Novel Application of Query-Based Qualitative Predictors for Characterization of Solvent Accessible Residues in Conjunction with Protein Sequence Homology

Proceedings of the 22nd International Workshop on Database and Expert Systems Applications (2011), 70-74

Daniel A. Rose*, Reecha Nepal*, Radhika Mishra*, Robert Lau†, Shabnam Gholizadeh# and Brooke Lustig*
*Department of Chemistry, San Jose State University, San Jose, CA 95192-0101
#Department of Chemical and Materials Engineering, San Jose State University, San Jose, CA 95192-0082
†Department of Biology San Jose State University, San Jose, CA 95192-0100
email: brooke.lustig@sjsu.edu

Abstract—**Prediction of relative solvent accessibility (RSA) is a standard first-approach in predicting three-dimensional protein structures. Here we have applied linear regression methods that include various sequence homology values for each residue as well as query residue qualitative predictors, corresponding to each of the twenty canonical amino acids. We fit the 268-protein learning set with a variety of sequence homology terms, including 20 and 6-term sequence entropy, and residue qualitative predictors. Then estimated RSA values are subsequently generated for the 215-protein Manesh test set. The qualitative predictors describe the actual query residue type (e.g. Gly) as opposed to the measures of sequence homology for the aligned subject sequences. This is consistent with our framework of modeling a limited set of discrete and/or physically intuitive predictors. Initial calculations involving normalized RSA values were considered as a likely first attempt, incorporating the notion of fitting an explicit binary characterization of individual residues, either as buried or accessible. Interestingly, the utilization of qualitative predictors showed significant prediction accuracy. Subsequent calculations using the original RSA values gave estimated values that, upon binary classification, indicated accuracies comparable to other first stage methods. Development of a second stage methodology is of current interest.**

Keywords-hydrophobicity, sequence entropy, buried residues, surface accessibilities, qualitative predictors

I. INTRODUCTION

Historically, the details of protein structure and their corresponding description of protein function have required high-resolution three-dimensional structures, typically involving x-ray crystallography. However with the advent of extensive databases involving various aspects of protein structure and function, the elucidation of function is not necessarily related to the difficult task of detailed protein structure determination [1][2]. This can include the screening of large numbers of sequences and related characterization of possible function. The need for prediction from sequence and/or sequence homology is a result of the current state of affairs, where significant difficulties remain for determining structures derived via X-ray [3]. This has been ameliorated to some extent by the use of NMR structures as well as comparative and homology modeling, but a large number of protein sequences remain without corresponding reliable three-dimensional structures [4].

One key descriptor of function is the characterization of solvent accessible surfaces, the results of which are useful in many applications in protein design and structural biology [5][6]. Notably these include identifying catalytic and other key functional residues, including those found on protein surfaces. For the greatest coverage of the proteome, this would not necessarily require inputs of high-resolution three-dimensional structure. And with the advent of proteomics, there is considerable interest in such surface prediction calculations as applied to characterizing protein-protein interactions [7]. Moreover, prediction of surface accessible residues from just sequence is a reasonable first-approach for the important goal in structural biology of modeling three-dimensional protein structure. Methods using protein sequence information, including first-generation machine learning approaches, typically have shown percent accuracy on the order of 70-75% [8]-[10]. Two-stage and related regression approaches are reported to have somewhat better results for certain proteins [10]-[13]. The most applicable of these includes improvements by Meller and coworkers in their versions of SABLE [14].

Other structural features that may prove amenable to characterization from sequence include specifically the identification of key core e.g. strongly hydrophobic residues [15][16]. Such residues can describe key constraints in modeling a protein's folding and may help design modifications for proteins. The calculation of Shannon entropy has been put forth as one of several methods for scoring amino acid conservation in proteins [17][18]. Shannon entropies for protein sequences have been shown to correlate with configurational entropies calculated from local physical parameters, including backbone geometry [19]. However, such sequence entropy by itself does not appear a unique identifier of structural features [20][21].

Sequence entropy has shown some potential to characterize protein-protein interfaces [22]. We have previously shown two regions for sequence entropy and hydrophobicity of individual residues when plotted with respect to the inverse of their respective Cα packing density [23]. The second region (associated with less than

11 Cα per 9Å radius) is essentially flat and consistent with the most flexible residues, typically showing significant exposure to solvent.

Here we propose to apply relevant sequence homology parameters, including sequence entropy, as predictors of residue solvent accessibility. They are used in conjunction with qualitative predictors in the form of query (i.e. directly from sequence) amino acid type (e.g. Gly). Though their application here is somewhat novel, qualitative predictors have been successfully utilized in economics, social sciences and biology [24]. The qualitative predictors are separate terms with respect to ones involving homology-based values and can be treated as such. This allows consideration for query residue type and/or measures of the conservation for that residue. The key is to utilize only a limited set of discrete and/or physically intuitive terms. This makes it easier to note any intrinsic factors, including limitations, involved in predicting solvent accessible residues.

II. Methods and Results

A. Homology Based Parameters

The first key practical step is to characterize those homology-based parameters that are likely to prove useful in the prediction of solvent accessible residues. To this purpose, sets of subject protein sequences are aligned to query residues of the relevant protein sequences. For the learning set we utilized a diverse 268-protein list [25]. Sequence alignment involved a straightforward and non-biased standard application of BLASTP and PSI-BLAST to a non-redundant database (Genbank) [26]. Our test set used the standard 215-protein list from Naderi-Manesh et al. [27]. Sequence homology parameters for Lustig and coworkers [23][25][28] include 20-term (E20) and 6-term sequence entropy (E6) [18][29][30]. Here, standard 20-term sequence entropy i.e. E20 at some residue position k is expressed as

$$S_k = -\sum_{j=1,\,20} P_{jk} \log_2 P_{jk}\,, \qquad (1)$$

where probability P_{jk} at amino acid sequence position k is derived from the frequency for an amino acid type j for N aligned residues. And alternatively for 6-term entropy, j is indexed over 6 classes of amino acid. Other sequence homology parameters are the fraction of aligned residues (corresponding to the query residues) that are strongly hydrophobic (FSHP) and small i.e. Gly or Ala (FSR). Shown in Fig. 1 is the correlation plot of these parameters with respect to inverse Cα packing density.

B. RSA Calculations

Note the general description of two regions (see Fig. 1), including for Region II corresponding to packing densities less than 11 Cα in a standard packing radius less than 9Å. Characterization of Region II residues indicates RSA values consistent with exposure to solvent. The RSA values were determined from the corresponding query X-ray structures using NACCESS [31]. The fairly consistent characterization of a solvent accessible region for the homology-based parameters suggests their utility in predicting residues accessible to solvent.

C. Qualitative Predictors

An initial simple calculation illustrates the application of query-related qualitative predictors for RSA prediction [24]. Here we utilize just two such predictors, strongly hydrophobic residues i.e. SHP (VLIFYMW) and the remaining 13 non-strongly hydrophobic residues (NSHP). Note we fit our 73,675 residue RSA (relative surface accessibility) values to the variable X_{i1} corresponding to the E6 value at each residue i, and two qualitative predictors SHP (X_{i2} is 0) and NSHP (X_{i2} is 1) with following expression

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \varepsilon_i. \qquad (2)$$

The generalized response function can be written as

$$E\{Y\} = \beta_0 + \beta_1 X_1 + \beta_2 X_2, \qquad (3)$$

with two respective response functions for fitting SHP and NSHP

$$E\{Y\} = \beta_0 + \beta_1 X_1 \text{ where } X_2 \text{ is } 0; \qquad (4)$$

$$E\{Y\} = (\beta_0 + \beta_2) + \beta_1 X_1 \text{ where } X_2 \text{ is } 1. \qquad (5)$$

The R derived fit is show in Fig. 2 for both plots of the same slope of 8.280 and intercepts β_0 and $(\beta_0 + \beta_2)$ of 5.10 and 24.93, respectively.

D. Results for 11 Models

It was decided for the first extensive set of regressions to renormalize RSA values as being in the interval from 0 to 1. This approach [32], used in the social sciences and related fields, was first considered here as an intermediate "step" between SVM binary classification and regression involving original RSA values.

Here we initially assume a fairly conservative threshold that amino acids with less than 20% relative exposure to solvent i.e RSA are to be classified as buried versus being surface (and solvent) accessible [33]. The accuracy for the test set of both buried and surface accessible residues is then calculated by the standard expression of Richardson & Barlow [9]. We first do linear regression for various models on the learning set by fitting normalized RSA values (with respect to a range 0 to 1) to various sequence homology parameters (determined exclusively from just aligned residues), notably the sequence entropies and a set of twenty

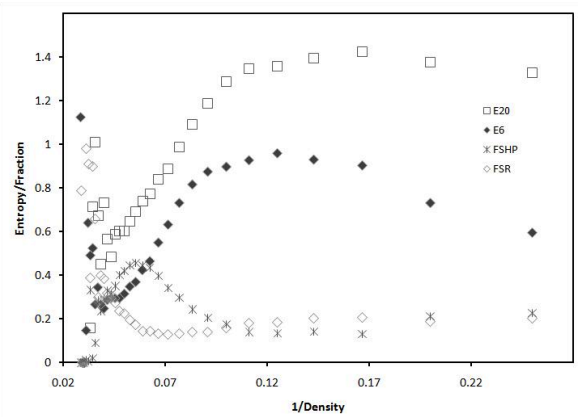

Figure 1. Combined aggregate correlation plots of sequence entropy and other homology-based parameters for the 268-protein list, calculated with gaps excluded for a total of 235,138 BLASTP alignments. Packing density is the number of Cα within a 9Å radius and excluded here is the portion of Region II with packing densities less than 4 (<1% of all residues). Average sequence entropy, E20 (open-square, ordinate) and E6 (closed-diamond) are calculated by averaging the respective values for 73,727 query residues for each inverse packing density value (abscissa). Fraction of strongly hydrophobic residues (asterisk) and fraction of small residues (open-diamond) are calculated and averaged over a total of 7.12E7 aligned residues, plotted against inverse packing density. Average values for all the homology-based parameters are determined by averaging their respective values within the same packing density interval. Note that the standard deviations for E20 and E6 are comparable (typically 0.3), while typically 0.1 for FSHP and FSR.

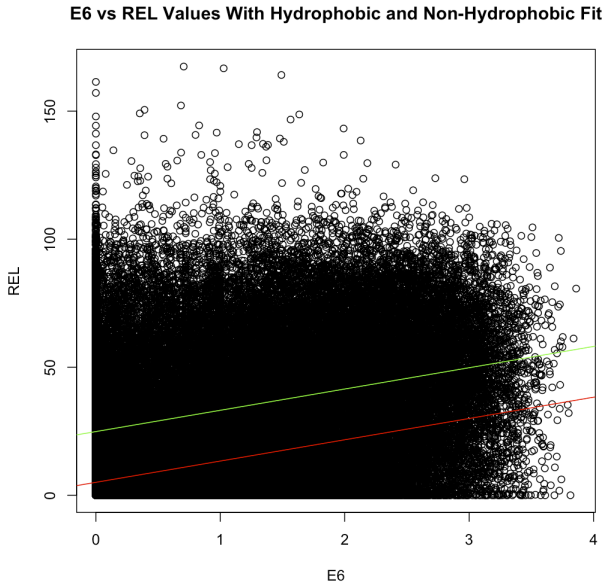

Figure 2. Sample regression fit for 73,675 query residues from the learning 268-protein list. Here we fit to original RSA (REL) values to a variable term X_{i1} as E6 and the qualitative predictor term having two values, where X_{i2} is 0 for SHP query residues and X_{i2} is 1 for NSHP query residues. The slope (8.280) corresponding to the variable term is the same for the two plots, while the intercepts are 5.10 and 24.93 for β_0 and $(\beta_0 + \beta_2)$, respectively. The extrapolation of RSA values to the 50,635 residues of the 215-Manesh-test set gives an estimated prediction accuracy of 0.638 as estimated prediction accuracy. This accuracy assumes a modified threshold of 23% or greater for classifying model Manesh RSA values as being on the surface.

qualitative predictors (AA-set) describing query residue type (e.g. Gly) [24][32]. Then, we analyze the test set with those parameters to estimate normalized RSA values for those 11 models (see Table 1). Here any predicted renormalized value greater or equal to 0.5 is classified as surface accessible. The addition of qualitative predictors looks promising, especially noting relatively significant first-stage prediction accuracies that involve no or few other predictors.

There was a small increase in prediction accuracy of up to 0.745 noted for test set proteins classified by including assignments of secondary structure sub-class. However, this was only as a result of keeping in-class learning and test sets.

TABLE I. SUMMARY OF REGRESSION ACCURACY FOR 215-TEST SET

Models = REL	**Accuracy**		**Models = REL**	**Accuracy**	
	Un-norm.	Norm.		Un-norm.	Norm.
E20	0.631	0.627	E20+E6+ AA	0.729	0.734
E6	0.670	0.667	E20+E6	0.674	0.670
FSHP	0.673	0.670	E20+FSR +FSHP+ (AA)	0.734	0.733
AA	0.706	0.703	E6+FSR +FSHP+ (AA)	0.732	0.733
E20+ AA	0.721	0.731	E20+E6+ FSR+ FSHP+ (AA)	0.724	0.733
E6+AA	0.731	0.725			

Moreover, the normalized results for both BLAST and PSI-BLAST derived data sets are comparable as shown for the eleven models. So we repeated the calculations without normalizing the learning and test set RSA values for the 11 models of interest, using just the BLAST-derived data (Table 1). Calculating RSA values directly has proven a viable alternative [34][35] especially for second stage methods [11]. All regression related calculations involved version 2.12.2 of R.

I. DISCUSSION

Two major regions are noted for the sequence homology parameters when plotted against inverse Cα packing density, here defined as the number of Cα within a 9Å radius. Region II corresponding to packing densities less than 11, is consistent with the corresponding query residues generally being accessible to solvent. Region I residues, with packing densities greater than or equal to 11, are typically considered buried. Interestingly, substitutions of small residues are disproportionately indicated in the most densely packed regions, consistent with earlier observations for original sequence i.e. query residues [36]. A very limited number of sequence homology parameters in conjunction

with our AA-set qualitative predictors can well predict the likelihood of a residue being buried as part of the binary classification. The optimal set of predictors included the E6, FSR, FSHP and AA-set, and alternatively E, E6, FSR, FSHP and AA-set.

The direct introduction as qualitative predictors of secondary sub-class information did not improve prediction results. However, limiting learning and test sets [37] to their respective sub-classes did show some improvement. In this regard it might be useful to determine if an experimental method, such as circular dichroism (CD), can allow us to independently classify protein secondary structure sub-class classifications for such purposes. Moreover there may be utility in using computationally designed proteins as learning sets, given secondary structure prediction appears to significantly improve by the use of such learning sets [38]. Moreover, it is surprising that E6 entropy by itself is so useful, suggesting other alternative types of entropy that need to be explored. Noteworthy are pair-wise entropies that may allow filtering of possible tertiary structure contact pairs, which are known to be largely in the buried region [39] of proteins.

Multiple stages have been shown to be of some advantage with respect to predicting residue solvent accessibility with SVM and regression approaches [10][11]. Others have calculated solvent accessible surface areas values explicitly, rather than RSA values [40]. But even with a range of accessibility thresholds for prediction optimization, accuracy remains at about 74-79%. Indeed our calculations compare favorably to existing single-stage calculations [41]. A second-stage implementation with our first-stage qualitative predictor methodology is in development.

ACKNOWLEDGMENT

This work was supported by Intel Grant 32553 and the California State University General Research Grant-2009.

REFERENCES

[1] A. R. Panchenko, F. Kondrashov, and S. Bryant. “Prediction of funtional sites by analysis of sequence and structure conservation”. Prot Sci., vol. 13, 4, pp. 884-892, 2004.

[2] J. D. Watson, R. A. Laskowski, and J. M. Thornton. “Predicting protein function from sequence and structural data”. Curr Opin Struct Biol., vol. 15, 3, pp. 275-284, 2005.

[3] G. E. Dale, C. Oefner, and A. d’Arcy. “The protein as a variable in protein crystallization”. J Structr Biol., vol. 142, 1, pp. 88-97, 2003.

[4] U. Pieper, et al. “A database of annotated comparative protein structure models, and associated resources”. Nucl Acids Res., vol. 39, 1, pp. D465-D474, 2011.

[5] N. V. Petrova and C.H. Wu, “Prediction of catalytic residues using Support Vector Machine with selected protein sequence and structural properties”. BMC Bioinformatics., vol. 7, 4, pp. 313-323, 2006.

[6] F. K. Pettit, A. Tsai, and J. U. Bowie. “A statistical approach to finding biologically relevant features on protein surfaces”. J Mol Biol., vol. 369, 3, pp. 863-879, 2007.

[7] A. Porollo and J. Meller. “Prediction-Based fingerprints of protein-protein interactions”. Proteins, vol. 66, 3, pp. 630-645, 2007.

[8] B. Rost and C. Sander. “Conservation and prediction of solvent accessiblity in protein families”. Proteins, vol. 20, 3, pp. 216-226, 1994.

[9] C. J. Richardson and D. J. Barlow. “The bottom line for prediction of residue solvent accessibility”. Prot Engr., vol. 12, 12, pp. 1051-1054, 1999.

[10] M. N. Nguyen and J. C. Rajapakse. “Prediction of protein relative solvent accessibility with a two-stage SVM approach”. Proteins, vol. 59, 1, pp. 30-37, 2005.

[11] M. N. Nguyen and J. C. Rajapakse. “Two-stage support vector regression approach for predicting accessible surface areas of amino acids”. Proteins, vol. 63, 3, pp. 543-550, 2006.

[12] R. Adamczak, A. Porollo, and J. Meller. “Accurate prediction of solvent accessibility using neural networks–based regression”. Proteins, vol. 56, 4, pp. 753-767, 2004.

[13] J.Y. Wang, H. M. Lee, and S. Ahmad. “SVM-Cabins: Prediction of solvent accessibility using accumulation cutoff set and support vector machine”. Proteins, vol. 68, 1, pp. 82-91, 2007.

[14] Solvent Accessibilities2 (SABLE2) (2008-2009). [Online]. Available: http://sable.cchmc.org/

[15] A. M. Poupon and J.P. Mornon. “ Predicting the protein folding nucleus from a sequence”. FEBS Lettr., vol. 452, 3, pp. 283-289, 1999.

[16] I. N. Berezovsky and E. N. Trifonov. “Van der Waals locks: Loop-n-lock structure of globular proteins”. J Mol Biol., vol. 307, 5, pp. 1419-1426, 2001.

[17] P. S. Shenkin, B. Erman, L. D. Mastrandrea. “Information-theoretical entropy as a measure of sequence variability”. Proteins, vol. 11, 4, pp. 297-313, 1991.

[18] W. S. J. Valdar. “Scoring residue conservation”. Proteins, vol. 48, 2, pp. 227-241, 2002.

[19] P. Koehl and M. Levitt. “Protein topology and stability define the space of allowed sequences”. Proc Natl Acad Sci USA., vol. 99, 3, pp. 1280-1285, 2002.

[20] L. Oliveira, A. C. M. Paiva, and G. Vriend. “Correlated mutation analyses on very large sequence families”. ChemBioChem., vol. 3, 10, pp. 1010-1017, 2002.

[21] C. Yan, et al. “Predicting DNA-binding sites of proteins from amino acid sequence”. BMC Bioinformatics., vol. 7, 4, pp. 262-271, 2006.

[22] M. Guharoy and P. Chakrabarti. “Conservation and relative importance of residues across protein-protein interfaces”. Natl Acad Sci USA., vol. 102, 43, pp. 15447-15452, 2005.

[23] H. Liao, W. Yeh, D. Chiang, R. L. Jernigan, and B. Lustig. “Protein sequence entropy is closely related to packing density and hydrophobicity”. Prot Engr., vol. 18, 2, pp. 59-64, 2005.

[24] M. H. Kutner, C. J. Nachshelm, J. Neter. “Applied Linear Statistical Models”, 5th ed, NY, McGraw-Hill, 2004, Chapter 8, pp.294-335.

[25] R. Mishra, “Characterization of Protein Residue Structural Accessibility Using Sequence Entropy”, MS Thesis, Chem, SJSU, CA, 2010.

[26] GenBank (NCBI) (2009-2010). [Online]. Available: http://www.ncbi.nlm.nih.gov/genbank/

[27] H. Naderi-Manesh, M. Sadheghi, S. Araf, and A. A. M. Movahedi. “Predicting of protein surface accessibility with information theory”. Proteins, vol. 42, 4, pp. 452-459, 2001.

[28] S. Do, H. Lakkaraju, S. Potluri, K. Pham, K. Kantardjieff, and B. Lustig . “Protein Sequence Homology Parameters Applied to the Prediction of Solvent Accessible Residues”. 52nd Biophysical Society Meeting Abstract, (2008) 52nd Biophysical Society Meeting Abstracts. Biophys. J. 94, Supplement, 3280P.

[29] M. Gerstein and R. B. Altman. “Average core structures and variability measures for protein families: Application to the immunoblobulins”. J Mol Biol., vol. 251, 1, pp. 161-175, 1995.

[30] L. A. Mirny and E. I. Shakhnovich. “Universally conserved positions in protein folds: Reading evolutionary signals about stability, folding kinetics and function”. J Mol Biol., vol. 291, 1, pp. 177-196, 1999.

[31] S. J. Hubbard and J. M. Thornton. NACCESS, Computer Program, Department of Biochemistry and Molecular Biology, University College London. 1993.

[32] O. Hellevik. “Linear Versus logistic regression when the dependent variable is a dichotomy”. Qual Quant., vol. 43, 1, pp. 59-74, 2009.

[33] O. Carugo. “Predicting residue solvent accessibility from protein sequence by considering the sequence environment”. Protein Eng., vol. 13, 9, pp. 607-609, 2000.

[34] S. Ahmad, M. M. Gromiha, and A. Sarai. “Real value prediction of solvent accessibility from amino acid sequence”. Proteins, vol. 50, 4, pp. 629-635, 2003.

[35] M. Wagner, R. Adamczak, A. Porollo, and J. Meller. “Linear regression models for solvent accessibility prediction in proteins”. JComput Biol., vol. 12, 3, pp. 355-369, 2005.

[36] M. Eilers, et al. “Internal packing of helical membrane proteins”. Proc Natl Acad Sci., vol. 97, 11, pp. 5796-5801, 2000.

[37] L. A. Kurgan and L. Homaeian. “L. Prediction of structural classes for protein sequences and domains—Impact of prediction algorithms, sequence representation and homology, and test procedures on accuracy”. Pattern Recogn., vol. 39, 12, pp. 232-2343, 2006.

[38] R. Bondugula, A. Wallqvist, and M. S. Lee. “Can computationally designed protein sequences improve secondary structure prediction?”. Prot. Engr., vol. 24, 5, pp. 455-461, 2011.

[39] H. Kim and H. Park. “Prediction of protein relative solvent accessibility with support vector machines and long-range interaction 3D local descriptor”. Proteins, vol. 54, 3, pp. 557-562, 2004.

[40] Z. Yuan and B. Huang. “Prediction of protein accessible surface areas by support vector regression”. Proteins, vol. 57, 3, pp. 558-564, 2004.

[41] G. Gianese, F. Bossa, and S. Pascarella. “Improvement in prediction of solvent accessibility by probability profiles”. Prot. Engr., vol. 16, 12, pp. 987-992, 2003.