Available online at www.sciencedirect.com
SCIENCE DIRECT®
Journal of Chromatography A, 1075 (2005) 133–143
JOURNAL OF CHROMATOGRAPHY A
www.elsevier.com/locate/chroma

Is it possible to predict the average surface hydrophobicity of a protein using only its amino acid composition?

J. Cristian Salgado [a,*], Ivan Rapaport [b], Juan A. Asenjo [a]

[a] *Centre for Biochemical Engineering and Biotechnology, Department of Chemical Engineering and Biotechnology, Millennium Institute for Advanced Studies in Cell Biology and Biotechnology, University of Chile, Beauchef 861, Santiago, Chile*
[b] *Department of Mathematical Engineering, Centre for Mathematical Modeling, University of Chile, Blanco Encalada 2120, Santiago, Chile*

Received 30 November 2004; received in revised form 24 March 2005; accepted 6 April 2005
Available online 28 April 2005

Abstract

Hydrophobicity is one of the most important physicochemical properties of proteins. Moreover, it plays a fundamental role in hydrophobic interaction chromatography, a separation technique that, at present time, is used in most industrial processes for protein purification as well as in laboratory scale applications. Although there are many ways of assessing the hydrophobicity value of a protein, recently, it has been shown that the average surface hydrophobicity (ASH) is an important tool in the area of protein separation and purification particularly in protein chromatography. The ASH is calculated based on the hydrophobic characteristics of each class of amino acid present on the protein surface. The hydrophobic characteristics of the amino acids are determined by a scale of aminoacidic hydrophobicity. In this work, the scales of Cowan–Whittaker and Berggren were studied. However, to calculate the ASH, it is necessary to have the three-dimensional protein structure. Frequently this data does not exist, and the only information available is the amino acid sequence. In these cases it would be desirable to estimate the ASH based only on properties extracted from the protein sequence. It was found that it is possible to predict the ASH from a protein to an acceptable level for many practical applications (correlation coefficient > 0.8) using only the aminoacidic composition. Two predictive tools were built: one based on a simple linear model and the other on a neural network. Both tools were constructed starting from the analysis of a set of 1982 non-redundant proteins. The linear model was able to predict the ASH for an independent subset with a correlation coefficient of 0.769 for the case of Cowan–Whittaker and 0.803 for the case of Berggren. On the other hand, the neural model improved the results shown by the linear model obtaining correlation coefficients of 0.831 and 0.836, respectively. The neural model was somewhat more robust than the linear model particularly as it gave similar correlation coefficients for both hydrophobicity scales tested, moreover, the observed variabilities did not overcome 6.1% of the mean square error. Finally, we tested our models in a set of nine proteins with known retention time in hydrophobic interaction chromatography. We found that both models can predict this retention time with correlation coefficients only slightly inferior (11.5% and 5.5% for the linear and the neural network models, respectively) than models that use the information about the three-dimensional structure of proteins.

Keywords: Mathematical modeling; Hydrophobicity; Protein hydrophobicity prediction; Neural network; Proteins; Aminoacidic composition

1. Introduction

Hydrophobicity is one of the most important physicochemical properties of proteins. This property is so essential that it is considered as one of the fundamental components that govern protein folding [1]. Moreover, the hydrophobic characteristics of a protein perform a fundamental role defining its behavior in solution and how the protein relates to other biomolecules. In fact, this property plays a fundamental role in hydrophobic interaction chromatography (HIC), a separation technique that, at present time, it is being used in most industrial processes for protein purification as well as in laboratory scale applications.

The hydrophobicity value of a protein can be assigned by many different methodologies which can be experimental or

* Corresponding author. Tel.: +56 2 6784716; fax: +56 2 6991084.
E-mail address: jsalgado@ing.uchile.cl (J.C. Salgado).

0021-9673/$ – see front matter © 2005 Elsevier B.V. All rights reserved.
doi:10.1016/j.chroma.2005.04.014

theoretical. However, a great number of the main properties of a protein are determined by the features of its surface. For example, protein functions such as catalysis or molecular recognition occur predominantly on or near the protein surface. Also, it has been observed that the superficial amino acid composition is well correlated with the cellular localization of the protein [2]. Thus, it is natural that the estimation of the hydrophobicity will be related with the analysis of the protein surface. A method for establishing the hydrophobicity of a protein consists on considering the relative contribution of each one of the amino acids presents on the surface, defining by this way an average surface hydrophobicity (ASH) [3]. In this case, the contribution of each amino acid will be determined by its abundance and by its hydrophobic characteristics. The choice in how the aminoacidic hydrophobicity is quantified determines, in definitive, the protein hydrophobicity and the practical application of this value. For example, Berggren and collaborators showed that it is possible to predict the protein behavior in aqueous two-phase system knowing the value of ASH. The value of ASH was calculated based on a scale of aminoacidic hydrophobicity specially developed for that purpose [3]. On the other hand, Lienqueo and collaborators found that the ASH is correlated satisfactorily with the retention times in HIC [4]. In this case, one of the aminoacidic hydrophobicity scales that best modeled the behavior was proposed by Cowan and Whittaker [5].

However, to calculate the ASH, it is necessary to have the three-dimensional protein structure. Frequently this data does not exist, and the only information available is the amino acid sequence. In these cases, to estimate the surface composition of the protein it is necessary to start with the construction of three-dimensional models, usually using the methodology of comparative modeling [6] or maybe in some cases through the developing of ab initio models. These methodologies are quite complex. The question that then arises is: is it possible to carry out an estimation of the ASH based on simpler features like the aminoacidic composition? It has been pointed out that some features of the proteins can be predicted based on their amino acid composition. For example, it has been reported that the prediction of the protein's secondary structural content [7], and the protein structural class [8] can be carried out successfully from its amino acid composition only.

Keeping this in mind, the main objective of this paper is to show if it is possible to predict the ASH of a protein based on its amino acid composition and also in investigating possible mathematical models that could be suitable for this purpose.

2. Materials and methods

2.1. Data selection

The protein three-dimensional data used in this study was obtained from the PDB database [9]. This study used the database available until March 2004, when, the number of structures stored in PDB was nearly 25,000. In order to reduce data redundancy and to guarantee a minimum resolution in the considered structures, the working database was built using one of the subsets published by Hobohm et al. [10]. The available selection corresponds to that published in December 2003 and had a sequence identity cutoff of 25%. This selection can be found in the website http://homepages.fh-giessen.de/~hg12640/pdbselect/recent.pdb_select25. Also, the database was filtered eliminating membrane proteins, which were 60. The search was carried out searching directly on the PDB files with the following text patterns: "membrane", "transmembrane", "trans-membrane", "fiber" and "fibrous" and analyzing the results manually. The proteins that truly corresponded to membrane proteins were eliminated from the database. Finally, after all these operations, the working database was conformed by 1982 proteins.

2.2. Determination of the protein average surface hydrophobicity

The protein average surface hydrophobicity was computed assuming that each amino acid in the protein surface contributes, proportionally to its abundance, with the properties associated to the protein surface [3]. According to the previous hypothesis, the ASH can be calculated by the following equation:

$$\langle \Phi_{\text{surface}} \rangle = \sum_{i \in A} \hat{r}_i \phi_i \tag{1}$$

where $\langle \Phi_{\text{surface}} \rangle$ is the average superficial hydrophobicity for a given protein, A is the collection of the 20 possible amino acids and ϕ_i is the hydrophobicity of the amino acid of type i. The hydrophobicity of each amino acid was assigned according to the scale of Cowan–Whittaker [5] or according to the scale of Berggren [3] depending on the desired application. These hydrophobicity scales are detailed in Table 1. The fraction of superficial area $\hat{r}_i$ occupied by the amino acid i is defined by:

$$\hat{r}_i = \frac{S_i}{\sum_{j \in A} S_j} \tag{2}$$

where, S_i is the sum of the accessible superficial area (ASA) for all the amino acids of type i. The value of ASA was calculated using the software STRIDE [11].

2.3. Linear model

The feasibility of modeling the ASH by means of the data provided by the amino acid composition of the protein sequence was studied. The model I, is the simplest model considered in this work, and it is defined by the following equation:

$$\text{ASH}^{\text{I}} = c_0 + \sum_{i=1}^{20} c_i \hat{a}_i^{\text{I}} + c_{21} \hat{l} \tag{3}$$

Table 1
Hydrophobicity scales used in this work

aa	Hydrophobicity scales[a]	
	Cowan–Whittaker	Berggren
Ala	0.660	0.169
Arg	0.176	0.000
Asn	0.306	0.257
Asp	0.433	0.099
Cys	0.763	0.169
Gln	0.323	0.257
Glu	0.467	0.099
Gly	0.557	0.109
His	0.000	0.035
Ile	1.000	0.264
Leu	0.998	0.264
Lys	0.061	0.000
Met	0.846	0.169
Phe	0.983	0.796
Pro	0.768	0.169
Ser	0.401	0.169
Thr	0.494	0.169
Trp	0.914	1.000
Tyr	0.682	0.870
Tyr	0.682	0.870

[a] The scales are scaled in the interval [0; 1].

where, ASH^{I} is the ASH value for a protein given for model I, c_i corresponds to the parameters of the linear model obtained by the least squares procedure, $\hat{l}$ is the ratio between the length of the protein sequence and the maximum length observed in the working database. The value $\hat{a}_i^{I}$ corresponds to the fraction of the maximum accessible surface of the amino acids of type i when they are totally exposed, defined by:

$$\hat{a}_i^{I} = \frac{n_i S_{\max,i}}{\sum_{j \in A} n_j S_{\max,j}} \quad (4)$$

where, n_i is the number of amino acids of class i in the protein and $S_{\max,i}$ is the maximum possible value of ASA, obtained when arranging the amino acid of class i in a extended conformation tripeptide G–X–G [12]. The values of $S_{\max}$ in $Å^2$ are 113 (ala), 241 (arg), 158 (asn), 151 (asp), 140 (cys), 189 (gln), 183 (glu), 85 (gly), 194 (his), 182 (ile), 180 (leu), 211 (lys), 204 (met), 218 (phe), 143 (pro), 122 (to Be), 146 (thr), 259 (trp), 229 (tyr), 160 (val).

The previous model considers that each amino acid is in its maximum exposition state. An extension to this model was to incorporate a correction factor that takes into account the general tendency of each amino acid to be exposed or not to the solvent. The quantification of this trend can be performed in multiple ways. However, the options analyzed in this study were built starting from the analysis of the relative accessible superficial area (RASA). Then, the alternatives for the exposition factor considered in this study were:

(a) Average and median of the RASA of all the amino acids of class i in the database.
(b) Probability that one amino acid of class i has a RASA superior to a certain threshold μ.

The RASA of an amino acid k in a protein is defined as the ratio between their ASA value (s_k) and their maximum ASA ($S_{\max}$). Then, the model II is obtained when incorporating the following Eq. (5) into the model described by Eq. (3):

$$\hat{a}_i^{II} = \frac{n_i S_{\max,i} \alpha_i}{\sum_{j \in A} n_j S_{\max,j} \alpha_j} \quad (5)$$

where, α_i is the exposition factor for the amino acid of class i.

Finally, the model III establishes a linear relationship among the ASA S_i for all the amino acid of class i and the maximum possible ASA defined for $n_i S_{\max,i}$. In this case, $\hat{a}_i^{III}$ would be given by:

$$\hat{a}_i^{III} = \frac{n_i S_{\max,i} \beta_i + \eta_i}{\sum_{j \in A} (n_j S_{\max,j} \beta_j + \eta_j)} \quad (6)$$

where β_i and η_i are the coefficients of the linear model between S_i and $n_i S_{\max,i}$ calculated for all the amino acids of class i present in the complete database using the least squares procedure.

By definition, the sum of coefficients $\hat{a}_i$ is one, so these coefficients conform a linear dependent system. Therefore, the models analyzed in this work do not consider the data of the least hydrophobic amino acid, provided by histidine for the models associated to the scale of Cowan–Whittaker, and arginine, in the case the models associated to the Berggren scale.

2.4. Using linear models to predict ASH

The possibility of using the linear models described previously as predictors of the ASH for proteins with unknown three-dimensional structure was evaluated. For that purpose, the working database was divided in two subsets: train and test subsets in a ratio 2:1 (1321/661). The train subset was used to adjust the parameters of the models using the least square method. The test subset was used to evaluate the performance of the models like prediction tools. The construction of the train and test subsets was repeated 100 times in a random way. For each repetition, the effectiveness of the models was evaluated; the average on all the repetitions was finally reported.

2.5. Design of a neural network predictor

A predictor tool based on a neural network model was also considered as these models have shown to be robust and effective tools to solve this type of problem in a comprehensive spectrum of applications [13–16]. The design of the neural network was carried out considering as inputs the main variables of the models I–III. Those are, the variables $\hat{a}_i^{I}$, $\hat{a}_i^{II}$, $\hat{a}_i^{III}$ (inputs type I–III) and $\hat{l}$, these variables defined a group of 20 inputs. The components associated to histidine, in the case of Cowan–Whittaker and arginine for Berggren, were not considered. Two preprocessing techniques were

tested for the inputs and outputs of the network: scaling min/max, which makes a scaling of the entries so that they fall in the interval [−1; 1]; and normalization avg/std, which transforms the entries so that they will have zero mean and unity standard deviation.

The architecture of the network considered an input layer with 20 components, an output layer with one neuron and a hidden layer with a variable number of neurons. The size of the hidden layer was selected between 1 and 15 depending on the performance of the network. The activation function used in the network depended on the preprocessing of the input data. When the data were preprocessed using the scaling min/max or the normalization avg/std, the activation function was chosen to be the hyperbolic tangent function since it maps its entry to the interval [−1; 1]. On the other hand, when no preprocessing was done, the activation function was a sigmoid function which maps its entry to the interval [−1; 0].

The training algorithm selected was a commercial implementation of the Levenberg–Marquardt algorithm. This algorithm is more memory intensive than the traditional algorithms, however, it presents very efficient and fast learning features. The chosen parameters for the training algorithm in all the experiments were: $\mu_{LM} = 1.0$, $\mu_{DEC} = 0.8$ and $\mu_{INC} = 1.5$. The training style selected was Batch-training, in this modality all the data are presented to the network before making any change in the network parameters. In general, this training style allows obtention of softer and most continuous learning profiles.

As the main interest in this model is to use its prediction capabilities, it was necessary to train the network so it guarantees good generalization. To accomplish this objective, a widely used methodology, called early-stopping or stopped training, was used. Basically, this method requires the division of the data set into three subsets: train, validation and test. During the training, the update of the network parameters is made based only on the data provided by the train subset, while simultaneously, the performance of the network is monitored in the validation subset. The process of training stops when the error in the validation subset increases continually for a specified number or epochs. Finally, the prediction capabilities of the network are measured, in the test subset. In this study, the database was divided according to the ratio 2:1:1 (991/496/495) for the train, validation and test subsets, respectively.

In this case, it was of interest to know the robustness of the model to changes in the subsets of train, validation and test. To test this, each network was trained with 20 groups of train, validation and test subsets built randomly. Additionally, for each one of these repetitions the network was initialized with 20 sets of random different parameters to avoid reaching a local minimum in the error surface.

2.6. Performance measurement for the models

The performance of the models developed in this work was compared by mean of four fundamental parameters: The mean square error MSE, the mean absolute error MAE, the standard deviation of the MAE, the maximum value of MAE and the Pearson's correlation coefficient. The mean absolute error and the Pearson's correlation coefficient utilized in this work were calculated by means of the following expressions:

$$\mathrm{MAE} = \frac{100}{N} \sum_{k=1}^{N} \frac{|x_k - y_k|}{x_k} \tag{7}$$

$$\mathrm{Pearson} = \frac{N \sum_{k=1}^{N} (x_k y_k) - \sum_{k=1}^{N} x_k \sum_{k=1}^{N} y_k}{\sqrt{N \sum_{k=1}^{N} (x_k)^2 - \left(\sum_{k=1}^{N} x_k \right)^2} \sqrt{N \sum_{k=1}^{N} (y_k)^2 - \left(\sum_{k=1}^{N} y_k \right)^2}} \tag{8}$$

where x_k represents the ASH of the protein k; y_k, the prediction of the ASH for the protein k; and N, the number of proteins in consideration. The maximum MAE corresponds to the biggest percentage discrepancy observed in the set in consideration.

3. Results and discussion

3.1. Analysis of the database

The average surface hydrophobicity defined in Eq. (1) was calculated for all the proteins in the database using the hydrophobicity scales of Cowan–Whittaker and Berggren as discussed by Berggren and Lienqueo [3,4]. Fig. 1 shows a histogram of the ASH, as well as a scatter plot between ASH calculated by means of both scales. Both plots show that, in general, the value of the ASH for a protein calculated using the scale of Cowan–Whittaker is larger than the same calculated using Berggren. The average value observed in the ASH of Cowan–Whittaker is almost 2.5 times the observed in the case of Berggren. Also, the range for the ASH of Cowan–Whittaker is 1.7 times the range of Berggren, therefore, it is possible to state that the ASH calculated according to the scale of Berggren is less sensitive to changes in the composition of the protein surface, this being confirmed by the special features present in the scale of Berggren (the hydrophobic amino acids are well separated from the hydrophilic ones). It was determined that the correlation coefficient among the ASH calculated using both scales is slightly greater than 0.6, indicating that both measures are not easily interchangeable. Therefore, it is necessary to study both scales separately.

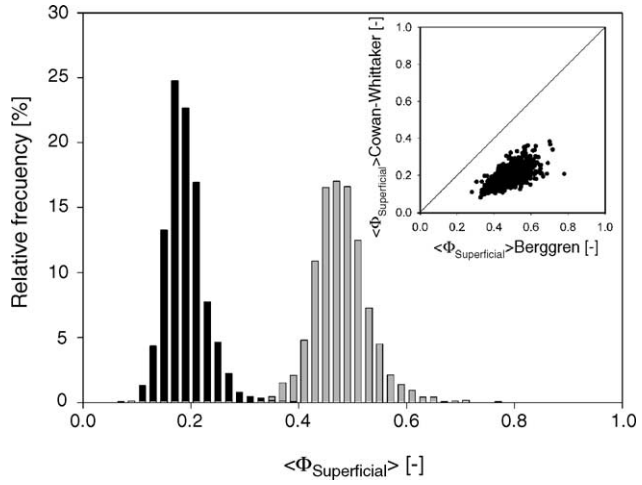

Fig. 1. Histogram of average surface hydrophobicity (ASH) in the whole database. (▨) ASH calculated using the scale of Cowan–Whittaker (<ASH> = 0.479 ± 0.051, ASH_{min} = 0.280, ASH_{max} = 0.779). (■) ASH calculated using the scale of Berggren (<ASH> = 0.189 ± 0.036, ASH_{min} = 0.080, ASH_{max} = 0.382). Inset: scatter plot between the ASH calculated by means of the scales of Cowan–Whittaker and Berggren (slope = 0.863, intercept = 0.315, Pearson = 0.623).

The aminoacidic composition of the database was investigated as shown in Fig. 2 and it was observed that the values obtained followed a reasonable distribution. Near 30% of the database was conformed by leucine (8.8%), alanine (7.8%), glycine (7.1%) and valine (6.9%) which correspond to some of the amino acids found with higher frequency in the proteins and are usually considered more as amino acids with a structural function than as components with some special role in the protein. On the other hand, the scarcest amino acids were histidine (2.4%), methionine (2.1%), cysteine (1.9%) and tryptophan (1.5%), which correspond to amino acids commonly found in active sites or binding sites of proteins, and therefore, are less frequent.

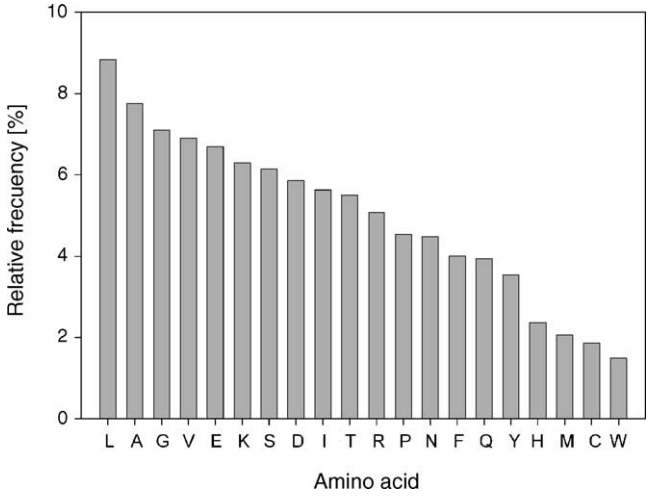

Fig. 2. Relative frequency of the amino acids in the working database.

3.2. Linear models

The linear models I–III were built using all the data in the database by means of the least squares methodology. Before using model II, it was necessary to calculate the exposition factors associated to each class of amino acid. In total, 12 factors were calculated, obtained from the analysis of the whole database. These coefficients are listed in Table 2. In a few words, the values in this table correspond quite well with what one would expect based on the hydrophobic characteristic associated to each amino acid. This way, the most hydrophilic amino acids present high exposition factors α_i indicating their predisposition to be exposed to the solvent. On the contrary, the hydrophobic amino acids possess the lowest exposition factors.

Model III required the determination of the parameters of the linear functions that associate the exposed area to the solvent S_i and the maximum possible area $n_i S_{max,i}$. The parameters of these functions together with the correlation coefficient associated to each one of these are listed in Table 3. In the same way as in the model II, the hydrophobic nature of each amino acid determined the general trend observed in Table 3. The correlation coefficients of the hydrophilic or amphipathic amino acids, for instance: lys (0.967), glu (0.963), asp (0.947), gln (0.931), asn (0.928) and arg (0.917), are very high. On the other hand, the smallest correlation coefficients were found for the hydrophobic amino acids: phe (0.671), ile (0.682). Consequently, the confidence intervals associated to the slope of hydrophilic amino acids models were smaller than the hydrophobic ones.

Table 4 summarizes the results obtained by the three models in this study. In the discussion that follows, the reference for discussion is taken as the results obtained by model I. It was determined that in the case of both hydrophobicity scales, the best results reached by the model II were obtained when selecting an exposition factor equal to the probability that one amino acid of class *i* has a RASA greater than 60% (μ = 0.6 in Table 2). However, although this model incorporates a higher quantity of information, represented by the exposition factor to the solvent, it was only able to decrease the MSE in 4% as average in both scales. The increase of the correlation coefficients was even less appreciable, only corresponding to an increment of 1% on average. With regard to model III, in the case of the Cowan–Whittaker scale, this was able to decrease the MSE in a little more than 18% and to increase the correlation coefficient in 6%. These results are almost five times better that those obtained by model II. In the case of Berggren, the performance of model III was almost the same as the one observed in model I. The results obtained by model III are justified by the greater amount of information regarding the way in which the amino acids are distributed in the protein surface as a function of their abundance. Although the results obtained by model III represent a reasonable enhancement with regard to model I, this progress is not necessarily justified with the amount of the additional information available for the model.

Table 2
Exposition factors α_i for each class of amino acid obtained from the distribution of relative accessible superficial area (RASA) in the whole database: average RASA, median RASA and the relative frequency to find a RASA larger than a threshold μ

aa	<RASA>	Median RASA	Relative frequency of RASA > μ									
			$\mu = 0$	$\mu = 0.1$	$\mu = 0.2$	$\mu = 0.3$	$\mu = 0.4$	$\mu = 0.5$	$\mu = 0.6$	$\mu = 0.7$	$\mu = 0.8$	$\mu = 0.9$
Ala	0.260	0.166	0.842	0.576	0.466	0.378	0.296	0.214	0.143	0.087	0.044	0.023
Arg	0.445	0.435	0.995	0.916	0.818	0.697	0.554	0.410	0.277	0.163	0.086	0.037
Asn	0.441	0.440	0.976	0.849	0.757	0.659	0.549	0.425	0.299	0.199	0.119	0.065
Asp	0.479	0.480	0.983	0.883	0.801	0.710	0.603	0.472	0.347	0.240	0.152	0.086
Cys	0.149	0.073	0.836	0.446	0.282	0.176	0.108	0.063	0.036	0.021	0.012	0.005
Gln	0.448	0.458	0.986	0.889	0.801	0.700	0.580	0.441	0.292	0.169	0.088	0.036
Glu	0.510	0.522	0.991	0.925	0.860	0.777	0.668	0.534	0.385	0.248	0.137	0.065
Gly	0.343	0.293	0.885	0.700	0.593	0.493	0.395	0.305	0.223	0.155	0.090	0.046
His	0.342	0.305	0.968	0.776	0.640	0.506	0.382	0.281	0.184	0.110	0.056	0.022
Ile	0.161	0.064	0.847	0.431	0.297	0.210	0.148	0.093	0.051	0.027	0.014	0.006
Leu	0.186	0.091	0.871	0.484	0.344	0.248	0.174	0.117	0.071	0.040	0.020	0.008
Lys	0.505	0.510	0.997	0.961	0.902	0.805	0.675	0.515	0.349	0.205	0.099	0.035
Met	0.241	0.146	0.881	0.563	0.437	0.330	0.244	0.176	0.120	0.079	0.050	0.031
Phe	0.176	0.086	0.891	0.473	0.323	0.220	0.155	0.102	0.063	0.036	0.018	0.006
Pro	0.402	0.398	0.965	0.814	0.713	0.607	0.497	0.383	0.267	0.163	0.080	0.030
Ser	0.380	0.356	0.941	0.763	0.658	0.559	0.451	0.343	0.243	0.165	0.099	0.054
Thr	0.337	0.313	0.942	0.747	0.629	0.516	0.398	0.278	0.170	0.102	0.056	0.027
Trp	0.202	0.134	0.950	0.573	0.383	0.253	0.166	0.110	0.067	0.033	0.017	0.006
Tyr	0.246	0.185	0.956	0.656	0.473	0.329	0.226	0.153	0.095	0.054	0.027	0.011
Val	0.171	0.073	0.833	0.452	0.325	0.231	0.160	0.099	0.056	0.030	0.015	0.006

With regards to model coefficients it was observed that, in general, the coefficients associated to the amino acids are positive, while the coefficients associated to the chain length are negative. In the case of Cowan–Whittaker the models included the contribution of almost the total of the available variables, except in the case of model I, which deleted the variable associated to lysine. This is reasonable, since lysine has the second greatest hydrophobicity drop in value in the scale of Cowan–Whittaker. In the case of Berggren, the models deleted a major quantity of variables, coinciding with the features characteristic of the scale, where hydrophobic amino acids are relatively far from the hydrophilic amino acids.

Table 3
Parameters, their confidence intervals (95%) and Pearson's correlation coefficients of the linear functions ($S_i = \beta_i n_i S_{\max,i} + \eta_i$) used to model the relationship between the observed ASA (S_i) and the maximum possible ASA ($n_i S_{\max,i}$) for each amino acid

aa	Pearson	η_I	β_I
Ala	0.855	90.7 ± 10.1	0.198 ± 0.005
Arg	0.917	163.2 ± 17.6	0.363 ± 0.007
Asn	0.928	89.3 ± 9.9	0.364 ± 0.007
Asp	0.947	113.4 ± 11.2	0.403 ± 0.006
Cys	0.725	7.0 ± 3.4	0.132 ± 0.006
Gln	0.931	73.4 ± 10.6	0.388 ± 0.007
Glu	0.963	100.8 ± 14.1	0.461 ± 0.006
Gly	0.890	101.2 ± 7.1	0.243 ± 0.006
His	0.840	43.1 ± 8.4	0.284 ± 0.008
Ile	0.682	94.9 ± 10.5	0.103 ± 0.005
Leu	0.723	181.1 ± 15.9	0.116 ± 0.005
Lys	0.967	96.1 ± 14.5	0.461 ± 0.006
Met	0.686	59.9 ± 7.2	0.155 ± 0.007
Phe	0.671	93.0 ± 10.0	0.110 ± 0.005
Pro	0.906	80.5 ± 9.4	0.326 ± 0.007
Ser	0.895	90.9 ± 10.6	0.308 ± 0.007
Thr	0.892	93.2 ± 10.2	0.265 ± 0.006
Trp	0.749	30.8 ± 5.8	0.152 ± 0.006
Tyr	0.799	98.7 ± 10.3	0.171 ± 0.006
Val	0.743	110.8 ± 9.9	0.109 ± 0.004

3.3. About the predictive performance of the linear models

In this section, the predictive capability of the linear models analyzed in the previous section is studied. Due to this, the performance of the model in the prediction of the ASH in a standalone test subset that was not used to adjust the parameters of each model was studied. The results obtained in this section don't differ substantially from those obtained previously when the complete set of data was modeled as could be reasonably expected.

Tables 5 and 6 detail the results for the calculated ASH based on the scales of Cowan–Whittaker and Berggren, respectively. The content of these tables allows us to affirm that remarkable changes did not take place in the comparative performance of the models I–III: in the case of Cowan–Whittaker, the tendency maintained model III as the best one, presenting a reduction of the MSE observed in the test group of 18% with regard to model I; in the same way, for the scale of Berggren, changes were not observed with regard to the previous section, where big discrepancies among the performance of each of the models were not found.

In general, the results obtained in the test group had inferior quality than those observed in the train subset, because the test group was not used in the construction of the models. The discrepancies among the value of obtained MSE between

Table 4
Performance indicators for the least square adjustment of the average surface hydrophobicity (ASH) and linear models I–III for the whole database: mean square error (MSE), mean absolute error (MAE), standard deviation of MAE (MAE std) and correlation coefficient (Pearson)

Performance index	Cowan–Whittaker			Berggren		
	Model I	Model II	Model III	Model I	Model II	Model III
MSE × 10^3 (–)	1.028	0.981	0.835	0.464	0.449	0.466
MAE (%)	5.3	5.2	4.8	9.1	8.9	9.1
MAE std (%)	4.1	4.0	3.9	7.0	7.0	7.5
MAE max (%)	28.7	27.8	38.6	48.7	46.3	70.8
Pearson (–)	0.776	0.788	0.823	0.810	0.816	0.809

both subsets were less than 5% for the two scales in study, in the case of model I discrepancies of around 3% were observed for both scales. For Cowan–Whittaker and Berggren the minimum value of maximum MAE was obtained in model II, reaching 22.4% in the case of Cowan–Whittaker and almost twice as much for Berggren (42.3%).

All the models demonstrated an appreciable robustness, only small standard deviations among all the repetitions were observed. In general, the maximum variability in the results was observed in the scale of Berggren, except per the measures of maximum MAE and the correlation coefficient. The standard deviation in the MSE of the test group was less than 7% in the case of Cowan–Whittaker and 8% in the case of Berggren. The biggest variability was found in the value of maximum MAE that was less than 24% in both cases. The variability of this measure associated with the scale of Berggren was inferior to that observed in the case of Cowan–Whittaker. All this indicates that the models associated with the scale of Cowan–Whittaker can be considered slightly more robust.

With regards to the parameters of the models: a great variability is not observed in the parameters found. In fact, great discrepancy is not observed between these parameters and the obtained in the previous section. This confirms the robustness of the models found.

Additionally, we tested this models on the set of nine proteins used by Lienqueo et al. [4]. The same equation proposed by Lienqueo et al. was used in order to correlate the predictions of the Cowan–Whittaker's ASH, carried out by models I–III, with the retention times in hydrophobic interaction chromatography of these proteins. The correlation coefficients were: 0.850 ± 0.01, 0.844 ± 0.01 and 0.954 ± 0.01 for models I, II and III, respectively. Lienqueo et al. gave a correlation coefficient of 0.96. It is important to observe that the standard deviations on the repetitions are small, being consistent with the previous observations.

Table 5
Average performance indicators and their standard deviation considering all repetitions for the linear models used as predictors of ASH in the case of Cowan–Whittaker scale: mean square error (MSE), mean absolute error (MAE), standard deviation of MAE (MAE std) and correlation coefficient (Pearson)

Performance index	Cowan–Whittaker					
	Model I		Model II		Model III	
	Train	Test	Train	Test	Train	Test
MSE × 10^3 (–)	1.025 ± 0.022	1.055 ± 0.069	0.974 ± 0.019	1.020 ± 0.061	0.828 ± 0.017	0.869 ± 0.053
MAE (%)	5.3 ± 0.1	5.4 ± 0.2	5.2 ± 0.1	5.3 ± 0.2	4.8 ± 0.0	4.9 ± 0.1
MAE std (%)	4.1 ± 0.0	4.1 ± 0.1	4.0 ± 0.1	4.0 ± 0.2	3.9 ± 0.1	4.0 ± 0.2
MAE max (%)	26.6 ± 2.8	23.6 ± 4.2	26.2 ± 2.7	22.4 ± 3.9	37.0 ± 3.1	30.9 ± 7.0
Pearson (–)	0.777 ± 0.007	0.769 ± 0.021	0.789 ± 0.006	0.779 ± 0.019	0.824 ± 0.004	0.816 ± 0.014

Table 6
Average performance indicators and their standard deviation considering all repetitions for the linear models used as predictors of ASH in the case of Berggren scale: mean square error (MSE), mean absolute error (MAE), standard deviation of MAE (MAE std) and correlation coefficient (Pearson)

Performance index	Berggren					
	Model I		Model II		Model III	
	Train	Test	Train	Test	Train	Test
MSE × 10^3 (–)	0.462 ± 0.011	0.477 ± 0.035	0.446 ± 0.009	0.464 ± 0.027	0.465 ± 0.011	0.474 ± 0.035
MAE (%)	9.0 ± 0.1	9.2 ± 0.3	8.9 ± 0.1	9.0 ± 0.3	9.1 ± 0.1	9.2 ± 0.3
MAE std (%)	7.0 ± 0.1	7.1 ± 0.3	6.9 ± 0.1	7.0 ± 0.3	7.5 ± 0.1	7.6 ± 0.5
MAE max (%)	47.2 ± 2.1	43.2 ± 5.0	45.2 ± 1.5	42.3 ± 4.4	69.9 ± 3.9	58.2 ± 13.7
Correlation coefficient (–)	0.811 ± 0.005	0.803 ± 0.015	0.817 ± 0.005	0.810 ± 0.015	0.809 ± 0.005	0.804 ± 0.014

Table 7
Optimum size of the hidden layer of the neural predictor of ASH found considering network inputs of types I–III and different preprocessing techniques

Preprocessing	Cowan–Whittaker (input type)			Berggren (input type)		
	I	II	III	I	II	III
None	3	3	4	3	2	3
Min/max	2	2	3	2	3	3
Avg/std	2	2	2	2	2	2

3.4. About the predictive performance of the neural network models

In this section the results of using a neural network as a predictor model of the ASH are shown. To determine the most appropriate architecture of the network as well as the effect of preprocessing the input data, all the permutation of inputs (I–III), preprocessing techniques (min/max, avg/std, none), and number of neurons in the hidden layer was evaluated. For each pair composed by the input type and their preprocessing, 15 neural networks were evaluated, varying the number of neurons in the hidden layer. Each valuation was repeated 20 times with train, validation and test subsets generated randomly. In turn, each one of these repetitions was trained 20 times with sets of different initial weights. Therefore, a total of 54.000 different configurations were evaluated for each hydrophobicity scale.

The valuation of these configurations allowed the determination of the optimum network architecture for each pair: input type/preprocessing type. The determination of the optimum number of neurons in the hidden layer was accomplished comparing the MSE average on the test subset obtained in all the repetitions. A summary of these results is shown in Table 7. In this table, it is observed that, in all cases, the optimum size of the network hidden layer was found between two and four neurons, that corresponds to a small number of neurons in the hidden layer in relation to the size of the input layer. The number of parameters to determine in the training of each network is much larger than the linear model. In each configuration, the number of inputs is constant and equal to 20 since the data provided by the histidine was removed in the case of Cowan–Whittaker and arginine, in the case of Berggren. Therefore, for a network with 20 inputs, two neurons in the hidden layer and one neuron in the output layer, the number of parameters to determine is 45 ($20 \times 2 + 2 + 2 \times 1 + 1$), while when the size of the hidden layer is increased to four neurons, it is 89 ($20 \times 4 + 4 + 4 \times 1 + 1$). The number of parameters in a linear model with 20 variables and one constant is 21. Then, the architecture of an optimum network has, at least, a little more than double the parameters than the equivalent linear model.

It was found that the preprocessing technique does not affect the performance of the neural model. In general, it is well known that preprocessing the input data for a neural network is determinant in its performance. However, in this application, in particular, Fig. 3 shows that the absence of preprocessing in the inputs allowed the obtention of the best results in all cases analyzed. The data are already in the interval [0; 1] and although they do not necessarily cover completely this range, this did not have a substantial effect in the results. In the measurements presented in Fig. 3, the min/max scaling technique showed slightly inferior results to those obtained without preprocessing. On the other hand, the avg/std preprocessing gave the worst results, deviating more than 40% in the case of the scale of Cowan–Whittaker and more than 16% in the case of Berggren. The variability of these results was consistent with the previous observations of the biggest values in the case of the normalization avg/std (more than 10%). Finally, it was observed that the addition of extra information to the aminoacidic composition, as in the case of the neural model based on the inputs type II and III, did not improve notably the results. In the cases of Cowan–Whittaker and Berggren, the discrepancy between the models based on the inputs of type II and III with those of type I was less than 2%, therefore, the discussion will consider only the neural model with input of type I.

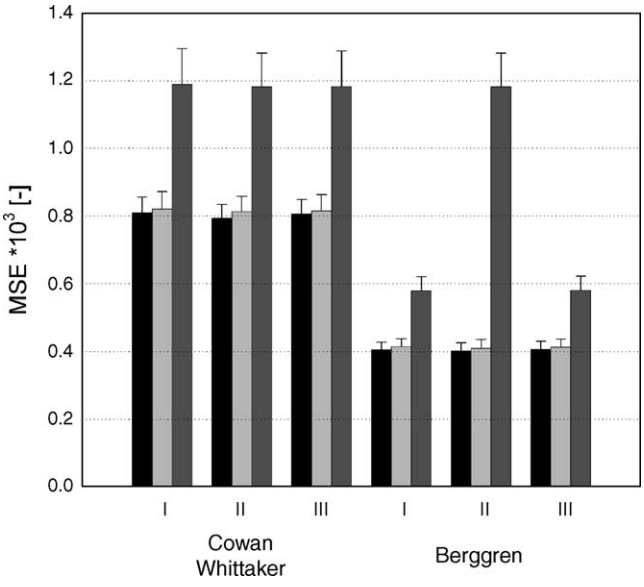

Fig. 3. Mean square error (MSE) of the prediction using the neural predictor of ASH on the test subsets for the network inputs type I–III. On the left side, the scale of Cowan–Whittaker; on the right side, the scale of Berggren. The color of the bar represents: (■) none input preprocessing, (▒) min/max input preprocessing, (▓) avg/std input preprocessing. The error bars represent the standard deviation of the MSE for all repetitions.

The previous analysis allows us to conclude that the optimum network architecture for the prediction of the ASH starting from the aminoacidic composition, is composed of only three neurons in the hidden layer, and that the preprocessing of the input data by means of the techniques analyzed is unnecessary. With this information the training of the neural model was repeated but considering an outline where had been increased the number of repetitions to 100 and the number of initials conditions to 40. The results of retraining the neural model with the conditions specified are shown in Table 8. As with the linear models, the performance

Table 8
Average performance indicators and their standard deviation considering all repetitions for the neural network predictor of ASH: mean square error (MSE), mean absolute error (MAE), standard deviation of MAE (MAE std) and correlation coefficient (Pearson)

Performance index	Cowan–Whittaker			Berggren		
	Train	Validation	Test	Train	Validation	Test
MSE × 10^3 (–)	0.730 ± 0.031	0.825 ± 0.054	0.800 ± 0.044	0.359 ± 0.017	0.410 ± 0.028	0.405 ± 0.025
MAE (%)	4.5 ± 0.1	4.7 ± 0.2	4.7 ± 0.1	8.1 ± 0.2	8.6 ± 0.3	8.5 ± 0.2
MAE std (%)	3.5 ± 0.1	3.8 ± 0.2	3.7 ± 0.2	6.3 ± 0.2	6.7 ± 0.4	6.7 ± 0.3
MAE max (%)	23.2 ± 2.8	25.1 ± 6.4	25.7 ± 5.4	43.4 ± 4.1	43.2 ± 6.0	43.5 ± 6.0
Correlation coefficient (–)	0.846 ± 0.010	0.829 ± 0.014	0.831 ± 0.014	0.857 ± 0.009	0.833 ± 0.015	0.836 ± 0.014

of the neural model in the test subset was inferior to that observed in the train subset. In the test subset and in the case of Cowan–Whittaker, the neural model was able to decrease the MSE obtained by the linear model I in 24.2%, in turn, increasing the correlation coefficient by 8.1%. For Berggren, the decrease of the MSE was inferior only corresponding to 15.1% associated to an increment of 4.1% in the correlation coefficient. The neural model was also robust to changes in the train, validation and test subsets. The variability in the results experienced a small decrease with regards to the linear model, inferior to 15% in most of the cases.

The results presented so far show that the predictions of the calculated ASH based on the scale of Cowan–Whittaker present a smaller MAE than the same ones based on the scale of Berggren. In fact, Table 8 establishes that the MAE average for the case of Berggren is almost twice that of Cowan–Whittaker. Fig. 4 shows a plot with the accumulative frequency of the MAE in a test group chosen randomly. By means of this plot we can confirm that the frequency of obtaining a prediction with high error is different for both scales, being more frequent in the case of Berggren than in the case of Cowan–Whittaker. For Cowan–Whittaker, the frequency of having a MAE greater than 10% is less than 8%, whereas in the case of Berggren, it is 22%. Although these values are dependent on the selected test subset, the general behavior should not change too much due to the small variabilities observed in all the actual repetitions of experiments.

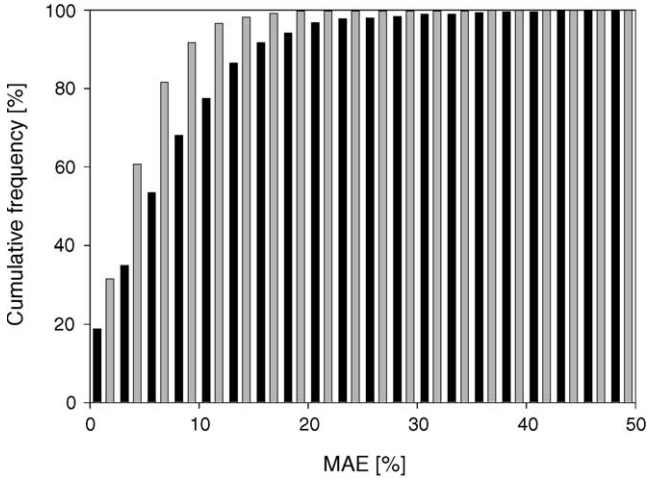

Fig. 4. Cumulative frequency of the mean absolute error (MAE) of the neural predictor of ASH on a test subset chosen randomly. The color of the bar represents: ASH calculated using (▒) the scale of Cowan–Whittaker and (■) the scale of Berggren.

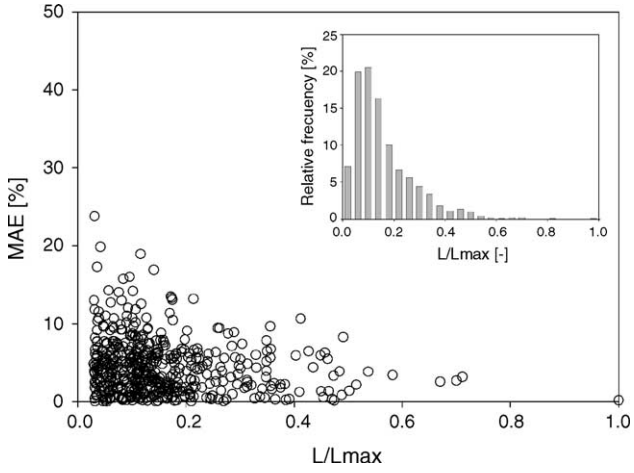

Fig. 5. Mean absolute error (MAE) of the neural predictor of ASH on a test subset in function of the normalized length of the protein. The test subset was chosen randomly and the ASH was calculated using the scale of Cowan–Whittaker. Inset: distribution of the length in the whole database.

The relationship between the predicted error and the length of the protein studied is shown in Figs. 5 and 6. A clear re-

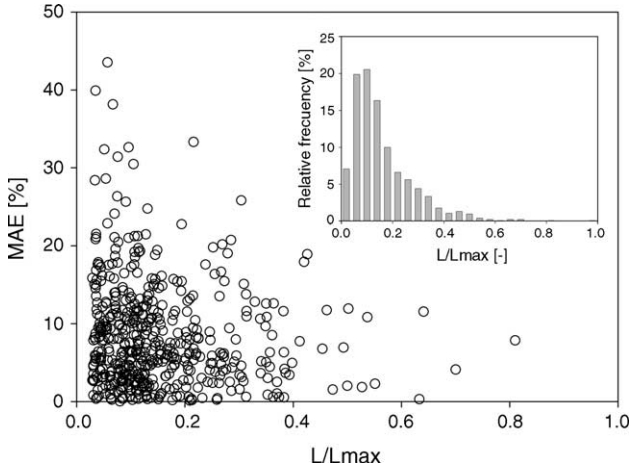

Fig. 6. Mean absolute error (MAE) of the neural predictor of ASH on a test subset in function of the normalized length of the protein. The test subset was chosen randomly and the ASH was calculated using the scale of Berggren. Inset: distribution of the length in the whole database.

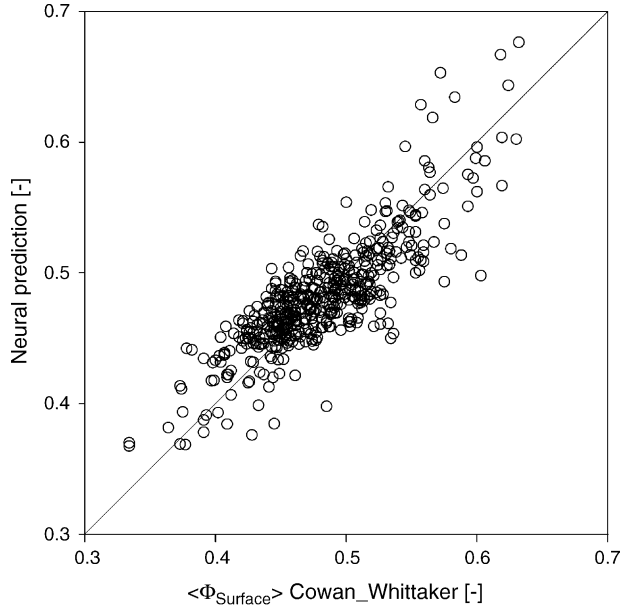

Fig. 7. Scatter plot between the ASH calculated using the scale of Cowan–Whittaker and the prediction of the neural network in a test subset chosen randomly.

lationship is not observed between the MAE of the prediction and the normalized protein length, unless the variance of this indicator diminishes as the protein length increases. This is explained by the distribution of the protein length in the database: 90% of the proteins in the database presented a length inferior to 30% of the maximum observed length (1014 amino acids) and therefore, the database is composed in its majority of proteins of medium size.

A direct relation between the error in the prediction and the value of ASH of the protein was not observed. The scatter plots in Figs. 7 and 8 show that the error extends regularly in the whole range of ASH, except toward both ends of the range. Those values have a higher associated error than those observed in the center of the range. The simplest explanation is that the available data to build the models concentrates in the center of the ranges (see Fig. 1), and therefore, the models present a better performance under those conditions. This can be useful to understand why most of the proteins analyzed in this study present medium hydrophobicity values.

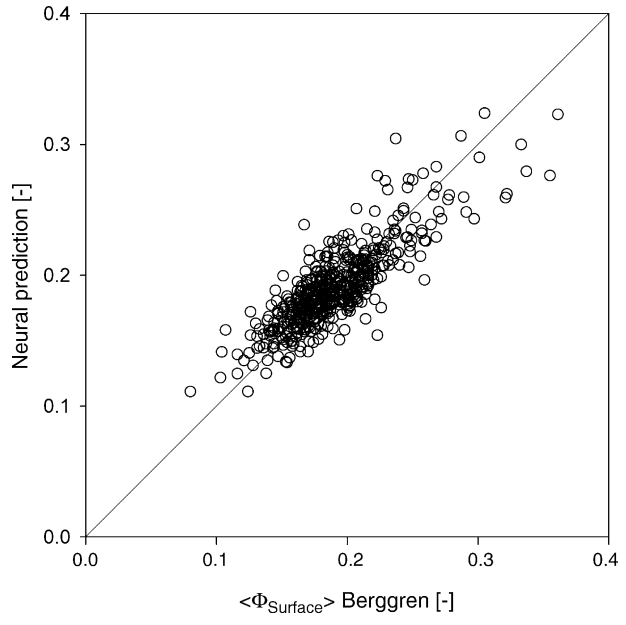

Fig. 8. Scatter plot between the ASH calculated using the scale of Berggren and the prediction of the neural network in a test subset chosen randomly.

Finally, we used this neural network model (with input of type I) to predict the Cowan–Whittaker's ASH on the set of nine proteins used by Lienqueo et al. [4]. These predictions were used to estimate the retention time in hydrophobic interaction chromatography of these proteins, in the same way that in the case of the linear models, giving, in this case, a correlation coefficient of 0.907 ± 0.02. Lienqueo et al. gave a correlation coefficient of 0.96.

4. Conclusions

The question this paper addresses is: Is it possible to predict the average superficial hydrophobicity of a protein using only their amino acid composition? The results analyzed in this study showed that indeed it is possible, although, the quality of the prediction can be subject to some considerations.

The simpler model was the linear model I based only on the aminoacidic protein composition. This model has 21 parameters and was able to predict the ASH for a standalone test subset with a correlation coefficient of 0.769 for the case of Cowan–Whittaker and 0.803 for the case of Berggren. In all the cases where it was evaluated, it gave a low variability in its performance.

A model based on a neural network was also evaluated. This model used the same inputs as the linear model I. It was observed that for this problem, the optimum configuration for a neural model with a single hidden layer considered three neurons in the hidden layer and with no preprocessing of the inputs. This model has 67 parameters and improved the results shown by the linear model in a little more than 24% for the case of Cowan–Whittaker and 15% for Berggren. The correlation coefficients obtained by this model were 0.831 and 0.836, respectively. The neural model was shown to be slightly more robust that the linear one. In both cases, the observed variabilities were not greater than 6.2% of the mean square error.

Additionally, it was determined that the addition of data about the aminoacidic exposition tendencies to the solvent was not translated in a substantial enhancement in the results obtained by means of both predictors. Clear interrelation was not observed between the quality of the prediction and the protein length or the range of ASH of the protein being studied. Distortions were only observed depending on the characteristics of the database distribution. Finally, although the neural model gave better results than the linear

model in all cases reviewed, these improvements did not justify, necessarily, the considerable increase in the parameters and complexity of the model.

Finally, we tested our models in the set of nine proteins used by Lienqueo et al. [4] where they reported a correlation coefficient of 0.96 when they used the ASH based on the scale of Cowan–Whittaker to predict the retention time in hydrophobic interaction chromatography of these proteins. The linear model I, proposed in this work, obtained a correlation coefficient only 11.5% inferior. On the other hand, the neural network model obtained a correlation coefficient only slightly inferior (5.5%). Both models showed very low standard deviations. These preliminary results show that both models have great potential for practical applications.

Acknowledgements

This work was supported by the Fondef project 011031, the Fondap project CMM II, the postgraduate scholarship of CONICYT and the Millennium Institute for Advance Studies in Cell Biology and Biotechnology (ICM-P99-031). We wish to thank Dr. Maria Elena Lienqueo for facilitating the retention times of the proteins used in this study and Dr. Barbara Andrews for critically reviewing the manuscript.

References

[1] W. Kauzmann, Adv. Protein Chem. 14 (1959) 1.
[2] M. Andrade, S. O'Donoghue, B. Rost, J. Mol. Biol. 276 (1998) 517.
[3] K. Berggren, A. Wolf, J.A. Asenjo, B.A. Andrews, F. Tjerneld, Biochim. Biophys. Acta 1596 (2002) 253.
[4] M.E. Lienqueo, A. Mahn, J.A. Asenjo, J. Chromatogr. A 978 (2002) 71.
[5] R. Cowan, R.G. Whittaker, Peptide Res. 3 (1990) 75.
[6] M.E. Lienqueo, A. Mahn, A. Olivera, J. Chromatogr. A (submitted for publication).
[7] T. Piližota, B. Lučić, N. Trinajstić, J. Chem. Inf. Comput. Sci. 44 (2004) 113.
[8] R.Y. Luo, Z.P. Feng, J.K. Liu, Eur. J. Biochem. 269 (2002) 4219.
[9] H.M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T.N. Bhat, H. Weissig, I.N. Shindyalov, P.E. Bourne, Nucleic Acids Res. 28 (2000) 235.
[10] U. Hobohm, M. Scharf, R. Schneider, C. Sander, Protein Sci. 1 (1992) 409.
[11] D. Frishman, P. Argos, Proteins 23 (1995) 566.
[12] S. Miller, J. Janin, A.M. Lesk, C. Chothia, J. Mol. Biol. 196 (1987) 641.
[13] K. Kaur, G.P. Raghava, Bioinformatics 20 (2004) 2751.
[14] M. Keil, T.E. Exner, J. Brickmann, J. Comput. Chem. 25 (2004) 779.
[15] J. Taskinen, J. Yliruusi, Adv. Drug Deliv. Rev. 55 (2003) 1163.
[16] Y.H. Xiang, M.C. Liu, X.Y. Zhang, R.S. Zhang, Z.D. Hu, B.T. Fan, J.P. Doucet, A. Panaye, J. Chem. Inf. Comput. Sci. 42 (2002) 592.