

QBES: Predicting Real Values of Solvent Accessibility from Sequences by Efficient, Constrained Energy Optimization

Zhigang Xu,¹ Chi Zhang,¹ Song Liu,¹ and Yaoqi Zhou^{1,2*}

¹Howard Hughes Medical Institute Center for Single Molecule Biophysics, Department of Physiology & Biophysics, State University of New York at Buffalo, New York

²The Key Laboratory of Molecular Engineering of Polymers, Department of Macromolecular Science, Fudan University, Shanghai, People's Republic of China

ABSTRACT Solvent accessibility, one of the key properties of amino acid residues in proteins, can be used to assist protein structure prediction. Various approaches such as neural network, support vector machines, probability profiles, information theory, Bayesian theory, logistic function, and multiple linear regression have been developed for solvent accessibility prediction. In this article, a much simpler quadratic programming method based on the buriability parameter set of amino acid residues is developed. The new method, called QBES (Quadratic programming and Buriability Energy function for Solvent accessibility prediction), is reasonably accurate for predicting the real value of solvent accessibility. By using a dataset of 30 proteins to optimize three parameters, the average correlation coefficients between the predicted and actual solvent accessibility are about 0.5 for all four independent test sets ranging from 126 to 513 proteins. The method is efficient. It takes only 20 min for a regular PC to obtain results of 30 proteins with an average length of 263 amino acids. Although the proposed method is less accurate than a few more sophisticated methods based on neural network or support vector machines, this is the first attempt to predict solvent accessibility by energy optimization with constraints. Possible improvements and other applications of the method are discussed. *Proteins* 2006;63:961–966. © 2006 Wiley-Liss, Inc.

Key words: solvent accessibility; amino acid residues; proteins

INTRODUCTION

It has been well established that hydrophobic residues tend to be buried inside a soluble protein, whereas hydrophilic residues tend to be on its surface. However, predicting the solvent accessibility of a residue is challenging because a given hydrophobic (hydrophilic) residue may well be on the surface (inside the core) due to the constraints such as bond connectivity, structural, or functional requirements. One possible method for predicting solvent accessibility is homology modeling. However, this is less accurate than the prediction of secondary structure because solvent accessibility is less conserved than secondary structure.¹ Other approaches have been developed for solvent accessibility prediction. Many are based on neural

network^{1–8} or support vector machines.^{9–11} Additional methods include probability profiles,¹² information theory,^{13,14} Bayesian theory,¹⁵ logistic function,¹⁶ and multiple linear regression.⁴ Although early methods focused on the classification of solvent accessibility into a few states such as exposed, buried, or intermediate, more recent methods predict the real values of solvent accessibilities.^{7,8,10,17,18}

Statistical methods mentioned above make prediction by learning from training sets. Each predictor is often treated as a black box with little understanding of numerous parameters used in prediction. For a better understanding of parameters involved in prediction and the physics behind the method, a physical-based approach is more desirable. In this article, we propose, to our knowledge, the first physical-based approach for solvent-accessibility prediction by making constrained energy optimization through a quadratic programming technique implemented by Matlab.¹⁹

In this approach, we employ a simple energy function based on a recent finding that the contribution of a residue to the stability of a protein, derived from a statistical energy function,²⁰ is linearly dependent on its buried solvent-accessible surface area (BASA).²¹ The slope of the linear dependence, called buriability, provides a quantitative measure of the tendency of the residue to be buried or exposed.²¹ This buriability parameter set has been used in the fold recognition method SPARKS for sequence-to-structure threading.^{22,23}

We call this method QBES (Quadratic programming and Buriability Energy function for Solvent accessibility prediction), which predicts solvent accessibility by performing constrained optimization of the protein stability upon burial of amino acid residues. Our results show that this simple method provides a reasonably accurate prediction

Grant sponsor: HHMI to SUNY Buffalo and by the Keck Center for Computational Biology at SUNY Buffalo; Grant numbers: NIH R01 GM 966049, R01 GM 068530. Grant sponsor: National Science Foundation of China; Grant number: 20340420391.

*Correspondence to: Yaoqi Zhou, Howard Hughes Medical Institute Center for Single Molecule Biophysics, Department of Physiology & Biophysics, State University of New York at Buffalo, 124 Sherman Hall, Buffalo, NY 14214. E-mail: yqzhou@buffalo.edu

Received 19 July 2005; Revised 25 November 2005; Accepted 6 December 2005

Published online 2 March 2006 in Wiley InterScience (www.interscience.wiley.com). DOI: 10.1002/prot.20934

TABLE I. The SASA at the Unfolded State and the Buriability Parameters of 20 Amino Acid Residues

Residue	A_i^{0a} (\AA^2)	Buriability ^b (cal/mol/ \AA^2)	Residue	A_i^{0a} (\AA^2)	Buriability ^b (cal/mol/ \AA^2)
TRP	264	23.5	GLU	187	6.6
TYR	237	18.8	THR	151	9.5
PHE	221	23.3	ASN	161	6.8
ILE	193	19.9	LYS	214	5.5
LEU	197	20.4	ALA	124	12.8
MET	216	15.1	PRO	149	8.9
VAL	169	19.0	ASP	154	7.3
HIS	201	10.0	CYS	94	22.5
ARG	244	7.6	SER	126	7.4
GLN	192	7.6	GLY	89	6.2

^aThe SASA of each type of amino acid at the unfolded state was presented by Shrake and Rupley.²⁴

^bThe buriability parameters of 20 amino acid residues are obtained from ref. ²¹.

of the real values of solvent accessibilities of amino acid residues. Possible further improvements and other applications of this method are discussed.

METHODS

QBES: Objective Function and Quadratic Programming

According to a statistical energy function based on a distance-scaled finite ideal-gas reference state (DFIRE),²⁰ each residue's contribution to stability is proportional to its buriability and BASA.²¹ That is, the stability of a protein, ΔG , satisfies,

$$\Delta G = \sum_{i=1}^N P(I_i) B_i, \quad (1)$$

where $P(I_i)$ and B_i are the buriability parameter for residue type I and BASA for a given residue i , respectively, and N is the number of amino acid residues. The objective is to maximize stability by optimizing BASA of all the residues. However, the buriability term alone is not enough because buriability parameters are all positive (pro-burial) for 20 amino acid residues²¹ and the optimal solution would be the complete burial for all amino acid residues. To avoid this trivial solution, we add a naive bond-connectivity term to the buriability term. The objective function is given by

$$\min \left(-w_1 \sum_{i=1}^N P(I_i) B_i + \sum_{j=1}^n \sum_{i=1}^{N-j} (B_i - B_{i+j})^2 \right), \quad (2)$$

where w_1 is a to-be-determined weight factor. Minimization is used here due to the negative sign in the first term. The second term in Equation (2) is based on the approximation that neighboring residues (within n residues in sequence positions) have similar BASA.

The above objective function subjects to the constraints that each BASA is less than or equal to the maximum possible solvent-accessible surface area (SASA) for a given residue, A_i^0 , that is,

$$0 \leq B_i \leq A_i^0, \quad (3)$$

and the sum of BASA should be less than the sum of maximum possible SASA, that is,

$$\sum_{i=1}^N B_i \leq w_2 \sum_{i=1}^N A_i^0, \quad (4)$$

where w_2 (<1) is a to-be-determined weight factor. Equation (4) is used to ensure that at least a certain fraction of residues are exposed to solvent.

The objective function in Equation (2) together with constraints [Eqs. (3)–(4)] are optimized for BASA by using the quadratic programming tool QUADPROG in Matlab.¹⁹ The SASA of unfolded state, A_i^0 , is obtained from Shrake and Rupley,²⁴ and the buriability parameter $P(I)$ is from ref. 21 (see Table I). Once the BASA, B_i , is obtained, the SASA is simply equal to $(A_i^0 - B_i)$. The actual SASA values of residues for a protein are calculated by McConkey's method based on the Voronoi procedure.²⁵

Training and Test Sets

We use a training set of 30 proteins and a test set of 185 proteins, which are identical to those used by Ahmad and Gromiha⁶ in the program NETASA for predicting solvent accessibility. The total 215 high-resolution structures of proteins (Manesh-215) with $<25\%$ sequence identity were originally collected by Manesh et al.¹³ The training set is used to determine the number of neighboring residues n and two weight parameters w_1 and w_2 . Thirty proteins in the training set are 1aba, 1abrB, 1bdo, 1beo, 1bib, 1bmfG, 1bncA, 1btmA, 1btn, 1cem, 1ceo, 1cewI, 1cfyA, 1chd, 1chkA, 1cyx, 1deaA, 1delA, 1dkzA, 1dosA, 1fua, 1gai, 1gpl, 1gsa, 1gtmA, 1havA, 2i1b, 2sns, 3grs, and 3mdda.

The test set of 185 proteins is made of 119l, 1edg, 1knyA, 1pea, 1sluA, 1wba, 2mtaC, 153l, 1edt, 1kptA, 1pex, 1smeA, 1whi, 2nacA, 1afra, 1erv, 1kte, 1pgs, 1smpl, 1who, 2pgd, 1afwA, 1esc, 1kuh, 1phe, 1sra, 1wsyB, 2phlA, 1amm, 1exnB, 1lba, 1php, 1std, 1xgsA, 2phy, 1amp, 1ezm, 1lcl, 1pioA, 1stfl, 1xnb, 2pia, 1aocA, 1fds, 1lki, 1plc, 1svpA,

1xvaA, 2pspA, 1atl, 1fjmA, 1lkkA, 1pmi, 1tadC, 1xyza, 2rn2, 1atnA, 1ftpA, 1ltsA, 1pne, 1tdx, 1yasA, 2rspB, 1axn, 1gcb, 1mai, 1poa, 1tfe, 1ysc, 2scpA, 1bbpA, 1gdoA, 1maz, 1poc, 1tfr, 1ytw, 2sil, 1bfg, 1gggA, 1mbd, 1pot, 1thv, 256bA, 2tysA, 1bgc, 1gnd, 1mkaA, 1ppn, 1thx, 2abk, 3chy, 1bhmB, 1gotB, 1mldA, 1pud, 1tib, 2arcA, 3cox, 1chmA, 1gpc, 1mml, 1pytA, 1tml, 2ayh, 3minB, 1cmkE, 1hfc, 1molA, 1qapA, 1tupC, 2bbvC, 3nll, 1cnv, 1hgxA, 1nar, 1ra9, 1tys, 2cae, 3sdhA, 1cseE, 1h1b, 1nbaB, 1rcf, 1ubi, 2cba, 5p21, 1csgA, 1hsbA, 1nox, 1rec, 1uby, 2ccyA, 5ptp, 1csn, 1htp, 1nozA, 1rgs, 1udiI, 2chsA, 6gsvA, 1dfjI, 1ida, 1ofgA, 1rnl, 1uxy, 2ctc, 6pfkA, 1dhr, 1ido, 1onrA, 1rro, 1vcaA, 2end, 7rsa, 1dktB, 1lfc, 1opr, 1rsy, 1vhh, 2gdm, 8atcB, 1dxy, 1irk, 1ospO, 1rvaA, 1vhrA, 2hft, 1eceA, 1itg, 1pbc, 1sbp, 1vid, 2hhmA, 1ecpA, 1jkw, 1pda, 1sftB, 1vin, 2hpdA, 1ede, 1knb, 1pdo, 1sig, 1vls, and 2liv.

To further test the QBES method without additional training, other data sets of protein structures have been used. These data sets include 126 protein data set (RS-126) of Rost and Sander,¹ 338 monomeric protein data set (Carugo-338) used by Carugo,¹⁴ and 513 protein data set (CB-513) developed by Cuff and Barton.²⁶ These data sets are also made of protein sequences with <25% homology.

Assessment of Prediction Accuracy

Solvent accessibility (%) is defined as the ratio of the SASA of a residue in a protein to its value at the unfolded state. For a two-state prediction, a residue is in a buried (an exposed) state if its solvent accessibility is smaller (greater) than a threshold. The accuracy of prediction is defined as the fraction of residues with correctly predicted states in a given protein. For a real-value prediction, the accuracy is assessed by the correlation coefficient between the predicted SASA values and the actual SASA values of residues obtained from the experimental structures and by Mean Absolute Error (MAE), which is defined as the absolute difference between the predicted and actual (desired) values of relative SASA (i.e., normalized by SASA in the unfolded state), per residue:

$$\text{MAE} = \frac{1}{N} \sum |(SASA\%)_{\text{pred}} - (SASA\%)_{\text{exp}}|, \quad (5)$$

where summation is carried out for all the residues in a protein sequence and N is the sequence length.¹⁷

RESULTS

Because there are only three adjustable parameters (w_1 , w_2 , and n), optimization can be achieved by a simple grid search. Figures 1 and 2 show how the average correlation coefficient changes as w_1 or w_2 changes at $n = 5$. With w_2 fixed, the correlation coefficient increases as w_1 increases. After w_1 is >300, the correlation coefficient begins to reduce gradually. When w_1 is fixed, the average correlation coefficient is less sensitive to the value of w_2 between 0.5 and 0.8. However, the accuracy based on a two-state definition has a maximum at $w_1 = 0.7$. In this article, we use $w_1 = 300$ and $w_2 = 0.8$ based on the maximum in average correlation coefficient. The fact that $w_1 \gg 1$ highlights the relative importance of the buriability term

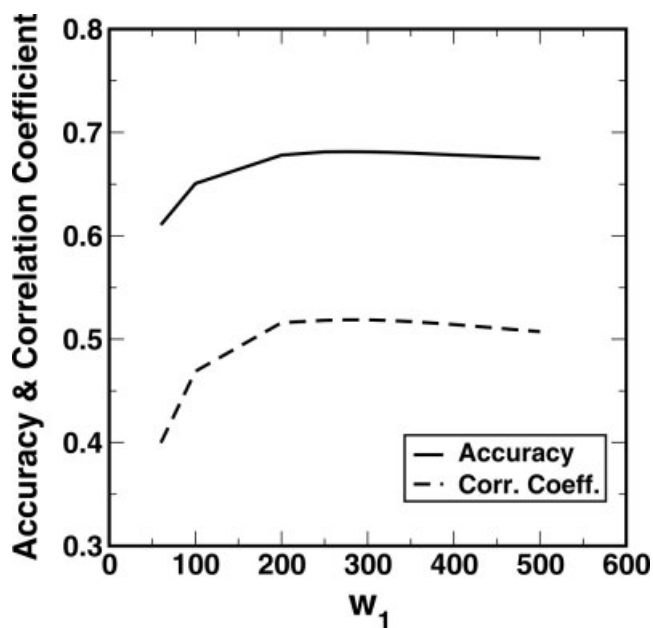


Fig. 1. The average accuracy based on a two-state definition (25% as threshold) and the average correlation coefficient between predicted and actual solvent-accessible surface area (SASA) values for the 30 training proteins as a function of w_1 at $w_2 = 0.8$.

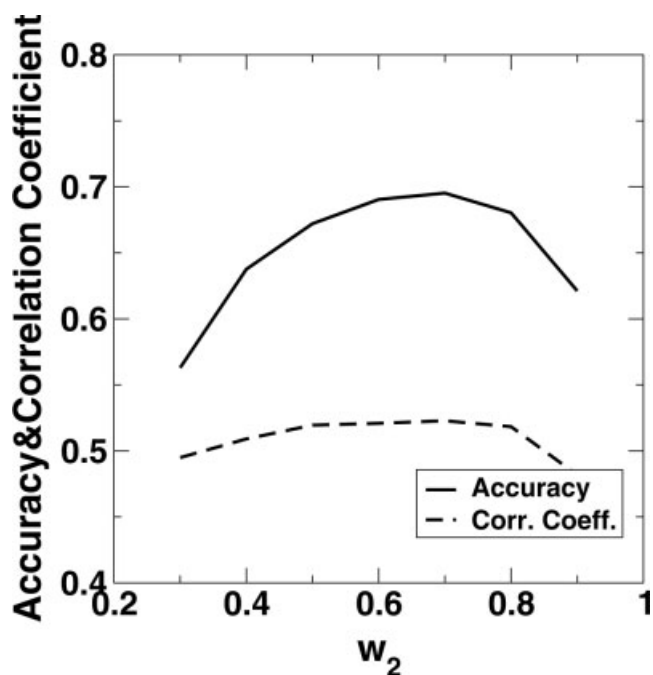


Fig. 2. As in Figure 1 but as a function of w_2 at $w_1 = 300$.

over the bond-connectivity term. We also varied the number of neighbors from 1 to 10. We found that QBES with $n > 4$ are more accurate than QBES with $n = 4$ and there is no significant difference between the results with $n > 4$. We take the small number, that is, $n = 5$, to simplify the quadratic terms, so that it can save computing time and memory cost. We also test QBES without the second term and the average correlation coefficient is only 0.32 for the

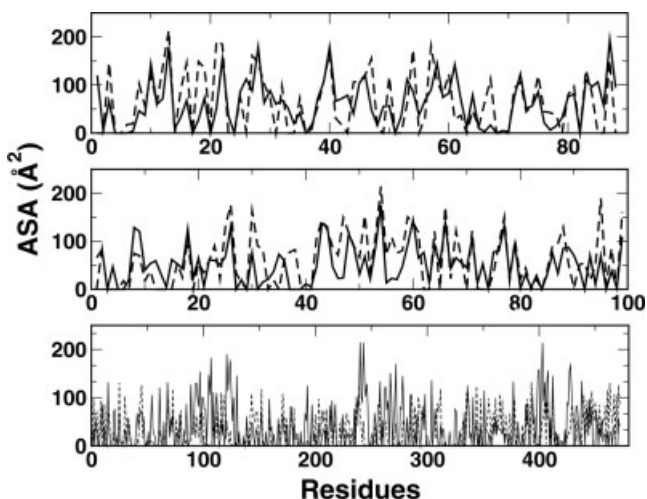


Fig. 3. The solvent-accessible surface areas (SASA) of residues in three proteins: 1aba (top), 1plc (middle), 1gai (bottom), where solid lines and dashed lines denote the actual and predicted values, respectively.

30 training proteins. Thus, the parameter set of $w_1 = 300$, $w_2 = 0.8$, and $n = 5$ is used throughout of this paper (i.e., for all test sets).

The QBES method with the w_1 , w_2 , and n parameters obtained above is tested by an independent set of 185 proteins. Figure 3 shows the results of three proteins, for which the correlation coefficients are 0.57 (1aba), 0.69 (1plc), and 0.20 (1gai), respectively. There is a reasonable agreement between structure-derived and experimental SASA values for 1aba and 1plc. If the accuracy is defined based on a two-state classification with 25% as the threshold, the prediction accuracy of the above three proteins are 66% for 1aba, 65% for 1plc, and 65% for 1gai, respectively. Thus, even for a low correlation coefficient of 0.20, the accuracy based on a two-state classification remains reasonable. This reflects a fact that a real-value prediction is more difficult than a two-state prediction.

For the test set of 185 proteins, the average correlation coefficient is 0.52, similar to that of the training set. The distribution of the correlation coefficients for 185 proteins is shown in Figure 4. There are more than 130 (70%) proteins whose correlation coefficients between predicted and actual SASA values are greater than 0.5. This can be compared to an average correlation coefficient of 0.47 and 60 (32%) proteins with the correlation coefficient greater than 0.5 if SASA is assigned with average values from the 30 training proteins for the same 185 proteins.

Figure 5 shows the MAE of amino acid residues as a function of their buriabilities. It shows that, in general, absolute errors are smaller for residues with higher buriabilities. That is, SASA of hydrophobic residues are predicted more accurately.

Table II compares the results of a neural network-based method RVP-Net¹⁷ with that of QBES for the test set of 185 proteins averaged over proteins with different sizes (the number of amino acid residues). It is interesting that both prediction accuracies (based on a two-state definition,

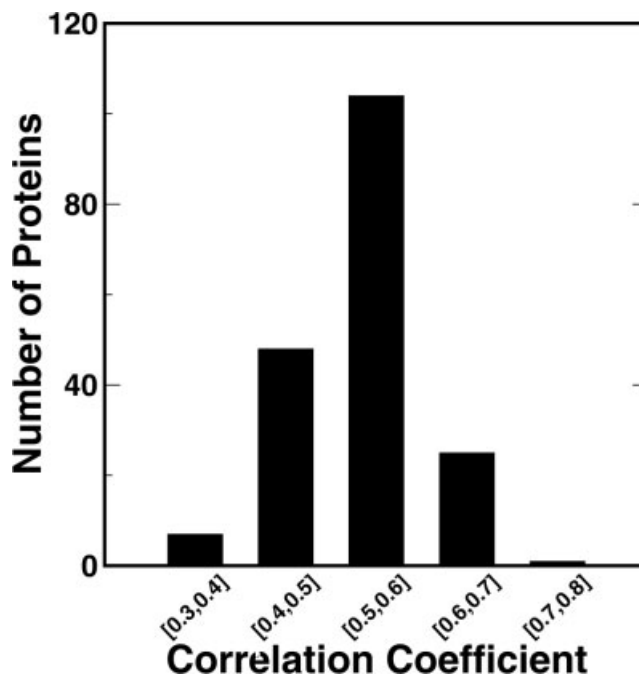


Fig. 4. The distribution of correlation coefficients between predicted and the experimental values of solvent-accessible surface area (SASA) for 185 test proteins. The average is 0.52 with the standard deviation of 0.10.

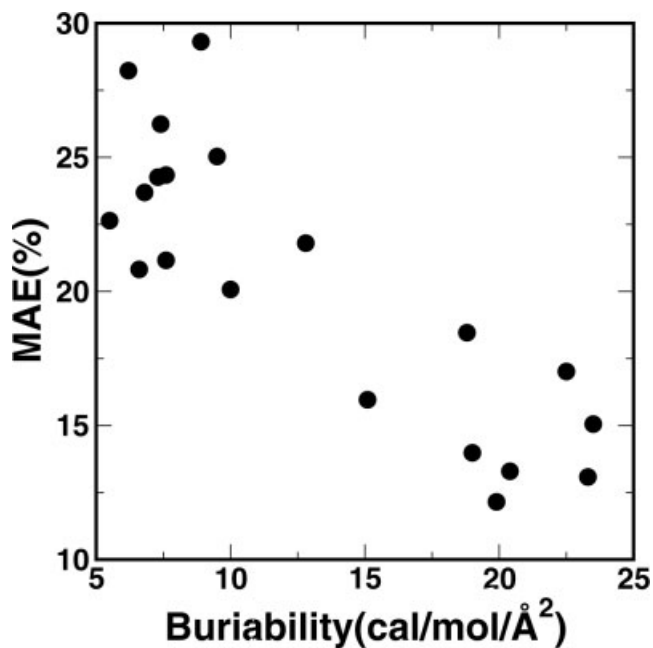


Fig. 5. The relationship between the mean absolute error (MAE) and Buriability for 20 residues for 185 test proteins.

threshold of 25%) increase for larger proteins. The average correlation coefficient between predicted and actual SASA values, however, is essentially the same for QBES for any sizes of proteins. Compared to RVP-Net, QBES gives a slightly higher average correlation coefficient, but also with a slightly higher average MAE.

TABLE II. Prediction Results of RVP-Net and QBES for 185 Test Proteins Classified Based on the Size (the Number of Residues) of Proteins

Protein length (# of proteins)	Accuracy (%) ^a		Corr. Coeff.		MAE (%)	
	RVP-Net ¹⁷	QBES	RVP-Net ¹⁷	QBES	RVP-Net ¹⁷	QBES
0–100 (9)	55.47	61.92	0.5154	0.5220	23.81	24.89
100–200 (74)	59.23	65.13	0.5115	0.5263	21.19	22.25
200–300 (51)	61.68	67.94	0.5083	0.5287	19.26	20.33
over 300 (51)	63.82	68.86	0.4982	0.5214	18.15	19.43

^aThe accuracy of a two-state model with 25% as threshold.

TABLE III. Correlation Coefficients Given by the Average Assignment Method, RVP-Net, SVR, SNNS, and QBES for Different Data Sets

Data set ^a	Average assignment ^c	RVP-Net ^d	SVR ^e	SNNS ^f				QBES ^g
				S	S + II	M	M + II	
RS-126	0.41	0.48	N/A	N/A	N/A	N/A	N/A	0.48
Manesh-215 (185)	0.47	0.50	N/A	0.53	0.61	0.63	0.67	0.52
Carugo-338	0.45	0.49	N/A	N/A	N/A	N/A	N/A	0.51
CB-513/502 ^b	0.41	0.48	0.52	0.52	0.60	0.62	0.65	0.49

^aData set RS-126 was first used by Rost and Sander,¹ 185 test proteins in Manesh-215 was collected by Manesh et al.,¹³ Carugo-338 was given by Carugo,¹⁴ and CB-513 was developed by Cuff and Barton.²⁶

^bCB-502 is a subset of the data set CB-513. Here results of RVP-Net, SVR, and SNNS are based on CB-502, while QBES and Average Assignment are applied on CB-513.

^c“Average Assignment” refers to the baseline method mentioned by Richardson and Barlow.²⁸ The assignment parameters are based on the statistics on the 30 training proteins from Manesh-215.¹³

^dAll results for RVP-Net are obtained from ref. ¹⁷.

^eResults for SVR here are from Yuan and Huang.¹⁰

^fResults for SNNS are from Raghava et al., where “S,” “M,” and “II” denote the use of single sequence, multiple sequences, and secondary structure information, respectively.⁷

^gThis work.

QBES is further tested on several additional data sets. They are RS-126,¹ Carugo-338,¹⁴ and CB-513.²⁶ The results are shown in Table III. The average correlation coefficients are obtained as 0.48, 0.51, and 0.49, respectively. By comparison, the average correlation coefficients by an average assignment method are 0.41, 0.45 and 0.41, respectively. Although the correlation coefficients given by QBES are similar to those given by the average assignment method, the fractions of proteins whose correlation coefficients exceed 0.5 differ significantly. For QBES, they are 49.2% for RS-126, 58.3% for Carugo-338, and 52.4% for CB-513, respectively. By comparison, the corresponding fractions for the average assignment method are 22.2, 28.4, and 19.3%, respectively.

DISCUSSION

We have developed a simple method for predicting SASA of proteins with their sequence information as the only input. The method has only three adjustable parameters, which were optimized for a training set of 30 proteins. Similar average correlation coefficients (about 0.5) were obtained for both the training and several test sets (as well as subsets of the test sets). This accuracy is not as good as some more sophisticated methods. For example, a method based on neural networks-based regression achieved an average correlation coefficient between 0.64–0.67 for different control sets.¹⁸ A method based on support vector regression achieved an average correlation coefficient of

0.52–0.67 depending on training sets.¹⁰ Inclusion of multiple sequence alignment and predicted secondary structures improves the accuracy of accessibility prediction.⁷ As shown in Table III, the accuracy of QBES is close to that of an early real-value neural-network method RVP-Net,¹⁷ which achieved a correlation coefficient of 0.50 for the same data set Manesh-215,¹³ and comparable correlations for the other data sets.

The simplicity of QBES is built on several approximations. First, the optimization of total buried energy makes QBES applicable only to globular proteins. Second, the connectivity condition, which is based on approximation that neighboring residues have similar BASA, ignores the difference in sizes of different residues and the possibility of neighboring residues having different solvent exposures. Third, surface flexibility is not taken into account. More critically, the water-mediated interaction between amino acid residues is far more complex than the buriability alone. These inherent limitations are likely the cause for the limited accuracy of QBES.

To our knowledge, QBES is the first method that attempts to predict SASA by energy optimization. The method is efficient through the use of quadratic programming. For example, optimization of 30 proteins with an average length of 263 amino acids takes only 20 min in a regular PC (Pentium IV 1.6 GHz CPU and 384 MB memory). A more time-consuming approach is to directly optimize model proteins with a reasonable energy function

by Monte Carlo and molecular dynamics simulations or other energy minimization tools.

The uniqueness of the method QBES makes it possible that an optimal combination of QBES with other methods may further improve the prediction. The result from QBES can be either used as the initial result to be refined by other methods or as one of the methods for a consensus prediction from multiple prediction results. The methodology developed here can be easily extended to predict protein contact maps. This can be done by constrained minimization of total energies using parameters of residue–residue contact energies.²⁷ This approach is attractive because it takes in the whole sequence and thus may improve the chance to predict important nonlocal contacts. The work in this area is in progress.

ACKNOWLEDGMENTS

We thank Dr. Dahai Xu for his valuable comments and discussion and Professor Shandar Ahmad for providing us his standalone copy of RVP-Net.

REFERENCES

- Rost B, Sander C. Conservation and prediction of solvent accessibility in protein families. *Proteins* 1994;20:216–226.
- Hobrook S, Mushal S, Kim S. Predicting surface exposure of amino acids from protein sequence. *Protein Eng* 1990;3:659–665.
- Pascarella S, De Persio R, Bossa F, Argos P. Easy method to predict solvent accessibility from multiple protein sequence alignments. *Proteins* 1999;32:190–199.
- Li X, Pan X.-M. New method for accurate prediction of solvent accessibility from protein sequence. *Proteins* 2001;42:1–5.
- Pollastri G, Baldi P, Fariselli P, Casadio R. Prediction of coordination number and relative solvent accessibility in proteins. *Proteins* 2002;47:142–153.
- Ahmad S, Gromiha M. M. NETASA: neural network based prediction of solvent accessibility. *Bioinformatics* 2002;18:819–824.
- Garg A, Kaur H, Raghava G. Real value prediction of solvent accessibility in proteins using multiple sequence alignment and secondary structure. *Proteins* 2005;61:318–324.
- Raih M, Ahmad S, Zheng R, R. M. Solvent accessibility in native and isolated domain environments: general features and implications to interface predictability. *Biophys Chem* 2005;114:63–69.
- Yuan Z, Burrage K, Mattick J. S. Prediction of protein solvent accessibility using support vector machines. *Proteins* 2002;48:566–570.
- Yuan Z, Huang B. Prediction of protein accessible surface areas by support vector regression. *Proteins* 2004;57:558–564.
- Kim H, Park H. Prediction of protein relative solvent accessibility with support vector machines and long-range interaction 3D local descriptor. *Proteins* 2004;54:557–562.
- Gianese G, Bossa F, Pascarella S. Improvement in prediction of solvent accessibility by probability profiles. *Protein Eng* 2003;16:987–992.
- Naderi-Manesh H, Sadeghi M, Arab S, Movahedi AAM. Prediction of protein surface accessibility with information theory. *Proteins* 2001;42:452–459.
- Carugo O. Predicting residue solvent accessibility from protein sequence by considering the sequence environment. *Protein Eng* 2000;13:607–609.
- Thompson M, Goldstein R. Predicting solvent accessibility: higher accuracy using bayesian statistics and optimized residue substitution classes. *Proteins* 1996;25:38–47.
- Mucchielli M. H, Tuffery P, Hazout S. PredAcc: prediction of solvent accessibility. *Bioinformatics* 1999;15:176–177.
- Ahmad S, Gromiha M. M, Sarai A. Real value prediction of solvent accessibility from amino acid sequence. *Proteins* 2003;50:629–635.
- Adamczak R, Porollo A, Meller J. Accurate prediction of solvent accessibility using neural networks-based regression. *Proteins* 2004;56:753–767.
- Mathworks Matlab. Natick, MA: Mathworks, Inc.; 1999.
- Zhou H, Zhou Y. Distance-scaled, finite ideal-gas reference state improves structure-derived potentials of mean force for structure selection and stability prediction. *Protein Sci* 2002;11:2714–2726.
- Zhou H, Zhou Y. Quantifying the effect of burial of amino acid residues on protein stability. *Proteins* 2004;54:315–322.
- Zhou H, Zhou Y. Single-body residue-level knowledge-based energy score combined with sequence-profile and secondary structure information for fold recognition. *Proteins* 2004;55:1005–1013.
- Zhou H, Zhou Y. SPARKS 2 and SP³ servers in CASP 6. *Proteins* 2005;(Suppl 7) 152–156.
- Shrake A, Rupley J. A. Environment and exposure to solvent of protein atoms. lysozyme and insulin. *J Mol Biol* 1973;79:351–371.
- McConkey B, Sobolev V, Edelman M. Quantification of protein surfaces, volumes and atom–atom contacts using a constrained voronoi procedure. *Bioinformatics* 2002;18:1365–1373.
- Cuff J, Barton G. Application of multiple sequence alignment profiles to improve protein secondary structure prediction. *Proteins* 2000;40:502–511.
- Miyazawa S, Jernigan R. Self-consistent estimation of inter-residue protein contact energies based on an equilibrium mixture approximation of residues. *Proteins* 1999;34:49–68.
- Richardson C, Barlow D. The bottom line for prediction of residue solvent accessibility. *Protein Eng* 1999;12:1051–1054.