Predicting the behaviour of proteins in hydrophobic interaction chromatography
2. Using a statistical description of their surface amino acid distribution

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

[a] *Centre for Biochemical Engineering and Biotechnology, Department of Chemical and Biotechnology Engineering, University of Chile, Beauchef 861, Santiago, Chile*

[b] *Department of Mathematical Engineering, Centre for Mathematical Modelling, University of Chile, Blanco Encalada 2120, Santiago, Chile*

Abstract

This paper focuses on the prediction of the dimensionless retention time (DRT) of proteins in hydrophobic interaction chromatography (HIC) by means of mathematical models based on the statistical description of the amino acid surface distribution. Previous models characterises the protein surface as a whole. However, most of the time it is not the whole protein but some of its specific regions that interact with the environment. It seems much more natural to use local measurements of the characteristics of the surface. Therefore, the statistical characterisation of the distribution of an amino acid property on the protein surface was carried out from the systematic calculation of the local average of this property in a neighbourhood placed sequentially on each of the amino acids on the protein surface. This process allowed us to characterise the distribution of this property quantitatively using three main statistics: average, standard deviation and maximum. In particular, if the property considered is a hydrophobicity scale, these statistics allowed us to characterise the average hydrophobicity and the hydrophobic content of the most hydrophobic cluster or hotspot, as well as the heterogeneity of the hydrophobicity distribution on the protein surface. We tested the performance of the DRT predictive models based on these statistics on a set of 15 proteins. We obtained better predictive results with respect to the models previously reported. The best predictive model was a linear model based on the maximum. This statistic was calculated using an index of the mobilities of amino acids in chromatography. The predictive performance of this model (measured as the Jack Knife MSE) was 26.9% better than those obtained by the best model which does not consider the amino acid distribution and 19.5% better than the model based on the hydrophobic imbalance (HI). In addition, the best performance was obtained by a linear multivariable model based on the HI and the maximum. The difference between the experimental data and the prediction carried out by this model was smaller than those observed previously. In fact, this model obtained better predictive capacities than a previous linear multivariable model decreasing the Jack Knife MSE in 8.7%. In addition, this model allowed us to diminish the number of variables required, increasing, in this way, the degrees of freedom of the model.

Keywords: Mathematical modelling; Hydrophobic interaction chromatography; Hydrophobicity; Retention time prediction; Proteins; Protein surface distribution; Statistics

1. Introduction

The surface characteristics of a protein determine to a great extent its main properties. For example, protein functions such as catalysis or molecular recognition occur predominantly on or near the protein surface. In addition, these properties define the protein behaviour in purification stages of great importance in industry and at the laboratory scale, such as aqueous two-phase systems and hydrophobic interaction chromatography (HIC). In fact, it has been pointed out that the rational design of industrial protein purification processes normally requires an HIC stage [1].

Clearly, the characteristics of the protein surface will be defined by its topology and by the properties of the amino acids located on the surface as well as how those are distributed on it. Particularly, in the case of HIC, it has been shown that the dimensionless retention time (DRT) can be correlated with the average surface hydrophobicity (ASH) calculated as the average of the

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

hydrophobicity of each amino acid on the surface corrected by its abundance [2]. It is important to note that this model did not consider the amino acid distribution on the protein surface. However, theoretical studies, using simplified protein models, have shown that proteins with a heterogenous hydrophobicity distribution on their surface can establish stronger interactions with a hydrophobic ligand than those with a homogenous distribution [3].

The model of Lienqueo et al. [4] based on the ASH has proven to be effective on the DRT prediction for several proteins. However, there are proteins whose surface distribution prevents their correct handling by this model. Mahn et al. [5] reported four proteins for which the model of Lienqueo is deficient. In order to deal with these proteins two models were proposed: one of experimental nature [5] and another theoretical [6]. The experimental model used a hydrophobic contact area (HCA) determined through a thermodynamic model that combined electrostatic, hydrophobic interactions and data determined in laboratory [7]. The theoretical model used a local hydrophobicity (LH) calculated considering only the amino acids located inside the most probable interaction zone between protein and stationary matrix, which was determined using molecular docking simulations. The main disadvantage of both methodologies is that they are very expensive in human and computational resources.

On the other hand, we previously introduced a vector called hydrophobic imbalance (HI) [8]. This vector, obtained from the characteristics of the protein surface, represents the displacement of the surface geometric centre of the protein when the effect of a certain amino acid hydrophobicity scale is considered. Therefore, the HI was used as a simple measurement of the characteristics of the hydrophobicity distribution on the protein surface. The interpretation of the HI is not trivial because the number of effects that could take part in the calculation of its magnitude prevent a direct interpretation. Nevertheless, using the HI we obtained correlation coefficients remarkably better (at least 67%) than models based on the local hydrophobicity and the hydrophobic contact area. In addition, the linear combination of the HI and other parameters allowed the development of a multivariable model which improved the predictive performance (quantified by the Jack Knife cross validation mean square error, MSE) by 24.9% with respect to the best model based on HI only and 31.8% with respect to the model based on ASH only. The correlation coefficient obtained for the multivariable model was 0.899.

In this article we propose the statistical description of the surface amino acid distribution to predict the DRT of proteins in HIC in a similar approach to those used by Jönsson et al. [3]. Jönsson et al., using very simple models for polymers and proteins, showed that the statistical quantification of the heterogeneity degree of the protein surface can be related to its adsorption on polymers. In fact, a strong correlation between the adsorption ability and the degree of heterogeneity of several protein models was found [3].

Keeping this in mind, the main objective of this article is to investigate if the statistical description of the protein surface, as a way to incorporate information about the amino acid surface distribution, allows the development of simple and computationally inexpensive mathematical models which can improve the performance of the prediction of DRT reported previously.

2. Methodology

2.1. Local average surface property

Let S be the surface of a protein. We code S by a set of points. Each point $k \in S$ is, for us, a particular amino acid. For each of these amino acids $k \in S$, ASA(k) corresponds to its accessible surface area. We also define $\varphi(k)$ as the value of an intrinsic aminoacidic property of k. The value of $\varphi(k)$ is given by an amino acid property vector (APV) (for instance, APV could be a hydrophobicity scale). The average surface property (ASP) of a protein is given by:

$$\mathrm{ASP} = \frac{\sum_{k \in S} \mathrm{ASA}(k)\varphi(k)}{\sum_{k \in S} \mathrm{ASA}(k)} \tag{1}$$

If the APV (from where the values of $\varphi(k)$ are taken) is simply a hydrophobicity scale, then the calculated ASP corresponds to the average hydrophobic contribution of each amino acid weighted by its accessible surface area. This quantity has been used to develop DRT predictive models previously [2,9,10]. The ASP of a protein is computed assuming that each amino acid on the protein surface contributes proportionally to its abundance to the ASP value [11]. Details of the ASP calculation appear in a previous work [8].

The ASP characterises the protein surface as a whole. Nevertheless, most of the time it is not the whole protein but some of its specific regions that interact with the environment. The introduction of local measures of ASP seem much more natural. We define the local ASP for each amino acid k located on the surface as follows:

$$\mathrm{ASP}(k) = \frac{\sum_{i \in \mathcal{N}r(k)} \mathrm{ASA}(i)\varphi(i)}{\sum_{i \in \mathcal{N}r(k)} \mathrm{ASA}(i)} \tag{2}$$

where $\mathcal{N}_r(k)$ is a neighbourhood of radius r around the amino acid k. A neighbourhood $\mathcal{N}_r(k)$ is defined as the set of all amino acids located on the protein surface and inside a ball of radius r centered on the amino acid k. In order to simplify the calculations the location of each amino acid was chosen to be equal to the location of its β-carbon (except for glycine, where its α-carbon was used). We chose the location of the β-carbon (instead of α-carbon) since this atom gives a better idea of the amino acid orientation with respect to the protein backbone.

The local ASP was calculated for all the amino acids on the protein surface for different values of r. Fig. 1 shows an example. Clearly, if a protein with L amino acids on its surface is considered and a set of R neighbourhood sizes is used, the number of times that the local ASP must be calculated is $L \times R$. In this way, the distribution of the local ASP on the surface depends on the value of r, the size of the neighbourhood considered in its calculation.

From each local ASP distribution we extracted three main statistics: the average $\mathrm{ASP}_{\mathrm{avg}}$, the standard deviation $\mathrm{ASP}_{\mathrm{std}}$ and the maximum $\mathrm{ASP}_{\mathrm{max}}$. Two linear combinations were also con-

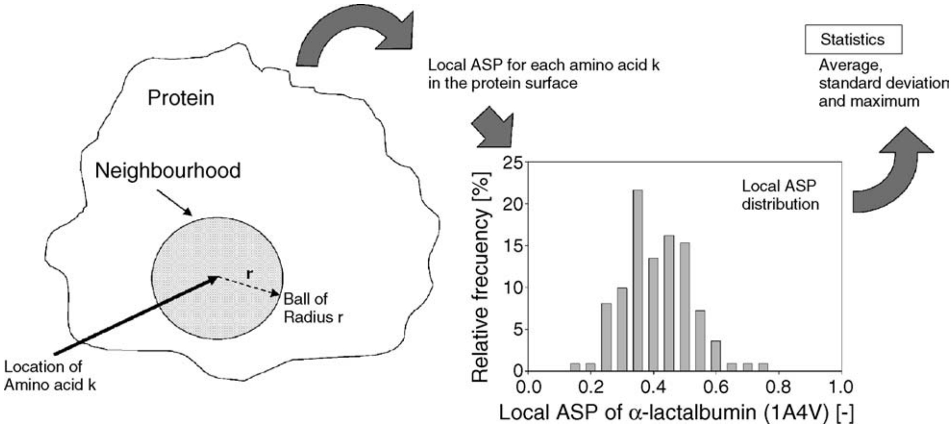

Fig. 1. Characterisation of the distribution of an amino acid property on the protein surface. The distribution was determined from the study of local ASP calculated on a neighbourhood of radius r $\mathcal{N}_r(k)$ around each amino acid k. The local ASP was calculated for all the amino acids on the protein surface allowing the quantitative determination of the surface distribution of this property for a particular neighbourhood radius. In particular, the figure shows the quantification of the distribution of the APV of Aboderin [15] on the surface of α-lactalbumin (1A4V) when a neighbourhood of radius 11 Å is considered.

sidered: $ASP_{max\text{-}min}$ and $ASP_{max\text{-}avg}$. Thus, each of the R local ASP distributions was characterized by five variables.

Following the previous analysis, if a set of R neighbourhood sizes is used, then each protein in the dataset will be represented by an $R \times 5$ matrix, where each row of the matrix contains the statistics calculated for a given radius. Again, if the APV is a hydrophobicity scale, then these statistics allow us to characterise the hydrophobicity distribution on the protein surface. ASP_{avg} and ASP_{max} give the average hydrophobicity and the hydrophobic content of the most hydrophobic cluster or hotspot, respectively. On the other hand, ASP_{std}, $ASP_{max\text{-}min}$ and $ASP_{max\text{-}avg}$, quantified the heterogeneity of the hydrophobicity distribution on the protein surface.

A synthesis of the procedure used for the determination of the statistics written in pseudocode follows:

```
P = [$P_1$, $P_2$, $P_3$, … , $P_{15}$]
r = [5, 6, 7, … , 25] Å
Pick a protein P from the database
For each radius r in the set of neighbourhood sizes do
    For each amino acid k located on the protein surface s.t. ASA(k) > 0 do
        Determine the amino acids located inside $\mathcal{N}_r(k)$
        Calculate the ASP(k) using equation (2)
    Build the local ASP distribution for radius r using the ASP(k) values
    Compute the $ASP_{avg}$, $ASP_{std}$ and $ASP_{max}$ from the local ASP distribution
End
```

Our approach is similar to that of Jönsson et al. [3]. However, Jönsson et al., used very simple protein models, corresponding to spheres where the amino acids have a binary hydrophobicity and equal level of accessible surface area. Additionally, the determination of the hydrophobicity distribution on the protein surface used by Jönsson et al. was made through a random sampling and not through exhaustive analysis as in our case.

2.2. Protein set

Fifteen proteins with known dimensionless retention time and known three-dimensional structure were used: Cytochrome *C* (1HRC), Myoglobin (1YMB), Conalbumin (1OVT), Ovoalbumin (1OVA), Lysozyme (2LYM), Thaumatin (1THV), Chymotrypsinogen A (2CHA), β-lactoglobulin (1CJ5), α-amylase (1BLI), α-chymotrypsin (4CHA), α-lactalbumin (1A4V), Ribonuclease S (1RBC), Ribonuclease A (1AFU), Ribonuclease T1 wild type (1RGC) and Ribonuclease T1 variant Y45W/W59Y (1TRP).

The three-dimensional structures were obtained from the PDB database [12] and the ASA was calculated using the software STRIDE from the protein three-dimensional structure [13]. DRT data correspond to those used by Lienqueo et al. [2] and Mahn et al. [6] and they are the DRTs observed in a hydrophobic interaction column, calculated as described in a previous work [8].

2.3. Collection of aminoacidic property vectors (APVs)

A collection of 74 APVs was used. This collection covered a wide spectrum of physical, chemical and biological aminoacidic characteristics. Amongst them: molecular weight, bulkiness, hydrophobicity scales, average solvent accessibility, secondary structure preferences, codon numbers, etc. [11,14–55]. All members in the APVs collection were numerically scaled in the interval [0; 1]. This scaling procedure was carried out so that values 0 and 1 were associated to the minimum and maximum values in the original scale, respectively. The hydrophilicity scales were transformed into hydrophobicity scales assigning 0 to the most hydrophilic amino acid and 1 to the most hydrophobic (the values for the rest of the amino acids were determined linearly). Vectors not associated to hydrophobicity scales were not modified.

2.4. Measurement of the performance of the predictive models

Our goal is to quantify the performance of the statistics $\mathrm{ASP}_{\mathrm{avg}}$, $\mathrm{ASP}_{\mathrm{std}}$, $\mathrm{ASP}_{\mathrm{max}}$, $\mathrm{ASP}_{\mathrm{max\text{-}min}}$ and $\mathrm{ASP}_{\mathrm{max\text{-}avg}}$ as predictors of the dimensionless retention time. This performance was evaluated by means of three parameters: the mean square error, the correlation coefficient (Pearson) and the Jack Knife cross validation mean square error ($\mathrm{MSE}_{\mathrm{JK}}$). These parameters were calculated using the equations and methodology presented in the previous study [8].

3. Results and discussion

In this section the results obtained when using the statistical description of the protein surface characteristics as a tool to predict its dimensionless retention time in hydrophobic interaction chromatography are described.

3.1. Calculation of the statistics using simple hydrophobicity scales

As in our previous work [8], we started considering only very simple hydrophobicity scales. Three scales were used:

- *Hard binary scale*: It assigns a value of 1 to the amino acids widely accepted as hydrophobic (Ala, Ile, Leu, Phe, Pro, Val) and 0 to the rest.
- *Soft binary scale*: As the previous one but it also considers the amphipathic amino acids (Lys, Met, Thr, Trp, Tyr) as hydrophobic (assigning a value of 1 to them).
- *Trinary scale*: It assigns a value of 0.5 to the amphipathic amino acids, 1 to the hydrophobic, and 0 to the rest.

Table 1
Correlation coefficients (Pearson) between the dimensionless retention time (DRT) and the statistics considered in this study determined on the 15 protein set

Statistic	Pearson	Radius (Å)	Hydrophobicity scale
$\mathrm{ASP}_{\mathrm{max\text{-}min}}$	0.701	11	HBS
$\mathrm{ASP}_{\mathrm{max}}$	0.675	11	HBS
$\mathrm{ASP}_{\mathrm{max\text{-}avg}}$	0.652	11	HBS
$\mathrm{ASP}_{\mathrm{std}}$	0.644	11	HBS
$\mathrm{ASP}_{\mathrm{avg}}$	0.574	10	HBS

The statistics were calculated for the 15 proteins set considering neighbourhoods with radii between 5 and 25 Å, with 1 Å steps. The best correlation coefficient between the dimensionless retention time and these characteristics are shown in Table 1. These results indicate that the statistics considered in this work are better correlated to the DRT than the parameters used in previous studies as, for instance, the average surface hydrophobicity which is the ASP calculated using a hydrophobicity APV. Actually, $\mathrm{ASP}_{\mathrm{avg}}$ presents a correlation coefficient 14.1% greater than those obtained for the ASP calculated using the hard binary scale (HBS). Both magnitudes are very similar. In fact, the correlation coefficient between $\mathrm{ASP}_{\mathrm{avg}}$ and ASP is 0.962. However, the way in which $\mathrm{ASP}_{\mathrm{avg}}$ is calculated could allow a slight correction when the protein has regions with very low average hydrophobicity. This fact explains the somewhat better results shown by $\mathrm{ASP}_{\mathrm{avg}}$ with respect to the traditional ASP.

The sign of the correlation coefficients is positive for all the statistics listed in Table 1. This observation is coherent with those reported in the literature. $\mathrm{ASP}_{\mathrm{avg}}$ and $\mathrm{ASP}_{\mathrm{max}}$ measure the average and the maximum hydrophobicity on the protein surface. In most cases, the greater the global hydrophobicity the greater the DRT. In addition, it has been reported that the presence of clusters with high hydrophobicity on the protein surface favours the interaction of the protein with the HIC stationary matrix [3,5,6]. In fact, $\mathrm{ASP}_{\mathrm{max}}$ quantifies the hydrophobicity in those zones. On the other hand, $\mathrm{ASP}_{\mathrm{std}}$, $\mathrm{ASP}_{\mathrm{max\text{-}min}}$ and $\mathrm{ASP}_{\mathrm{max\text{-}min}}$ measure degree of heterogeneity of the surface hydrophobicity distribution. A high value of these parameters indicates a high heterogeneity. It has been reported that a big hydrophobic patch accessible to the hydrophobic matrix favours the interaction with the matrix and thus a high retention time in HIC would be expected [5].

It is interesting to note that the better correlation coefficients shown in Table 1 were found mainly in neighbourhoods with radii between 10 and 11 Å. This fact suggests a certain level of coherency in the amount of information required for the calculation of these parameters. A neighbourhood of 11 Å contains 19.0 ± 1.9 amino acids with ASA > 0. Certainly, this number corresponds only to a basic reference, since the contribution to the protein hydrophobicity of each amino acid in this neighbourhood will be very different. In fact, the contribution of some amino acids will be insignificant due to their small ASA. The model chose a medium size neighbourhood. Smaller neighbourhoods could introduce an excessive amount of noise in the parameters making them too sensitive to local disturbances on the surface hydrophobicity distribution.

Table 2
Correlation coefficients between the statistics considered in this study determined on the 15 protein set and calculated using the hard binary scale

	$ASP_{max\text{-}min}$	ASP_{max}	$ASP_{max\text{-}avg}$	ASP_{std}	ASP_{avg}
$ASP_{max\text{-}min}$	1	**0.994**	**0.979**	**0.890**	0.714
ASP_{max}		1	**0.978**	**0.848**	0.738
$ASP_{max\text{-}avg}$			1	**0.836**	0.582
ASP_{std}				1	0.597
ASP_{avg}					1

Correlation coefficients greater than 0.8 have been highlighted in bold.

Table 3
Correlation coefficients (Pearson) between the dimensionless retention time (DRT) and the average surface hydrophobicity (ASH), local hydrophobicity (LH), hydrophobic contact area (HCA), hydrophobic imbalance (HI) and statistics

Parameter	Pearson
ASH	−0.528
LH[a]	0.557
HCA[a]	0.483
HI[b]	−0.940
$ASP_{max\text{-}min}$	0.908
ASP_{max}	0.933
$ASP_{max\text{-}avg}$	0.938
ASP_{std}	0.624
ASP_{avg}	−0.309

The HI was reported in [8]. The ASH, LH and HCA were reported in [5,6]. The HI and statistics were calculated using the hard binary scale. The calculations only considered the following proteins: Ribonuclease S (1RBC), Ribonuclease A (1AFU), Ribonuclease T1 wild type (1RGC) and Ribonuclease T1 variant Y45W/W59Y (1TRP).

[a] Mahn et al. [6].

[b] Salgado et al. [8].

Additionally, all features in Table 1 preferred the hard binary scale to represent the amino acid hydrophobicity. This observation confirms results found previously [8] and it indicates that the best results were obtained when the hydrophobicity of the amphipathic amino acids was defined as hydrophilic (0.0); the hydrophobic ones as 1.0 and the hydrophilic ones as 0.0. This fact stresses the need for a more complex hydrophobicity scale.

The relationships between the statistics considered in this study are shown in Table 2. In this case, the relationship between two parameters was measured using the correlation coefficient between these magnitudes calculated for the 15 proteins for the hard binary scale. This table shows that ASP_{avg} is considerably different from the other variables. On the other hand, the rest of the variables display quite high correlations. It is important to note that in some cases $ASP_{max\text{-}min}$ will be very similar to ASP_{max} because ASP_{min} can be zero. On the other hand, the high correlation between ASP_{max} and ASP_{std} is notorious, and it means that both variables must be dealt with carefully in a multivariable model.

Table 3 shows the correlation coefficient between DRT and the statistics for a small protein set that contains only four proteins with similar average surface hydrophobicity and very different DRTs. These proteins were the same as those used by Mahn et al. [5]: Ribonuclease S (1RBC), Ribonuclease A (1AFU), Ribonuclease T1 wild type (1RGC) and Ribonuclease T1 variant Y45W/W59Y (1TRP). The Ribonuclease T1 variant has two surface amino acids interchanged, altering, in this way, the distribution of hydrophobic amino acids without changing the average surface hydrophobicity. Additionally, the correlation coefficients between DRT and LH or HCA reported by Mahn et al. [6] and amongst the DRT and the hydrophobic imbalance reported by Salgado et al. [8] are also shown in Table 3. The correlation coefficients obtained by the statistics in this small protein set are almost twofold those obtained for the LH and HCA and slightly smaller than those obtained for the HI. The results obtained by the statistics justify a further study of these parameters.

3.2. *Calculation of the statistics using the collection of aminoacidic property vectors (APVs)*

The prediction of the DRT by means of the statistics now calculated using the 74 APVs was tackled. This APV collection covered a wide spectrum of physical, chemical and biological aminoacidic characteristics. Amongst them: molecular weight, bulkiness, hydrophobicity scales, average solvent accessibility, secondary structure preferences, codon numbers, etc. The predictors were constructed using a linear model on the statistics. The predictive capacity of these models was characterised by means of the determination of the Jack Knife cross validation mean square error (MSE_{JK}) on the set of 15 proteins. The results from these experiments are shown in Tables 4 and 5.

The performance of the statistics in ascending order with respect to the MSE_{JK} is shown in Table 4. This table indi-

Table 4
Performance indices of the linear model based on the statistics on the prediction of the experimental DRT of 15 proteins

Statistic	APV	Description	Radius (Å)	MSE × 10^3	Pearson	MSE_{JK} × 10^3
ASP_{std}	Zimmerman [14]	Polarity	18	8.118	0.919	**12.337**
$ASP_{max\text{-}avg}$	Bhaskaran and Ponnuswamy [17]	Average flexibility index	19	15.045	0.844	**21.822**
ASP_{max}	Aboderin [15]	Mobilities of amino acids on chromatography paper	19	16.530	0.827	**22.712**
$ASP_{max\text{-}min}$	Bhaskaran and Ponnuswamy [17]	Average flexibility index	11	20.457	0.780	**28.945**
ASP_{avg}	Lifson and Sander [18]	Conformational preference for total β strand (antiparallel + parallel)	11	19.749	0.788	**29.061**

The best model for each feature along with the aminoacidic property vectors (APV) selected for it are listed in ascending order with respect to the Jack Knife cross validation mean square error (MSE_{JK}). The correlation coefficient (Pearson), the mean square error (MSE) and the neighbourhood radii are also shown. MSE_{JK} values have been highlighted in bold.

Table 5
Performance indices of the linear model based on the statistics on the prediction of the experimental DRT of 15 proteins

Statistic	APV	Description	Radius (Å)	MSE × 10^3	Pearson	MSE_{JK} × 10^3
ASP_{max}	Aboderin [15]	Mobilities of amino acids on chromatography paper	19	16.530	0.827	**22.712**
ASP_{avg}	Browne [19]	Retention coefficient in TFA	25	22.423	0.755	**29.368**
$ASP_{max-min}$	Wertz and Scheraga [20]	Fraction of buried amino acid on 20 proteins	11	27.327	0.690	**35.113**
$ASP_{max-avg}$	Bull and Breese [21]	Hydrophobicity (free energy of transfer to surface in kcal/mol)	21	26.234	0.705	**35.763**
ASP_{std}	Guy [22]	Hydrophobicity scale based on free energy of transfer (kcal/mol)	13	28.941	0.668	**36.478**

Only models with a positive slope and which use an APV related directly to measurements of the amino acids hydrophobicity were included. The best model for each feature along with the aminoacidic property vectors (APV) selected are listed in ascending order with respect to the Jack Knife cross validation mean square error (MSE_{JK}). The correlation coefficient (Pearson), the mean square error (MSE) and the neighbourhood radius are also shown. MSE_{JK} values have been highlighted in bold.

cates that the best parameter for the prediction of DRT was ASP_{std}. The linear model associated to that parameter was: $(1.800 \pm 0.319) - (17.276 \pm 4.441)ASP_{std}$ and the APV used was the APV of Zimmerman et al. [14] which quantifies the amino acid polarity. The ASP_{std} coefficient in this model was negative indicating that a protein with a larger ASP standard deviation on its surface, and hence a higher surface heterogeneity, will have a smaller DRT than another homogenous one. This behaviour was in opposition to that observed in the previous section and to that reported in the literature. These facts forced us to discard the models which use the APV of Zimmerman as a measurement of the amino acid hydrophobicity. The opposite behaviour observed in the models that use the APV of Zimmerman can be explained by the way in which this vector quantifies the amino acid polarity. The Zimmerman polarity scale assigns an extremely high value to the charged amino acids (Arg, Asp, Glu, His and Lys), being this value, approximately, one order of magnitude greater than the rest of the hydrophilic or amphipathic amino acids. This fact has the consequence that the value of the polarity index of a substantial part of the hydrophilic or amphipathic amino acids is very similar to the ones assigned to the hydrophobic amino acids.

In agreement with the previous discussion, Table 5 was constructed. Only models with a positive slope and which use an APV related directly to measurements of the amino acid hydrophobicity were included in the table. Clearly, this operation of selection modified the order of the variables observed in Table 4. In this case, the best model is based on the variable ASP_{max} followed by the models constructed on the basis of the variables ASP_{avg} and $ASP_{max-min}$. The ASP_{max} model selected the APV of Aboderin [15] which is an index of the mobilities of amino acids in chromatography. The model based on ASP_{max} presented an MSE_{JK} 26.9% better than the one obtained by the best ASP model and 19.5% better than the model based on the hydrophobic imbalance, both values reported in [8].

In addition, the ASP_{avg} probed to be slightly better than its global counterpart ASP improving the MSE_{JK} in only 5.4%. The radius selected by this model was the upper limit of the neighbourhood sizes (25 Å). When relaxing this upper limit, a minimum MSE_{JK} at the radius of 39 Å was found. In these conditions the predictive capacity was improved by 10.8% with respect to the one obtained by the ASP. This model used the APV of Meek [16] which quantifies retention coefficients in HPLC at pH 2.1. These results confirm only a part of the observations made in the previous section. The difference between the predictive capacity obtained by ASP_{avg} and the one by ASP is similar to those obtained in the previous section. Nevertheless, in this case the radius selected by the model is almost four times larger. A neighbourhood of 39 Å contains an average of 158.1 ± 74.2 amino acids on the surface (ASA > 0). So, the large number of amino acids included in this neighbourhood indicates that for medium sized proteins in the database (length ≈ 150 aa) the ASP_{avg} will be almost equal to the ASP. In fact, 9 of 15 proteins in the database have an average of 99% of their amino acids inside this neighbourhood. Consequently, in the case of this model the ASP_{avg} will be different from ASP only in the case of bigger proteins, such as: 1BLI, 1OVA, 1OVT, 1THV, 2CHA and 4CHA. This behaviour can be explained by the fact that in bigger proteins, and therefore with larger surfaces, there is a higher probability of finding a larger hydrophobic heterogeneity on the protein surface. In those cases, the ASP_{avg} would be useful.

A significant amount of difference between the radii selected by the models was observed. The ASP_{max} model selected a radius of 19 Å, whereas the ASP_{avg} model a radius of 39 Å. A neighbourhood of 19 Å contains an average of 63.7 ± 11.4 amino acids on the surface. The neighbourhood size difference can be explained on the basis of the nature of the variables. For instance, the ASP_{max} needs a medium size neighbourhood to be able to detect a hydrophobic hotspot or cluster on the protein surface.

3.3. Calculation of the statistics using the collection of aminoacidic property vectors (APVs) in the hydrophobic hemisphere

The effect in the predictive capacity of linear models when the statistics were calculated in the hydrophobic hemisphere was investigated. The hydrophobic hemisphere was defined as the subset of amino acids located in the protein hemisphere pointed out by the hydrophobic imbalance vector [8]. Briefly, the HI vector obtained from the characteristics of the protein surface, represents the displacement of the surface geometric centre of the protein when the effect of a certain amino acid

Table 6
Performance indices of the linear multivariable models based on the prediction of the experimental DRT of 15 proteins

No.	Statistics	APV	Description	Radius (Å)	MSE × 10^3	Pearson	MSE_{JK} × 10^3	DF	R^2_{adj} (%)
1	HI, ASP_{max}	Aboderin [15]	Mobilities of amino acids on chromatography paper	19	12.606	0.871	**19.329**	12	0.718
2	HI, ASP_{max}	Grantham [23]	Polarity	16	14.454	0.850	**22.682**	12	0.677
3	HI, ASP_{max}	Meek [16]	Retention coefficient in HPLC, pH 7.4	11	13.920	0.856	**23.313**	12	0.689
4	ASP_{avg}, ASP_{max}	Aboderin [15]	Mobilities of amino acids on chromatography paper	19	15.983	0.833	**23.650**	12	0.643

Only models with a positive slope and which use an APV related directly to measurements of the amino acid hydrophobicity were included. The four best models in ascending order with respect to the Jack Knife cross validation mean square error (MSE_{JK}) are listed. The correlation coefficient (Pearson), the mean square error (MSE), the neighbourhood radii, the degrees of freedom (DF) and the adjusted determination coefficient (R^2_{adj}) are also shown. MSE_{JK} values have been highlighted in bold.

hydrophobicity scale is considered. The results show that the restriction of the amino acids to only those considered inside the hydrophobic hemisphere has a negative effect on the predictive capacity of these variables. In fact, the performance of the models based on ASP_{max} and $ASP_{max\text{-}min}$ was worse. In the case of ASP_{max} its MSE_{JK} increased by 30% with respect to the value found in the previous section. For the rest of the variables the MSE_{JK} decreased, but at the cost of selecting a very small radius (5 Å) indicating that these improvements correspond to model artifacts.

The results obtained in this section indicate that the performance of the statistics as DRT predictors is related directly to the amount of information used for their calculation. These statistics require all of the available information for their determination.

3.4. Multivariable models based on the statistics

In this section the results obtained using linear combinations of the hydrophobic imbalance, average surface properties and the statistics to predict the DRT are described. The objective is to find out whether the linear combination of these variables is able to improve the results obtained in Section 3.2 by the model based on ASP_{max}.

All the combinations of these variables were systematically tested. Nevertheless those models which considered APVs not related directly to hydrophobicity, as well as those whose coefficients did not present the expected sign (for example, a negative coefficient for ASP_{std}), were eliminated. The results obtained in this operation appear in Table 6.

It is interesting to note that the best models were constituted by HI and ASP_{max}, only differing in the APV and in the radii selected. These results confirm the importance of HI in the prediction of DRT reported in a previous study [8]. In addition, the presence of ASP_{max} in all models confirms the results obtained in this paper. The best model used the APV of Aboderin [15] which quantifies mobilities of amino acids in chromatography. This APV was the same selected by the best linear model based on ASP_{max} determined in the previous section. Also the radius selected by the model was kept.

The relative importance of HI and ASP_{max} in the multivariable model is shown in Fig. 2. This figure shows the changes in the predictive capacity of the model (quantified as its MSE_{JK}) when removing each one of the variables. Clearly, the most important variable in the model is ASP_{max}. Its removal implies an increase of almost 3.4-fold the original value of MSE_{JK}. On the other hand, the removal of HI only produces an increase of the MSE_{JK} of 17.5%. Even though this decrease in the predictive quality of the model cannot be disregarded, it is significantly smaller than the one observed when removing ASP_{max}.

The use of linear multivariable models allowed the improvement of the results obtained in the previous section. In fact, the best multivariable model improve the previous results by decreasing the MSE_{JK} in 14.9%.

3.5. Final discussion

The best DRT predictive model found in this work was the linear multivariable model that follows:

$$\begin{aligned} DRT = &-(1.748 \pm 0.827) - (0.164 \pm 0.185) \times HI \\ &+ (5.937 \pm 2.179) \times ASP_{max} \end{aligned} \tag{3}$$

where, DRT is the dimensionless retention time, HI is the hydrophobic imbalance and ASP_{max} is the greater ASP value observed in a neighbourhood of radius equal to 19 Å. HI and ASP_{max} were calculated using the APV of Aboderin [15], which is shown in Table 7.

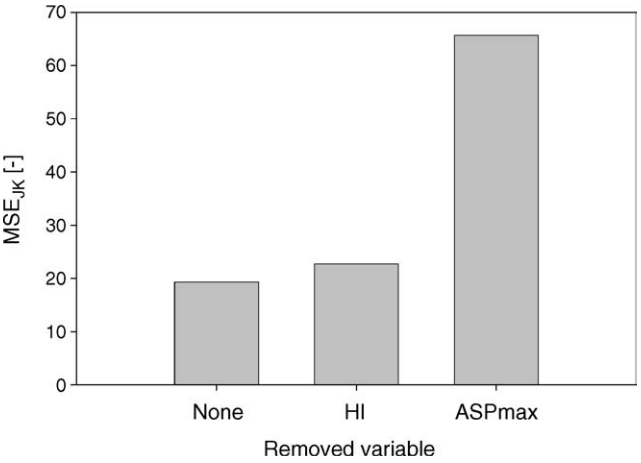

Fig. 2. Effect of the removal of each one of the variables of the multivariable model in its predictive capacity, measured as the observed value of Jack Knife cross validation mean square error (MSE_{JK}) in the set of 15 proteins.

Table 7
Amino acidic property vector (APV) of Aboderin [15]

aa	Original	Scaled to (0; 1)
Ala	5.10	0.51
Arg	2.00	0.20
Asn	0.60	0.06
Asp	0.70	0.07
Cys	0.00	0.00
Gln	1.40	0.14
Glu	1.80	0.18
Gly	4.10	0.41
His	1.60	0.16
Ile	9.30	0.93
Leu	10.00	1.00
Lys	1.30	0.13
Met	8.70	0.87
Phe	9.60	0.96
Pro	4.90	0.49
Ser	3.10	0.31
Thr	3.50	0.35
Trp	9.20	0.92
Tyr	8.00	0.80
Val	8.50	0.85

The confidence intervals at 95% determined for the parameters of the model did not exceed a 50% of their nominal values, with exception of HI. In fact, the uncertainty in the determination of the HI coefficient was the highest, reaching 113% in relation to the nominal value. Nevertheless, the *p*-value associated to HI was 0.077, since this value being less than 0.1, that term is statistically significant at a 90% confidence level. Given the data characteristics, this level of significance is still acceptable. In addition, it is interesting to highlight that the sign of the coefficient for HI is negative, maintaining therefore the behaviour observed previously [8].

Fig. 3 shows the scatter plots between the experimental DRT and the predictions carried out by the ASP model, a linear multivariable model reported in a previous work (A) [8] and the linear multivariable model developed in this work (B). This plot shows that, in general, the difference between the experimental value and the prediction carried out by the models is smaller in the case of the linear multivariable model developed in the present article (B). In fact, model B obtained better predictive capacities than model A, decreasing the MSE_{JK} by 8.7%. Nevertheless, in the case of model A, the error is distributed in a more uniform way than in model B, being observed in that case an outlier with DRT $\approx$ 0.8. This is clearly indicated by the distribution of the residual error for the predictive models shown in Fig. 4. The outlier in model B corresponds to the protein RNAse S (1RBC). The unusual behaviour of this protein was reported previously by Mahn et al. [5] and attributed to its great flexibility. However, if we took into consideration only the four ribonucleases reported by Mahn et al. which show unusual behaviour when modelling DRT by ASP only, the correlation coefficient for the multivariable model B was 0.901, slightly inferior to the one observed in the model A. On the other hand, this correlation coefficient was 75.4% and 102.3% greater than the correlation coefficients of the models based on the LH and HCA, respectively.

Finally, a direct relation was not observed between the residual magnitude and the protein length or with the value of the DRT, in fact, the correlation coefficient between these magnitudes were inferior to 0.300 in both cases.

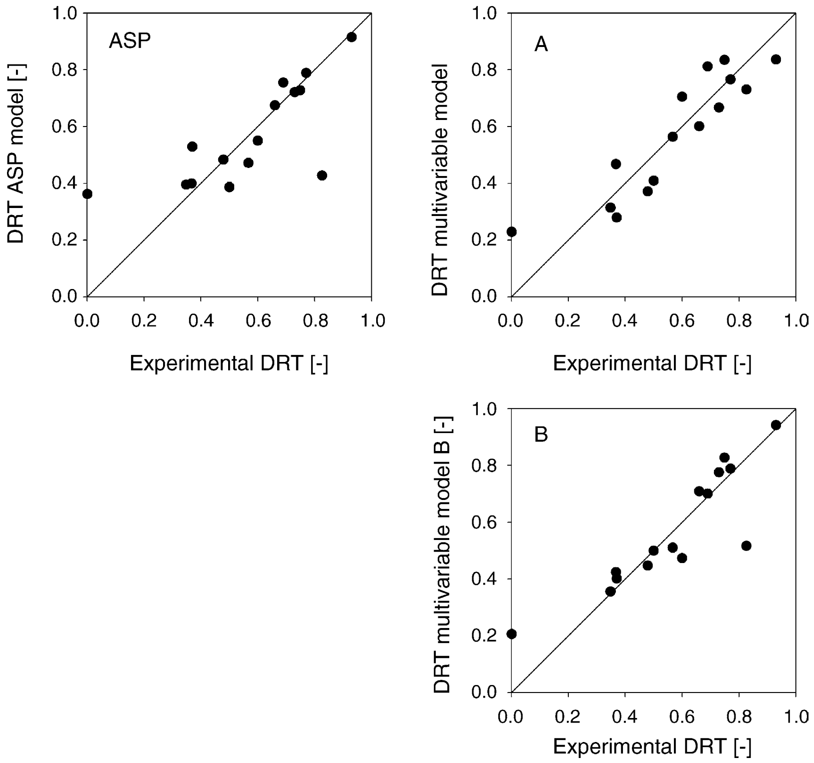

Fig. 3. Scatter plots between the experimental dimensionless retention time (DRT) and DRT estimated by the ASP model and two multivariable models. The ASP model and the multivariable model A were described previously [8] and the multivariable model B was developed in this work.

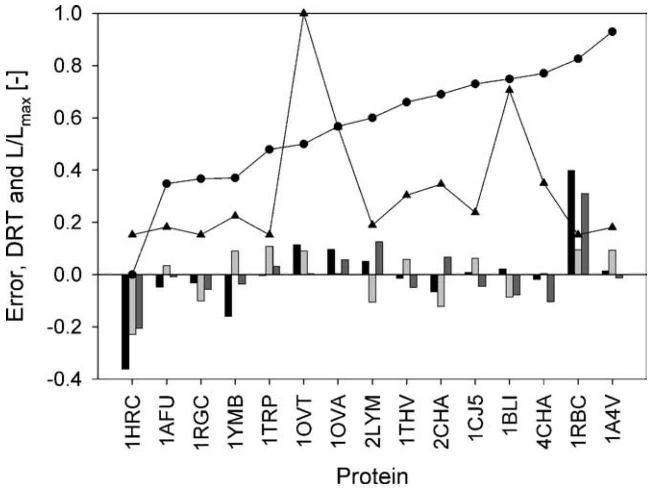

Fig. 4. Plot of the residual error between the experimental dimensionless retention time (DRT) and DRT estimated by the ASP model (■), the multivariable model A (□) and the multivariable model B (▩). The experimental DRT (●), and the dimensionless length (▲) are also shown. The multivariable model A was described previously [8] and the multivariable model B was developed in this work. The proteins are arranged in ascending order with respect to their DRT.

4. Conclusions

In this article the use of a statistical description of the surface amino acid distribution in order to predict the behaviour of proteins in hydrophobic interaction chromatography was investigated. The statistics obtained from the statistical characterisation of the amino acid surface distribution were used to model the DRT of four ribonucleases reported in [5] with similar ASP and very different DRTs and therefore with a DRT hard to predict using only the ASP. These calculations were carried out using simple hydrophobicity scales. The correlation coefficients obtained in this way are almost twofold those obtained for the models based on the local hydrophobicity [6] and the hydrophobic contact area [5] and slightly smaller than the one obtained for the hydrophobic imbalance [8].

The DRT predictive capacity of linear models constructed on the basis of the statistics was also analysed. In this case the statistics were calculated using a collection of 74 aminoacidic property vectors. The results obtained by these models were in general superior to the ones reported previously. The best linear model was obtained with ASP$_{max}$ calculated using the APV of Aboderin [15] which is an index of the mobilities of amino acids in chromatography. This model gave an MSE$_{JK}$ 26.9% better than the one obtained by the best ASP model and 19.5% better than the model based on the hydrophobic imbalance, both values were reported previously [8]. This result is in agreement with those previously reported: the presence of clusters with high hydrophobicity in the protein surface favours the interaction of the protein with the HIC stationary matrix [3,5,6] and ASP$_{max}$ quantifies directly the hydrophobicity in those zones indeed.

The best performance was obtained by a linear multivariable model based on HI and ASP$_{max}$ calculated using the APV of Boderin. The difference between the experimental value and the prediction carried out by this model was smaller than those observed in previous models. In fact, this model obtained better predictive capacities than a previous linear multivariable model [8] decreasing the MSE$_{JK}$ by 8.7%. In addition, this model allowed a decrease in the number of variables required from three to two increasing in this way the degrees of freedom of the model.

We found that the statistical characterisation of the amino acid surface distribution allows the prediction of the dimensionless retention time of proteins with an acceptable level for many practical applications (correlation coefficients >0.8). The best predictive model developed in this article was a multivariable model, such as in our previous work [8]. Although, the reduction of the degrees of freedom (from 13 to 12) and the increase in the complexity of the model with respect to the linear model based on HI is moderate. The improvement of the predictive capacity is not particularly important.

Acknowledgements

We wish to thank Dr. Maria Elena Lienqueo for facilitating the dimensionless retention times of the proteins used in this study.

This work was supported by the Fondef project 011031 and the postgraduate scholarship of CONICYT.

References

[1] J.A. Asenjo, B.A. Andrews, J. Mol. Recognit. 17 (2004) 236.
[2] M.E. Lienqueo, A. Mahn, J.A. Asenjo, J. Chromatogr. A 978 (2002) 71.
[3] M. Jönsson, M. Skepö, F. Tjerneld, P. Linse, J. Phys. Chem. B 107 (2003) 5511.
[4] M.E. Lienqueo, A. Mahn, L. Vásquez, J.A. Asenjo, J. Chromatogr. A 1009 (2003) 189.
[5] A. Mahn, M.E. Lienqueo, J.A. Asenjo, J. Chromatogr. A 1043 (2004) 47.
[6] A. Mahn, G. Zapata-Torres, J.A. Asenjo, J. Chromatogr. A 1066 (2005) 81.
[7] W. Melander, D. Corradini, Cs. Horváth, J. Chromatogr. 317 (1984) 67.
[8] J.C. Salgado, I. Rapaport, J.A. Asenjo, J. Chromatogr. A 1107 (2006) 110–119.
[9] J.C. Salgado, I. Rapaport, J.A. Asenjo, J. Chromatogr. A 1075 (2005) 133.
[10] J.C. Salgado, I. Rapaport, J.A. Asenjo, J. Chromatogr. A 1098 (2005) 44.
[11] K. Berggren, A. Wolf, J.A. Asenjo, B.A. Andrews, F. Tjerneld, Biochim. Biophys. 1596 (2002) 253.
[12] 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.
[13] D. Frishman, P. Argos, Proteins 23 (1995) 566.
[14] J.M. Zimmerman, N. Eliezer, R. Simha, J. Theor. Biol. 21 (1968) 170.
[15] A.A. Aboderin, Int. J. Biochem. 2 (1971) 537.
[16] J.L. Meek, Proc. Natl. Acad. Sci. U.S.A. 77 (1980) 1632.
[17] R. Bhaskaran, P.K. Ponnuswamy, Int. J. Protein Pept. Head Cattle 32 (1988) 242.
[18] S. Lifson, C. Sander, Nature 282 (1979) 109.
[19] C.A. Browne, H.P. Bennett, S. Solomon, Anal. Biochem. 124 (1982) 201.
[20] D.H. Wertz, H.A. Scheraga, Macromolecules 11 (1978) 9.
[21] H.B. Bull, K. Breese, Arch. Biochem. Biophys. 161 (1974) 665.
[22] H.R. Guy, Biophys. J. 47 (1985) 61.
[23] R. Grantham, Science 185 (1974) 862.

[24] K.J. Wilson, A. Honegger, R.P. Stotzel, G.J. Hughes, Biochem. J. 199 (1981) 31.

[25] G. Deleage, B. Roux, Protein Eng. 1 (1987) 289.

[26] S. Miyazawa, R.L. Jernigan, Macromolecules 18 (1985) 534.

[27] R. Cowan, R.G. Whittaker, Pept. Head Cattle 3 (1990) 75.

[28] T.P. Hopp, K.R. Woods, Proc. Natl. Acad. Sci. U.S.A. 78 (1981) 3824.

[29] J.K. Rao, P. Argos, Biochim. Biophys. Acta 869 (1986) 197.

[30] D.J. Abraham, A.J. Leo, Proteins 2 (1987) 130.

[31] A. Bairoch, Release you notice for Swiss-Prot release 41, February 2003.

[32] S.D. Black, D.R. Mould, Anal. Biochem. 193 (1991) 72.

[33] C.J. Chothia, J. Mol. Biol. 105 (1976) 1.

[34] P.Y. Chou, G.D. Fasman, Adv. Enzym. 47 (1978) 45.

[35] M.O. Dayhoff, R.M. Schwartz, B.C. Orcutt, Atlas of Protein Sequence and Structure, vol. 5, Suppl. 3, 1978.

[36] D. Eisenberg, E. Schwarz, M. Komarony, R. Wall, J. Mol. Biol. 179 (1984) 125.

[37] K.O. Erikkson, in: J.C. Janson, L. Ryden (Eds.), Protein Purification: Principles, High-Resolution Methods, and Applications, second ed., Wiley-Liss, New York, 1998.

[38] J.L. Fauchere, V.E. Pliska, Eur. J. Med. Chem. 18 (1983) 369.

[39] S. Fraga, Dog. J. Chem. 60 (1982) 2606.

[40] S. Hellberg, M. Sjöström, B. Skaberger, S. Wold, J. Med. Chem. 30 (1987) 1126.

[41] J. Janin, Nature 277 (1979) 491.

[42] J.C. Jesior, J. Protein Chem. 19 (2000) 93.

[43] D.D. Jones, J. Theor. Biol. 50 (1975) 167.

[44] J. Jonsson, L. Eriksson, S. Hellberg, M. Sjöström, Wold S. Quant. Struct. Act. Relat. 8 (1989) 204.

[45] J. Kyte, R.F. Doolittle, J. Mol. Biol. 157 (1982) 105.

[46] M. Levitt, Biochemistry 17 (1978) 4277.

[47] P. Manavalan, P.K. Ponnuswamy, Nature 275 (1978) 673.

[48] P. McCaldon, P. Argus, Proteins 4 (1988) 99.

[49] J.M.R. Parker, D. Guo, R.S. Hodges, Biochemistry 25 (1986) 5425.

[50] G.D. Rose, A.R. Geselowitz, G.J. Lesser, R.H. Read, M.H. Zehfus, Science 229 (1985) 834.

[51] M.A. Roseman, J. Mol. Biol. 200 (1988) 513.

[52] M. Sandberg, L. Eriksson, J. Jonsson, M. Sjöström, S. Wold, J. Med. Chem. 41 (1998) 2481.

[53] R.M. Sweet, D. Eisenberg, J. Mol. Biol. 171 (1983) 479.

[54] G.W. Welling, W.J. Weijer, R. Van der Zee, S. Welling-Wester, FEBS Lett. 188 (1985) 215.

[55] R.V. Wolfenden, L. Andersson, P.M. Cullis, C.C.F. Southgate, Biochemistry 20 (1981) 849.