, sorted and grouped according to their atomic identity and mean ASA values. Based on these mean values, classified graphs of mean ASA values are shown in Fig. 2. The highest diversity of ASA values was observed in carbon atoms. ASA values of side chain carbon atoms could be grouped into five categories viz. (a) highest exposure carbon atoms with mean ASA values higher than 15.0 Å^2 (CE of Lys and Met, CD of Lys, CG2 of Thr, CE1 of His, CG of Pro, CB of Ala, Ser and Pro fall in this category); (b) highly exposed carbon atoms with mean ASA values ranging from 10.0 to 15.0 Å^2 (consisting of CA of Gly, CD of Arg and

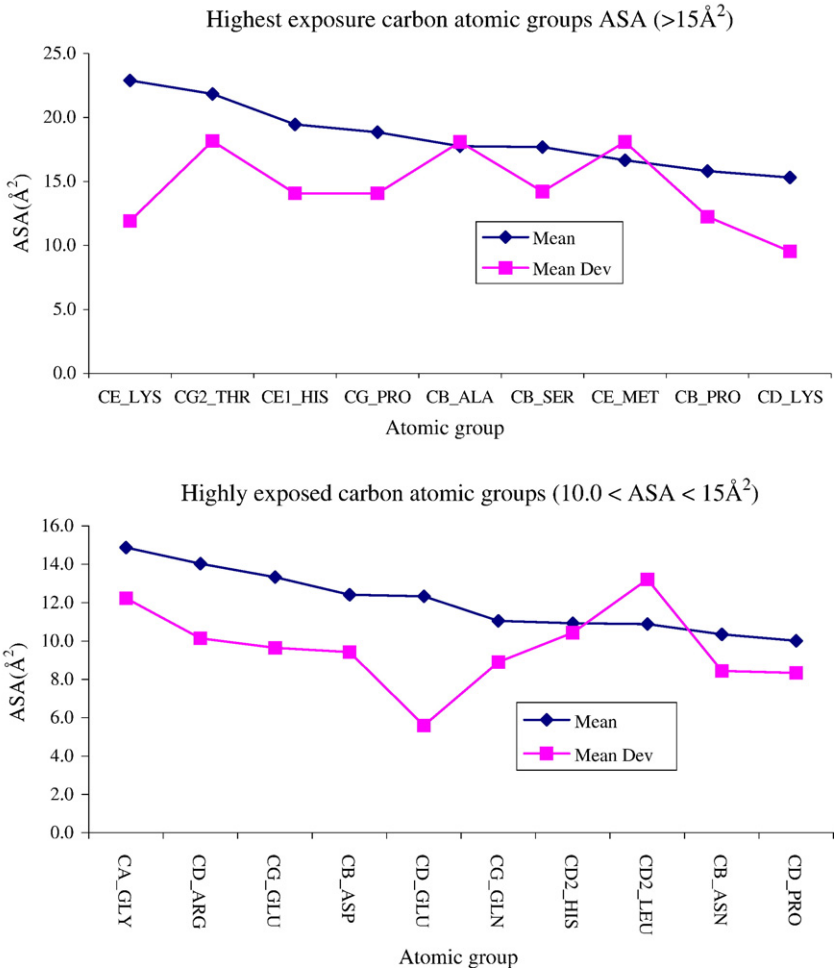

Fig. 2. Summary of mean and deviation in ASA values of atomic groups, clustered by their atom type and mean ASA.

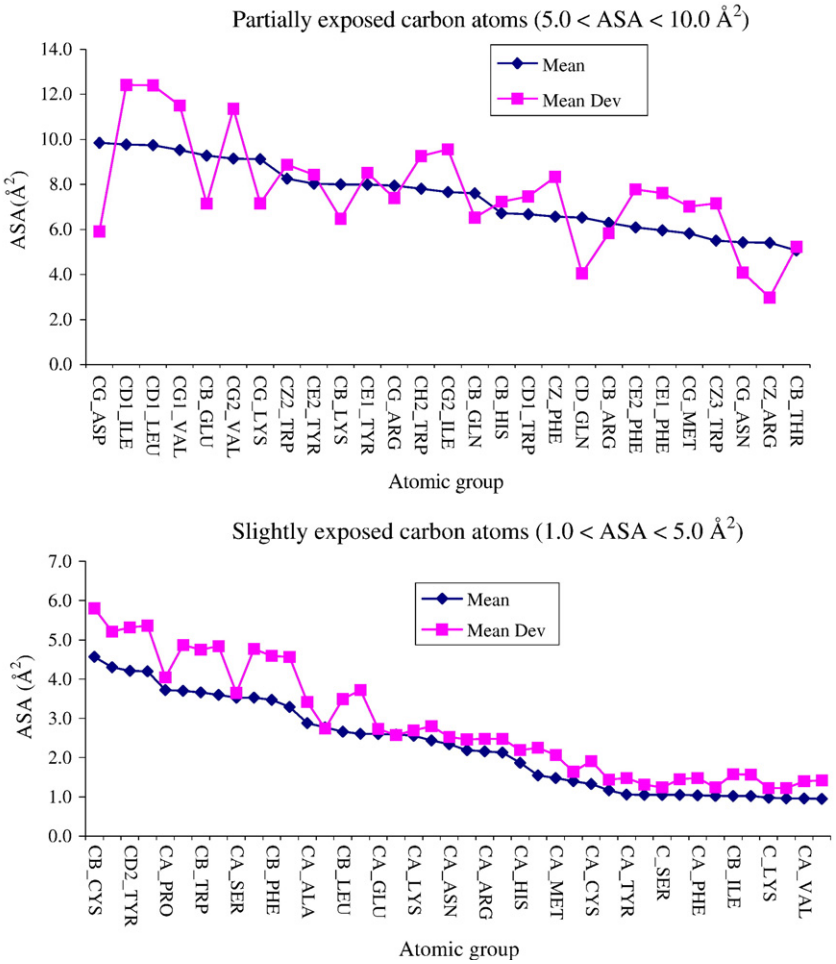

Fig. 2 (*continued*).

Pro, CD2 of His, Leu and Asn and CG of Gln); (c) partially exposed carbon atoms with mean ASA values ranging from 5.0 to 10.0 Å^2 (e.g., CB of Trp and Ile); (d) slightly exposed carbon atoms with mean ASA values ranging from 1.0 to 5.0 Å^2 (e.g., CA of Ala and Pro); and (e) totally buried carbon atoms with less than 1.0 Å^2 mean ASA values (most backbone C atoms and CA of Leu and Ile, and some side chain atoms of Trp and Tyr). Noteworthy among them are CG2 of Thr and CE of Lys, both of which have more than 20.0 Å^2 mean ASA value. Mean deviation in this class of atomic groups is usually less than their mean values, suggesting that very few of these residues will be in buried state. In most cases, carbon atoms with highest values of ASA reside on the farthest locations from the backbone.

Oxygen atoms were not found to have such a wide range of mean ASA values and could be conveniently grouped into two categories viz. backbone oxygen (usually buried) and side chain oxygen (usually exposed). Although the highest mean ASA value of oxygen atom was found for OE2 and OD2 of Glu and Asp, OE1 and OD1 of their neutral relatives Gln and Asn also show a high value of mean ASA. Thus, highly exposed ASA indicates a clear preference of these two oxygen atoms in all four residues to be on the protein surface. All the backbone oxygen atoms, on the other hand, are usually buried, with the highest mean ASA approaching in Gly, Lys and Pro. Of all the well-known hydrophobic residues, Pro residue is the only one in which backbone oxygen has a mean ASA greater than 5.0 Å^2. Apart from this exception, most backbone oxygen atoms in all residues show a simple correspondence to the well-known hydrophobicity of the residue to which they belong. The exception of Pro may probably be due to the unusual ring structure of its side chain, allowing greater exposure to its backbone oxygen, perhaps by shifting CB atoms away and, hence, at the cost of ASA of the backbone nitrogen (average ASA less than 1.0 Å^2, being the least exposed of all nitrogen atoms in any amino acid).

Nitrogen atoms are very clearly divided into three ranges of their mean ASA values. The most exposed nitrogen atoms are NZ of Lys and NH2 of Arg (both basic residues). Close to these maximum values are NE2 of Gln and ND2 of Asn. NH1 of Arg also falls in this category of the highest mean ASA values although its mean ASA (22.0 Å^2) is somewhat lower than NZ of Lys (average ASA = 38.5 Å^2). Frequency histogram in Supplementary Figure 1 reveals that NZ is rarely in the buried state and almost always has a high ASA value, making its average higher than others. All other side chain atoms viz. NE2 of His, NE of Arg, NE1 in Trp and ND1 of His have a moderately high mean ASA values (5.0–10.0 Å^2). Without exception, all backbone nitrogen atoms have much smaller ASA than side chain nitrogen atoms, and about half of them have their mean ASA

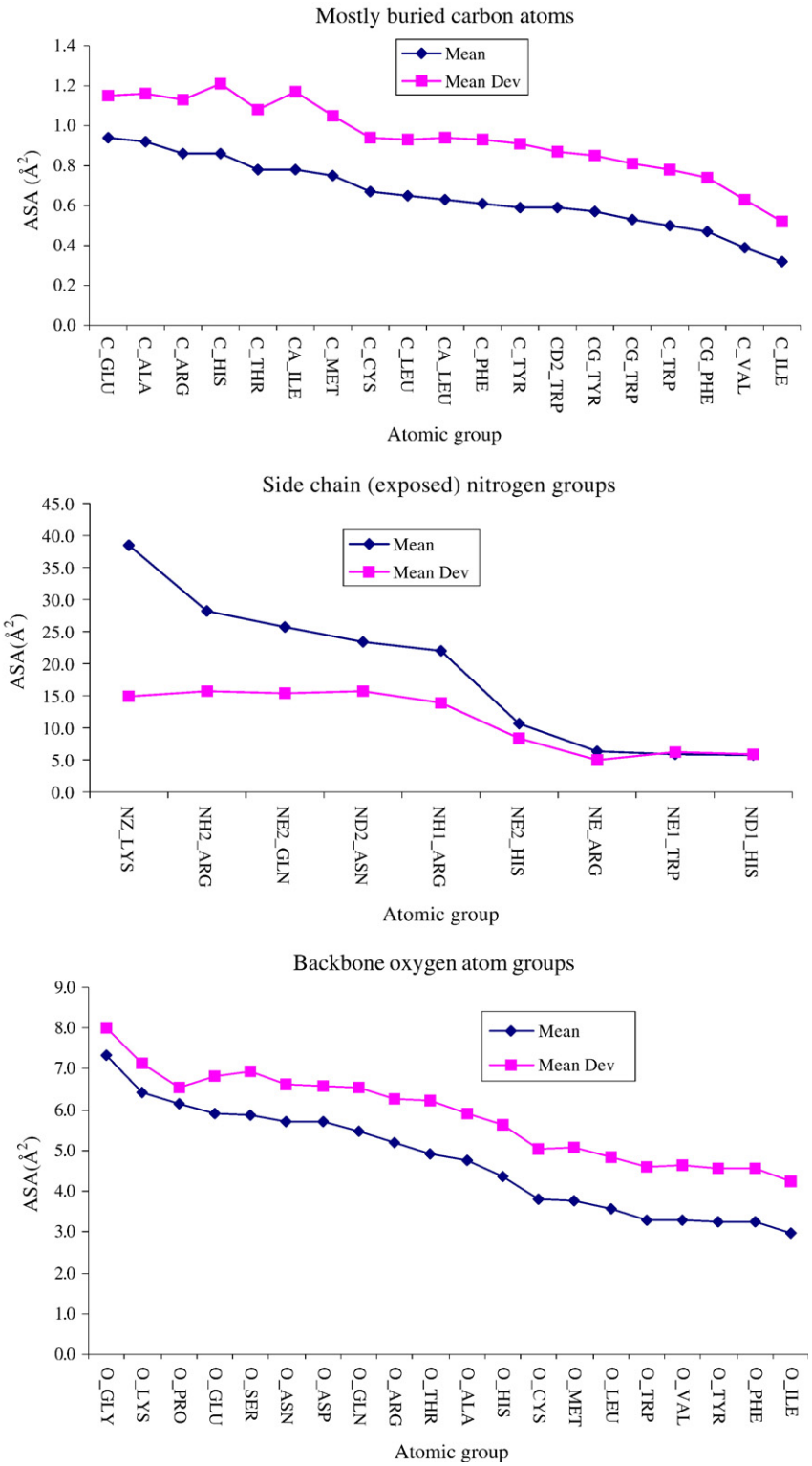

Fig. 2 (*continued*).

less than 1.0 Å^2. Backbone nitrogen in Gly and Met are the only exceptions with their mean ASA value greater than 2.0 Å^2. However, the mean value is not so drastically higher than other nitrogen atoms in the backbone.

Other prominent features of the normalized frequency plots (Supplementary Figure 1) are as follows:

3.1. CA atoms

In most residue types ASA of CA has a strong peak near 0 Å^2. In general, the frequency of data with more than 10.0 Å^2 is very low and almost no CA atoms in any residue have more than 12.0 Å^2 ASA. CA of Gly is the only exception to

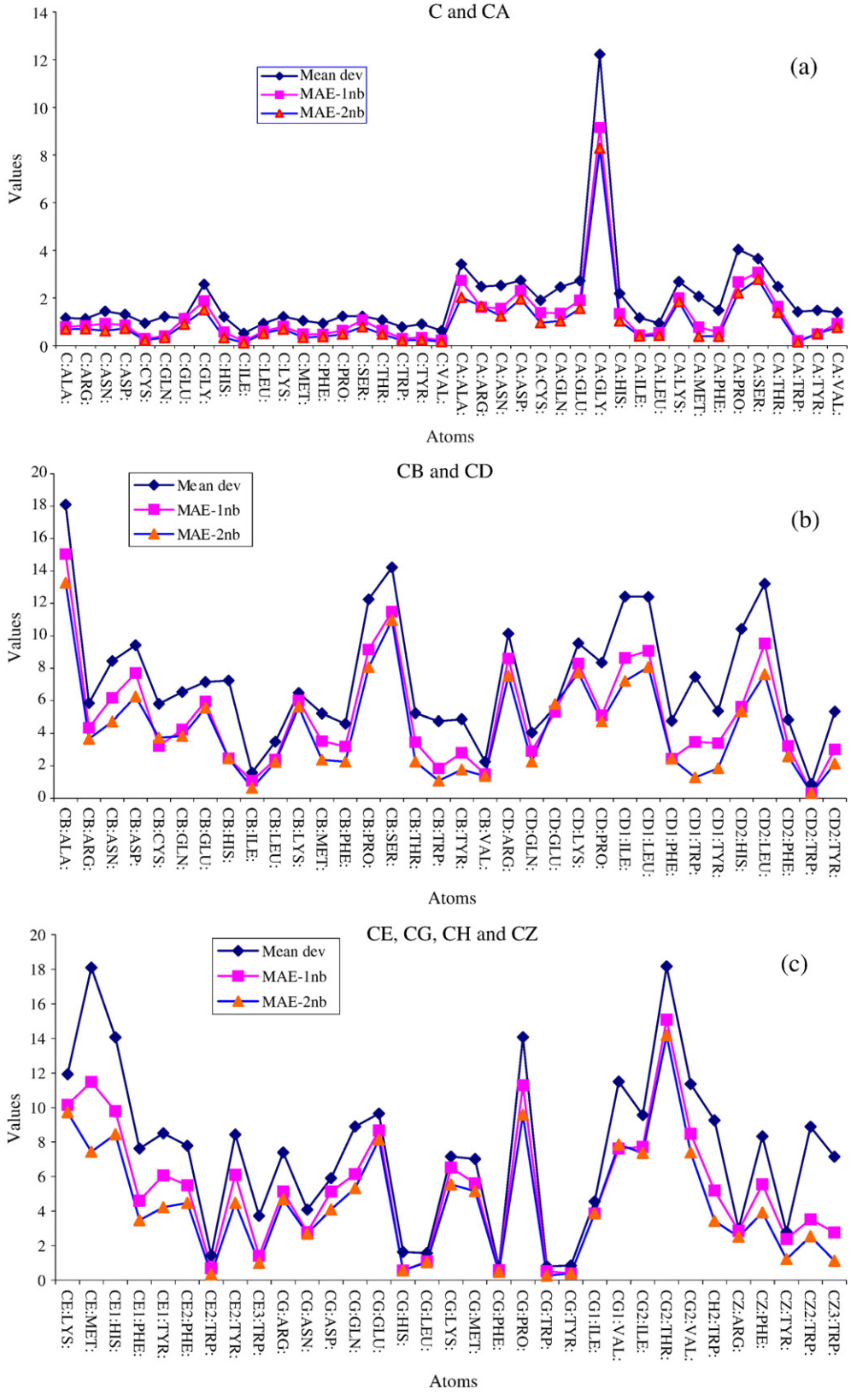

Fig. 3. Summary of mean deviation and neural network-based prediction MAE of one and two neighbor in ASA values.

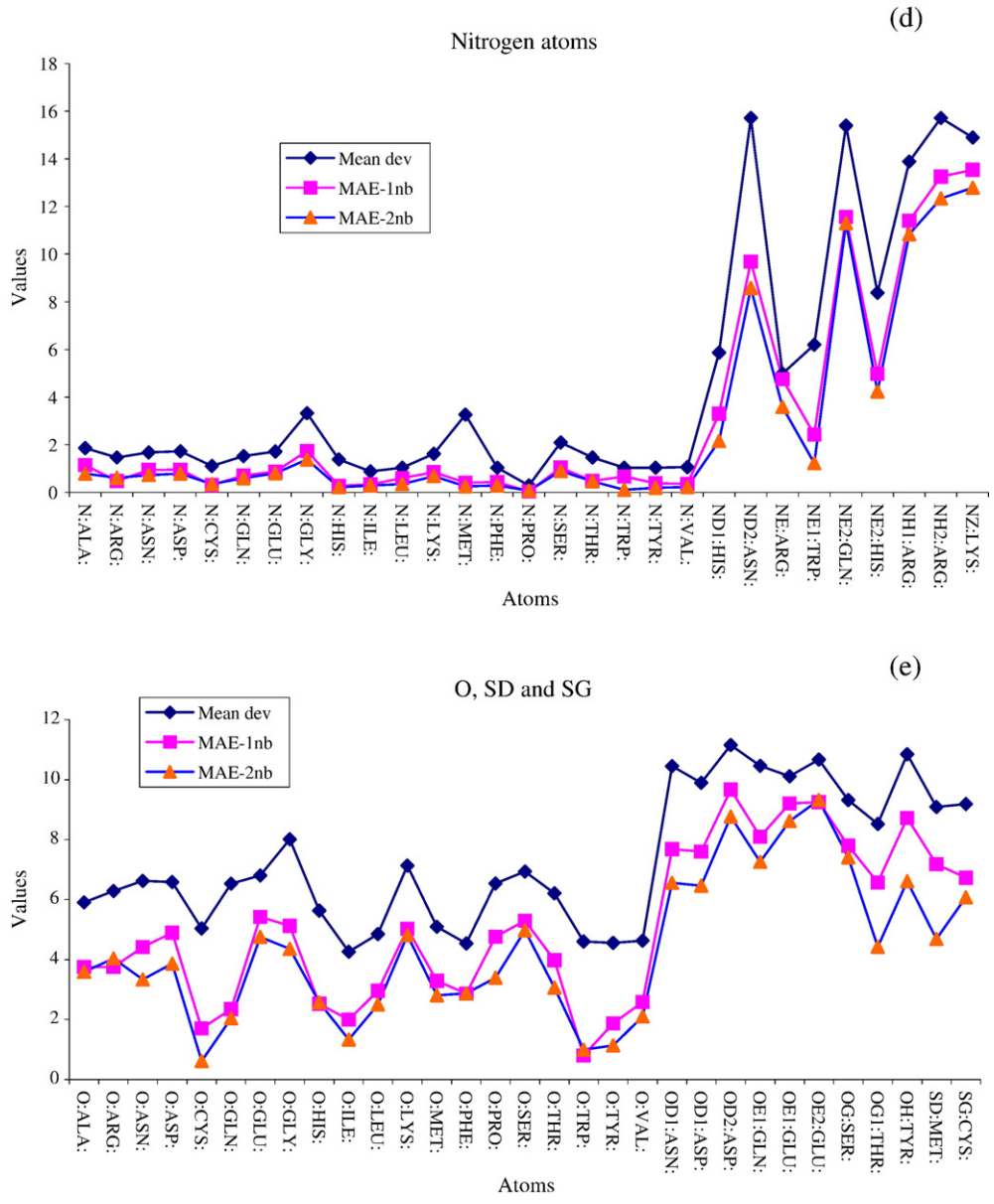

Fig. 3 (*continued*).

this rule, most likely because of its small size and absence of (heavy atoms in its) side chain. Pro tends to have a second peak near 10.0 Å^2, which could be due to its unusual ring structure.

3.2. CB atoms

CB atoms have larger exposure than CA. Pro, Ser and Thr are the most prominent residues in which many CB atoms have high ASA values. In particular ASA of some CB atoms in Ser and Pro could be as high as 30–35 Å^2.

3.3. CD, CD1 and CD2 atoms

Despite being further away from the backbone, CD1 has a quite small ASA (mostly less than 5.0 Å^2). CD2 atoms in Trp have a small second peak, again suggesting two structural conformations, which could be energetically favorable.

3.4. CE, CE1 and CE2 atoms:

Most of the CE, CE1 and CE2 atoms remain in the buried state.

3.5. Deviation statistics, average assignment and neural network predictions

Mean deviation statistics of ASA have been plotted along with their mean values in Fig. 2. Detailed values of standard deviation and maximum are also included in Supplementary Table 1. Mean deviation has been particularly selected for analysis as this shows how effectively we can predict ASA of an

atom, simply by knowing its identity. Thus, the mean deviation corresponds to the MAE of prediction if all the atomic groups are assigned their mean values from the database (no cross-validation is used in this assessment). Now, if the atomic and residue identities are known, MAE is a prediction without the knowledge of its sequence neighbors. To asses the extent to which neighboring residues constraint the ASA values of these atoms, we trained 167 neural networks each representing an atomic group as above, and training them with single neighbor and two neighbor information. The neural network used in this case is similar to what we used in our earlier works [10,15]. Fig. 3 shows the results of such neural network predictions. Almost all atomic groups showed improvement in their MAE values with the inclusion of the information about their neighbor, albeit to different extents. Fig. 3 shows the mean deviation, neural network predictions with one neighbor and two neighbors. Detailed numerical values corresponding to these graphs can be seen in Supplementary Table 2. The most prominent atomic groups showing highest improvements in their prediction by including neighbor information are oxygen atoms (both backbone and side chain). CD1 of Ile and Leu, CE and N of Met, ND2 of Asn also show systematic improvement in prediction with neighbor information. Of these CE of Met and OH of Tyr show significant improvements with the first and then again with the second neighbor. In most other cases, the improvement with second neighbor is significantly less than that of the first neighbor. Looking at these graphs for each atom, one observation is that the ASA of the backbone CA atoms remains only negligibly affected by the sequence neighbors. Overall fold of the protein backbone may be more important in determining ASA of CA atoms compared to the sequence neighbors. The other backbone atoms C, N and O do not share this property so strongly. For example, backbone N in Met could be predicted with much better accuracy (smaller MAE) by a neighbor-dependent neural network than its neighbor-independent prediction accuracy measured by the mean deviation. This indicates that ASA values of Met N-terminal atoms are sensitive to their neighbors. In case of terminal residues, the information about the absence of a neighbor will be helpful in predicting the ASA of these atoms. This does not apply to CA atoms of Met because the exposure of CA atoms in the terminal residues does not differ so much from those inside the chain, as CA is not a terminal atom of an amino acid. Most of the other side chain atoms follow a general pattern of improvement in prediction from residue neighbor information.

4. Conclusion

We have carried out a large-scale analysis on the solvent accessibility of each of the 167 atomic groups in 20 amino acid residues using a data set of protein domains. We observed that the carbon and sulfur atoms are usually buried and the positively charged atoms are typically exposed to solvent. An average assignment method is expected to yield a prediction MAE equivalent to the mean deviation in the corresponding atomic group. To improve these predictions by making use of residue neighbor information, we have developed neural network methods for predicting atomic level ASA. Predictability of each atomic ASA does depend–although not exclusively–on the distance of the corresponding atom from the backbone and other factors. We observed that the ASA of backbone atoms are more conserved than the side chain atoms. The results obtained in this work can be useful for protein design, structure prediction and for understanding the folding and stability of protein molecules.

References

[1] B. Rost, C. Sander, Conservation and prediction of solvent accessibility in protein families, Proteins 20 (1994) 216–226.
[2] M.J. Thompson, R.A. Goldstein, Predicting solvent accessibility: higher accuracy using Bayesian statistics and optimized residue substitution classes, Proteins 25 (1996) 38–47.
[3] S. Pascarella, R. De Persio, F. Bossa, P. Argos, Easy method to predict solvent accessibility from multiple protein sequence alignments, Proteins 32 (1998) 190–199.
[4] M.H. Mucchielli-Giorgi, S. Hazout, P. Tuffery, PredAcc: prediction of solvent accessibility, Bioinformatics 15 (1999) 176–177.
[5] C.J. Richardson, D.J. Barlow, The bottom line for prediction of residue solvent accessibility, Protein Eng. 12 (1999) 1051–1054.
[6] O. Carugo, Predicting residue solvent accessibility from protein sequence by considering the sequence environment, Protein Eng. 13 (2000) 607–609.
[7] J.A. Cuff, G.J. Barton, Application of multiple sequence alignment profiles to improve protein secondary structure prediction, Proteins 40 (2000) 502–511.
[8] H. Naderi-Manesh, M. Sadeghi, S. Arab, A.A. Moosavi Movahedi, Prediction of protein surface accessibility with information theory, Proteins 42 (2001) 452–459.
[9] X. Li, X.M. Pan, New method for accurate prediction of solvent accessibility from protein sequence, Proteins 42 (2001) 1–5.
[10] S. Ahmad, M.M. Gromiha, NETASA: neural network based prediction of solvent accessibility, Bioinformatics 18 (2002) 819–824.
[11] Z. Yuan, K. Burrage, J.S. Mattick, Prediction of protein solvent accessibility using support vector machines, Proteins 48 (2002) 566–570.
[12] G. Pollastri, P. Baldi, P. Fariselli, R. Casadio, Prediction of coordination number and relative solvent accessibility, Proteins 47 (2002) 142–153.
[13] G. Gianese, F. Bossa, S. Pascarella, Improvement in prediction of solvent accessibility by probability profiles, Protein Eng. 16 (2003) 987–992.
[14] H. Kim, H. Park, Prediction of protein relative solvent accessibility with support vector machines and long-range interaction 3D local descriptor, Proteins 54 (2004) 557–562.
[15] S. Ahmad, M.M. Gromiha, A. Sarai, Real-value prediction of solvent accessibility from amino acid sequence, Proteins 50 (2003) 629–635.
[16] J.Y. Wang, S. Ahmad, M.M. Gromiha, A. Sarai, Look-up tables for protein solvent accessibility prediction and nearest neighbor effect analysis, Biopolymers 75 (2004) 209–216.
[17] R. Adamczak, A. Porollo, J. Meller, Accurate prediction of solvent accessibility using neural networks-based regression, Proteins 56 (2004) 753–767.
[18] Z. Yuan, B. Huang, Prediction of protein accessible surface areas by support vector regression, Proteins 57 (2004) 558–564.
[19] N.N. Minh, C.R. Jagath, Prediction of protein relative solvent accessibility with a two-stage SVM approach, Proteins 59 (2005) 30–37.
[20] D. Eisenberg, A.D. McLachlan, Solvation energy in protein folding and binding, Nature 319 (1986) 199–203.
[21] P.K. Ponnuswamy, M.M. Gromiha, On the conformational stability of folded proteins, J. Theor. Biol. 166 (1994) 63–74.
[22] A.G. Murzin, S.E. Brenner, T. Hubbard, C. Chothia, SCOP: a structural classification of proteins database for the investigation of sequences and structures, J. Mol. Biol. 247 (1995) 536–540.

[23] J.M. Chandonia, G. Hon, N.S. Walker, L. Lo Conte, P. Koehl, M. Levitt, S.E. Brenner, The ASTRAL Compendium in 2004, Nucleic Acids Res. 32 (2004) 189–192.
[24] S.F. Altschul, W. Gish, W. Miller, E.W. Myers, D.J. Lipman, Basic local alignment search tool, J. Mol. Biol. 215 (1990) 403–410.
[25] W. Kabsch, C. Sander, Dictionary of protein secondary structure: pattern recognition of hydrogen-bond and geometrical features, Biopolymer 22 (1983) 2577–2637.
[26] F. Eisenhaber, P. Argos, Improved strategy in analytical surface calculation for molecular system—handling of singularities and computational efficiency, J. Comp. Chem. 14 (1993) 1272–1280.
[27] M.M. Gromiha, M. Oobatake, H. Kono, H. Uedaira, A. Sarai, Role of structural and sequence information in the prediction of protein stability changes: comparison between buried and partially buried mutations, Protein Eng. 12 (1999) 549–555.
[28] M.M. Gromiha, M. Oobatake, H. Kono, H. Uedair, A. Sarai, Importance of mutant position in Ramachandran plot for predicting protein stability of surface mutations, Biopolymers 64 (2002) 210–220.
[29] K.A. Bava, M.M. Gromiha, H. Uedaira, K. Kitajima, A. Sarai, ProTherm, version 4.0: thermodynamic database for proteins and mutants, Nucleic Acids Res. 32 (2004) D120–D121.