Prediction and Evolutionary Information Analysis of Protein Solvent Accessibility Using Multiple Linear Regression

Jung-Ying Wang,[1,2] Hahn-Ming Lee,[1] and Shandar Ahmad[3,4]*

[1]*Department of Computer Science and Information Engineering, National Taiwan University of Science and Technology, Taipei, Taiwan*

[2]*Department of Multimedia and Game Science, Lunghwa University of Science and Technology, Taoyuan, Taiwan*

[3]*Department of Bioscience and Bioinformatics, Kyushu Institute of Technology, Iizuka, Fukuoka-ken, Japan*

[4]*Department of Bioscience, Jamia Millia Islamia University, New Delhi, India*

ABSTRACT **A multiple linear regression method was applied to predict real values of solvent accessibility from the sequence and evolutionary information. This method allowed us to obtain coefficients of regression and correlation between the occurrence of an amino-acid residue at a specific target and its sequence neighbor positions on the one hand, and the solvent accessibility of that residue on the other. Our linear regression model based on sequence information and evolutionary models was found to predict residue accessibility with 18.9% and 16.2% mean absolute error respectively, which is better than or comparable to the best available methods. A correlation matrix for several neighbor positions to examine the role of evolutionary information at these positions has been developed and analyzed. As expected, the effective frequency of hydrophobic residues at target positions shows a strong negative correlation with solvent accessibility, whereas the reverse is true for charged and polar residues. The correlation of solvent accessibility with effective frequencies at neighboring positions falls abruptly with distance from target residues. Longer protein chains have been found to be more accurately predicted than their smaller counterparts. Proteins 2005;61:481–491.**

Key words: solvent accessibility; multiple linear regression; structure prediction

INTRODUCTION

The prediction of protein structures directly from an amino acid sequence is one of the most challenging issues in biological research. Knowledge of the solvent accessibility or accessible surface area (ASA)[1] of each residue in the protein gives us valuable information for the prediction of tertiary structure and function of proteins. Given an amino acid sequence, the goal of ASA prediction is to estimate the accessibility of each residue in the sequence. Many pattern-recognition and machine-learning methods have been proposed to address this issue. Some typical approaches are

(i) neural networks,[2–8]
(ii) support vector machines,[9–11]
(iii) information theory,[12–14]
(iv) Bayesian analysis.[15]

Among these machine-learning methods, (simple and complex) neural networks and support vector machines have been shown to be the most effective for ASA prediction. Traditionally, ASA prediction proceeds by subdividing residues into two (exposed or buried) or three states (exposed, intermediate, or buried), based on different thresholds of relative exposed surface area. This requires the selection of an ASA value that can be used as a threshold to divide ASA states. However, in real protein structures, there are no such well-defined ASA states, a condition which leads to arbitrary choices, poor comparison, and loss of information. In light of this, we recently proposed a method for the direct prediction of real values of ASA instead of its arbitrarily defined states.[7] We also developed look-up tables for making these predictions based on patterns in two- or three-residue segments.[16] Meanwhile, Yuan et al.[11] have used support-vector regression, and Adamczak et al.[8] have used neural network-based regression to make real-value ASA predictions.

The multiple linear regression method has been shown to be quite useful in solving pattern-recognition problems. Li and Pan[17] have also used multiple linear regression to complete the solvent accessibility prediction in two states (exposed or buried). Here, we apply the method of multiple regression for the prediction of real-value solvent accessibility and carry out a thorough analysis of the information that leads to those predictions. Prediction performance is verified by a five-fold cross-validation, and the role of evolutionary information is studied by constructing a correlation matrix at different window positions.

Grant sponsor: National Science Council of Taiwan; Grant number: NSC 93-2213-E-011-045.

*Correspondence to: Shandar Ahmad, Kyushu Institute of Technology, Bioscience and Bioinformatics, Iizuka, Fukuoka-ken, 820-8502, Japan. E-mail: shandar@bse.kyutech.ac.jp

Received 30 November 2004; Revised 26 March 2005; Accepted 11 April 2005

Published online 16 September 2005 in Wiley InterScience (www.interscience.wiley.com). DOI: 10.1002/prot.20620

MATERIALS AND METHODS

Databases

We used three independent datasets for a fair comparison and wider applicability of analysis. These datasets consist of

(i) Barton502, the same non-redundant set of 502 proteins from our previous studies, which was originally compiled by Cuff and Barton[18] as a dataset of 513 proteins, and was slightly modified by us.[6,7,16] The values of ASA used here have also been directly taken from our previous work.[6]
(ii) Meller860, the dataset used by Adamczak et al.[8] The list of proteins used was provided by the authors, and ASA values were computed and normalized for analysis by us.
(iii) Yuan1277, the dataset used by Yuan and Huang.[11] The list of proteins used was obtained from their work, and ASA values were computed and normalized for analysis by us.

Most of the discussion in this manuscript draws from the first dataset (Barton502), unless otherwise noted. However, the results suggest that the choice among the three datasets does not significantly alter the major conclusions from this study (see later sections).

Coding Scheme

In order to develop any prediction model, amino acids at target and neighboring positions are encoded by a numerical vector. The most obvious way of coding amino acids is to use orthogonal unit vectors in 20 dimensions. In such a coding system, each residue is represented by 20 units, all of which are set to zero, except for one position that is unique to that residue whose value is set to 1.[6,7,19] An additional unit is sometimes added to indicate a null/terminal residue input, thus making the information of a single residue 21-dimensional. A complete residue environment is therefore represented by 21-dimensional representations of the target residue and its neighbor. Typically, a moving window of n neighboring residues is used, making the input a $[21 \times (2n + 1)]$-dimensional binary vector.

In our study, we first consider the orthogonal coding of a moving window with a size of 13 neighboring residues. The window was shifted residue by residue through the protein chain, thus yielding N data points for a chain with N residues. The prediction was made for the central residue in the window frame. In order to allow the moving window to overlap the amino- or carboxyl-terminal end of the protein, a null input was added for each protein chain. Therefore, each data point contained $13 \times 21 = 273$ values. Hence each data could be represented as a 273-dimensional vector. Note that the main dataset (Barton502) consists of 83,330 data points.

Multiple Linear Regression

The prediction model based on a linear regression function can be expressed as:

$$\mathrm{ASA}(i) = \alpha + \sum_{i=1}^{i=13} \sum_{j=1}^{j=21} \beta_{i,j} X_{i,j} \qquad (1)$$

In our study, we optimized the multiple linear regression model to solve for the unknown coefficients, α and $\beta_{i,j}$ ($1 + 13 \times 21 = 274$ coefficients), by performing a least squares fit. Note that of the 21 values of vector $\mathbf{X}$ (e.g. $\mathbf{X_{1,1}}, \ldots, \mathbf{X_{1,21}}$), only one has the value of 1, while all others are zero.

Evolutionary Information

We used the program BLASTPGP to generate multiple sequence alignments of proteins in our database, with the expectation value (E-value) of 0.01, and we chose the non-redundant protein sequence database (NCBI nr database) to search. The alignments were represented as profiles or position-specific substitution matrices (PSSMs). PSSMs provide the effective frequency of occurrence of all 20 amino acid residues at each position of the sequence, normalized by background frequencies obtained for an entire protein-sequence space (e.g. constituting a BLOSUM62 matrix). In case no alignment is observed at a given position, PSSM simply returns a row of the BLOSUM62 matrix representing the target residue. PSSM data obtained from BLASTPGP were directly used as inputs to our regression model. A prediction system based on PSSM therefore replaces the 20-bit orthogonal coding by the rows of PSSM data. Additional units are also added to represent indel and entropy data obtained from PSSM.

Assessment of Prediction Performance

Consistent with other ASA prediction methods, we used two standard measures of assessing prediction quality, namely the mean absolute error (MAE) and the correlation coefficient between the predicted and experimental ASA values. These scores have been sufficiently defined elsewhere.[8,11,16]

Analysis of Sequence and Evolutionary Information

Two methods were employed here to asses the role of neighbors in enhancing or reducing the tendency of residues to be on the surface. In the first method (applied to sequence-only based information), first the mean values of ASA for the entire dataset were calculated. Then, one-by-one a subset of the data were selected that had a particular residue in a given window position. New mean values of ASA were then computed and compared with the global average. The differences between these average values for the data subset and the global average provide a direct measure of the bias caused in the ASA values by the presence of an identified residue. These results were compiled as average ASA value tables.

PSSM information could not be analyzed in this way because here each residue position was not occupied by one residue. Instead, for each position, the probability of occurrence of a residue was written. Thus data subsets as defined above did not make much sense for PSSM data. We

TABLE I. Summary of the Real-Value Prediction Error (MAE %) and Correlation Coefficient (in brackets) Obtained from Five-Fold Cross-Validation on Different Datasets.

	Barton502	Yuan1277	Meller860
# Protein	502	1277	860
# Residue	83830	256656	212316
Orthogonal coding	18.9	19.1	18.7
%MAE (correlation)	(0.52)	(0.55)	(0.53)
PSSM coding	16.6	17.2	16.8
%MAE (correlation)	(0.63)	(0.63)	(0.62)
PSSM + composition + sequence length	16.2	16.4	16.2
%MAE (correlation)	(0.64)	(0.66)	(0.65)

therefore calculated the correlation coefficients between the probability of occurrence of a residue in a given position (PSSM matrix elements) and the window positions. These correlation tables contain information about the dependence of ASA (and not its prediction) on different effective frequency values.

Validation Method

Five-fold validation of results was carried out. The whole dataset was divided into four parts. The multiple linear regression model was developed on four-fifths of the data and tested on the remaining fifth. After running this process five times, an average over all the five test datasets was calculated and is listed in the results tables.

RESULTS AND DISCUSSION

Table I lists the main results of prediction using different datasets and information sources. These results show that using the orthogonal coding with single sequence, only the average MAE is 18.7-19.1%, which is competitive with previous studies based on more complex models such as neural networks.[8,16] If we use PSSM with two more features of insertion-deletion (indel) and entropy as input coding, then the MAE could be reduced to 16.6%, which is further reduced to 16.2% by the inclusion of a 21-dimensional vector (20 for amino-acid composition and one for the sequence length). As an example, Figure 1 shows the comparison between experimental and predicted values for each residue in protein 256B chain A, which has 106 amino acid residues. Except in those cases in which the experimental values are too high or too low, our method predicts a good linear relationship between the experimental and predicted values. A quick look at Table I and comparison with prediction accuracy obtained by other neural network models suggests that the solvent accessibility values of neighbors affect each other predominantly in a linear way, and adding a layer of non-linearity and complexity by way of hidden layers in neural networks only slightly improves prediction. For example, a neural network with one hidden layer using sequence-only information resulted in 18.8% MAE in our previous work. In a subsequent study, Adamczak et al.[8] could successfully reduce this MAE to 15.3% by using PSSM-derived evolutionary information. Our prediction MAE using simpler linear regression MAE is almost the same for the sequence-based prediction and slightly poorer for the PSSM-based method. We note that the evolutionary-information based non-linear models perform somewhat better than their linear-regression based methods. There are basically two ways in which a linear regression model differs from a neural network. First is the non-linearity of a neural network, introduced by the hidden layers. In this regard, it seems that the role of non-linearity in solvent accessibility predictions is small, and a linear regression model is almost as capable of capturing features of amino-acid environment as its non-linear counterpart. The second major difference lies in the training-validation method. A neural network is trained only as long as accuracy of prediction on the test data also shows simultaneous improvement in accuracy, whereas a linear regression using least square fit does not take into account the test data until the training is completed. This may lead to overfitting of the model to the training data, consequently showing poorer generalization for the test datasets. Due to this training-test-validation system of machine-learning methods and training-testing methods of multiple linear regressions, a direct comparison may look difficult. However, in our problem the prediction accuracy of the training and test sets were almost identical, suggesting that the data size is large enough so that it may not be possible to over-train all proteins in the training datasets. Also, to get robust values of accuracy from our previously reported neural network model, we repeated our calculations with different validation schemes, splitting data for three-fold, four-fold, and five-fold validation schemes. We did not find any significant difference in the prediction accuracy results obtained in all of these schemes.

Variation of Prediction Error with ASA Value Range

Table II shows the variation in MAE for linear regression-based prediction for different ranges of ASA. We found that 83.7% of residues with less than 60% exposure could be predicted within an average of 11.1% MAE. The prediction error increases linearly with increasing exposure percentage. When the exposure area exceeds 80% (only 5.9% of residues fall in this range), we observed that the prediction error was significantly higher. This is because our datasets have a high frequency of occurrence (coverage) with ASA less than 60% causing an overall bias of the model in that range. Similar results were also observed in our previous works based on neural networks.[7]

Residue-Specific Variation in Prediction Error

Table III and Figure 2 show the mean absolute error of 20 amino acids. As expected, Gly shows the highest MAE due to its flexibility, and other polar residues show similar behavior. Hydrophobic amino acids[20] (C, F, I, L, M, V, and W) are better predicted than less hydrophobic amino acids. These results are also in agreement with our neural network-based predictions.[7]

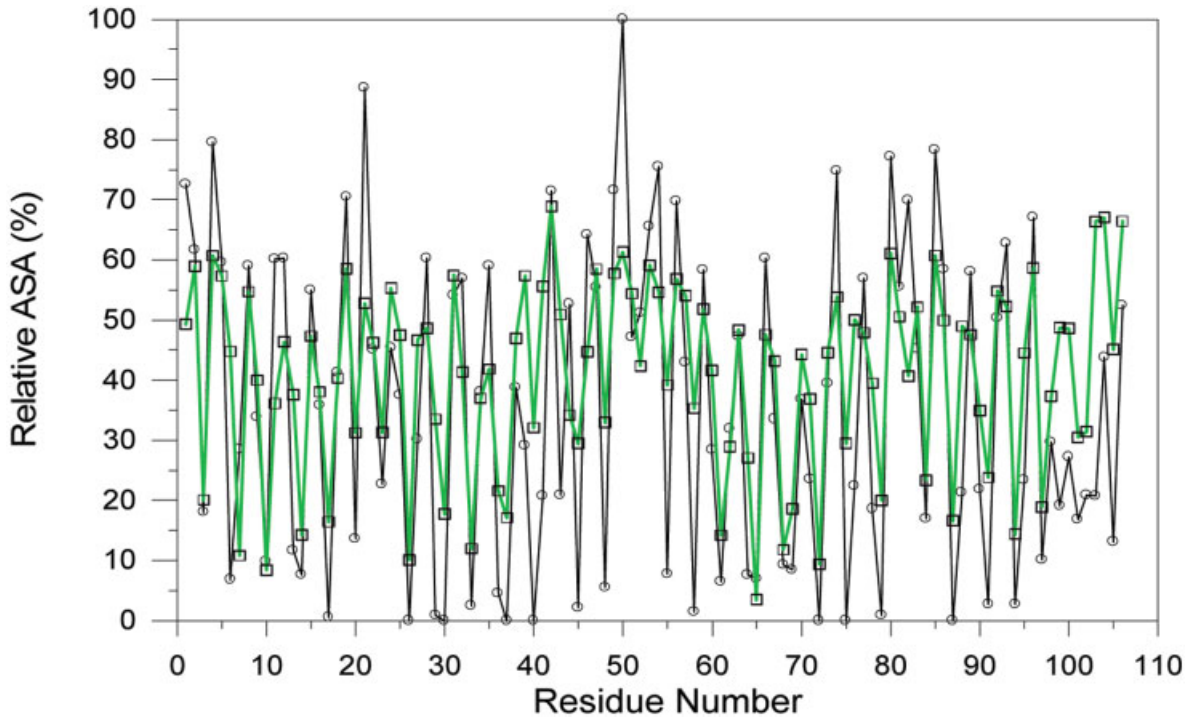

Fig. 1. A comparison of predicted (squares) and experimental (circles) values of ASA in protein 156B chain A, using PSSM in addition to amino-acid composition and sequence length information.

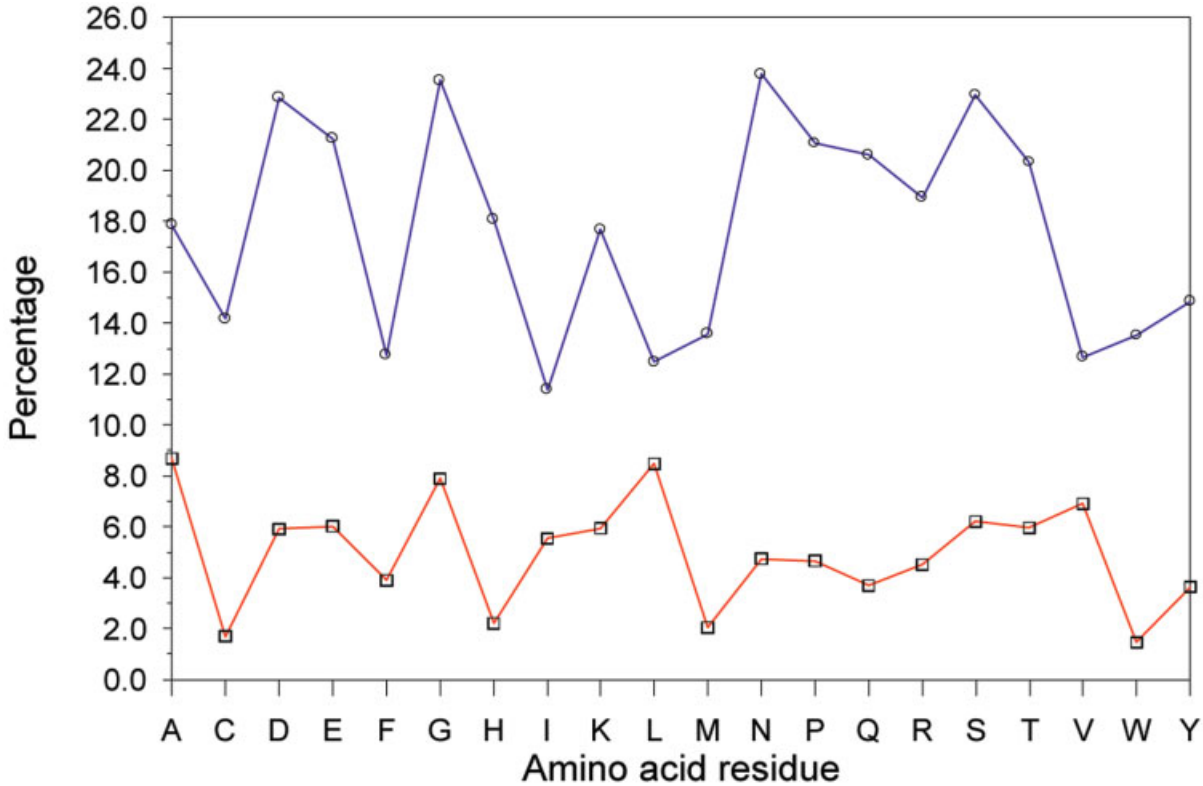

Fig. 2. The MAE of 20 amino acids (Barton502). Circles represent the prediction error, and squares show the corresponding data distribution of 20 amino acids.

TABLE II. Variation in Prediction Error for Different Ranges of ASA

Data set	Barton502		Meller860		Yuan1277		Average	
ASA range (%)	Data coverage	MAE (%)	Data coverage	MAE (%)	Data coverage	MAE (%)	Data coverage	MAE (%)
0–10	36.9	14.2	37.2	14.1	36.2	14.5	36.7	14.3
10–20	11.6	14.8	12.0	14.6	11.3	15.2	11.6	14.9
20–30	10.1	12.1	9.9	11.9	9.5	12.5	9.8	12.2
30–40	9.0	10.1	8.9	10.1	8.7	10.4	8.8	10.2
40–50	8.4	10.6	8.4	11.0	8.4	10.6	8.4	10.8
50–60	7.7	14.9	7.5	15.6	7.8	14.4	7.7	14.9
60–70	6.2	21.6	5.9	22.1	6.4	20.6	6.2	21.3
70–80	4.2	29.3	4.3	29.6	4.8	27.8	4.5	28.7
80–90	3.0	37.7	3.0	38.0	3.5	36.1	3.2	37.0
90–100	3.0	48.0	2.9	46.2	3.5	43.1	3.2	44.9

TABLE III. MAE for Different Amino-Acid Types[†]

Data	Barton502		Meller860		Yuan1277		Average	
Amino acid	Data coverage	MAE (%)	Data coverage	MAE (%)	Data coverage	MAE (%)	Data coverage	MAE (%)
A	8.7	15.6	8.3	15.2	8.5	15.8	8.4	15.5
C	1.7	14.2	1.5	12.6	1.6	13.1	1.6	13.1
D	5.9	20.8	5.8	20.8	6.0	21.0	5.9	20.9
E	6.0	19.3	6.6	19.9	6.5	19.8	6.4	19.8
F	3.9	11.9	4.0	12.4	3.9	12.4	4.0	12.3
G	7.9	21.1	7.5	20.7	7.8	20.8	7.7	20.8
H	2.2	15.7	2.3	16.4	2.3	16.5	2.3	16.4
I	5.5	10.6	5.7	10.7	5.4	10.9	5.5	10.8
K	5.9	16.3	6.0	17.0	6.0	17.0	6.0	16.9
L	8.5	11.6	8.7	11.6	8.4	11.9	8.5	11.7
M	2.0	12.9	2.3	12.6	2.2	13.6	2.2	13.1
N	4.7	21.0	4.4	20.5	4.5	20.9	4.5	20.8
P	4.7	18.2	4.6	18.5	4.7	18.5	4.7	18.5
Q	3.7	18.0	3.8	18.4	3.9	18.6	3.8	18.4
R	4.5	17.1	5.1	17.4	4.7	17.5	4.8	17.4
S	6.2	19.8	5.8	19.2	5.9	19.6	5.9	19.5
T	6.0	17.1	5.6	16.8	5.8	16.9	5.7	16.9
V	6.9	11.2	7.0	11.7	6.9	11.7	6.9	11.6
W	1.5	13.2	1.4	13.7	1.5	12.7	1.5	13.2
Y	3.6	13.3	3.6	13.6	3.5	14.1	3.6	13.8

[†]All values are on percentage scale.

Effect of Protein Chain Length on MAE

Table IV shows how protein chain length affects the MAE of predicted ASA values. Dependence of solvent accessibility prediction on chain length was not examined in our previous study, and this finding clearly establishes that the solvent accessibility prediction in longer-chain proteins is definitely better than in shorter ones. This may be attributed to the fact that the probability of distant-residue interaction is higher in longer proteins, reducing each residue's solvent accessibility. This leads to an overall lower value of solvent accessibility and hence an apparent fall in MAE for longer proteins in comparison to shorter ones, because lower ASA values are more accurately predicted.

Effect of Alignment Coverage and Number of Iterations

We tried to study the effect of two alignment parameters on the prediction MAE, so that an optimum procedure for generating alignments could be developed. The parameters are the number of alignment iterations for developing the PSSM and the E-value cutoff for compiling alignments. It was observed that three iterations of alignment cycles in BLASTPGP were enough, and no improvement in accuracy could be observed by increasing the number of iterations (data not shown). Similarly, an E-value of 0.01 was found to be the best, and no improvement in accuracy was observed by increasing the alignment coverage (i.e., by raising the E-values).

Analysis of Sequence and Evolutionary Information

The effect of including sequence neighbor information on prediction accuracy has been analyzed before.[1,3,12,14] However, each of these approaches gives a cumulative effect of neighbors on the accuracy of prediction. Thus, if one wants to see how the information at a particular window position

TABLE IV. Effect of Protein Chain Length on MAE†

Datasets Chain length	Barton502 MAE (%)	Meller860 MAE (%)	Yuan1277 MAE (%)	Average MAE (%)
Less than 100 residue	18.2 (#139)	19.2 (#183)	18.7 (#333)	18.7 (#655)
100–200 residue	16.5 (#219)	17.3 (#255)	17.8 (#502)	17.4 (#976)
200–300 residue	15.8 (#85)	16.3 (#149)	16.2 (#193)	16.1 (#427)
Greater than 300 residue	15.6 (#59)	15.7 (#273)	15.4 (#249)	15.5 (#581)

†# : Number of protein chains

TABLE V. Summary of the Information Present in Sequence Neighbors of Amino-Acid Residues in Proteins†

	Hydrophilic window positions	Hydrophobic window positions	Comments
Ala	—	0,+2,+3	Cooperative effect on nearby positions
Arg	0	+1,−2,+2	Opposite effect on nearby positions
Asn	0	−5,+5	Cooperative effect on distant residues
Asp	0,+1	−5	Cooperative effect similar to aliphatic counterpart (Asn)
Cys	−2,−5,+5	—	Hydrophilic effect on neighbors
Gln	0	—	No effect on neighbors
Glu	0,+1,−3,−4	−2,+2,+4	Different effects on N- and C-terminal residues
Gly	0,+1	−3,−4,−5,+5	Hydrophobic effect on +5 and −5 positions, perhaps due to a turn conformation
His	—	−1,−2,+2,−3,+3,−4,+4,−5	Hydrophobic effect falling gradually
Ile	—	0,−1,+1,−3,−4	Stronger effect on N-terminal neighbors
Leu	+5	0,+1,+3,−4,+4	Perhaps helical effect on +4 and −4 positions
Lys	0,−1,−3,+3,−4,+4	—	Cooperative hydrophilic effect
Met	—	0,−1,+1,−3,+3,−4,+4	Long-range hydrophobic effect
Phe	—	0,−1,−4,+4	Hydrophobic effect perhaps due to helix formation
Pro	0,−1,+2,+3,+4	+1,−5	−1, +2 positions are found to be more important than even 0 position, suggesting strong hydrophilic effect on neighbors
Ser	0	—	No effect on neighbors
Thr	0	−1	−1 position is more significant than itself
Trp	—	0,−1,−2,+3,−4	—
Tyr	—	0,−1,−2,+3,−4	—
Val	+5	0,−1,+1,−2,−3,−4	Cooperative hydrophobic effect on +5 position. Possible helical effect on −4 position.

†Hydrophilic (hydrophobic) window position means the sequence neighbor position that causes an increase (decrease) in ASA of a residue, as observed by a positive (negative) value of correlation. If a window position is found to have a hydrophobic (hydrophilic) effect on a known hydrophobic (hydrophilic) residue, the effect is termed as cooperative. Only the higher values of correlation have been considered.

affects the predictability of ASA, it can be easily achieved by comparing the predictions including and not including that window information. How the ASA values themselves change due to the occurrence of a residue has not been studied in great detail. This point may be illustrated by the following example. We observed in our previous work[7] that a neural network using one-neighbor information can predict ASA with 23.7% MAE. Including one more neighbor brings prediction MAE to nearly 19.0%, suggesting a 4% improvement in prediction of ASA from the first neighbor. This result helps us to understand how ASA predictability is constrained by immediate neighbors but says nothing about how neighboring residues affect the ASA. A residue in first-neighbor position may increase or decrease the ASA, both of which would lead to better prediction results. Thus, we developed the average ASA and correlation tables (see Materials and Methods), which give a direct measure of the role played in increasing or decreasing the ASA by any of the 20 residue types at different window positions. Detailed tables so obtained are shown in Appendix A. Here, Table V shows the major conclusions drawn from these data. Most significant contributions to ASA values can be classified as hydrophobic (decreasing ASA) or hydrophilic (increasing ASA). A hydrophilic effect on a well-known hydrophilic residue could be considered cooperative and vice versa. We observe that most of the effects are cooperative (e.g., hydrophobicity of a residue is enhanced by similar neighbors). Correlation and average ASA tables show that the residues immediately after the first neighbor play very small individual roles in biasing the ASA of a target residue. However, the cumulative effect of these small contributions is quite strong, as can be seen by deleting certain neighbor information and making prediction models.

CONCLUSIONS

In this study, we used a multiple linear regression method to predict the real-value solvent accessibility. Our method has comparable performance to models using the regression of neural networks and support-vector ma-

chine. Our best models of linear regression can predict ASA with about 16.2% MAE, which is close to the accuracy values obtained by other methods of higher complexity. An analysis has been completed to explicate the role of a residue's effective frequency (in evolution) at specific locations in the determination of solvent accessibility. This analysis leads to the conclusion that the role of sequence neighbors in determining solvent accessibility falls very rapidly with distance from the target residue. Some interesting observations about the role of each neighbor have been made.

ACKNOWLEDGMENTS

This work was supported in part by the National Science Council of Taiwan.

REFERENCES

1. Rost B, Sander C. Improved prediction of protein secondary structure by using sequence profiles and neural networks. Proc Natl Acad Sci USA 1993;90:7558–7562.
2. Holbrook SR, Muskal SM, Kim SH. Predicting surface exposure of amino acids from protein sequences. Protein Eng 1990;3:659–665.
3. Rost B, Sander C. Conservation and prediction of solvent accessibility in protein families. Proteins 1994;20:216–226.
4. Fariselli P, Casadio R. RCNPRED: prediction of the residue co-ordination numbers in proteins. Bioinformatics 2001;17:202–204.
5. Pollastri G, Baldi P, Fariselli P, Casadio R. Prediction of coordination number and relative solvent accessibility in proteins. Proteins 2002;47:142–153.
6. Ahmad S, Gromiha MM. NETASA: neural network based prediction of solvent accessibility. Bioinformatics 2002;18:819–824.
7. Ahmad S, Gromiha MM, Sarai S. Real value prediction of solvent accessibility from amino acid sequence. Proteins 2003;50:629–635.
8. Adamczak R, Porollo A, Meller J. Accurate prediction of solvent accessibility using neural networks-based regression. Proteins 2004;56(4):753–767.
9. Yuan Z, Burrage K, Mattick JS. Prediction of protein solvent accessibility using support vector machines. Proteins 2002;48(3):566–570.
10. Kim H, and Park H. Prediction of protein relative solvent accessibility with support vector machines and long-range interaction 3D local descriptor. Proteins 2004;54(3):557–562.
11. Yuan Z, Huang B. Prediction of protein accessible surface areas by support vector regression. Proteins 2004;57:558–564.
12. Naderi-Manesh H, Sadeghi M, Arab S, Movahedi AA. Prediction of protein surface accessibility with information theory. Proteins 2001;42:452–459.
13. Richardson CJ, Barlow DJ. The bottom line for prediction of residue solvent accessibility. Protein Eng 1999;12:1051–1054.
14. Carugo O. Prediction residue solvent accessibility from protein sequence by considering the sequence environment. Protein Eng 2000;13:607–609.
15. Thompson MJ, Goldstein RA. Prediction solvent accessibility: higher accuracy using Bayesian statistics and optimized residue substitution classes. Proteins 1996;25:38–47.
16. Wang J-Y, Ahmad S, Gromiha MM and Sarai A. Look-up tables for protein solvent accessibility prediction and nearest neighbor effect analysis. Biopolymers 2004;75:209–216.
17. Li X, Pan X-M. New method for accurate prediction of solvent accessibility from protein sequence. Proteins 2001;42:1–5.
18. Cuff JA, Barton GJ. Application of multiple sequence alignment profiles to improve protein secondary structure prediction. Proteins 2000;40:502–511.
19. Qian N, Sejnowski T. Predicting he secondary structure of globular proteins using neuron netwok models. J Mol Biol 1988;202(4):865–884.
20. Betts MJ, Russell RB. Amino acid properties and consequences of substitutions. In: Barnes MR, Gray IC, editors. Bioinformatics for geneticists. New York: Wiley; 2003. p 297–299.

APPENDIX

TABLE VI. Effect of Occurrence of a Residue in Different Window Positions on Mean Solvent Accessibility of a Target Residue†

(a) Barton502 dataset

Window position residue	0	−1	+1	−2	+2	−3	+3	−4	+4	−5	+5
A	−6.63%	−0.22%	−0.91%	−1.15%	−2.14%	−0.56%	−2.08%	−0.49%	−1.14%	−1.33%	−1.50%
R	7.79%	−1.14%	−2.31%	−3.70%	−2.31%	−0.31%	0.29%	0.45%	0.18%	−0.74%	−1.37%
N	11.71%	−0.65%	1.10%	1.21%	−1.43%	−0.60%	0.01%	−1.57%	−0.39%	−2.13%	−2.44%
D	15.42%	−0.19%	2.02%	0.01%	−1.81%	−0.76%	0.65%	−0.58%	0.48%	−3.65%	−1.58%
C	20.87%	1.20%	0.11%	3.36%	0.77%	−0.70%	0.86%	0.57%	0.69%	2.99%	2.74%
Q	11.12%	−0.24%	1.07%	−2.35%	−1.46%	1.86%	1.62%	1.80%	1.59%	−1.81%	−1.13%
E	18.11%	0.85%	2.65%	−3.30%	−2.26%	3.05%	1.05%	3.47%	2.21%	−1.94%	−1.25%
G	2.20%	−0.21%	4.92%	−0.91%	0.93%	−4.03%	−1.54%	−3.30%	−4.05%	−3.87%	−4.32%
H	−0.75%	−4.24%	−0.52%	−3.16%	−4.05%	−2.64%	−4.17%	−3.79%	−2.50%	−3.08%	−1.21%
I	−17.95%	−3.96%	−3.91%	−1.66%	−0.69%	−3.23%	−1.42%	−2.93%	−1.14%	1.65%	1.89%
L	−16.46%	−1.53%	−2.15%	1.89%	−0.07%	−1.84%	−3.69%	−2.09%	−3.10%	1.55%	2.04%
K	17.54%	2.31%	1.60%	−1.52%	1.14%	2.85%	4.08%	2.41%	2.77%	−0.56%	−1.20%
M	−14.03%	−3.01%	−2.25%	−0.63%	−1.45%	−3.60%	−4.93%	−2.93%	−5.36%	−0.07%	−0.92%
F	−16.00%	−2.32%	−1.65%	−0.78%	0.11%	−0.73%	−2.12%	−2.49%	−2.36%	1.33%	1.92%
P	4.92%	7.04%	−3.05%	0.18%	5.87%	0.66%	5.23%	−0.11%	2.83%	−2.94%	−0.69%
S	5.35%	1.33%	0.79%	0.86%	0.91%	−0.58%	−0.28%	−0.39%	0.24%	−1.45%	−1.45%
T	1.45%	−2.52%	−1.45%	0.60%	1.75%	−1.09%	−0.09%	−1.46%	1.23%	−1.78%	−0.68%
W	−14.43%	−4.79%	−1.99%	−1.37%	−2.09%	−1.68%	−2.10%	−2.93%	−2.32%	1.06%	0.52%
Y	−9.80%	−3.87%	−1.79%	−2.77%	−1.42%	−0.90%	−2.66%	−2.00%	−1.82%	0.85%	1.79%
V	−15.66%	−4.49%	−3.36%	−2.87%	−1.59%	−3.52%	−0.49%	−2.33%	−0.23%	1.01%	2.16%

†A value *dx* in a window position *n* for residue *X* means that the mean okay of all residues (relative units) is higher by *dx* units if *X* is observed at *n* neighbor position, compared to the overall average ASA value. These values use only sequence and no PSSM information.

(b) Meller860 Dataset

Window position residue	0	−1	+1	−2	+2	−3	+3	−4	+4	−5	+5
A	−7.39%	−0.73%	−1.18%	−1.60%	−2.63%	−1.18%	−2.47%	−1.41%	−1.39%	−1.43%	−1.50%
R	8.92%	−0.71%	−1.62%	−2.29%	−1.92%	0.12%	0.54%	0.86%	0.60%	−0.41%	−1.04%
N	11.24%	−0.74%	2.27%	1.59%	−1.60%	−0.60%	0.06%	−1.29%	−0.49%	−2.31%	−2.66%
D	14.89%	−0.33%	1.99%	0.41%	−1.29%	0.14%	0.76%	−0.70%	0.81%	−3.25%	−1.75%
C	13.46%	−0.41%	−2.06%	0.89%	−0.40%	−1.19%	−2.28%	0.40%	−1.71%	1.25%	1.50%
Q	11.17%	−0.08%	0.87%	−2.58%	−1.17%	1.47%	1.97%	1.67%	1.50%	−2.12%	−1.54%
E	18.41%	1.00%	2.10%	−2.80%	−2.06%	3.21%	1.88%	3.41%	2.98%	−1.88%	−1.41%
G	1.91%	−0.28%	4.20%	−1.11%	1.07%	−3.54%	−0.97%	−3.30%	−4.04%	−3.85%	−4.43%
H	0.39%	−3.08%	−1.45%	−3.03%	−2.55%	−2.01%	−2.94%	−1.75%	−2.27%	−2.24%	−1.48%
I	−17.44%	−3.75%	−3.87%	−1.44%	−0.92%	−2.88%	−2.25%	−2.60%	−2.13%	1.29%	1.68%
L	−15.45%	−1.23%	−1.91%	1.71%	0.29%	−1.98%	−3.49%	−1.75%	−3.40%	1.85%	1.41%
K	18.20%	3.02%	2.22%	−1.02%	1.13%	3.42%	4.14%	3.24%	3.24%	0.12%	−0.91%
M	−12.49%	−2.23%	−1.48%	−0.79%	−2.22%	−3.47%	−4.54%	−3.32%	−4.23%	−1.00%	−0.73%
F	−15.34%	−3.18%	−2.63%	−0.11%	−0.47%	−1.99%	−2.77%	−2.52%	−2.00%	0.90%	1.96%
P	4.45%	6.61%	−2.71%	0.86%	5.61%	0.08%	4.56%	0.64%	2.13%	−1.27%	−0.23%
S	4.21%	1.46%	1.19%	0.68%	0.42%	−0.07%	0.08%	−0.57%	0.31%	−1.69%	−1.23%
T	0.94%	−1.66%	−1.29%	0.29%	1.53%	−0.95%	−0.41%	−1.91%	0.53%	−1.55%	−0.03%
W	−12.49%	−3.94%	−1.99%	−0.09%	−1.24%	−1.36%	−2.75%	−2.01%	−1.41%	1.17%	0.64%
Y	−9.49%	−4.05%	−1.96%	−2.80%	−1.93%	−1.70%	−3.04%	−2.47%	−1.75%	0.09%	0.70%
V	−14.87%	−3.52%	−2.76%	−2.00%	−1.11%	−2.89%	−0.83%	−2.28%	−0.54%	0.94%	1.71%

TABLE VI. Continued

(c) Yuan1277 Dataset

Window position residue	0	−1	+1	−2	+2	−3	+3	−4	+4	−5	+5
A	−7.58%	−0.90%	−0.71%	−1.69%	−2.82%	−0.93%	−2.95%	−1.31%	−1.32%	−1.59%	−1.52%
R	10.30%	−0.47%	−1.67%	−2.52%	−1.76%	0.25%	0.81%	1.09%	0.27%	−0.18%	−1.74%
N	11.07%	−1.20%	1.52%	0.51%	−1.72%	−1.41%	−0.61%	−1.79%	−1.44%	−2.75%	−2.78%
D	14.80%	−0.33%	1.86%	0.59%	−1.42%	0.19%	0.51%	−0.99%	−0.02%	−3.47%	−2.22%
C	15.11%	−0.75%	−1.47%	1.58%	−0.94%	−0.53%	−2.24%	0.27%	−0.79%	2.83%	1.14%
Q	11.70%	−0.05%	0.21%	−2.19%	−1.93%	1.53%	1.97%	1.74%	1.29%	−1.45%	−0.97%
E	19.29%	0.59%	2.81%	−2.94%	−1.56%	3.12%	1.84%	3.33%	3.12%	−1.95%	−0.74%
G	1.51%	−0.22%	5.01%	−1.62%	0.68%	−3.96%	−1.45%	−3.78%	−4.37%	−4.74%	−4.71%
H	−0.55%	−3.30%	−1.03%	−2.95%	−2.82%	−2.17%	−3.35%	−2.98%	−1.46%	−2.41%	−1.26%
I	−18.74%	−4.07%	−3.74%	−1.54%	−1.14%	−3.37%	−2.22%	−2.95%	−2.01%	1.41%	1.97%
L	−16.38%	−1.05%	−1.88%	2.29%	0.74%	−1.77%	−3.70%	−1.58%	−3.23%	1.98%	1.82%
K	18.87%	3.06%	1.87%	−0.83%	1.43%	3.32%	4.43%	3.28%	3.49%	−0.20%	−1.08%
M	−11.21%	−1.57%	0.55%	−0.39%	−1.28%	−3.23%	−5.00%	−3.15%	−4.52%	−0.71%	−1.44%
F	−16.52%	−3.26%	−3.42%	−1.00%	−0.94%	−2.57%	−3.08%	−2.98%	−2.30%	0.63%	1.79%
P	5.38%	7.37%	−3.10%	0.53%	6.37%	−0.24%	5.18%	0.11%	1.92%	−2.31%	−0.71%
S	4.77%	1.68%	0.96%	0.69%	0.06%	−0.35%	0.22%	−0.66%	0.18%	−1.47%	−1.66%
T	0.58%	−2.39%	−2.59%	0.37%	0.93%	−1.13%	−0.25%	−2.10%	0.35%	−1.80%	−0.86%
W	−14.66%	−4.96%	−2.66%	−1.56%	−2.08%	−2.40%	−3.45%	−3.01%	−1.83%	1.40%	0.85%
Y	−9.98%	−4.68%	−2.64%	−2.77%	−2.52%	−1.53%	−3.40%	−2.46%	−1.93%	−0.47%	0.63%
V	−16.29%	−4.21%	−3.39%	−2.92%	−1.25%	−3.49%	−0.71%	−2.50%	−0.56%	0.50%	1.89%

TABLE VII. Coefficients of Correlation Between the Occurrence of a Residue at a Given Position (in Aligned Sequences) and Solvent Accessibility[†]

(a) Barton502 Dataset

Window position residue	0	−1	+1	−2	+2	−3	+3	−4	+4	−5	+5
A	0.033	0.029	0.007	−0.015	−0.039	0.023	−0.014	0.026	0.008	−0.010	−0.022
R	0.327	0.070	0.048	0.002	0.020	0.082	0.092	0.091	0.083	0.001	−0.020
N	0.394	0.072	0.109	0.048	0.027	0.061	0.063	0.040	0.047	−0.046	−0.049
D	0.444	0.089	0.118	0.043	0.028	0.085	0.085	0.065	0.079	−0.048	−0.028
C	−0.196	−0.009	−0.029	0.021	0.006	−0.017	−0.021	−0.003	−0.020	0.036	0.025
Q	0.395	0.076	0.087	0.003	0.024	0.104	0.100	0.103	0.095	−0.016	−0.015
E	0.465	0.094	0.112	0.008	0.023	0.122	0.104	0.119	0.111	−0.023	−0.014
G	0.210	0.063	0.101	0.018	0.035	−0.003	0.028	−0.002	−0.012	−0.053	−0.063
H	0.201	0.023	0.062	0.002	0.008	0.052	0.026	0.033	0.036	−0.019	−0.009
I	−0.338	−0.042	−0.054	0.012	0.013	−0.024	−0.017	−0.014	−0.003	0.077	0.074
L	−0.329	−0.016	−0.040	0.043	0.015	−0.002	−0.037	−0.001	−0.027	0.079	0.069
K	0.456	0.121	0.096	0.027	0.061	0.115	0.140	0.117	0.120	−0.002	−0.012
M	−0.261	−0.035	−0.047	0.013	−0.007	−0.015	−0.046	−0.011	−0.036	0.052	0.038
F	−0.285	−0.037	−0.019	0.018	0.012	−0.002	−0.034	−0.013	−0.022	0.059	0.061
P	0.242	0.148	0.069	0.058	0.110	0.050	0.120	0.042	0.082	−0.019	0.013
S	0.279	0.075	0.050	0.041	0.019	0.045	0.039	0.040	0.045	−0.033	−0.036
T	0.123	0.013	0.012	0.039	0.040	0.030	0.034	0.015	0.055	−0.006	0.005
W	−0.135	−0.018	0.008	0.017	0.011	0.015	−0.005	0.010	0.001	0.043	0.030
Y	−0.148	−0.028	0.003	0.003	0.009	0.020	−0.016	0.011	0.001	0.046	0.046
V	−0.318	−0.047	−0.056	−0.003	0.006	−0.027	−0.005	−0.012	0.003	0.067	0.071
Indel	−0.222	−0.091	−0.095	−0.054	−0.057	−0.094	−0.089	−0.090	−0.091	−0.027	−0.025
Entropy	−0.065	−0.071	−0.070	−0.073	−0.072	−0.072	−0.071	−0.071	−0.071	−0.071	−0.071

[†]This information is derived from PSSM tables obtained by PSI BLAST.

(b) Meller860 Dataset

Window position residue	0	−1	+1	−2	+2	−3	+3	−4	+4	−5	+5
A	0.015	0.019	0.004	−0.021	−0.040	0.019	−0.015	0.016	0.007	−0.016	−0.022
R	0.327	0.074	0.056	0.010	0.021	0.088	0.100	0.099	0.095	0.010	−0.009
N	0.378	0.068	0.109	0.047	0.029	0.066	0.067	0.047	0.056	−0.037	−0.039
D	0.424	0.083	0.119	0.042	0.030	0.091	0.088	0.068	0.087	−0.042	−0.025
C	−0.206	−0.016	−0.041	0.007	−0.001	−0.022	−0.032	−0.013	−0.034	0.026	0.022
Q	0.382	0.075	0.089	0.004	0.022	0.110	0.104	0.106	0.103	−0.007	−0.010
E	0.454	0.094	0.113	0.010	0.026	0.128	0.110	0.121	0.123	−0.015	−0.005
G	0.192	0.054	0.091	0.014	0.031	0.003	0.031	−0.002	−0.010	−0.050	−0.058
H	0.201	0.022	0.062	0.001	0.010	0.051	0.029	0.040	0.042	−0.015	−0.008
I	−0.332	−0.038	−0.051	0.013	0.014	−0.027	−0.022	−0.018	−0.011	0.068	0.073
L	−0.322	−0.016	−0.039	0.041	0.017	−0.010	−0.038	−0.008	−0.030	0.074	0.066
K	0.449	0.123	0.105	0.031	0.057	0.124	0.144	0.126	0.130	0.007	−0.002
M	−0.253	−0.034	−0.047	0.008	−0.006	−0.023	−0.048	−0.018	−0.041	0.042	0.034
F	−0.274	−0.037	−0.024	0.019	0.011	−0.012	−0.035	−0.018	−0.023	0.052	0.055
P	0.227	0.139	0.073	0.061	0.109	0.049	0.116	0.048	0.079	−0.007	0.021
S	0.261	0.074	0.053	0.036	0.020	0.050	0.045	0.039	0.047	−0.025	−0.029
T	0.106	0.017	0.018	0.032	0.036	0.031	0.034	0.012	0.052	−0.004	0.013
W	−0.129	−0.021	0.002	0.017	0.012	0.010	−0.009	0.009	0.003	0.037	0.034
Y	−0.142	−0.030	0.000	0.003	0.007	0.009	−0.018	0.003	0.002	0.035	0.042
V	−0.311	−0.044	−0.052	−0.002	0.009	−0.030	−0.010	−0.017	−0.001	0.060	0.071
Indel	−0.210	−0.089	−0.098	−0.053	−0.057	−0.098	−0.095	−0.089	−0.098	−0.028	−0.033
Entropy	−0.062	−0.069	−0.068	−0.071	−0.072	−0.070	−0.072	−0.069	−0.072	−0.071	−0.074

TABLE VII. Continued

(c) Yuan1277 Dataset

Window position residue	0	−1	+1	−2	+2	−3	+3	−4	+4	−5	+5
A	0.011	0.010	−0.002	−0.029	−0.052	0.015	−0.028	0.013	−0.003	−0.022	−0.032
R	0.326	0.068	0.049	0.007	0.023	0.084	0.095	0.097	0.092	0.009	−0.011
N	0.372	0.063	0.101	0.043	0.023	0.059	0.060	0.038	0.049	−0.043	−0.043
D	0.419	0.080	0.111	0.041	0.025	0.083	0.082	0.059	0.076	−0.046	−0.027
C	−0.206	−0.020	−0.037	0.010	−0.004	−0.020	−0.029	−0.009	−0.024	0.031	0.023
Q	0.386	0.071	0.085	0.008	0.024	0.107	0.102	0.109	0.102	−0.005	−0.004
E	0.456	0.091	0.111	0.012	0.025	0.126	0.108	0.121	0.120	−0.016	−0.004
G	0.186	0.048	0.091	0.008	0.024	−0.005	0.023	−0.012	−0.019	−0.058	−0.064
H	0.185	0.011	0.048	−0.005	0.000	0.041	0.016	0.028	0.031	−0.022	−0.012
I	−0.343	−0.045	−0.060	0.007	0.009	−0.032	−0.024	−0.019	−0.009	0.066	0.068
L	−0.328	−0.018	−0.042	0.039	0.017	−0.010	−0.040	−0.005	−0.025	0.075	0.064
K	0.450	0.121	0.102	0.031	0.062	0.123	0.142	0.128	0.129	0.011	−0.001
M	−0.254	−0.037	−0.048	0.008	−0.004	−0.022	−0.049	−0.016	−0.036	0.044	0.035
F	−0.288	−0.045	−0.034	0.013	0.005	−0.015	−0.041	−0.022	−0.023	0.051	0.053
P	0.228	0.142	0.070	0.058	0.110	0.050	0.116	0.047	0.075	−0.012	0.014
S	0.269	0.072	0.051	0.040	0.017	0.052	0.041	0.041	0.048	−0.025	−0.031
T	0.106	0.004	0.002	0.028	0.031	0.026	0.031	0.011	0.052	−0.009	0.009
W	−0.143	−0.032	−0.008	0.005	0.002	0.003	−0.021	−0.001	−0.004	0.029	0.031
Y	−0.155	−0.040	−0.010	−0.005	−0.002	0.004	−0.025	−0.004	−0.002	0.030	0.040
V	−0.322	−0.053	−0.063	−0.010	0.001	−0.036	−0.013	−0.019	0.001	0.057	0.068
Indel	−0.191	−0.068	−0.075	−0.039	−0.041	−0.082	−0.075	−0.078	−0.086	−0.017	−0.020
Entropy	−0.129	−0.136	−0.135	−0.137	−0.138	−0.136	−0.137	−0.134	−0.136	−0.133	−0.136