J Biomol NMR (2015) 62:387–401
DOI 10.1007/s10858-015-9957-0

ARTICLE

Accessible surface area from NMR chemical shifts

Noor E. Hafsa[1] · David Arndt[1] · David S. Wishart[1,2]

Received: 3 April 2015 / Accepted: 9 June 2015 / Published online: 16 June 2015

Abstract Accessible surface area (ASA) is the surface area of an atom, amino acid or biomolecule that is exposed to solvent. The calculation of a molecule's ASA requires three-dimensional coordinate data and the use of a "rolling ball" algorithm to both define and calculate the ASA. For polymers such as proteins, the ASA for individual amino acids is closely related to the hydrophobicity of the amino acid as well as its local secondary and tertiary structure. For proteins, ASA is a structural descriptor that can often be as informative as secondary structure. Consequently there has been considerable effort over the past two decades to try to predict ASA from protein sequence data and to use ASA information (derived from chemical modification studies) as a structure constraint. Recently it has become evident that protein chemical shifts are also sensitive to ASA. Given the potential utility of ASA estimates as structural constraints for NMR we decided to explore this relationship further. Using machine learning techniques (specifically a boosted tree regression model) we developed an algorithm called "ShiftASA" that combines chemical-shift and sequence derived features to accurately estimate per-residue fractional ASA values of water-soluble proteins. This method showed a correlation coefficient between predicted and experimental values of 0.79 when evaluated on a set of 65 independent test proteins, which was an 8.2 % improvement over the next best performing (sequence-only) method. On a separate test set of 92 proteins, ShiftASA reported a mean correlation coefficient of 0.82, which was 12.3 % better than the next best performing method. ShiftASA is available as a web server (http://shiftasa.wishartlab.com) for submitting input queries for fractional ASA calculation.

Keywords Nuclear magnetic resonance · Chemical-shifts · Machine learning · Accessible surface area · Protein

Introduction

Accessible surface area is a concept first introduced and popularized by Dr. Frederic M. Richards and co-workers in the early 1970s (Lee and Richards 1971; Richards 1974, 1977). It grew from the observation that certain parts of a folded protein seemed to be impenetrable to water while other parts were highly exposed. This differential exposure seemed to be driven by the hydrophobicity or hydrophilicity of individual amino acid side chains, the 3D structure of the protein and the influence that the hydrophobic effect had on the overall protein folding process. Richards and colleagues also pointed out that water molecules are not infinitely small point particles and that the surface of a protein that was water accessible was not equal to the van der Waals surface area but rather could be calculated by rolling a ball of finite size (roughly the size of an oxygen atom of 1.4 Å) over the entire van der Waals surface of a protein. The resulting, "smoothed-surface" defined the water accessible area or the accessible surface area (ASA). ASA is a quantifiable property measured in square Angstroms (Å^2). It can be determined for

Electronic supplementary material The online version of this article (doi:10.1007/s10858-015-9957-0) contains supplementary material, which is available to authorized users.

✉ David S. Wishart
david.wishart@ualberta.ca

[1] Department of Computing Science, University of Alberta, Edmonton, Canada

[2] Department of Biological Sciences, University of Alberta, Edmonton, Canada

entire proteins or for individual residues or even atoms. ASA can also be re-cast as a fractional accessible surface area (fASA) that reports the percentage of ASA relative to a fully exposed protein (or residue). This concept can be carried further to a relative accessibility, or RSA, which is a more qualitative measure of surface accessibility. With the RSA concept, residues are considered buried (B), partially buried (P) or exposed (E) based on their fASA. Typically buried residues have a fASA of <0.25, partially buried have a fASA between 0.25 and 0.5 and exposed residues have a fASA of >0.50.

Since its first description, the concept of ASA has proven to be extremely useful for assessing the quality of protein folds and for scoring protein structure predictions (Benkert et al. 2008), for assessing conformational changes upon protein or ligand binding, for calculating protein folding energies, for determining protein–ligand binding constants and for calculating protein enthalpy and entropy changes (Lavigne et al. 2000). More recently, indirect measurements of residue-specific ASAs through targeted chemical modification or partial proteolysis have been used to provide constraints for low-resolution protein structure determination efforts by mass spectrometry (Serpa et al. 2014). Indeed since its first description some 40 years ago, the concept of ASA has probably been proven to be among the most useful concepts for understanding, comparing and evaluating protein folds and protein functions.

Quantitative ASA measurements can only be determined from protein coordinate data (i.e. solved structures). However, given the utility of ASA measurements as structural constraints or for evaluating structural/thermodynamic properties of proteins, there has been a growing interest in finding ways of predicting ASA, fASA or RSA from sequence data alone. As a result there have been a number of published studies that describe methods for predicting accessible surface area and relative surface accessibility from sequence (Ahmad and Gromiha 2002; Ahmad et al. 2003; Wagner et al. 2005; Petersen et al. 2009; Nguyen and Rajapakse 2005; Li and Pan 2001; Pollastri et al. 2002; Chen and Zhou 2005; Naderi-Manesh et al. 2001; Thompson and Goldstein 1996; Rost and Sander 1994; Garg et al. 2005; Yuan and Huang 2004; Holbrook et al. 1990; Adamczak et al. 2004). The majority of these prediction systems rely on using multiple sequence alignments, pairwise residue assessments and the predictive power of machine-learning algorithms. The best performance reported by these sequence-only methods using a two-state (Buried, Exposed) and a three-state RSA measure (Buried, Partially Buried, Exposed) yielded Q_2 and Q_3 scores of 88 and 63 % respectively (Ahmad and Gromiha 2002). For real-value ASA predictions, the best performance so far reported used PSSM matrices from PSI-BLAST (Altschul et al. 1997) profiles in a two-stage support-vector regressor to achieve a correlation coefficient between observed and calculated ASA of 0.68 (Nguyen and Rajapakse 2005).

While these sequence-only results are promising, Rost and Sander (1994) pointed out that surface accessibility is less conserved in structural homologs than secondary structure and therefore ASA would be predicted less accurately from homology modeling. The Rost et al. study also showed that the correlation coefficient of relative solvent accessibility between 3D homologues (by structural alignment) is only 0.77, whereas prediction of accessibility by homology modeling (sequence alignment) resulted in a correlation coefficient of about 0.68. This suggests that the upper limit of ASA prediction that could be achieved by sequence-only methods would yield a correlation of 0.70–0.75.

Over the last two decades it has been observed that a number of experimentally measurable properties in proteins correlate reasonably well with accessible surface areas. For instance, folding and unfolding free energies as measured through calorimetry appear to correlate quite well with ASA or fASA (Myers et al. 1995). Protease cleavage sites or protease susceptibility along with chemical modification susceptibility also appears to map with solvent accessibility (RSA or ASA) (Croy et al. 2004). Hydrogen exchange, as measured by mass-spectrometry (MS) or NMR also allows the identification of buried and exposed residues in proteins (Huyghues-Despointes et al. 1999). NMR Chemical shifts also appear to be influenced by ASA effects. The first evidence of such a phenomenon was reported in 1994 (Wishart and Sykes 1994). Nearly a decade later Avbeli et al. (2004) studied the effect of secondary structure and solvent exposure on backbone chemical shifts. They demonstrated that proton secondary shifts have a different chemical shift distribution for solvent exposed residues, particularly in smaller peptides. In a later study by Vranken and Rieping (2009), the effect of secondary structure and solvent exposure on chemical shift assignments was re-examined on a large database of proteins for which both reported atomic coordinates and chemical shift values were available. There were two major findings from this study. First, they found that non-polar atoms have significantly larger chemical shift dispersion and a somewhat different chemical shift distribution compared to polar atoms. Secondly those atoms with greater atomic ASA, exhibited chemical shift values that tended towards random coil values. The relationship between chemical shifts and ASA was actually used to develop a significantly improved structure-based chemical shift prediction algorithm, called ShiftX2 in 2011 (Han et al. 2011). Most recently, Berjanskii and Wishart (2013) proposed a simple formula to calculate per-residue fASA from side-chain chemical shifts and observed a correlation of more than 70 % with the observed fASA values over a subset of 15 proteins.

These studies demonstrate that both sequence and chemical shift information can be used individually to estimate the ASA values with reasonable accuracy. Now the question is: Can one develop more accurate fASA estimation by more intelligently combining sequence AND chemical shift information? Here we report the development of a machine-learning based method that can be used to accurately estimate per-residue fractional ASA of water-soluble proteins using sequence and chemical shifts. After training on a set of 30 fully assigned proteins, the performance of the resulting model, called ShiftASA, was compared with other sequence-based and chemical-shift based methods over a test set of 65 proteins. For this test set ShiftASA achieved a mean correlation coefficient of 0.79 compared to correlation coefficients of 0.73 and 0.59 found for sequence-only and chemical shift-only methods respectively. On a separate test set of 92 proteins, ShiftASA attained a correlation coefficient of 0.82. A number of other statistical measures were also used to prove that this method shows a consistently better performance than any existing method.

Materials and methods

Dataset

A set of 30 proteins with complete experimental NMR chemical shift assignments and available high-resolution X-ray structures was chosen for training purposes. The list of proteins along with their PDB and BMRB identifiers is provided on the ShiftASA website. Note that the number of training proteins was varied to examine any enhancement in training and test performance. However no (or very little) improvement was observed with an increased number of proteins. Two separate sets of 65 and 92 proteins with available experimental chemical shifts and high-resolution X-ray structures were used as independent test sets. Henceforth we shall refer to the training data set and two test data sets as TRAIN, TEST1 and TEST2, respectively. The list of the TEST1 and TEST2 proteins along with their PDB and BMRB identifiers is provided on the ShiftASA website. No two proteins shared more than 40 % sequence identity in the TRAIN set. Similarly, no two proteins shared more than 40 % sequence identity in the TEST1 and TEST2 sets. The TRAIN proteins had ~92 and ~83 % of their backbone and side-chain chemical shifts assigned, respectively. The TEST1 proteins had on average ~90 % (max = 100 %, min = 49 %) and ~60 % (max = 91 %, min = 0 %) of their backbone and side-chain chemical shifts assigned while those in TEST2 had an average of ~97.50 % (max = 100 %, min = 85 %) and ~83.5 % (max = 89 %, min = 53 %) of their backbone and side-chain chemical shifts assigned. Note that no attempt was made to handle missing assignments in either the training or the test data sets. The TRAIN proteins had ~49 % of their resides in regular secondary structure while the TEST1 and TEST2 proteins had ~63 and ~44 % (respectively) of their residues in regularly secondary structure as assessed by STRIDE (Frishman and Argos 1995).

Computation of observed fractional ASA

Most predictive studies associated with ASA prediction have focused on generating RSA or binary/ternary class predictions. However, in the majority of cases, real-valued or fractional ASA is more informative than the binary/ternary classification of residues into buried or exposed states. This is because the threshold for classifying residues in a protein into two or three exposure classes is subjective and often depends on the mean ASA over all the residues in a particular protein (Ahmad et al. 2003). In the absence of a universal threshold for categorical prediction of buried and exposed states, fractional ASA (fASA) is considered to be more reliable or useful estimation of residue-specific solvation status. Therefore for this study we focused on developing a predictor for fASA. The fractional ASA of a residue is defined as the ratio between absolute ASA (aASA) calculated within a three-dimensional structure and that is observed for a central residue location in an extended tri-peptide (*Ala-X-Ala*) conformation, denoted as mASA:

$$fASA_i = \frac{aASA_i}{mASA_i}$$

Hence, fASA values range between 0.0 and 1.0, with 0.0 corresponding to a fully buried and 1.0 to a fully exposed residue, respectively. Absolute ASA values were calculated using the Dictionary of Secondary Structure Prediction (DSSP) (Kabsch and Sander 1983) program. The values of the extended state ASAs for all 20 residues were extracted from Eisenhaber and Argos (1993).

Mapping fractional ASA prediction as a regression task

Given a protein with a length of *n* amino acids, the task is to estimate the fASA at each residue. We initially mapped the estimation problem as a regression task and then employed a Stochastic Gradient Boosting Tree model to solve the regression problem as outlined by Ridgeway (2007) and Trevor et al. (2001). To map the problem as a regression task we defined an error function as the square of the difference between the observed per-residue fASA values and the predicted per-residue fASA values over the length of the training set sequences. The predicted per-residue fASA was calculated from a set of features (see below) and expressed as function $f^*(x)$, of amino acid position or sequence length. In stochastic gradient boosting, the method approximates the function $f^*(x)$, in an iterative fashion through fitting the solution tree in each step that maximally reduces the expectation of the error function. The gradient step in each iteration m ($m = 1\ldots T$, where T = total number of iterations), updates the model according to a learning rate or a shrinkage parameter that controls the rate at which the boosting algorithm descends upon the error surface. For each iteration, only a fraction p of the N training observations is randomly sampled (without replacement) and the next solution tree is grown with that subsample. The solution tree that is generated for each boosting iteration is a K-terminal node regression tree.

After mapping the fractional ASA prediction problem into a Stochastic Gradient Boosted Tree Model (SGBM), the model was optimized on the protein data in the training set. The "GBM" package (Ridgeway 2007), written in R (R Development Core Team 2008) was used for optimizing the training model.

Feature set

To use or develop machine-learning algorithms it is necessary to extract a set of input features from the training data that will be used to infer or calculate the desired output (i.e. the fractional ASA). Features can either be the raw data (i.e. sequence, NMR chemical shifts, etc.) or derived data (i.e. estimated hydrophobicity) that is calculated from the raw data. We derived a set of five different feature types from our chemical shift and sequence data. The features included: (1) residue specific hydrophobicity, (2) chemical shift-derived three-state secondary structure probability, (3) random coil index values relating to flexibility using backbone and side-chain chemical shifts (Berjanskii and Wishart 2005, Berjanskii and Wishart 2013), (4) multiple sequence alignment derived residue conservation score (Valdar 2002; Mayrose et al. 2004), and (5) SABLE predicted ASA (Adamczak et al. 2004). These features are explained in more detail below.

Residue specific hydrophobicity

Hydrophobicity is a widely used physico-chemical characteristic of amino acids that is used to measure their relative water aversion. Hydrophobicity scales are numeric scales that define the relative hydrophobicity of amino acid residues. In general terms, the more positive the number, the more hydrophobic the amino acid, and consequently the more buried it is likely to be. Over the past few decades, a number of different hydrophobicity scales have been developed. We investigated six different hydrophobicity scales to see which gave the best prediction results on a validation set when combined with other features. The scales we examined included Janin's scale (Janin 1979), Kyte and Doolittles's scale (Kyte and Doolittle 1982), Eisenberg's scale (Eisenberg et al. 1984), Engelman's scale (Engelman et al. 1986), Hopp and Woods scale (Hopp and Woods 1981) and Manavalan's scale (Manavalan and Ponnuswamy 1978). The best correlation was achieved using Janin's hydrophobicity values (data not shown). Interestingly Janin's scale was developed by analyzing the relative surface accessibility of all 20 amino acid residues from solved protein structures. In this regard Janin's scale is more a solvent accessibility scale than a hydrophobicity scale.

Two different approaches were examined regarding how to use hydrophobicity as a feature: (1) single-residue hydrophobicity and (2) a running average of hydrophobicity over a 3-residue window. The first approach exhibited a comparatively better correlation than the second one (data not shown) and so this was incorporated in our feature set.

Chemical-shift derived secondary structure probability

The secondary structure probability of a residue is derived from the secondary chemical shift value of its constituent atoms. The secondary chemical shift (Δδ) is defined as the difference between the absolute chemical shift (δ_{abs}) and the corresponding random coil (δ_{rc}) shift (Wishart 2011).

$$\Delta\delta = \delta_{abs} - \delta_{rc}$$

The probability of a residue being in one of the three states "α-helix", "β-strand" or "coil" is derived from its six backbone atom secondary chemical shifts, as described in Wang and Jardetzky (2002). For each backbone atom, a Gaussian probability distribution was assumed, where the two parameters of the distribution corresponded to (1) the average secondary chemical shift value for each of three different secondary structure states and (2) the standard deviation of the distribution. These statistical parameters were derived from the "RefDB" database (Zhang et al. 2003). A more detailed description of the secondary structure probability method is given by Wang and Jardetzky (2002).

Random coil index

The random coil index (RCI) is a technique that can be used to determine the flexibility of an amino acid residue in a polypeptide chain from its backbone and side-chain chemical shifts (Berjanskii and Wishart 2005, 2013). Both the backbone and side-chain RCI quantitatively trace the relative amount to which a protein backbone and side-chain's chemical shifts match with the random coil values. These features were calculated using the RCI equations provided in the original RCI papers.

Residue conservation score

Residue conservation is a measure of how often a given residue is seen at an equivalent position, in an equivalent protein, across different species. Generally highly conserved residues are buried within the protein core, while less conserved residues are generally exposed or found in loops (albeit with some exceptions). The conservation score for each residue position is calculated as described by Valdar (2002). First, a PSI-BLAST (Altschul et al. 1997) search with three iterations for query sequence is done on UniREf90 clustered database (UniProt Consortium. 2010). Then a multiple sequence alignment is performed using ClustalOmega (Sievers et al. 2011). The conservation score for each column in the alignment (each residue in the target sequence) is then calculated using Shannon's entropy formula as described below,

$$s(x) = \lambda \sum_{a}^{K} p_a \log p_a$$

where, p_a is the probability of observing the ath amino acid and, λ is the scaling factor and defined as,

$$\lambda = [\log(\min(N, K))]^{-1}$$

where N = number of sequences in the alignment, K = length of amino acid alphabet. The probability of observing ath amino acid is the summed weight of sequences having the symbol a in the position x in the sequence which is defined as,

$$p_a = \sum w_i$$

where, w_i is the weight of the ith sequence. w_i is defined as,

$$w_i = \frac{1}{L} \sum_{x}^{L} \frac{1}{k_x n_x}$$

where, L = length of the alignment, k_x = the number of amino-acid types present at the xth position, n_x = the number of times the ath amino acid occurring in the ith sequence at the xth position.

SABLE-predicted ASA

To further improve the performance of ShiftASA we supplemented our method with another sequence-only ASA prediction tool called SABLE (Adamczak et al. 2004). SABLE is a pure sequence-based method for predicting real-valued relative solvent accessibilities of amino acid residues in proteins. It was initially developed using neural network based regression models and later refined using other linear regression models (Wagner et al. 2005). It has a reported correlation coefficient between predicted and experimental values of 0.64–0.67 on various test sets. Because SABLE's correlation coefficient was comparable to the reported correlation of shift-based ASA estimations, it was expected that including sequence estimated ASA would enhance the performance of ShifASA. Therefore the SABLE predicted real valued ASA for each residue was included in the ShiftASA feature vector for the training and test data points.

Local residue interactions

To take into account the local-residue interaction in the protein structure, a 3-residue window feature set was used throughout this study. Accounting for nearby residue-interactions provides important information about local geometry and the local environment that is accessible/non-accessible to solvent.

Training the prediction model

The prediction model parameters were optimized so as to obtain an estimator that minimized the (absolute) difference between actual output and predicted ASA values. The model was also optimized to achieve a better correlation between the observed and response (i.e. predicted) variables. With those two objectives in mind, a repeated tenfold cross-validation (CV) was performed to estimate the optimal number of iterations (*n.trees*, *T*) and interaction depth of each regression tree (*interaction.depth*, *K*) for our SGBM. This was done after the model had been initially fit on the set of 30 sample observations.

Optimization using tenfold repeated cross-validation (CV) suggested that the optimal number of iterations should be 180. That is, the final regression model best approximates the response value after 180 gradient steps. The second parameter estimated by the optimization process was the optimal depth of interaction among the predictor variables in each regression tree. The optimal depth of interaction was found to be eight (8). Specifically, the loss function was minimized when eight predictor variables were split in each regression tree.

Analysis of feature influence

During the optimization of ShiftASA, an analysis of the feature influence was performed as a part of the boosting

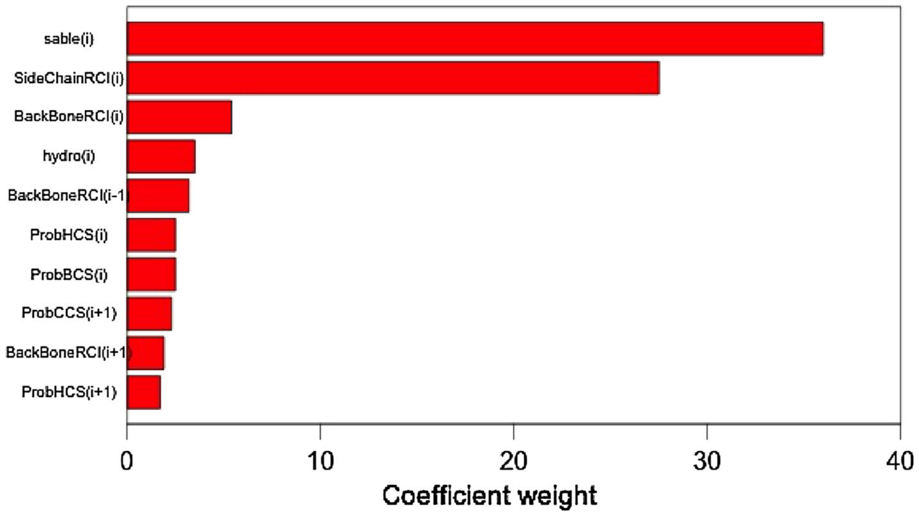


Fig. 1 Top ten relevant features in the SGBM method. The importance of these predictor variables or features is normalized to a scale of 1–100. The predictor variable names are shown on the *vertical axis*

process. The top ten features are shown in Fig. 1. The influence of the predictor variable (∼X) indicates the relative importance or contribution of that variable in predicting the response (∼Y) and can be estimated by the weighting coefficient associated with that variable in the method formulation. This analysis helped to identify those variables that had the most significant influence on the response. The weighting coefficients of all features are described in Table S2 (see Supplementary Information).

Evaluation

The performance of ShiftASA was evaluated using several different metrics. This was done to more completely ascertain its performance against other methods as well as to better assess the effects brought on by using different weighting protocols. Specifically the following metrics were used:

1. Root mean square error (RMSE)—RMSE is a statistical measure that calculates the difference between the values predicted by an estimator model and the actual observed values. RMSE is the square root of the average squared deviation between predicted and actual values, and thus gives larger deviations more weight. A smaller value indicates a better model performance;
2. R^2 or the coefficient of determination—R^2 is a statistical measure that indicates how well a set of data points fit to a regression line or curve;
3. Spearman's rank correlation coefficient (SRCC)—SRCC is a non-parametric measure of the monotonic relationship between two variables, irrespective of whether their relationship is linear;
4. Mean absolute error (MAE)—MAE is the average of the absolute errors in a prediction i.e. the absolute difference between predicted and true values in a set of outcomes. Unlike other measures, larger deviations are not given additional weight;
5. Mean squared error (MSE)—MSE measures the average of the square of the "error" or deviation of the estimator from the quantity being estimated. MSE tends to heavily weight outliers.

Results and discussion

Training performance and feature importance

During the optimization process, a tenfold repeated cross-validation protocol yielded the lowest RMSE (0.18) and the best R-squared values (0.65) for the training data. The weighting coefficients of all features are described in Table S2. These data indicate that the SABLE (Adamczak et al. 2004) estimated ASA at the central (*i*)th residue is the most informative ASA predictor. The side-chain random coil index, backbone random coil index and hydrophobicity, were found to be next three most influential variables in our fASA estimation. The next most important feature was the random coil index value of the (*i*−1)th residue followed by helix propensity of the (*i*)th and β-strand propensity of the (*i*)th residue. The helix and β-strand propensities of the central residue have comparatively higher importance they often indicate that this residue is buried as buried residues have a higher propensity to form α-helices and β-sheets in proteins and have a tendency to interact with the residues in the core region. Our analysis shows that central residue features carry the most information content (occupying six of the top seven positions), with exceptions of the flexibility information of neighboring residues [the RCI value of the (*i* − 1)th residue]. Although the SABLE estimation is found to be the most relevant feature, the chemical shift features also provide significant contribution a, roughly equal to the SABLE feature at central residue position. Residue hydrophobicity also carries useful information to estimate the fASA. Other than the SABLE ASA estimation and hydrophobicity, eight of the top ten features are chemical-shift features, and have collectively larger weights in the final formulation. It is notable that residue conservation scores are not present among ten most relevant features, which indicates their somewhat smaller contribution to the feature set.

Test performance

The final parametric regression tree model generated by repeated cross-validated optimization of the TRAIN set was used to predict the fractional ASA values for proteins in the TEST1 and TEST2 data sets. The Spearman correlation coefficient was calculated between the actual fASA and the predicted fASA using both our ShiftASA method and five other models. The results are shown in Tables 1 and 2 (Table S1 lists the individual performance of all TEST1 proteins). As seen in Table 1, the mean correlation coefficient for the predicted fASA values for ShiftASA of 0.79. This corresponds to an 8.2 % improvement over the best sequence-only and a 22 % improvement over chemical shift-only prediction methods. The prediction accuracy of the different methods was also evaluated using other statistical metrics and is shown in the same table.

Table 1 also shows that ShiftASA reports the highest mean prediction accuracy among all five methods that were evaluated. The mean absolute error was decreased from 0.20 (the best MAE among other methods) to 0.14 Å^2 with ShiftASA, which is a 26 % improvement over the best sequence-only method. The mean squared error also decreased to 0.03 from 0.07 Å^2 as measured over all TEST1 proteins. Moreover, ShiftASA shows the lowest deviation in Spearman's rank correlations. These data indicate that ShiftASA is not only the most accurate, but also the most consistent among the five methods. Bar plots exhibiting the mean Spearman's rank correlations and the corresponding standard deviations reported by the five methods are shown in Fig. 2.

The Spearman correlation coefficients for the TEST2 proteins as well as other statistical measures from predictions derived by ShiftASA, SABLE (Adamczak et al. 2004) and Side-chain RCI (Berjanskii and Wishart 2013) are shown in Table 2. As seen in this table, ShiftASA estimation has a correlation of 0.82, whereas SABLE's and Side-chain RCI's correlations are 0.67 and 0.73 respectively. The mean absolute error is also significantly decreased (0.14 compared to 0.31 and 0.14 compared to 0.23).

Examples of the per-residue correlation for the predicted fASA values of two protein chains, the double-sided ubiquitin binding of Hrs-UIM (PDB ID: 2D3G(B)) and a putative dinitrogenase iron-molybdenum cofactor from *Thermotoga maritima* (PDB ID: 1O13(A)) are displayed in Figs. 3 and 4 respectively. The first example shows a strong correlation (0.82) between SHIFTASA and the observed fASA. For the second example, a stronger correlation (0.86) is evident, compared to the correlations (0.72, 0.62 and 0.67) reported by three other methods, namely SABLE (Adamczak et al. 2004), RVPNet (Ahmad et al. 2003), and Side-chain RCI (Berjanskii and Wishart 2013). As seen in Figs. 3 and 4, ShiftASA yields better agreement in matching the observed ASA amplitude, which certainly contributes to its higher correlation coefficients.

Buried-exposed and buried-intermediate-exposed classification

Categorical ASA measures are still commonly used in the field of ASA prediction and evaluation. However, no universal threshold for categorical prediction of buried and exposed states exists and so fractional ASA (fASA) is

Table 1 The mean absolute error (MAE), mean squared error (MSE), root mean squared error (RMSE), mean Spearman's correlation, and the standard deviation of Spearman's correlation for all six fASA prediction methods (including ShiftASA, with and without SABLE) evaluated over the TEST1 set (65 test proteins) [64 proteins for side-chain RCI]

Evaluation metric	ShiftASA (w SABLE)	ShiftASA (w/o SABLE)	SABLE (Seq.)	RVPNet (Seq.)	SARpred (Seq.)	Side-chain RCI (Chem. shift)
MAE	0.14	0.16	0.19	0.20	0.24	0.20
MSE	0.03	0.04	0.07	0.07	0.09	0.07
RMSE	0.19	0.02	0.26	0.26	0.31	0.26
Minimum Spearman correlation	0.72	0.70	0.54	0.47	0.21	0.22
Mean Spearman correlation	0.79	0.76	0.73	0.60	0.38	0.59
Maximum Spearman correlation	0.86	0.83	0.82	0.70	0.67	0.77
SD (Spearman correlation)	0.04	0.03	0.07	0.05	0.15	0.12

Table 2 The mean absolute error (MAE), mean squared error (MSE), root mean squared error (RMSE), mean Spearman correlation coefficient, and the standard deviation of Spearman correlation coefficient for ShiftASA (with and without SABLE), SABLE (Adamczak et al. 2004) and Side-chain RCI (Berjanskii and Wishart 2013) evaluated over the TEST2 set

Evaluation metric	ShiftASA (w SABLE)	ShiftASA (w/o SABLE)	SABLE (Seq.)	Side-chain RCI (Chem. shift)
MAE	0.14	0.15	0.31	0.23
MSE	0.03	0.04	0.16	0.09
RMSE	0.18	0.02	0.41	0.30
Minimum Spearman correlation	0.67	0.76	0.20	0.27
Mean Spearman correlation	0.82	0.79	0.67	0.73
Maximum Spearman correlation	0.89	0.88	0.85	0.84
SD (Spearman correlation)	0.04	0.03	0.12	0.07

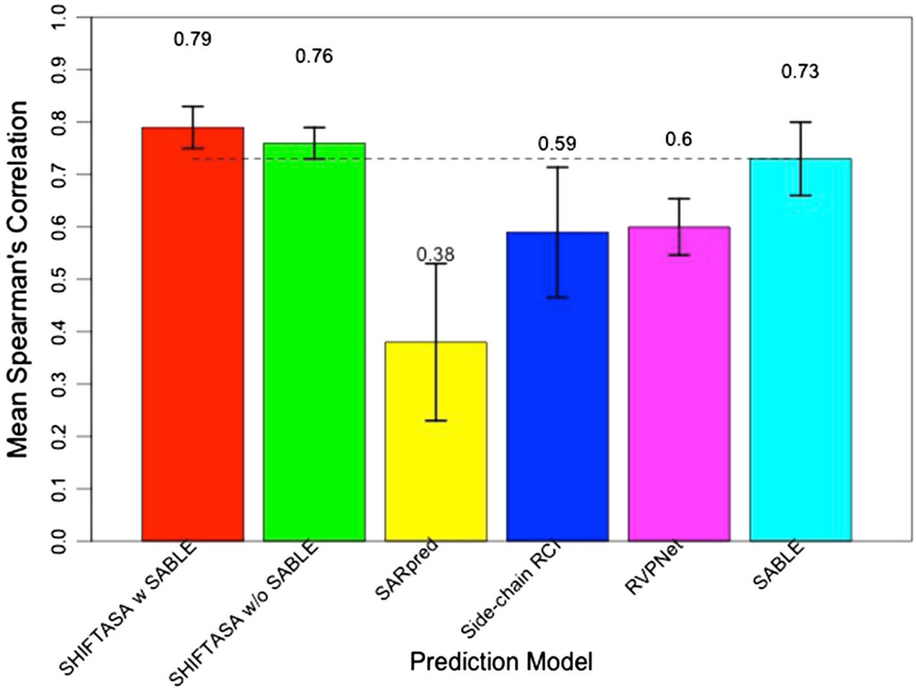


Fig. 2 Mean Spearman correlation coefficient and the standard deviation of correlations of all five fASA prediction models (including ShiftASA) are shown. The performance is measured over the TEST1 data set. The mean correlation associated with each method is shown at the *top of each bar* diagram

generally considered to be a more reliable estimation of residue-specific solvation status. Nevertheless, we performed a detailed evaluation of ShiftASA's performance for categorical ASA prediction. Two-state and three-state classification of residue fractional ASA values for different threshold systems were calculated based on the real-value fASA predictions by ShiftASA, SABLE (Adamczak et al. 2004) and RVPNet (Ahmad et al. 2003). The number of residues in each (Exposed, Intermediate or Buried) class using different threshold cutoffs is described in Table 3. The accuracy and precision of classification results are shown in Table 3.

The performance of ShiftASA for two-state and three-state classification of real-value solvent accessibility using different threshold values was found to be comparable or higher (in most cases) than that of RVPNet (Table 3). ShiftASA also showed consistently high accuracy (≥80 %) for all threshold values in the two-state classification. Three-state classifications (buried-intermediate-exposed) were challenging for the current method (although ShiftASA reports better accuracies than RVPNet). The probable reason might be the lower estimation accuracies associated with more exposed residues (fASA range ≈0.6–1.0—see "Discussion" section for more details).

Discussion

The performance of ShiftASA is clearly superior to other methods for fASA prediction. Obviously the inclusion of

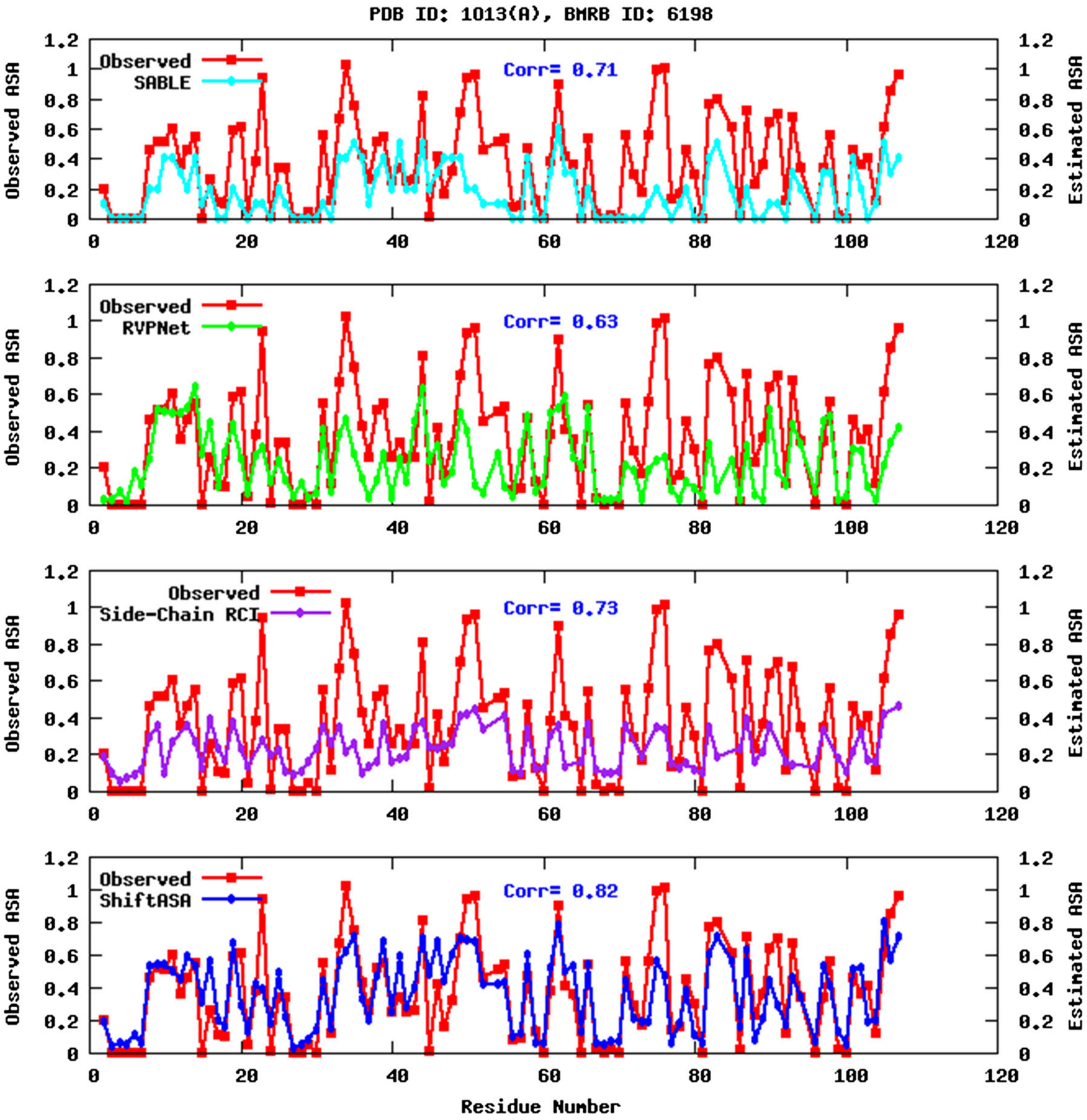


Fig. 3 Agreement between predicted and observed residue-specific fASA values by SABLE, RVPNet, Side-chain RCI and ShiftASA for the putative dinitrogenase iron-molybdenum cofactor from *Thermotoga maritima* (PDB ID: 1O13, chain A). The corresponding BMRB ID is 6198. The Spearman correlation coefficient is shown in the *centre* of each graph

experimental information (i.e. NMR chemical shifts) means that additional information, beyond sequence data, is being exploited in the prediction process. However, unlike most other sequence-based fASA predictors, ShiftASA also makes use of residue-specific hydrophobicity to help with its prediction. This is based on the fact that there is a strong relationship between residue-specific hydrophobicity and solvent exposure (Manavalan and Ponnuswamy 1978). Indeed, several hydrophobicity scales have been derived by calculating the solvent accessible surface area for residues in solved proteins or by employing empirical solvation parameters derived from calculated surface areas (Chothia 1976; Biswas et al. 2003). Because ShiftASA employs both chemical shifts and residue-specific hydrophobicity, it would be of interest to analyze the predictive ability of sequence alone and chemical shift alone to estimate fractional ASA values. In addition it would also be useful to explore how ShiftASA's prediction accuracy varies as the fraction of complete shift assignment changes. In the following subsections, we investigated these two issues along with other issues based on the performance of the TEST1 proteins.

Sequence and chemical-shift based prediction: combined versus alone

To address the issue of predictive accuracy for sequence-only versus shift-only versus combined, another two

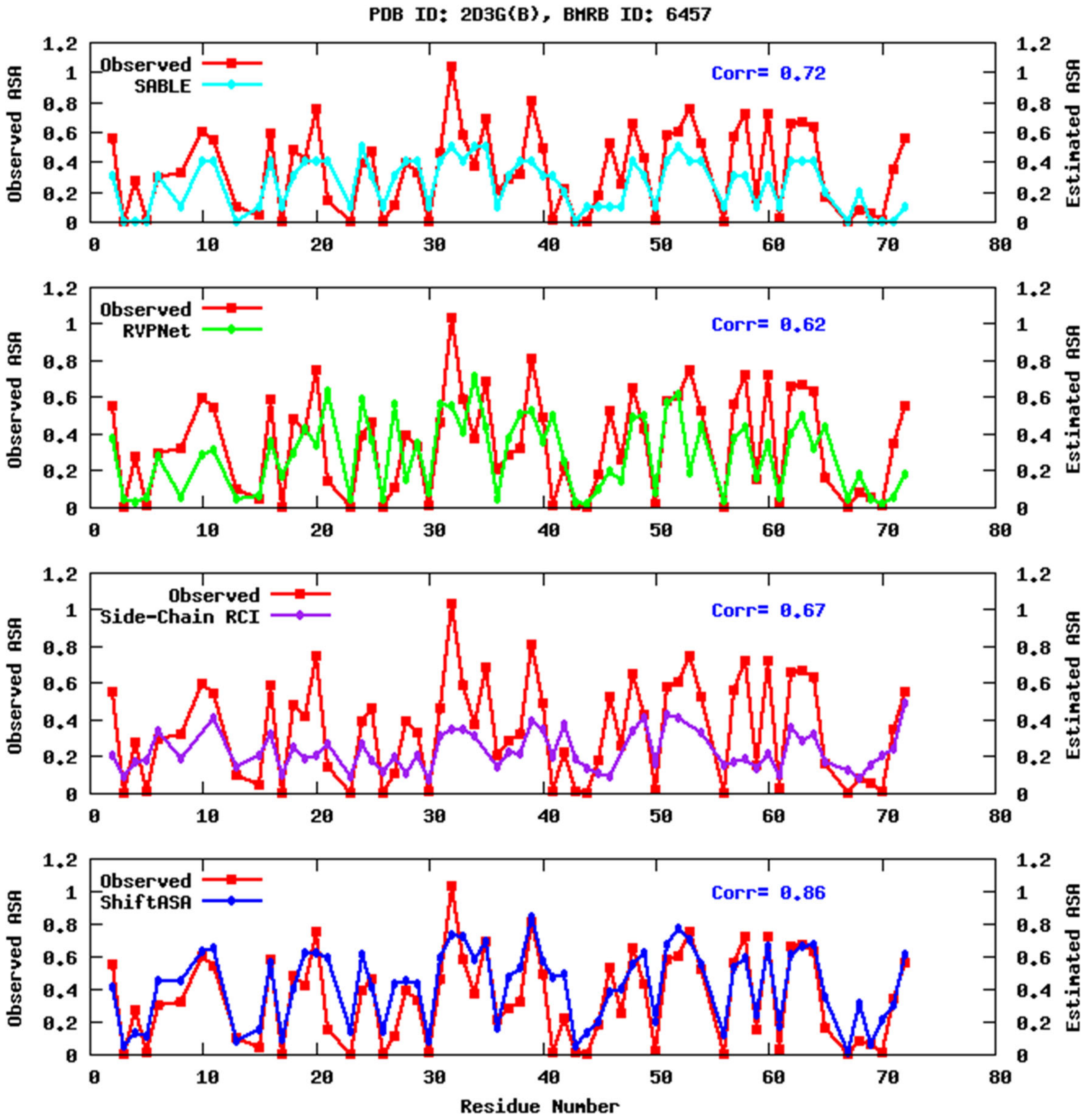


Fig. 4 Agreement between the predicted and observed residue-specific fASA values for SABLE, RVPNet, Side-chain RCI and ShiftASA for ubiquitin bound to Hrs-UIM (PDB ID: 2D3G, chain B). The corresponding BMRB ID is 6457. The Spearman correlation coefficient is shown on the *right* of each graph

stochastic gradient boosting regression tree models were developed and trained using sequence-only and chemical shift-only training features for each residue in a 3-residue window. Parameter optimization indicated the optimal “number of trees” as 150 and 225 and the optimal “interaction depth” as six (6) and eight (8) respectively for the final regression trees of these sequence-only and chemical shift-only models. The optimized models were then evaluated on TEST1 proteins. Figure S1 shows the correlation between the actual and predicted fASA values using the sequence-only and chemical shift-only prediction models. For comparative purposes, the correlations of ShiftASA’s predictions are also shown. The graph clearly shows a significant performance difference. Note that, the mean correlation for our sequence-only prediction is 0.60 (for the 65 protein test set). The green line in the graph shows the correlation between chemical shift derived parameters and fASA values, which is 0.46. These results show that the performance improvement seen for ShiftASA was not just achieved through the use of sequence-derived parameters, but also by the sensible use of chemical shift data.

It is also interesting to compare this sequence-only method with the four sequence-only fASA prediction systems we evaluated in this study, namely, SABLE (Adamczak et al. 2004), RVPNet (Ahmad et al. 2003) and SARpred (Garg et al. 2005). SABLE, RVPNet and

Table 3 Two-state and three-state mean classification accuracy and precision of fASA for the TEST1 set reported by ShiftASA, SABLE (Adamczak et al. 2004) and RVPNet (Ahmad et al. 2003) for different threshold systems

Threshold system	ShiftASA		SABLE		RVPNet	
	Accuracy	Precision	Accuracy	Precision	Accuracy	Precision
0 % (2-state) #B = 823, #E = 6711	0.90	0.89	0.80	0.80	0.89	0.90
5 % (2-state) #B = 1672, #E = 5862	0.81	0.80	0.84	0.82	0.79	0.80
10 % (2-state) #B = 2202, #E = 5332	0.82	0.78	0.76	0.75	0.77	0.77
15 % (2-state) #B = 2603, #E = 4931	0.82	0.80	0.78	0.78	0.76	0.76
25 % (2-state) #B = 3420, #E = 4114	0.81	0.78	0.76	0.73	0.74	0.74
50 % (2-state) #B = 5246, #E = 2288	0.81	0.80	0.70	0.66	0.72	0.73
10–20 % (3-state) #B = 2202, #I = 808, #E = 4524	0.71	0.68	0.72	0.62	0.67	0.67
15–25 % (3-state) #B = 2603, #I = 817, #E = 4114	0.72	0.69	0.66	0.67	0.65	0.65
25–50 % (3-state) #B = 3420, #I = 1826, #E = 2288	0.66	0.62	0.58	0.50	0.54	0.55

The number of buried (#B), intermediate (#I) and exposed (#E) residues for each threshold system are shown in the first column

SARpred all are neural network-based prediction systems. SABLE uses feed-forward neural networks that estimate real value RSA's based on information derived from the PSI-BLAST (Altschul et al. 1997) position specific scoring matrix (PSSM), hydrophobicity, volume, entropy and secondary structure propensity of amino acids in a running window of 11 residues. The second method (RVPNet) uses only sequence data with adjacent neighbor information encoded in a binary (0/1) sequence, and the last method (SARpred) method feeds multiple sequence alignments into a two-stage neural network to predict fASA. For the TEST1 proteins, SABLE achieved a mean correlation of 0.73 and RVPNet achieved a correlation of 0.60, whereas the correlation of SARpred was found to be 0.38. With the exception of SABLE, the other two methods (RVPNet and SARpred) appear to be either comparable or significantly worse than our sequence-only approach.

Prediction error versus complete shift assignments

Recent studies (Marsh 2013; Berjanskii and Wishart 2013) have demonstrated a correlation between fASA and local flexibility as well as global flexibility. Marsh (2013) found a mean correlation of 0.61 between RCI-predicted local flexibility and residue-specific fractional ASA over a set of monomeric proteins. Likewise, Berjanskii and Wishart (2013) found a correlation of 0.74 between the side chain RCI (a chemical shift-derived parameter) and residue specific fASA over a set of 15 proteins. However, one of the limitations of these RCI-based methods is that a complete or near-complete chemical shift assignment is required to achieve relatively moderate prediction accuracy. In the present study, we found the mean Spearman correlation of the side-chain RCI method over protein in TEST1 to be 0.59, which was somewhat less than what was originally reported (albeit using a different set of proteins). It was also found that the side-chain RCI method was particularly sensitive to missing or incomplete assignments. This is reflected in the spread of up to 12 % in the Spearman correlation coefficient distributed over the TEST1 set. Fortunately, one of the strengths of ShiftASA is the fact that it is not solely dependent on side-chain chemical shifts but also on the relatively more easily measured backbone chemical shifts. Furthermore, when all of the described sequence and chemical-shift derived features (see "Materials and methods" section) are combined, ShiftASA's accuracy does not vary significantly in the absence of complete shift assignments. This invariance is shown using the line connected by red diamonds in Figure S2. We believe the robustness that ShiftASA exhibits to missing chemical shifts is due to the redundancy in information that is available from both sequence and

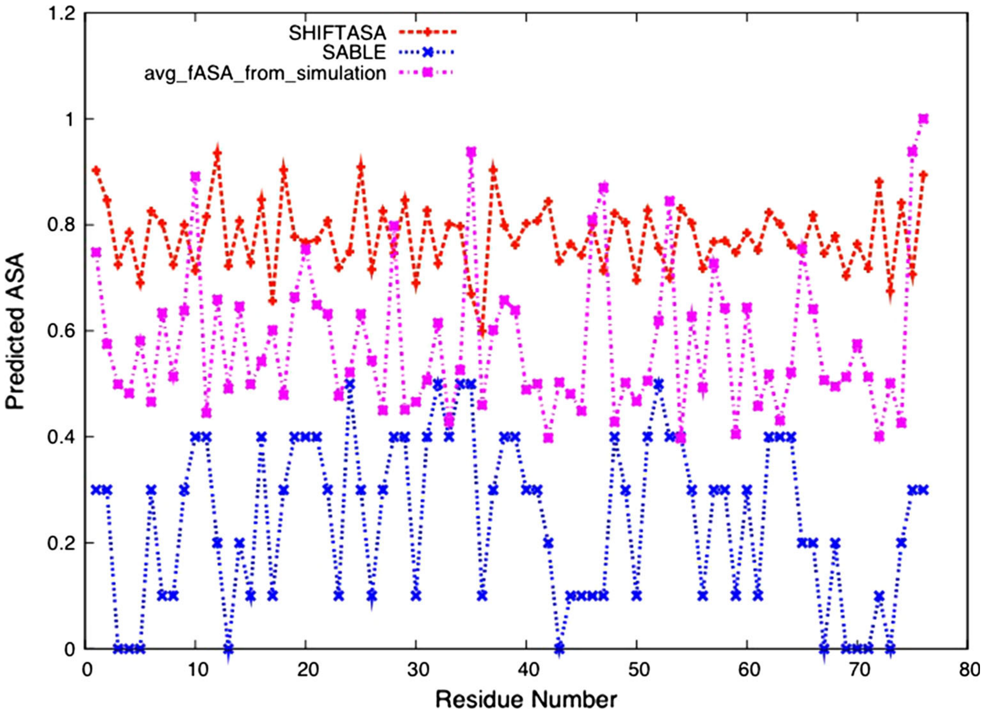


Fig. 5 Per-residue fASA values for unfolded ubiquitin (BMRB ID: 4357). The *red*, *green* and *blue line* indicates the estimated fASA by ShiftASA, the average fASA from 10,000 simulated unfolded structures of ubiquitin and the estimated fASA by SABLE respectively

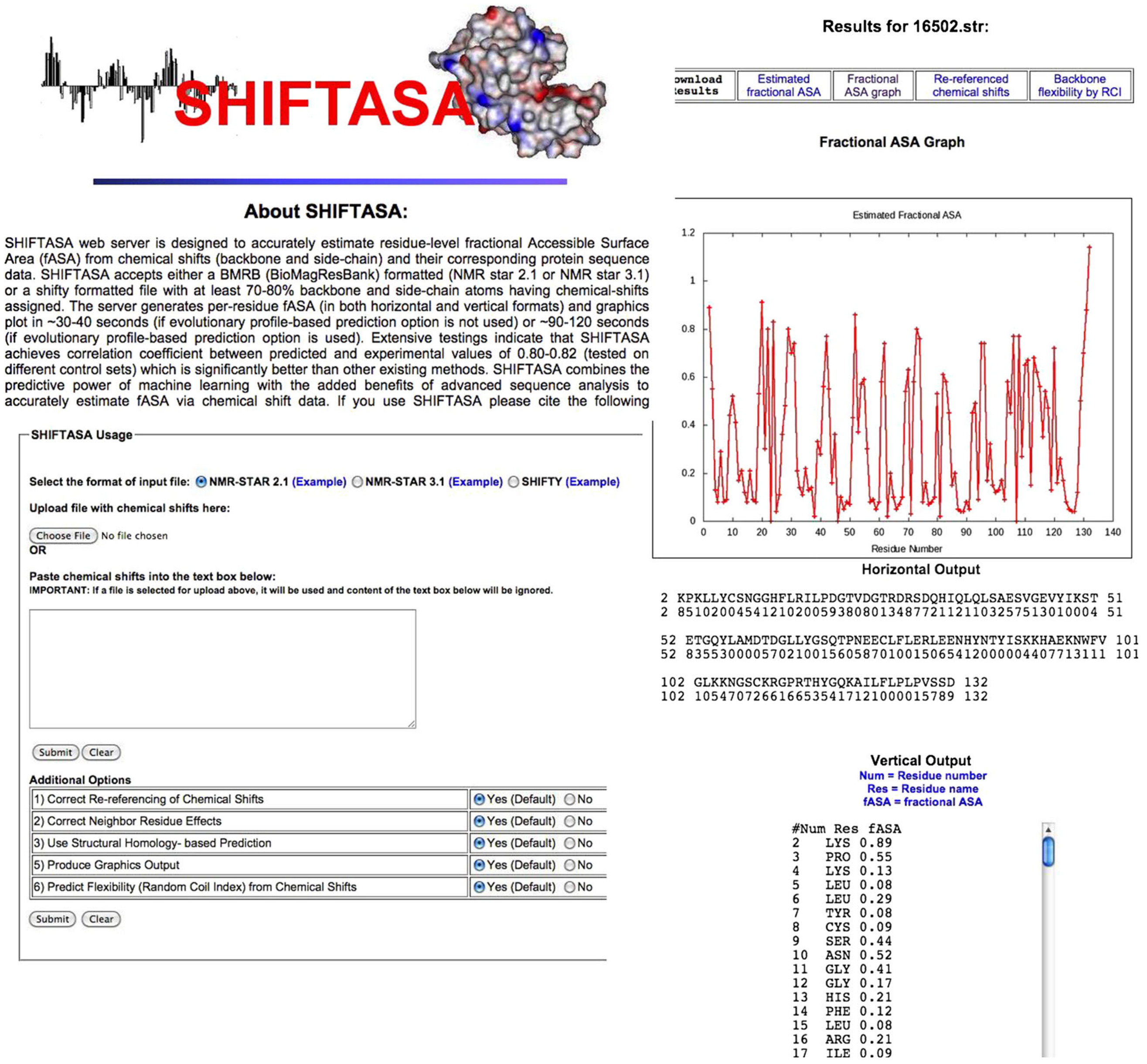
neighboring residue chemical shift data. On the other hand, chemical shift-only estimation performance varies with the amount of complete shift assignments and shows a Pearson correlation coefficient of 0.75 (depicted by a line connected by blue rectangles in Fig. S2).

Prediction error versus residue specific variance in test set ASA distribution

Figure S3 illustrates the relationship between the standard deviation of the fASA value for each of the 20 different amino acids in the test set and the corresponding fASA prediction error. The variance in the fASA values shows a relatively good agreement with the prediction error (MAE), yielding a Spearman correlation coefficient of 0.93. In general, the prediction error in fASA values for exposed residues is higher than for buried residues. Frequently buried and partially buried residues such as CYS, ILE, VAL, PHE, TYR, TRP and LEU have comparatively lower variability in the observed fASA values, leading to the lower associated prediction errors. Among these residues, CYS, ILE and VAL have less than a 10 % mean prediction error, while others are within a 10-13 % error range. This might be because buried residues generally have a more conserved fASA distribution, as can be seen in Fig. S3. In contrast, exposed and partially exposed residues such as ASP, GLU, PRO and GLY have a much higher (≥17 %) mean estimation error. ASN, GLN, SER, HIS and ALA fall into the medium range of prediction errors (15–16 %). These increased prediction errors might be a consequence of the high fASA variability seen in exposed residues (see Fig. S3). The most difficult residue to predict is ASP, which produces the highest mean prediction error of 19.3 %. All aromatic residues (PHE ~11 %, TRP ~12 %, TYR ~13 %) are within a 13 % error limit, which again confirms their relatively buried nature or their affinity to associate with residues in buried regions. Overall our data show that buried and partially buried residues are predicted with relatively higher accuracy than exposed, partially exposed or charged residues. More exposed residues tend to have fewer assignments due to their higher mobility, higher overlap, and lower importance to researchers.

ASA range versus prediction error versus training point fractions

The error distribution with the fASA value range and the corresponding sample training size revealed some interesting and unexpected trends. These are shown in Fig. S4. The training fraction curve reveals that there is a relative abundance of chemical shift (and ASA) data for buried and partially buried regions of proteins, which facilitates higher prediction accuracies in those regions. As training fractions slowly decrease for higher fASA ranges (partially exposed and fully exposed residues), so does the prediction accuracy for those residues. This trend partially explains why ShiftASA performs somewhat differently in estimating the accessible surface area of buried, partially buried, partially exposed and fully exposed residues in proteins.



About SHIFTASA:

SHIFTASA web server is designed to accurately estimate residue-level fractional Accessible Surface Area (fASA) from chemical shifts (backbone and side-chain) and their corresponding protein sequence data. SHIFTASA accepts either a BMRB (BioMagResBank) formatted (NMR star 2.1 or NMR star 3.1) or a shifty formatted file with at least 70-80% backbone and side-chain atoms having chemical-shifts assigned. The server generates per-residue fASA (in both horizontal and vertical formats) and graphics plot in ~30-40 seconds (if evolutionary profile-based prediction option is not used) or ~90-120 seconds (if evolutionary profile-based prediction option is used). Extensive testings indicate that SHIFTASA achieves correlation coefficient between predicted and experimental values of 0.80-0.82 (tested on different control sets) which is significantly better than other existing methods. SHIFTASA combines the predictive power of machine learning with the added benefits of advanced sequence analysis to accurately estimate fASA via chemical shift data. If you use SHIFTASA please cite the following

SHIFTASA Usage

Select the format of input file: ● NMR-STAR 2.1 (Example) ○ NMR-STAR 3.1 (Example) ○ SHIFTY (Example)

Upload file with chemical shifts here:

Choose File No file chosen

OR

Paste chemical shifts into the text box below:

IMPORTANT: If a file is selected for upload above, it will be used and content of the text box below will be ignored.

Submit Clear

Additional Options

1) Correct Re-referencing of Chemical Shifts	● Yes (Default) ○ No
2) Correct Neighbor Residue Effects	● Yes (Default) ○ No
3) Use Structural Homology- based Prediction	● Yes (Default) ○ No
5) Produce Graphics Output	● Yes (Default) ○ No
6) Predict Flexibility (Random Coil Index) from Chemical Shifts	● Yes (Default) ○ No

Submit Clear

Results for 16502.str:

ownload esults | Estimated fractional ASA | Fractional ASA graph | Re-referenced chemical shifts | Backbone flexibility by RCI

Fractional ASA Graph

Horizontal Output

```
2 KPKLLYCSNGGHFLRILPDGTVDGTRDRSDQHIQLQLSAESVGEVYIKST 51
2 85102004541210200593808013487721121103257513010004 51

52 ETGQYLAMDTDGLLYGSQTPNEECLFLERLEENHYNTYISKKHAEKNWFV 101
52 83553000057021001560587010015065412000004407713111 101

102 GLKKNGSCKRGPRTHYGQKAILFLPLPVSSD 132
102 1054707266166535417121000015789 132
```

Vertical Output

Num = Residue number
Res = Residue name
fASA = fractional ASA

```
#Num Res fASA
2    LYS 0.89
3    PRO 0.55
4    LYS 0.13
5    LEU 0.08
6    LEU 0.29
7    TYR 0.08
8    CYS 0.09
9    SER 0.44
10   ASN 0.52
11   GLY 0.41
12   GLY 0.17
13   HIS 0.21
14   PHE 0.12
15   LEU 0.08
16   ARG 0.21
17   ILE 0.09
```

Fig. 6 A montage of the ShiftASA webserver showing the home page (*left*) and several screenshots of the output pages (*right*)

SABLE improves prediction performance

In ShiftASA, we tried to incorporate as much information as available both from sequence and chemical shifts in order to achieve optimal performance. Because of the excellent performance of the sequence-only method SABLE (Adamczak et al. 2004) we decided to include its sequence-based prediction into the ShiftASA algorithm. Indeed, this addition led to an increase of mean correlation coefficients between predicted and experimental values from 0.76 to 0.79 (TEST1) and 0.79 to 0.82 (TEST2). This improvement is statistically significant ($p < 0.001$). Evidently SABLE's sequence-driven structural homology and evolutionary profile based prediction provides additional information that helps to accurately estimate the buried/exposed states of residues.

ShiftASA accurately estimates fractional ASA of "unfolded" proteins

We also investigated the performance of ShiftASA for estimating fractional ASA values for a completely unfolded protein (i.e. unfolded ubiquitin in 8 M urea—BMRB 4375). As a substitute for observed fASA values, an average per-residue fASA value is calculated from 10,000 unfolded structures of ubiquitin generated using the computer program Flexible Meccano (Ozenne et al. 2012). As seen in Fig. 5, ShiftASA was able to estimate the exposed state of this protein with a moderate accuracy. In contrast to the sequence-only method, SABLE (Adamczak et al. 2004) most of the protein was estimated to contain a high proportion of buried regions. Because SABLE predicts the fractional ASA from sequence, it simply reported the ASA states of the folded ubiquitin structure retrieved by a PSI-BLAST (Altschul et al. 1997) search. However, because ShiftASA weighs both the experimental chemical shift information with sequence-derived features, its performance was not compromised.

The ShiftASA web server

A web server (http://shiftasa.wishartlab.com) has been developed that accepts a BMRB (NMR-Star 2.1 or NMR-Star 3.1) or SHIFTY-formatted chemical shift file and generates per-residue fractional ASA (in both horizontal and vertical formats) along with a fractional ASA plot. The server supports a number of user-selectable options including the choice of using sequence homology for the SABLE (Adamczak et al. 2004) prediction. The web server has been implemented as a Python CGI-script. In general, the web server takes <60 s (if homology is off) or >140 s (if homology is on). A screen shot of the ShiftASA web server and its output is shown in Fig. 6.

Conclusion

We have developed a method that accurately predicts the per-residue fASA of water-soluble proteins using a combination of both sequence and chemical shift data. Our prediction method, called ShiftASA, demonstrates superior performance relative to sequence-only or chemical shift-only methods in two independent test sets of 65 and 92 proteins (TEST1 and TEST2, respectively). In particular, with the TEST1 data set, ShiftASA showed a mean Spearman's rank correlation coefficient between predicted and experimental values of 0.79, which is a 8.2 % improvement over the best performing method. The mean absolute error was found to drop from 0.19 to 0.14 $\mathring{A}^2$ and the root mean squared error fell from 0.26 to 0.19 $\mathring{A}^2$ compared to its sequence-only and chemical shift-only counterparts. On the TEST2 set, ShiftASA attained a mean correlation coefficient of 0.82, a clear improvement over correlation coefficients of 0.67 and 0.73 reported by the best performing sequence-only and chemical-shift-only methods, respectively. In addition, the real-value fASA prediction by ShiftASA allows flexible, categorical prediction of binary or ternary ASA states. Overall, we believe that ShiftASA, with its improved prediction of ASA parameters, will not only facilitate protein fold recognition and de novo protein structure prediction methods, but as we will show in upcoming papers, contribute to the generation and refinement of protein structures by NMR and the calculation of useful thermodynamic parameters from chemical shift data.

Acknowledgments The authors would like to thank Dr. Mark Berjanskii for his helpful suggestions in preparing the ShiftASA program. Financial support from the Natural Sciences and Engineering Research Council (NSERC), the Alberta Prion Research Institute (APRI) and PrioNet is gratefully acknowledged.

References

Adamczak R, Porollo A, Meller J (2004) Accurate prediction of solvent accessibility using neural networks-based regression. Proteins Struct Funct Bioinform 56(4):753–767

Ahmad S, Gromiha MM (2002) NETASA: neural network based prediction of solvent accessibility. Bioinformatics 18(6):819–824

Ahmad S, Gromiha MM, Sarai A (2003) Real value prediction of solvent accessibility from amino acid sequence. Proteins Struct Funct Bioinform 50(4):629–635

Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 25(17):3389–3402

Avbelj F, Kocjan D, Baldwin RL (2004) Protein chemical shifts arising from alpha-helices and beta-sheets depend on solvent exposure. Proc Natl Acad Sci USA 101(50):17394–17397

Benkert P, Tosatto SC, Schomburg D (2008) QMEAN: a comprehensive scoring function for model quality assessment. Proteins Struct Funct Bioinform 71(1):261–277

Berjanskii MV, Wishart DS (2005) A simple method to predict protein flexibility using secondary chemical shifts. J Am Chem Soc 127(43):14970–14971

Berjanskii MV, Wishart DS (2013) A simple method to measure protein side-chain mobility using NMR chemical shifts. J Am Chem Soc 135(39):14536–14539

Biswas KM, DeVido DR, Dorsey JG (2003) Evaluation of methods for measuring amino acid hydrophobicities and interactions. J Chromatogr A 1000(1):637–655

Chen H, Zhou HX (2005) Prediction of solvent accessibility and sites of deleterious mutations from protein sequence. Nucleic Acids Res 33(10):3193–3199

Chothia C (1976) The nature of the accessible and buried surfaces in proteins. J Mol Biol 105(1):1–12

Croy CH, Koeppe JR, Bergqvist S, Komives EA (2004) Allosteric changes in solvent accessibility observed in thrombin upon active site occupation. Biochemistry 43(18):5246–5255

Eisenberg D, Weiss RM, Terwilliger TC (1984) The hydrophobic moment detects periodicity in protein hydrophobicity. Proc Natl Acad Sci 81(1):140–144

Eisenhaber F, Argos P (1993) Improved strategy in analytic surface calculation for molecular systems: handling of singularities and computational efficiency. J Comput Chem 14(11):1272–1280

Engelman DM, Steitz TA, Goldman A (1986) Identifying nonpolar transbilayer helices in amino acid sequences of membrane proteins. Annu Rev Biophys Biomol Struct 15(1):321–353

Frishman D, Argos P (1995) Knowledge-based protein secondary structure assignment. Proteins Struct Funct Bioinform 23(4):566–579

Garg A, Kaur H, Raghava GPS (2005) Real value prediction of solvent accessibility in proteins using multiple sequence alignment and secondary structure. Proteins Struct Funct Bioinform 61(2):318–324

Han B, Liu Y, Ginzinger SW, Wishart DS (2011) SHIFTX2: significantly improved protein chemical shift prediction. J Biomol NMR 50(1):43–57

Holbrook SR, Muskal SM, Kim SH (1990) Predicting surface exposure of amino acids from protein sequence. Protein Eng 3(8):659–665

Hopp TP, Woods KR (1981) Prediction of protein antigenic determinants from amino acid sequences. Proc Natl Acad Sci 78(6):3824–3828

Huyghues-Despointes BM, Langhorst U, Steyaert J, Pace CN, Scholtz JM (1999) Hydrogen-exchange stabilities of RNase T1 and variants with buried and solvent-exposed Ala → Gly mutations in the helix. Biochemistry 38(50):16481–16490

Janin J (1979) Surface and inside volumes in globular proteins. Nature 277:491–492

Kabsch W, Sander C (1983) Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. Biopolymers 22(12):2577–2637

Kyte J, Doolittle RF (1982) A simple method for displaying the hydropathic character of a protein. J Mol Biol 157(1):105–132

Lavigne P, Willard L, Sykes BD, Bagu JR, Boyko R, Holmes CE (2000) Structure-based thermodynamic analysis of the dissociation of protein phosphatase-1 catalytic subunit and microcystin-LR docked complexes. Protein Sci 9(2):252–264

Lee B, Richards FM (1971) The interpretation of protein structures: estimation of static accessibility. J Mol Biol 55(3):379–400

Li X, Pan XM (2001) New method for accurate prediction of solvent accessibility from protein sequence. Proteins Struct Funct Bioinform 42(1):1–5

Manavalan P, Ponnuswamy PK (1978) Hydrophobic character of amino acid residues in globular proteins. Nature 275:673–674

Marsh JA (2013) Buried and accessible surface area control intrinsic protein flexibility. J Mol Biol 425:3250–3263

Mayrose I, Graur D, Ben-Tal N, Pupko T (2004) Comparison of site-specific rate-inference methods: Bayesian methods are superior. Mol Biol Evol 21:1781–1791

Myers JK, Nick PC, Martin SJ (1995) Denaturant m values and heat capacity changes: relation to changes in accessible surface areas of protein unfolding. Protein Sci 4(10):2138–2148

Naderi-Manesh H, Sadeghi M, Arab S, Moosavi MAA (2001) Prediction of protein surface accessibility with information theory. Proteins Struct Funct Bioinform 42(4):452–459

Nguyen MN, Rajapakse JC (2005) Prediction of protein relative solvent accessibility with a two-stage SVM approach. Proteins Struct Funct Bioinform 59(1):30–37

Ozenne V, Bauer F, Salmon L, Huang JR, Jensen MR, Segard S, Blackledge M (2012) Flexible-meccano: a tool for the generation of explicit ensemble descriptions of intrinsically disordered proteins and their associated experimental observables. Bioinformatics 28(11):1463–1470

Petersen B, Petersen TN, Andersen P, Nielsen M, Lundegaard C (2009) A generic method for assignment of reliability scores applied to solvent accessibility predictions. BMC Struct Biol 9(1):51

Pollastri G, Baldi P, Fariselli P, Casadio R (2002) Prediction of coordination number and relative solvent accessibility in proteins. Proteins Struct Funct Bioinform 47(2):142–153

R Development Core Team (2008) R: a language and environment for statistical computing. R Foundation for Statistical Computing, ISBN 3-900051-07-0. http://www.R-project.org

Richards FM (1974) The interpretation of protein structures: total volume, group volume distributions and packing density. J Mol Biol 82(1):1–14

Richards FM (1977) Areas, volumes, packing and protein structure. Annu Rev Biophys Bioeng 6:151–176

Ridgeway G (2007) Generalized boosted models: a guide to the GBM package. R package vignette. http://CRAN.R-project.org/package=gbm

Rost B, Sander C (1994) Conservation and prediction of solvent accessibility in protein families. Proteins Struct Funct Bioinform 20(3):216–226

Serpa JJ, Makepeace KA, Borchers TH, Wishart DS, Petrotchenko EV, Borchers CH (2014) Using isotopically-coded hydrogen peroxide as a surface modification reagent for the structural characterization of prion protein aggregates. J Proteomics 100:160–166

Sievers F, Wilm A, Dineen D, Gibson TJ, Karplus K, Li W et al (2011) Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. Mol Syst Biol 7(1):539

Thompson MJ, Goldstein RA (1996) Predicting solvent accessibility: higher accuracy using Bayesian statistics and optimized residue substitution classes. Proteins Struct Funct Genet 25(1):38–47

Trevor H, Robert T, Friedman JJH (2001) The elements of statistical learning, vol 1. Springer, New York

UniProt Consortium (2010) The universal protein resource (UniProt) in 2010. Nucleic Acids Res 38(Suppl 1):D142–D148

Valdar WSJ (2002) Scoring residue conservation. Proteins Struct Funct Bioinform 48(2):227–241

Vranken W, Rieping W (2009) Relationship between chemical shift value and accessible surface area for all amino acid atoms. BMC Struct Biol 9(1):20

Wagner M, Adamczak R, Porollo A, Meller J (2005) Linear regression models for solvent accessibility prediction in proteins. J Comput Biol 12(3):355–369

Wang Y, Jardetzky O (2002) Probability-based protein secondary structure identification using combined NMR chemical-shift data. Protein Sci 11(4):852–861

Wishart DS (2011) Interpreting protein chemical shift data. Prog Nucl Magn Reson Spectrosc 58(1):62–87

Wishart DS, Sykes BD (1994) Chemical shifts as a tool for structure determination. Methods Enzymol 239:363–392

Yuan Z, Huang B (2004) Prediction of protein accessible surface areas by support vector regression. Proteins Struct Funct Bioinform 57(3):558–564

Zhang H, Neal S, Wishat DS (2003) RefDB: a database of uniformly referenced protein chemical shifts. J Biomol NMR 25:173–195