Theor Chem Acc (1999) 101:186–193
DOI 10.1007/s00214980m126

Theoretical
Chemistry Accounts
© Springer-Verlag 1999

Regular article

Prediction of solvent accessibility of amino acid residues: critical aspects*

M.H. Mucchielli-Giorgi, P. Tufféry, S. Hazout

Equipe de Bioinformatique Moleculaire, INSERM U155, Université Paris 7, case 7113, 2 place Jussieu, F-75251 Paris Cedex 05, France

Received: 4 June 1998 / Accepted: 17 September 1998 / Published online: 10 December 1998

Abstract. For predicting solvent accessibility from the sequence of amino acids in proteins, we use a logistic function trained on a non-redundant protein database. Using a principal component analysis, we find that the prediction can be considered, in a good approximation, as a monofactorial problem: a crossed effect of the burial propensity of amino acids and of their locations at positions flanking the amino acid of interest. Complementary effects depend on the presence of certain amino acids (mostly P, G and C) at given positions. We have refined the predictive model (1) by adding supplementary input data, (2) by using a strategy of prediction correction and (3) by adapting the decision rules according to the amino acid type. We obtain a best score of 77.6% correct prediction for a relative accessibility of 9% . However, compared to trivial strategy only based upon the frequencies of buried or exposed residues, the gain is less than 4%.

Key words: Solvent accessibility – Logistic function – Hydrophobicity – Burial index – Amino acids

1 Introduction

Since its introduction by Lee and Richards [1], the concept of solvent accessibility of protein amino acids has been employed in various contexts such as the identification of residues implied in protein function [2–4], targeting for site-directed mutagenesis [5] or the assessment of the correctness of determined protein structures [6–8]. The prediction of solvent accessibility of proteins from their sequences can also be expected to lead to structural prediction. Different approaches to predict solvent accessibility have been proposed: Holbrook and co-workers [9] trained a neural network starting from a set of 20 non-homologous protein structures, considering sequence windows of 11 residues to predict a two-state accessibility of the residues. Rost and Sander [10] employed a similar approach and showed that starting from multiple alignments of homologous sequences can improve the prediction. Thompson and Goldstein [11] used a strategy combining Bayesian statistics and multiple alignment within families of structural proteins under the form of residue substitution classes.

A first goal of the present study is to analyse what determinants are likely to contribute to solvent accessibility prediction as a starting point to refine the prediction. We analyse the contributions provided by the amino acids located at the different positions in a window surrounding a given residue. To achieve this, we used a logistic function as a model of prediction for binary categories (buried and exposed). This model uses a reduced number of parameters compared to the previous studies. We treat the estimated parameters of the logistic function by a principal component analysis [12] for extracting the factors controlling the accessibility prediction. We then assess the dependencies between these factors and the usual physicochemical properties of amino acids.

We then define, from the previous results, different strategies to improve the prediction of the accessibility:

1. By adding complementary informations relative to the sequence (such as the relative protein size and the relative amino acid frequencies).
2. By introducing a "prediction-correction" process.
3. By using a thresholding adapted to each type of amino acid.

Finally, several critical aspects are addressed: (1) comparison of the performance of the prediction compared to trivial approaches, and (2) distinction between the prediction of buried or exposed residues.

2 Materials and methods

2.1 Database of protein structures

We have selected 342 non-homologous protein structures of a non-redundant database [13, 14], a subset of

[*] Contribution to the Proceedings of Computational Chemistry and the Living World, 20–24 April 1998, Chambery, France

Correspondence to: M.H. Mucchielli-Giorgi the Protein Data Bank [15]. The selection was restricted to monomeric, single domain proteins since it has been shown that the residues on the surface of protein subunits or domains have different amino acid distributions than residues exposed to solvent in monomeric proteins [16]. A subset of 228 proteins was defined for the training phase. The remaining 114 proteins were used for the prediction assessing phase.

2.2 Residue solvent accessibility

Solvent accessibilities were calculated with the DSSP program [17]. Buried residues (labelled *bur*) are defined as those exhibiting a relative accessible surface less than a user-defined fraction (S) of a standard state exposure; accessible residues (labelled *exp*) correspond to non-buried residues. We considered S values of 9% , 16% and 25% for comparison with the results of previous studies.

For evaluating the prediction accuracy, we use two conventional measures. The coefficient Q_2 corresponds to the percentage of correctly predicted residues in two states (*bur*, *exp*). The Matthews coefficient C_M [18] corresponds to the Pearson correlation between the occurrences of the predicted and observed states.

2.3 A simple model for prediction: the logistic function

The method for predicting solvent accessibilities of amino acid residues is based on the use of a logistic function which is equivalent to a perceptron, i.e. a neural network without hidden layer.

In our case, the basic input of the logistic function is a boolean matrix $\mathbf{X}[X_{ij}]$ where locations within the window (or subsequence) of $2N+1$ residues centred around the amino acid of interest are indexed by i $(i = -N, \ldots, +N)$. The amino acids types are indexed by j $(j = 1, \ldots, 20)$. X_{ij} is 1 if an amino acid j is in position i, else X_{ij} is 0 (see Fig. 1). This information was conventionally used in the previous studies. We have used a window size of 9 (i.e. $N = 4$), since this size of window was found optimal (not shown).

The probability $P(\mathbf{X})$ that a given residue is buried is evaluated as:

$$P(\mathbf{X}) = 1 \Big/ \left(1 + exp\left(- w_0 - \sum_i \sum_j w_{ij} X_{ij}\right)\right)$$

where w_{ij} is the weight of the amino acid j located at position i of the window. Weights w_{ij} are assumed independent of the location of the central residue in the sequence; hence the prediction is only dependent on the flanking sequence of the residue of interest. The matrix of weights $\mathbf{W}[w_{ij}]$ is estimated by the maximum likelihood method, using the S-Plus software [19].

The decision rule is the following: for each residue, if its probability to be buried $P(\mathbf{X})$ is higher than a threshold P_s, the residue is predicted buried (i.e. $D = bur$); otherwise it is predicted to be exposed (i.e. $D = exp$). When the value P_s is fixed, the probability of errors R associated with the prediction of all residues of the database is expressed as:

$$\begin{aligned} R = {} & P(D = bur/E = exp) \times P(E = exp) \\ & + P(D = exp/E = bur) \times P(E = bur) \end{aligned}$$

The labels D and E denote the decision and the reality, respectively. The optimal decision rule is for the P_s value which minimizes the probability of error ($R_{\min}$). The percentage of correctly predicted accessibilities Q_2 is equal to $1 - R_{\min}$. The search for the optimal decision rule is shown in Fig. 2.

The percentage of correct prediction Q_2 is at least equal to the largest proportions between buried $P(bur)$

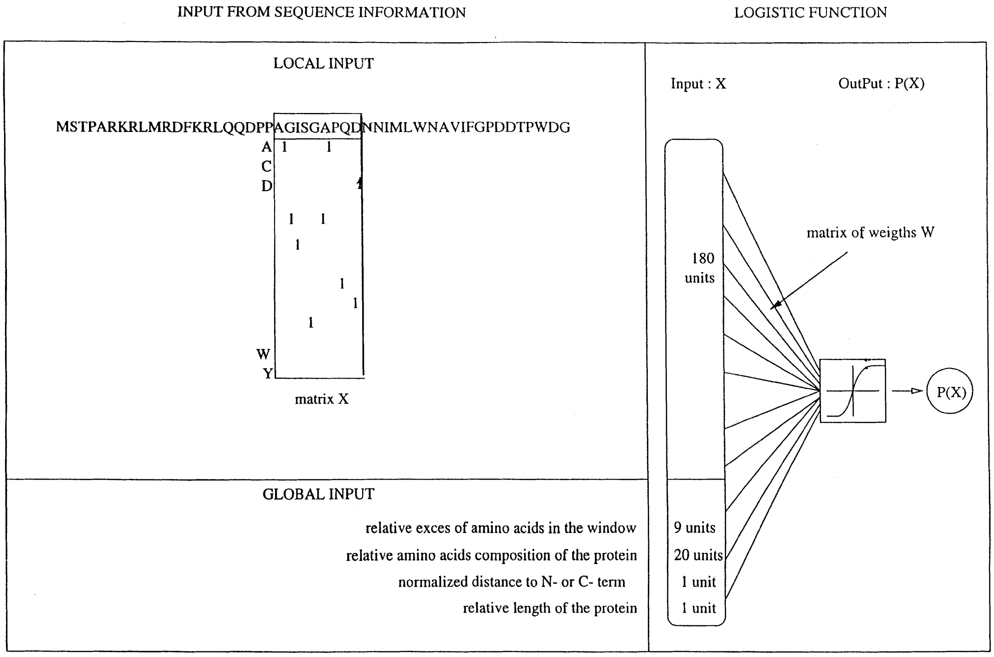
INPUT FROM SEQUENCE INFORMATION
LOGISTIC FUNCTION
LOCAL INPUT
MSTPARKRLMRDFKRLQQDPPAGISGAPQDNNIMLWNAVIFGPDDTPWDG
A
C
D
W
Y
1
matrix X
GLOBAL INPUT
relative exces of amino acids in the window
relative amino acids composition of the protein
normalized distance to N- or C- term
relative length of the protein
Input : X
OutPut : P(X)
180 units
9 units
20 units
1 unit
1 unit
matrix of weigths W
P(X)

Fig. 1. Prediction model of accessibility. *Local input*: the local input is a Boolean matrix giving the positions of the amino acids in a window of nine residues for centred on the residue for which the accessibility is predicted. In this example, we want to predict the accessibility of the residue G. *Global input*: the different global input data added to the local input in the prediction model are summarized (Sect. 2.3.1). *Logistic function*: this block of the figure indicates the number of parameters taken into account in the logistic function, the matrix of weights estimated by the logistic function and the probability to be buried, $P(\mathbf{X})$, given by the logistic function

and exposed residues $P(exp)$: $Q_2 \geq \max[P(exp), P(bur)]$. The simplest strategy, called "strategy of order 0", consists of predicting all the residues as buried if $P(bur) > 0.5$ or as exposed in the opposite case. Then, the percentage of correct prediction Q_2 is equal to $\max[P(bur), P(exp)]$.

2.3.1 Additional input

The input Boolean matrix was supplemented by:

1. The length of the protein sequence, since the percentage of buried residues is an increasing function of the length. For example, for a relative accessibility equal to 25% , the proportion of buried residues grows from 25% to 37% for protein lengths less than 190 residues and is slightly increasing (proportion close to 48% for protein lengths more than 190 residues). The relative length of the protein was introduced as: $ln(L_{\text{prot}}/L_{\text{mean}})$, where L_{prot} and L_{mean} are respectively the length of the protein and the mean length of proteins in the database.
2. The distance to the C- and N-terminal ends of the sequence. This parameter has been considered by different authors [9, 10]. In the present study, it was expressed as: $d_k(1 - d_k)$ with $d_k = (k - 1)/(L_{\text{prot}} - 1)$ where k is the residue location in the sequence.
3. The relative frequencies of the 20 amino acids in a given protein compared to those issued from the whole database. It was expressed as: $ln(F^j_{\text{prot}}/F^j_{\text{base}})$ where F^j_{prot} and F^j_{base} are respectively the frequency of the amino acid j in the protein and its mean frequency in the database.

2.3.2 "Prediction-correction" strategy

We have considered a "prediction-correction" learning. In this case, two logistic functions are used, and the probabilities estimated by the first logistic function are reintroduced as complementary inputs of a second logistic function. This leads to eight supplementary parameters corresponding to the differences between the estimated burial probabilities of residues of the window and of the central position.

Using such a procedure could be relevant since the estimated accessibilities of the residues neighbouring the residue of interest should interfere with the estimation of the accessibility of this central residue.

2.3.3 "Adapted" thresholding strategy

As some amino acids have a high propensity to be exposed or buried, the accessibility previously defined must be modulated by the amino acid type. Thus, we have considered the opportunity of using a thresholding adapted to each amino acid type. In such a case, 20 threshold values $P_s(i)$, $i = 1, \ldots, 20$, must be defined instead of the same value for the whole of the amino acids.

The thresholds are obtained by minimization of the probability of error for every amino acid type. The minimal global probability of error is the weighted mean of the individual probabilities of error:

$$R_{\min} = \sum_{j=1}^{20} q_j R^j_{\min}$$

where q_j and $R^j_{\min}$ are respectively the frequency of the amino acid type j and the corresponding minimal probability of error.

The proportion of residues correctly predicted Q_2 is $1 - R_{\min}$. The percentage Q_2 is higher or equal to the mean of the probabilities of error computed for the strategies of order 0, i.e.

$$Q_2 \geq \sum_{j=1}^{20} q_j \ \max[P_j(bur), \ P_j(exp)]$$

where $P_j(exp) = 1 - P_j(bur)$ and where $P_j(exp)$ and $P_j(bur)$ are the proportions of the exposed and buried amino acids j, respectively. The simple strategy of order

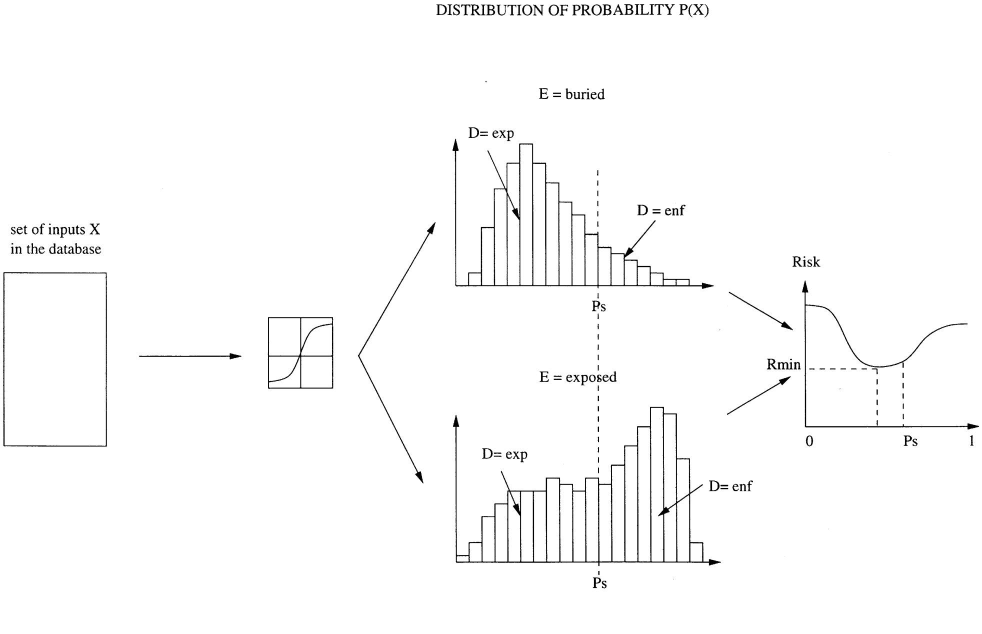

Fig. 2. Determination of the optimal cutoff P_s associated with a minimum probability of error $R_{\min}$. The values of the burial probability $P(\mathbf{X})$ are calculated by the logistic model for a set of residues. From the set of probabilities, we build two distributions of $P(\mathbf{X})$ according the truly residue state ($E =$ buried or $E =$ exposed). For each value of cutoff P_s of the decision rule, the probability of error is computed and the optimal value of P_s minimizing the probability of error is deduced

1 consists of predicting a residue buried when this amino acid is more often buried than exposed in the database, and reciprocally. In this case, Q_2 takes the minimum value of the previous formula.

2.4 A mean for extracting the most informative factors in the prediction: the principal components analysis

The simplest model of accessibility prediction corresponds to the independence between the weights attributed to amino acids and those attributed to the positions within the window. In this case, the matrix of weights $\mathbf{W}$ issued from the logistic model is deduced from a multiplicative effect of amino acid type and of window position and can be written as $\mathbf{W} = \mathbf{A}^1 \cdot {}^t\mathbf{F}^1$, where $\mathbf{A}^1[A_i^1]$ and $\mathbf{F}^1[F_j^1]$ are respectively the vector of weights associated with each position of the window and the contributions of the amino acid type to the burial. ${}^t\mathbf{F}$ denotes the transposed vector of $\mathbf{F}$. Such a matrix is of rank 1.

As the highest rank of $\mathbf{W}$ is the lower dimension of the matrix, i.e. 9, the general model which decomposes linearly the matrix $\mathbf{W}$ consists of adding 9 matrices of rank 1. So, $\mathbf{W} = \sum_{k=1}^{9} \mathbf{A}^k \cdot {}^t\mathbf{F}^k$ or $w_{ij} = \sum_{k=1}^{9} A_i^k F_j^k$. If we want the factors $\mathbf{F}^1, \mathbf{F}^2, \ldots, \mathbf{F}^9$ to bring a decreasing contribution in this order, the solution is to perform a principal component analysis where the factors $\mathbf{F}^k$ are the principal components and the weights $\mathbf{A}^k$ are the loadings. Each principal component, $\mathbf{F}^k$ takes into account a part of the variability of $\mathbf{W}$. Each variability is interpreted as the contribution of this component to the burial prediction.

The relationships between principal components and physicochemical properties of amino acids are assessed by correlations. We have considered the hydrophobicity coefficient [20], the propensity to be located in a coil, in a strand or in a helix [21], the flexibility [22], the accessibility coefficient [23], the residue burial [24], the accessible surface [25], the size [26], the polarity [26], the charge (encoded in three classes: negatively charged by -1, positively charged by $+1$ and the others by 0), the radius (radius of a sphere including at better a set of conformations of the side chain) and the "propensity to be in a buried contact".

We obtain this last property by determining contacts between side-chains in the bank of protein structures, and by counting it in two different matrices of occurrence (dimension 20×20): one matrix with contacts between two buried residues and one matrix with the other contacts. We have fitted within S-Plus software [27] a log-linear model [28] in order to estimate the joint effect between the amino acid type and its accessibility, called the "propensity to be in a buried contact".

3 Results and discussion

The main aspects tackled in the paper are:

1. To extract and interpret the factors controlling the burial of the central residue, associated with the type of amino acids and the locations in the window of analysis.
2. To assess the gains of the prediction accuracy provided by the different refinements of the initial model, and to discuss the improvements relative to the trivial models of order 0 or 1. The results will be given for three relative solvent accessibility cutoffs of 9%, 16% and 25%.

3.1 The solvent accessibility prediction appears to be mainly a mono factorial problem

The logistic function was trained using the basic Boolean input information (see Methods). The matrix of weights $\mathbf{W}$ of the logistic function was then decomposed by a principal component analysis. For a relative solvent accessibility of 25%, the first principal component explains the major part of the variability (89.1%) (respectively 89.5% and 90.5% for 16% and 9%) and the five first principal components 98.6% (98.2 for 16% and 9%). The components 2, 3, 4 and 5 correspond to 3.93%, 2.64%, 1.67% and 1.21% of the variability, respectively. Since the major part of the variability is explained by only one component ($\mathbf{W} \simeq \mathbf{A}^1 \cdot {}^t\mathbf{F}^1$), the model is clearly over-parametrized. This can be confirmed by performing the learning test procedure for restricted models. The logistic function based only on the first principal component gives 67.7% correct predictions for a relative solvent accessibility of 25% (respectively 70.5% and 75.2% for 16% and 9%), and 68.6% with the first five components (respectively 70.9% and 75.2%), values compared with 68.7% for the complete model (model with nine principal components) (respectively 71.2% and 75.8%). Thus for a model reduced from 180 to 45 parameters the accuracy loss is 0.1% (respectively 0.3% and 0.6%), and the loss is only 1% for the monofactorial model (9 parameters) (respectively 0.7% and 0.6%).

3.2 New indexes of amino acid burial

We have checked the correlations between these principal components and common physicochemical properties. No significant correlations could be derived for the principal components 2–5. Only 5 among the 13 amino acids' properties appear correlated to the first component. Two are negatively correlated: the polarity with a correlation of -0.737 ($p < 2 \times 10^{-4}$) and the accessibility coefficient with a correlation of -0.684 ($p < 9 \times 10^{-4}$). This is not surprising since these are measures of how residues might be exposed. Three are positively correlated: the tendency to be located in a β strand with a correlation of 0.789 ($p < 10^{-4}$), the hydrophobicity with a correlation of 0.775 ($p < 10^{-4}$) and the "tendency to be in a buried contact" with a correlation of 0.872 ($p < 10^{-8}$). This last property is the only one that exhibits a significant partial correlation [29] (i.e. the correlation with the first component is maintained when the other properties are fixed). This implies that this property explains some part of the variability which is not taken into account by the other properties.

Thus, coefficients associated with component 1 and the "tendency to be in a buried contact" (Table 1) can be considered as burial indexes. Even if the hydrophobic and hydrophilic tendencies are conserved, some differences can be noted between these scales and the scales of accessibility [23] and of burial [24]. The major differences concern tryptophan (W) and tyrosine (Y), which are more buried in our two indexes, and cysteine (C), which is less buried.

3.3 The effect of the neighbours of a residue on its burial prediction

3.3.1 Equivalent contributions of the central residue and its neighbours

Tables 1 and 2 list for a relative accessibility cutoff of 25% the three first principal components and the corresponding loadings. The contribution of an amino acid i in the position j is obtained by combining the results of Tables 1 and 2: the burial (or the accessibility) of the central residue is favoured when the product of the coefficients of i in Table 1 and of j in Table 2 is positive (or negative) for a given order k ($k = 1, 2, 3$) in the tables.

For the component 1, all residues except alanine (A) have a non-negligible effect. The loadings associated with the first component show that the central residue has the major effect, which is equivalent in magnitude to the effect of all other residues of the window (see Table 2). The hydrophobic and non-polar residues (or polar and charged) in positions ±1, ±3 and ±4 favour the burial (or the accessibility) of the central residue. One notes that such positions are compatible with what one could expect for an α-helix: when a residue is located on the buried side of an α-helix, then the residues located at these positions are not accessible, and conversely. Similar profiles were derived for 16% and 9% (not shown). However, some effects related to particular residues are accentuated: when the relative accessibility cutoffs decrease, the burial is more favoured by the presence of an isoleucine (I) (its coefficient is 1.137, 1.275 and 1.38 respectively for the relative accessibility cutoffs of 25%, 16% and 9%) and inversely for lysine (K) (−1.839, −2.119 and −2.346). Also, the contributions to the accessibility of proline (P) and threonine (T) (coefficients are respectively −0.804, −0.641 and −0.589 for P and −0.442, −0.4 and −0.232 for T) are lowered.

Table 1. Contributions of amino acids on the burial of the central residue[a]

Amino acid	Comp 1	Comp 2	Comp 3	Ind
F	1.440	−0.010	−0.117	0.405
W	1.245	0.230	0.021	0.322
L	1.187	−0.027	0.168	0.412
I	1.137	0.071	0.210	0.451
M	0.985	−0.049	0.058	0.372
C	0.983	−0.186	−0.574	0.307
Y	0.924	0.008	0.055	0.131
V	0.854	0.023	0.032	0.336
H	0.152	−0.208	0.162	0.049
A	0.019	0.117	−0.080	0.215
G	−0.350	−0.492	−0.006	0.223
T	−0.442	0.109	0.124	−0.083
S	−0.538	0.042	−0.039	−0.041
R	−0.730	0.100	0.232	−0.478
N	−0.803	−0.247	0.124	−0.274
P	−0.804	0.587	−0.139	−0.151
Q	−0.984	0.152	−0.079	−0.416
D	−1.136	−0.104	0.001	−0.342
E	−1.299	−0.042	−0.114	−0.511
K	−1.839	−0.074	−0.040	−0.916

[a] Only the three first principal components Comp 1, Comp 2 and Comp 3 (Sect. 2.4) and the index of the tendency to be in a buried contact (Ind) (Sect. 2.4) are given. Residues with positive (or negative) coefficients favour the burial (or the accessibility) of the central residue

Table 2. Effects of the positions of the flanking sequence of the central residue on its burial[a]

Position	Comp 1	Comp 2	Comp 3
−4	0.145	−0.129	0.150
−3	0.133	−0.220	0.270
−2	0.060	0.046	0.018
−1	0.167	−0.513	0.225
0	0.936	0.119	−0.295
+1	0.075	0.766	0.380
+2	0.092	−0.012	0.147
+3	0.162	−0.258	0.347
+4	0.118	0.041	0.692

[a] Positions with positive (or negative) coefficients favour the burial (or the accessibility) of the central residue

3.3.2 The rectifying effects of accessibility prediction are dependent on certain amino acids located at fixed positions relative to the central residue

Components 2 and 3 correspond to rectifying effects compared to component 1, since the effect of the central position appears much lower than those of the neighbouring positions (Table 2), and a smaller subset of amino acid types is involved compared to component 1 (respectively 11 and 10 amino acid types) (Table 1).

For the second component, glycine (G) and proline (P) have an effect roughly three times stronger than the other amino acids, and it is mostly the positions + 1 and −1 that exhibit strong effects. Glycine (or proline) at position −1 (or + 1) favours the burial (or the accessibility) of the central residue. Tryptophan (W), histidine (H) or asparagine (N) contribute to a lesser extent.

For the third principal component, the effect of the central position is unfavourable for burial. This is compensated by the effects associated with the other positions. Mostly, cysteines (C) contribute to this component. They show a tendency to be buried in the central position, while at other positions they favour the accessibility of the central residue.

Such results are consistent with the analysis performed by Holbrook and co-workers. In their study, the authors examined the weights matrix **W** and observed that the primary factor governing exposure of the residues V, W, M and E is the identity of the central residue itself. For residues P and G, they observed that the flanking sequence is more influential. They also observed that hydrophobic residues at position −3, −2 and +2, +3 favour the burial. In the present study, we have carried out a quantitative analysis of the contributions of the amino acid types associated with their positions in the window. The present results show some differences in the hydrophobic residues having the highest effect on the accessibility in the central position (i.e. F, W, L and I). We observe too a rectifying role of cysteine (C) which attenuates the effect of the central residue.

3.4 Refinements of the prediction

3.4.1 A maximum gain of 2% in the prediction is obtained with the default strategy compared to the trivial model

Table 3 gives the percentages of correct prediction Q_2 and the Matthews coefficients C_M for each model. For a relative accessibility of 25%, the simple strategy of order 0, predicting all the residues exposed, gives a correct prediction of 51.9%, equal to the frequency of the exposed residues in the database and a null correlation coefficient since the whole of the buried residues are incorrectly predicted. The simple strategy of order 1, predicting the residue buried (or exposed) when it is more often buried than exposed (or exposed than buried) in the database, gives a correct prediction of 66.8% and a C_M of 0.336. The results obtained with a such strategy are rather good because a few amino acids (A, G, H, S and T) have no particular tendency to be buried or exposed while the 15 other amino acids show stronger tendencies for accessibility or burial (more than 60%). Compared to these results, the model with the nine principal components gave a prediction accuracy of 68.7% with a C_M of 0.372, i.e. a gain of only 2% compared to the strategy of order 1. The better C_M value reflects the fact that the percentage of buried residues correctly predicted increases from 56.3% to 63.1% between the strategy of order 1 and the model with the nine principal components, while the percentage of exposed residues correctly predicted only decreases from 76.5% to 73.9%.

Similarly, the gains observed for relative accessibilities of 16% and 9% are of 1% and 1.5%, respectively. One remarks that the percentage of correct predictions increases when the relative accessibility cutoff decreases, and conversely for the correlation coefficient. Consequently, the prediction for the buried residues is less and less correct, however, as the frequency of buried residues in the database decreases, Q_2, increases.

3.4.2 Further refinements contribute to a maximum gain of 2.1% in the prediction

The addition of information relative to the protein (length and composition in amino acids of the protein, distance to the C-and N-terminal ends) leads to a gain of 0.8% for Q_2 and a C_M of 0.386, for relative accessibility of 25%. Using the prediction-correction strategy, the further gain in the percentage of correct prediction is 0.4% compared to the previous model and C_M increases to 0.396. Finally, using a decision rule dependent on the type of the central amino acid, the gain in the percentage of correct prediction is 0.8% and C_M increases to 0.417. For relative accessibilities of 16% and 9%, similar enhancements are obtained, with final gains of 2.1% and 1.7%, respectively.

Each of the additional refinements to the model with nine principal components allows a small gain. The effect of the prediction correction is slight. The best enhancements come from using a thresholding dependent on the type of amino acid. As shown in Table 4, the prediction is better for some amino acids. In fact, hydrophobic and charged (D, E, K) amino acids are the best predicted, with a percentage of correct prediction higher than 70%. The worse predictions are obtained for G, H, P, S and T, with a prediction accuracy lower than 65%. However, compared to the strategy of order 1, the prediction is much increased for some amino acids such as A, G, H, S, T and Y (gain more than 5%) and slightly for the other ones.

Table 3. Assessment of prediction accuracy for the different models

S	Model[a]	Q_2^b (%)	C_M^c
25%	Strategy of order 0	51.9	0
	Strategy of order 1	66.8	0.336
	Nine principal components	68.7	0.372
	+ length of the protein + composition in amino acid	69.5	0.386
	+ prediction correction	69.9	0.396
	+ thresholding by amino acid	70.7	0.417
16%	Strategy of order 0	63.7	0
	Strategy of order 1	70.2	0.331
	Nine principal components	71.2	0.355
	+ length of the protein + composition in amino acid	71.7	0.372
	+ prediction correction	71.7	0.369
	+ thresholding by amino acid	73.3	0.386
9%	Strategy of order 0	73.8	0
	Strategy of order 1	74.3	0.184
	Nine principal components	75.8	0.262
	+ length of the protein + composition in amino acid	76.4	0.302
	+ prediction correction	76.5	0.313
	+ thresholding by amino acid	77.6	0.339

[a] The different models are described in the text (Sect. 2.3)
[b] Percentage of correct prediction evaluated for different relative solvent accessibility cutoffs *S*
[c] Correlation coefficient of Matthews [18]

Table 4. Prediction accuracy by amino acid

Amino acid	Q_2[a] (%)	Δ[b] (%)	$P(D = bur/E = bur)$[c] (%)	$P(D = exp/E = exp)$[d] (%)
F	80.6	0.0	100	0
W	77.8	0.3	99.2	4.2
L	75.3	0.4	99.4	3.5
I	73.7	0.0	100	0
M	76.7	3.2	100	12.1
C	75.5	0.0	99.5	1.5
Y	82.6	10.9	87.8	69.5
V	70.1	1.0	98.5	6.6
H	60.6	5.1	60.4	60.8
A	65.3	12.2	68.9	61.1
G	65.4	9.7	58.9	70.6
T	65.5	8.2	44.3	81.3
S	64.7	6.3	46.7	77.7
R	66.6	3.0	15.5	95.8
N	68.2	2.8	20.6	93.4
P	63.7	1.9	13.4	94.7
Q	67.6	0.9	8.7	97.0
D	70.9	1.2	11.6	96.7
E	75.2	1.9	19.7	95.3
K	82.4	0.4	6.9	98.9

[a] Percentage of correct prediction (Sect. 2.3) evaluated by amino acid for a relative accessibility cutoff *S* of 25%

[b] Δ represents the gain in the prediction accuracy compared to the strategy of order 1 (Sect. 2.3.3). $\Delta = Q_2(\text{after refinements}) - Q_2(\text{strategy of order 1})$

[c] Percentage of correct prediction of truly buried residues

[d] Percentage of correct prediction of truly exposed residues

Although at 25% the refinements lead to a gain of only 1.6% in the percentage of correct prediction, the largest improvement lies in a better prediction of the exposed residues. In fact, the percentage of exposed residues correctly predicted increases from 73.9% to 76.4%, while that of buried residues correctly predicted increases from 63.1% to 64.4%.

These results, obtained with a simple model (one logistic function), are similar to those of previous studies. Compared to the results without multiple sequence alignment given in previous works, the prediction of the accessibility is slightly improved. For a relative accessibility cutoff of 20% the results are equivalent to the previous works: Holbrook et al. [9] obtained a Q_2 of 72% and Thompson and Goldstein [11] a Q_2 of 72.3%. Our values obtained for 25% and 16% (Q_2 of 70.7% and 73.3%) bracket these values. For 16% and 9%, Rost and Sander [10] obtained Q_2 values of 71.1% and 72.8%, respectively, while we obtain 73.3% and 77.6%. These improvements can be explained either by our refinements or by the use of a different database.

4 Conclusion and perspectives

As explained previously, the advantage of the method is to have an explicit model allowing the determination of the amino acids and of their locations in the flanking sequence which favour the accessibility of a given residue. We show that the problem of accessibility prediction is mainly monofactorial, since 89% of the variability can be explained by the crossed effect of one factor measuring the amino acids' propensity to be accessible to the solvent and of one factor associated with their locations in the window. However, certain amino acids in the flanking sequence of a given residue rectify their accessibility prediction. Moreover, we show that the first principal component is highly correlated with some amino acids properties and we give two new indexes: an index of the solvent accessibility and an index of the tendency to establish a buried contact, these two indexes being highly correlated. The prediction accuracy after different model refinements is improved: a maximum gain of 2.1% for three thresholds of relative solvent accessibility (9%, 16% and 25%).

Further work will be carried out to study first the improvement of the prediction by taking account of multiple alignment and, second, the possible effects of order 2 related to the co-occurrences of amino acids in the flanking sequence of a given residue.

References

1. Lee BK, Richards FM (1971) J Mol Biol 55:379
2. Eisenberg D, Weiss RM, Terwilliger TC (1984) Proc Natl Acad Sci USA 81:140
3. Cornette JL, Cease KB, Margalit H, Spouge JL, Berzofsky JA, DeLisis C (1987) J Mol Biol 195:659
4. Viari A, Soldano H, Ollivier E (1990) CABIOS 6:71
5. Vriend G, Eijsink V (1993) J Comput Aided Mol Des 7:367
6. Baumann G, Fromel C, Sander C (1989) Protein Eng 2:329
7. Sippl MJ (1993) 17:355
8. Sippl MJ (1993) J Comput Aided Mol Des 7:473
9. Holbrook SR, Muskal SM, Kim SH (1990) Protein Eng 3:659
10. Rost B, Sander C (1994) Proteins 20:216
11. Thompson MJ, and Goldstein RA (1996) Proteins 25:38
12. Mardia, Kent, Bibby (1979) Multivariate analysis. Academic Press, London
13. Hobohm U, Scharf M, Schneider R, Sander C (1992) Protein Sci. 1:409
14. Hobohm U, Sander C (1994) Protein Sci. 3:522
15. Bernstein FC, Koetzle TF, Williams GJ, Meyer EF, Brice MD, Rodgers JR, Kennard O, Shimanouchi T, Tasumi M (1977). J Mol Biol 112:535
16. Argos P (1988) Protein Eng 2:101
17. Kabsch W, Sander C (1983) Biopolymers 22:2577
18. Kendall M, Stuart A (1977). The advanced theory of statistics. Griffen, London
19. Statistical Sciences (1993) SPlus guide to statistical and mathematical analysis, version 3.2. StatSci, Seattle
20. Kyte J, Doolittle RF (1982) J Mol Biol 1957:105
21. Chou PY, Fasman G (1978) Adv Enzymol 47:145
22. Karplus PA, Schutz GE (1985) Naturwissenschaften 72:212

23. Janin J (1976) J Mol Biol 105:13
24. Chothia C (1976) J Mol Biol 105:1
25. Rose GD, Geselowitch AR, Lesser GJ, Lee RH, Zehfus MH (1985) Science 229:834
26. Grantham R (1974) Science 185:862
27. Statistical Sciences (1996) S-Plus guide to statistical and mathematical analysis. StatSci, Seattle
28. Agresti A (1990) Categorical data analysis. Wiley, New York
29. Weisberg (1985) Applied linear regression, 2nd edn. Wiley New York
30. Orengo CA, Michie AD, Jones S, Jones DT, Swindells MB, Thornton JM (1997) Structure 5:1093