Predicting the behaviour of proteins in hydrophobic interaction chromatography 1: Using the hydrophobic imbalance (HI) to describe 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 of proteins (DRT) in hydrophobic interaction chromatography (HIC) by means of mathematical models based on characteristics of the surface hydrophobicity distribution. We introduce a new parameter, called hydrophobic imbalance (HI), obtained from the three-dimensional structure of proteins. This parameter quantifies the displacement of the superficial geometric centre of the protein when the effect of the hydrophobicity of each amino acid is considered. This parameter is simpler and less expensive than those reported previously. We use HI as a way to incorporate information about the surface hydrophobicity distribution in order to improve the prediction of DRT. We tested the performance of our DRT predictive models in a set of 15 proteins. This set includes four proteins whose DRTs are known as very difficult to predict. By means of the variable HI, it was possible to improve the predictive characteristics obtained by models based on the average surface hydrophobicity (ASH) by 9.1%. Also, we studied linear multivariable models based on characteristics determined from the HI. By using this multivariable model, a correlation coefficient of 0.899 was obtained. With this model, we managed to improve the predictive characteristics shown by previous models based on ASH by 31.8%.

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

1. Introduction

Hydrophobic interaction chromatography (HIC) is a technique widely used for the purification of proteins. At present time, HIC is used in most industrial processes for protein purification as well as in laboratory scale applications. Commonly, it is used as a stage in the protein purification process following an ion exchange chromatography stage. It has been shown that the rational design of industrial protein purification processes normally requires an HIC stage [1].

Therefore, it is of great interest to have methodologies allowing us to carry out rational designs of these operations, both at laboratory scale and at industrial scale. Within this context, the availability of mathematical tools for simulating and predicting the behaviour of proteins in HIC is of main importance. Although the phenomenon of interaction between proteins and a stationary matrix is not entirely understood, several efforts have been in order to develop predictive mathematical models.

Lienqueo et al. found that the dimensionless retention time (DRT) correlate well (correlation coefficient ≈ 0.95) with the average surface hydrophobicity which was calculated considering the relative contribution of each one of the amino acids present on the surface [2]. These models have been validated for several standard proteins (α-amylase, ovalbumin, concanavalin A and β-lactoglobulin) and recombinant proteins from cell extracts (human superoxide dismutase from yeast and β-glucanase from *E. coli*), with satisfactory results [3]. The main

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

disadvantage of these models is that they do not consider the effect of the surface hydrophobicity distribution.

Mahn et al. [4,5] studied the effect of the surface hydrophobicity distribution in the prediction of the DRT of a set of four proteins where the previous predictive models obtain bad performances. In their first study, they proposed the use of an hydrophobic contact area (HCA) to predict the DRT. This HCA was determined through a thermodynamic model that combined electrostatic and hydrophobic interactions [6]. The adjustment of this model was carried out using variables measured at the laboratory. Then, it is possible to consider that the HCA has an experimental nature. In their second study, they defined a new parameter called local hydrophobicity (LH). The LH corresponded to the average surface hydrophobicity calculated considering only the amino acids located inside the most probable interaction zone between the protein and the stationary matrix. This interaction zone was determined using molecular docking simulations and, therefore, this parameter has a theoretical nature. The authors argue that by means of both parameters the DRT can be predicted with an acceptable performance.

However, the main disadvantage of these methodologies is that they are very expensive in human and computational resources. In the first case, the HCA determination requires a substantial amount of laboratory time. In the second case, the computational time needed to identify the most probable interaction zone is considerable, and consequently, it cannot be disregarded.

For these reasons, we introduce a new parameter called hydrophobic imbalance (HI). This parameter is simpler and less expensive than those reported previously and is obtained only from the characteristics of the protein surface. Briefly, it represents the displacement of the superficial geometric centre of the protein when the effect of a certain amino acidic hydrophobicity scale is considered. Only two related, but very different, tools has been reported in the literature previously: the hydrophobic helical moment, a vector with amplitude and direction which provides a measure of the amphiphilicity of a helix perpendicular to the helical axis [7] and the global hydrophobic moment which provides a measure of the degree and direction of the amphiphilicity or hydrophobic imbalance across the entire protein tertiary structure [8].

At the time of submission of this paper and the paper that follows it [9], a different methodology was proposed by the group of Steven Cramer [10]. This methodology uses quantitative structure retention relationship (QSRR) modelling to study and predict the retention and selectivity in HIC using molecular descriptors based on the three-dimensional structure of proteins, the primary structure information and a set of new hydrophobicity descriptors. Their results show that their models, based on Support Vector Machines (SVM) models, can predict the protein retention quite well, considering four different chromatographic systems (different pairs of ligands and media) and a set of 27 proteins. Although their models are more general than those proposed in this paper and the paper that follows, they require of a larger number of variables and consequently they show a greater mathematical complexity.

Therefore, the main objective of this paper is to investigate the use of the hydrophobic imbalance (HI) as a way to incorporate information about the surface hydrophobicity distribution in order to develop simple and computationally inexpensive mathematical models which can improve the performance of the prediction of DRT reported previously.

2. Materials and methods

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 [2,11,12]. Notice that 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 [13]. The ASA was calculated using the software STRIDE from the protein three-dimensional structure [14].

2.1. Hydrophobic imbalance (HI)

In this study, each amino acid in a protein is represented, basically, by its location in the space, its accessible surface area (ASA) and by its APV value given by φ. 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 superficial geometric centre $\mathrm{r_C}$ of a protein can be calculated using the following expression:

$$\mathrm{r_C} = \frac{\sum_{k \in S} \mathrm{ASA}(k) \cdot \mathrm{r}(k)}{\sum_{k \in S} \mathrm{ASA}(k)} \tag{2}$$

where the r(k) is the vector indicating the location of the amino acid k.

If we add to the previous equation the information given by vector φ, we get r_φ, which is the superficial geometric centre corrected by φ:

$$\mathrm{r}_\varphi = \frac{\sum_{k \in S} \mathrm{ASA}(k) \cdot \varphi(k) \cdot \mathrm{r}(k)}{\sum_{k \in S} \mathrm{ASA}(k) \cdot \varphi(k)} \tag{3}$$

In the case that φ corresponds to an hydrophobicity scale φ_H, the vector is denoted as $\mathrm{r_H}$. The hydrophobic imbalance (HI) is

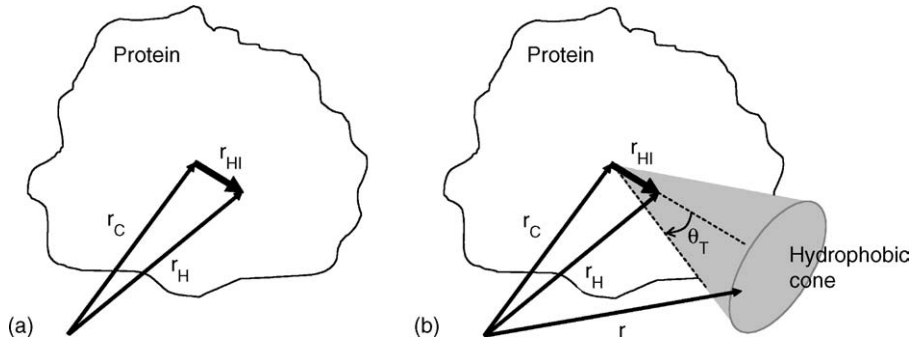

Fig. 1. The hydrophobic imbalance r_{HI} was defined as the substraction of r_H and r_C vectors and represents the displacement of the superficial geometric centre of the protein when the effect of a certain amino acidic property is considered. r_C is the superficial geometric centre of the protein and r_H is the superficial geometric centre corrected by an hydrophobicity scale (a). The hydrophobic cone (HC), shown in grey, is the set of all amino acids located inside a cone with vertex in r_C and with its axis parallel to the direction defined by r_{HI}. The volume of this cone is defined by the angle θ_T (b).

defined as the subtraction of r_H and r_C as shown in Fig. 1a and in the following equation:

$$r_{HI} = r_H - r_C \tag{4}$$

The r_{HI} vector represents the displacement of the superficial geometric centre of the protein when the effect of a certain amino acidic property is considered. Therefore, given φ, the magnitude of the r_{HI} vector can be interpreted as a measurement of the characteristics of the distribution of that property in the protein surface.

2.2. Hydrophobic hemisphere

The hydrophobic imbalance (HI) points towards the protein hemisphere of greater hydrophobicity, as it is possible to appreciate in Fig. 1b. We define the hydrophobic cone (HC) as the set of all amino acids located inside a cone with vertex located in r_C and with its axis parallel to the direction defined by r_{HI}. Clearly, the volume of this cone is defined by the angle θ_T. In particular, when $\theta_T = 90^\circ$, it corresponds to the hydrophobic hemisphere.

Any amino acid is included in the hydrophobic cone only if its position r (see Fig. 1b) satisfies the following inequality:

$$\cos^{-1}\left(\frac{(r - r_C)\cdot r_{HI}}{|r - r_C|\cdot|r_{HI}|}\right) \leq \theta_T \tag{5}$$

It is possible to calculate a "local" ASP considering only those amino acids located inside the HC. This ASP can be calculated by two ways (dividing by the total protein surface or dividing by the surface delimited by the HC):

$$ASP_{HC(\theta_T),T} = \frac{\sum_{k\in HC(\theta_T)} ASA(k)\cdot\varphi(k)}{\sum_{k\in S} ASA(k)} \tag{6}$$

$$ASP_{HC(\theta_T),P} = \frac{\sum_{k\in HC(\theta_T)} ASA(k)\cdot\varphi(k)}{\sum_{k\in HC(\theta_T)} ASA(k)} \tag{7}$$

where $ASP_{HC(\theta_T),T}$ is referred to the total protein surface while $ASP_{HC(\theta_T),P}$ is referred to the surface delimited by the HC.

2.3. Materials

2.3.1. Protein set and DRT

Fifteen proteins with known dimensionless retention time (DRT) 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 [15]. DRT values correspond to those used in [2,5] and they are the DRTs observed in a hydrophobic interaction column, calculated according to:

$$DRT = \frac{t_R - t_0}{t_f - t_0} \tag{8}$$

where t_R corresponds to the time where the peak of the chromatogram takes place, t_0 to the time when the salt gradient starts and t_f to the time when the salt gradient finishes. The DRT values used in this work were obtained in a 1 ml Phenyl-Sepharose Fast Flow column using 2 M ammonium sulphate as the eluent.

2.3.2. Collection of amino acidic property vectors (APV)

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. [13,16–57]. 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

The performance of the models was evaluated by means of three parameters: the mean square error (MSE), the correlation coefficient (Pearson) and the Jack Knife cross validation mean square error (MSE_{JK}). The MSE and the Pearson were calculated using the following expressions:

$$\mathrm{MSE} = \frac{1}{N}\sum_{k=1}^{N}(\mathrm{DRT}_k - \widehat{\mathrm{DRT}_k})^2 \tag{9}$$

$$\mathrm{Pearson} = \frac{N\sum_{k=1}^{N}(\mathrm{DRT}_k - \widehat{\mathrm{DRT}_k}) - \sum_{k=1}^{N}\mathrm{DRT}_k \cdot \sum_{k=1}^{N}\widehat{\mathrm{DRT}_k}}{\sqrt{N\sum_{k=1}^{N}(\mathrm{DRT}_k)^2 - (\sum_{k=1}^{N}\mathrm{DRT}_k)^2}\cdot\sqrt{N\sum_{k=1}^{N}(\widehat{\mathrm{DRT}_k})^2 - (\sum_{k=1}^{N}\widehat{\mathrm{DRT}_k})^2}} \tag{10}$$

where DRT_k is the DRT of protein k, $\widehat{\mathrm{DRT}_k}$ is the prediction of the DRT for protein k and N is the number of proteins with experimentally known DRT considered ($N = 15$).

The MSE_{JK} was used to estimate the prediction error of the models for proteins not considered in their determination. In this case, the size of the data set is modest. Hence, other techniques of re-sampling like k-folding cross validation or boostrap cannot be used. The Jack Knife re-sampling method (leave-one-out) is a well accepted methodology [58]. Actually, it is regarded as the most objective and effective tool for the evaluation of predictor models [59,60]. The mathematical principle and a comprehensive discussion about this can be found in [61]. Briefly, this method consists of repeating the fitting of the model as many times as the size of the data set, leaving in each occasion one element out of the calculations. Thus, in each step, the error of the model for the prediction of the element that was left out is calculated. At the end of the process, the final prediction error of the model is estimated as the average of the prediction error of each element that was left out. In other words, this process is carried out systematically so that in the kth adjustment, the kth element of the data is not considered. The model determined by means of the kth adjustment is used to calculate the prediction of the DRT of protein k, denoted by $\widehat{\mathrm{DRT}_k}^{-k}$, where $-k$ means that the kth element has been left out. Therefore, the MSE_{JK} is obtained calculating the average on the collection of N proteins as indicated in the following equation:

$$\mathrm{MSE_{JK}} = \frac{1}{N}\sum_{k=1}^{N}(\mathrm{DRT}_k - \widehat{\mathrm{DRT}_k}^{-k})^2 \tag{11}$$

3. Results and discussion

3.1. Models based only on the hydrophobic imbalance (HI)

This section details the results obtained when using the hydrophobic imbalance (HI) of a protein as an instrument to mathematically model its dimensionless retention time (DRT) in hydrophobic interaction chromatography (HIC). The HI represents the displacement of the surface geometric centre of the protein when it is corrected with the hydrophobic characteristics of the amino acids.

3.1.1. Calculation of HI using discreet hydrophobicity scales

With the aim to facilitate an initial analysis, three discrete scales of hydrophobicity 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.

Mahn et al. [4] observed that some proteins with similar average superficial hydrophobicity (ASH) had very different DRTs, these proteins were: 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 swaped, altering, in this way, the distribution of hydrophobic amino acids without changing the average surface hydrophobicity. The correlation coefficients (Pearson) between DRT and HI is shown in Table 1. The Pearsons between the DRT and local hydrophobicity (LH) and hydrophobic contact area (HCA) reported by Mahn et al. [5] are also shown in Table 1. The correlation coefficients obtained by the hard binary and trinary scales are almost twofold those obtained for the LH and HCA, justifying in this way the study of HI. The correlation coefficients for the hard binary and trinary scales are almost three times higher than the ones obtained for the soft binary scale. This indicates that the best results were obtained defining the hydrophobicity of the amphipathic amino acids with an intermediate value (0.5) or as hydrophilic (0.0).

Fig. 2 shows a simplified plot of β-carbons location for Ribonuclease A (1AFU). This plot indicates that the HI vector aims toward the protein hemisphere of greater hydrophobicity. In

Table 1
Correlation coefficients (Pearson) between the dimensionless retention time (DRT) and the average surface hydrophobicity (ASH), local hydrophobicity (LH), hydrophobic contact area (HCA) and hydrophobic imbalance (HI)

Parameter	Pearson
ASH	−0.528
LH	0.557
HCA	0.483
HI (HBS)	−0.940
HI (SBS)	−0.247
HI (TS)	−0.930

The HI was calculated using three discreet hydrophobicity scales: hard binary scale (HBS), soft binary scale (SBS) and trinary scale (TS). The ASH, LH and HCA were reported in [4,5]. The calculations only considered the proteins: Ribonuclease S (1RBC), Ribonuclease A (1AFU), Ribonuclease T1 wild type (1RGC) and Ribonuclease T1 variant Y45W/W59Y (1TRP).

fact, the hydrophobicity of this hemisphere (measured as the sum of the accessible area of all the hydrophobic amino acids) doubles the hydrophobicity of the opposite hemisphere. On the other hand, the sign of the correlation coefficients shown in Table 1 indicates that the HI calculated by means of the three discreet scales is inversely proportional to the DRT. Certainly, the HI magnitude is related to the surface hydrophobicity distribution, but its direct interpretation is difficult, due to the great amount of topological factors that participate in its determination.

The correlation coefficients between the HI calculated with the three discreet hydrophobicity scales and the DRT of the 15 proteins considered in this study were calculated. Correlation coefficients of −0.285, 0.297 *y* −0.074 for the hard binary, soft binary and trinary scale, respectively, were obtained. This correlation coefficients are very low (less of 30% in all cases) in comparison with the ones shown in Table 1. More likely, they are originated by an artifact in the model; therefore, it is necessary to optimise the hydrophobicity scale for the calculation of the HI to capture the complexity of the complete protein set.

3.1.2. Calculation of HI using the collection of aminoacidic property vectors (APV)

The HI was calculated using each one of the 74 aminoacidic property vectors (APV), which allowed us to determine the average and standard deviation of HI as indicated in Fig. 3. It is interesting to notice that the standard deviation of HI is similar for all the proteins and that it has great magnitude (ca. 70% of the average) as well. In the case of 1RGC-1TRP and 1AFU-1RBC proteins, the first pair presents similar averages with a difference of only 3%; however, in the second pair, this difference is greater, being of almost 13%.

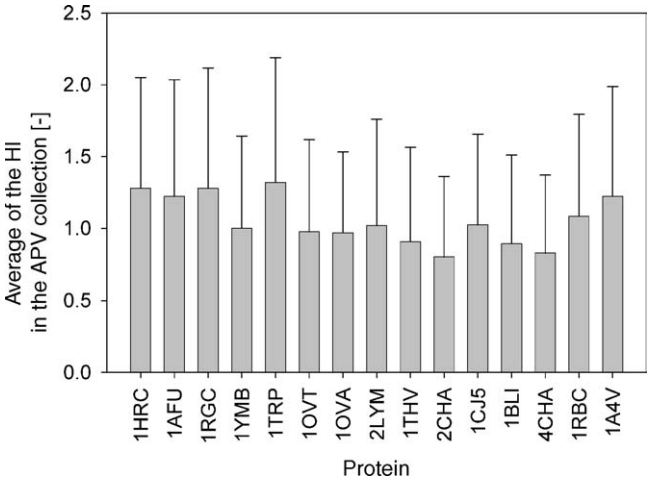

Fig. 3. Average and standard deviation of the hydrophobic imbalance in the collection of 74 amino acidic property vectors (APV), for each protein considered in the study in ascending order with respect to their DRT.

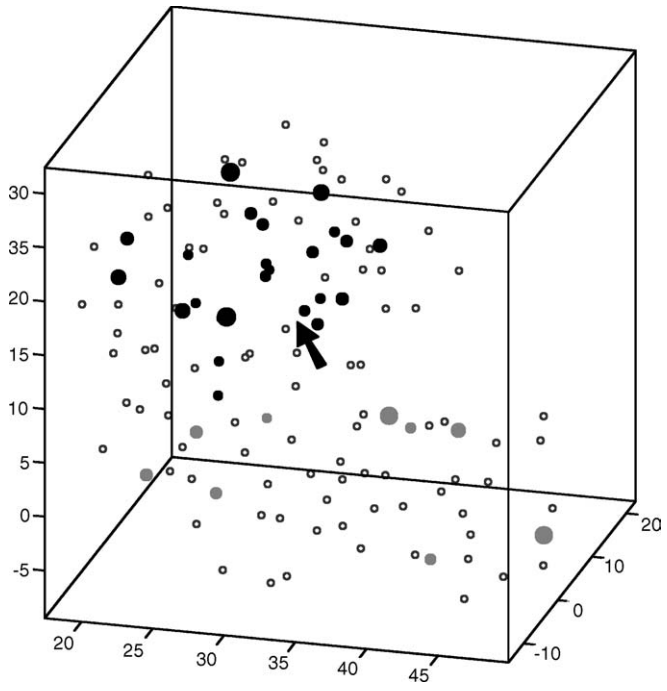

Fig. 2. Simplified plot of the Ribonuclease A (1AFU). The circles and the dots in the figure represent the location of the β-carbons of each amino acid in the protein. The dots (grey and black) represent the amino acids widely accepted as hydrophobic (Ala, Ile, Leu, Phe, Pro, Val) and the circles indicate the hydrophilic ones. In the case of the hydrophobic amino acids, the size of the dots are equivalent to the product between the ASA and the amino acid hydrophobicity (in this case 1 for all the hydrophobic amino acids). If the dots are located inside the hydrophobic hemisphere they are coloured black and if they are outside they are coloured grey. The hydrophobic imbalance vector r_{HI} has been drawn as a black arrow.

The prediction of the DRT by means of the HI calculated using the 74 APVs, in comparison with those obtained when using the ASP was investigated. For both predictors, a linear model was used. The predictive characteristics of the models were characterised by means of the determination of the Jack Knife cross validation mean square error (MSE_{JK}) in the set of 15 proteins. Tables 2 and 3 show the results of these calculations.

Table 2 shows the results obtained by the linear model based on the ASP sorted by the MSE_{JK} value. Interestingly enough, the three better models were constructed using APVs which correspond to hydrophobicity scales and those were determined from the protein behaviour in high-performance liquid chromatography (HPLC). The best model was constructed using the APV proposed by Browne [16]. The hydrophobicity scale of Browne was calculated from the protein retention coefficients observed in HPLC with trifluoroacetic acid (TFA). The results detailed in Table 3 indicate that the best linear model based on the HI has predictive characteristics slightly better than the one based on the ASP, shown as a decrease of 9.2% in the MSE_{JK}. In this case, the best APV was the one of Zimmerman [17], which quantifies the amino acids polarity. The others APVs in Table 3 correspond to hydrophobicity scales, in the same way that in Table 2. We must notice that the sign of the slope between HI and the DRT is negative ($DRT = (0.908 \pm 0.181) - (0.235 \pm 0.112) \times HI$), maintaining an inverse relationship between these magnitudes found previously.

The performance obtained by the models based on the HI, although better than the observed in the models based on the ASP, is still insufficient for practical applications (correlation coefficient > 0.8). In fact, the correlation coefficient for the

Table 2
Effect of the aminoacidic property vectors (APV) on the performance indexes of the linear model based on an average surface property (ASP) in the prediction of the experimental DRT of the 15 proteins

No.	APV	Description	MSE $\times 10^3$	Pearson	$MSE_{JK} \times 10^3$
1	Browne [16]	Retention coefficient in HPLC and TFA	23.267	0.745	**31.054**
2	Meek [49]	Retention coefficient in HPLC, pH 2.1	25.280	0.719	**34.288**
3	Parker [50]	Hydrophilicity scale derived from HPLC peptide retention times	27.384	0.690	**34.815**

The three best APV (out of 74) in ascending order with respect to the Jack Knife cross validation mean square error (MSE_{JK}) are listed. The correlation coefficient (Pearson) and the mean square error (MSE) are also shown. MSE_{JK} values have been highlighted in bold.

Table 3
Effect of the aminoacidic property vectors (APV) on the performance indexes of the linear model based on the hydrophobic imbalance (HI) in the prediction of the experimental DRT of the 15 proteins

No.	APV	Description	MSE $\times 10^3$	Pearson	$MSE_{JK} \times 10^3$
1	Zimmerman [17]	Polarity	20.168	0.784	**28.222**
2	Hopp [22]	Hydrophilicity	28.607	0.673	**40.877**
3	Cowan and Whittaker [21]	Hydrophobicity indexes at pH 7.5 determined by HPLC	28.520	0.674	**40.974**

The three best APV (out of 74) in ascending order with respect to the Jack Knife cross validation mean square error (MSE_{JK}) are listed. The correlation coefficient (Pearson) and the mean square error (MSE) are also shown. MSE_{JK} values have been highlighted in bold.

model based on HI was 0.784. Consequently, the models need the incorporation of more information to improve their performance.

3.2. Models based on the characteristics of the hydrophobic hemisphere

The ASP calculated in the hydrophobic hemisphere was determined by the amino acids located in the protein hemisphere indicated by the HI vector. This ASP was calculated in two ways: partially ($ASP_{HC(90^\circ),P}$) refers to the surface of the hydrophobic hemisphere; and, more generally ($ASP_{HC(90^\circ),T}$), refers to the total surface of the protein. In both cases the angle θ_T was equal to 90°.

The performance in the prediction of the DRT by both magnitudes was evaluated. Table 4 shows the results of the evaluation of both predictors sorted by their MSE_{JK}. The best results were obtained when using the $ASP_{HC(90^\circ),T}$ calculated using the APV of Wilson [18], this APV corresponds to hydrophobic constants derived from HPLC retention times for peptides. The MSE obtained by this model was located between the MSE given by the ASP and HI models. However, the predictive characteristics of the model were inferior to the ones observed previously. Nevertheless, the better APVs were associated to hydrophobicity scales again. It is necessary to highlight that the first appearance of a model based on the variable $ASP_{HC(90^\circ),P}$ is in the fifth place of the table. The MSE_{JK} of this model was 17.4% worse than the best model based on $ASP_{HC(90^\circ),T}$. This is explained by the fact that the calculation of the $ASP_{HC(90^\circ),T}$ considers all the amino acids on the protein surface, quantifying, in this way, the relation between the hydrophobic characteristics of the hydrophobic hemisphere and the total protein surface.

In addition, the effect of the reduction of the surface covered by the hydrophobic hemisphere in the prediction of the DRT was investigated. By doing that the total and partial ASP were calculated in a cone with an angle θ_T moving in the interval 15–90°, in 5° steps. The results of these experiments are shown in Fig. 4. The plot in Fig. 4 shows the minimum value of MSE_{JK} found when considering the models constructed using each one of 74 APVs. Also, the plot includes the average and standard deviation of the five better values of MSE_{JK}. It is observed that the main tendency corresponds to a smaller value of the MSE_{JK} as the angle θ_T increases. This is, the predictive characteristics of the models improve as the section of the hydrophobic hemisphere that is considered increases, and higher is the amount of information available for the model. It is important to notice that

Table 4
Effect of the aminoacidic property vectors (APV) on the performance indexes of the linear model based on the average surface property calculated in the hydrophobic hemisphere in the prediction of the experimental DRT of the 15 proteins: in partial form (P) referred to the surface of the hydrophobic hemisphere; and, in a total way (T), referred to the total surface of the protein

No.	Model	APV	Description	MSE $\times 10^3$	Pearson	$MSE_{JK} \times 10^3$
1	T	Wilson [18]	Hydrophobic constants derived from HPLC peptide retention times	22.981	0.749	**33.019**
2	T	Parker [50]	Hydrophilicity scale derived from HPLC peptide retention times	28.129	0.680	**35.125**
3	T	Browne [16]	Retention coefficient in HPLC with heptafluorobutyric acid (HFBA)	27.054	0.695	**35.781**
4	T	Hellberg [39]	Statistical analysis of amino acid properties z1	28.863	0.670	**36.208**
5	P	Bull and Breese [29]	Hydrophobicity (free energy of transfer to surface in kcal/mol)	27.293	0.692	**36.465**

The five best APV (out of 74) in ascending order with respect to the Jack Knife cross validation mean square error (MSE_{JK}) are listed. The correlation coefficient (Pearson) and the mean square error (MSE) are also shown. MSE_{JK} values have been highlighted in bold.

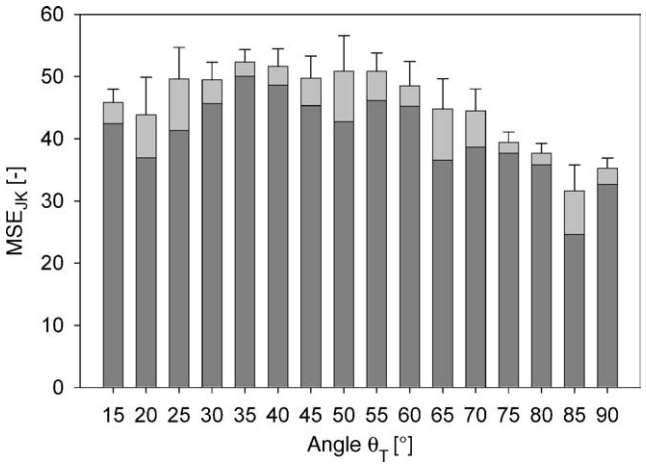

Fig. 4. Effect of the size of the hydrophobic cone on the performance indexes of the linear model based on the average surface property calculated in the hydrophobic cone at angle θ_T in the prediction of the experimental DRT of the 15 proteins: minimum Jack Knife cross validation mean square error (MSE_{JK}) observed in the APV collection (■) and average MSE_{JK} and standard deviation for the five best linear models (□).

the standard deviation of MSE_{JK} tends to decrease regularly as the angle θ_T is increased. This behaviour can be interpreted as a stabilisation of the performance of the predictive model as the amount of information available is increased.

On the other hand, the plot in Fig. 4 shows that there exists a remarkable lowering of the minimum MSE_{JK} at 85° of almost 25% with respect to the value observed at 90°. It is possible to notice that as an average amongst all proteins and considering to all the APVs, the band between 85 and 90° is equivalent to $7.4 \pm 4.1\%$ of the total surface of the hydrophobic hemisphere (90°). This behaviour could be explained as a steric adjustment of the model caused by a natural impossibility of the molecule to use all the surface available in its hydrophobic hemisphere for the interaction with the hydrophobic matrix. However, in this case, the APV associated to the minimum MSE_{JK} at 85° was not related to an hydrophobicity scale directly. The APV selected at 85° was the proposed by Deleage and Roux [19], which corresponds to a β-turns conformational parameter. This APV has low correlations with others APV related to hydrophobicity scales: −0.768 with Miyazawa and Jerningan [20]; −0.570 and −0.603 with those of Cowan and Whittaker [21]; −0.519 with the one of Wilson [18]; and −0.554 with the proposed by Hopp [22]; to mention a few. This fact and the great standard deviation observed in that point forces to disregard this observable fact. Nevertheless, the calculations that follows will include the determination of the predictors at 85°.

3.3. Multivariable models based on the ASP, HI and ASP_{HA}

As was shown, the predictive capacity of each one of the features presented in this work is not sufficient to capture the phenomenon complexity. Therefore, it is interesting to investigate if a linear or nonlinear combination of these features could improve the models performance. Multivariable models were constructed using the features described previously: average surface properties (ASP), hydrophobic imbalance (HI), partial and total ASP in the hydrophobic hemisphere and partial and total ASP in a cone at 85°. All the linear combinations were investigated as well as a few nonlinear ones, that consisted, basically, in the product of the variables.

Table 5 shows the performance of the five best models found by means of this procedure. In the first two places, we found models that use APVs associated to hydrophobicity measures. The rest of the models in the table are based on APVs related to the analysis of conformational properties of the amino acids, specially with respect to their relation with specific secondary structures. Although it is widely accepted that the hydrophobicity is one of the main factors that determine the behaviour of an amino acid in a protein, these models were discarded because they are not based on direct measurements of the hydrophobicity.

The best model in Table 5 (model 1) corresponded to a linear combination of the predictors: HI, ASP and $ASP_{HC(85^\circ),T}$. These variables were determined using the APV of Rao and Argos [23] that corresponds to a membrane buried helix parameter. The second model was constructed using the APV of Hopp [22] also based on a hydrophobicity scale. The $ASP_{HC(85^\circ),T}$ in the best model confirms the superiority of this variable with respect to the partial version and the fact that this variable was calculated for an angle of 85°, it confirms the validity of the approach discussed previously.

On the other hand, model 1 decreased the MSE_{JK} by 24.9% with respect to the best model found previously (linear model based on HI) and 31.8% with respect to the model based on the ASP. So, it is of interest to know the relative importance of each variable in the model. For that purpose, the plot in Fig. 5 was constructed. This plot shows the effect that has the removal of each one of the variables of the model 1 in its predictive capacity. It was measured as the observed value of MSE_{JK} in the set of 15 proteins. This plot indicates that the most important variable in model 1 is HI. The removal of HI from the model 1 meant an increase of the MSE_{JK} to 3.8 times the observed one in the complete model. HI was followed by $ASP_{HC(85^\circ),T}$ and ASP with an increase to 2.4 and 2.0 times, respectively. Clearly, this fact

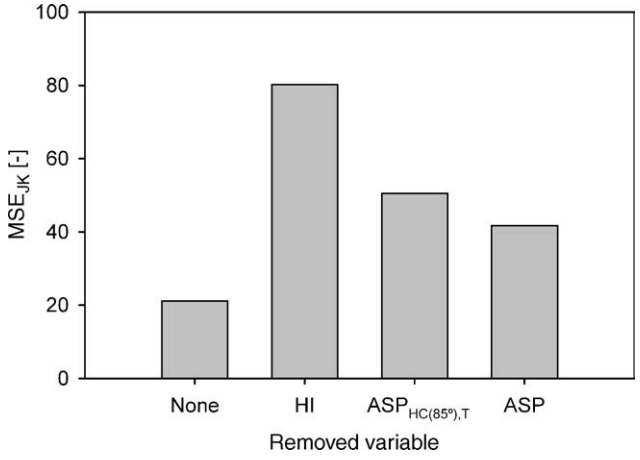

Fig. 5. Effect of the removal of each one of the variables of the multivariable model in its predictive capacity, measured as the observed value of Jack Knife cross validation mean square error (MSE_{JK}) in the set of 15 proteins.

Table 5
Performance indexes of the linear multivariable models based in the prediction of the experimental DRT of the 15 proteins

N°	Model	APV	Description	MSE × 10^3	Pearson	$MSE_{JK} \times 10^3$	θ_T (°)	DF	R^2adj (%)
1	HI, $ASP_{HC(85^\circ),T}$, ASP	Rao and Argos [23]	Membrane buried helix parameter	10.005	0.899	21.169	85	11	75.7
2	HI, $ASP_{HC(90^\circ),T}$, ASP	Hopp [22]	Hydrophilicity	13.814	0.858	27.085	90	11	66.4
3	$ASP_{HC(85^\circ),P}$, ASP	Chou and Fasman [31]	Conformational parameter for β-sheet (computed from 29 proteins)	18.595	0.803	27.571	85	12	58.5
4	$ASP_{HC(85^\circ),P}$, ASP	Chou and Fasman [31]	Conformational parameter for β-sheet (computed from 29 proteins)	18.954	0.799	27.665	90	12	57.7
5	$ASP_{HC(85^\circ),P}$, ASP	Meek [49]	Retention coefficient in HPLC, pH 2.1	16.137	0.832	27.817	90	12	64.0

The five 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 angle θ_T of the hydrophobic cone, the degrees of freedom (DF) and the adjusted determination coefficient (R^2adj) are also shown.

confirms the importance of the variable HI for the prediction of the DRT.

3.4. Final discussion

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

$$\begin{aligned} DRT = {} & (0.853 \pm 0.486) - (0.502 \pm 0.168) \times HI \\ & + (14.707 \pm 6.297) \times ASP_{HC(85^\circ),T} \\ & - (6.484 \pm 3.323) \times ASP \end{aligned} \quad (12)$$

where DRT is the dimensionless retention time of the protein, HI is its hydrophobic imbalance, $ASP_{HC(85^\circ),T}$ is its ASP referred to the total protein surface calculated in a cone at 85° and ASP is its average surface property. All these variables were calculated using the APV of Rao and Argos [23] which is shown in Table 6.

The confidence intervals at 95% determined for the parameters of the model did not exceed a 60% of their nominal values. The greatest uncertainty in the coefficients determination was observed in the case of the constant of the model, reaching a magnitude of 57%. On the contrary, the coefficient with the narrowest confidence interval was the associated to HI, with 33.4%. It is interesting to highlight that the sign of the coefficient for HI is negative maintaining therefore the behaviour observed previously. Also, it is remarkable that the coefficient associated to $ASP_{HC(85^\circ),T}$ be almost two-fold the obtained for ASP and that the sign of this last one is negative. In fact, we can approximate both coefficients so the Eq. (12) could be rewritten as $c_0 + c_1 \times HI + c_2 \times (2 \times ASP_{HC(85^\circ),T} - ASP)$. Clearly, the third component can be interpreted as the difference between the hydrophobicity of the hydrophobic cone and the rest of the protein, and then, as a measurement of the difference between the hydrophobicity of both hemispheres. On the other hand, the interpretation of the sign of the HI coefficient is not so clear. As it was discussed previously, this parameter quantifies the characteristics of the surface hydrophobicity distribution, but the amount of effects that could take part in the calculation of its magnitude prevent a clear interpretation. This suggests the necessity of a specific parameter, as for example a direct measurement of the homogeneity of the surface hydrophobicity distribution.

Table 6
Amino acidic property vector (APV) of Rao and Argos [23]

aa	Original	Scaled to (0;1)
Ala	1.360	0.858
Arg	0.150	0.041
Asn	0.330	0.162
Asp	0.110	0.014
Cys	1.270	0.797
Gln	0.330	0.162
Glu	0.250	0.108
Gly	1.090	0.676
His	0.680	0.399
Ile	1.440	0.912
Leu	1.470	0.932
Lys	0.090	0.000
Met	1.420	0.899
Phe	1.570	1.000
Pro	0.540	0.304
Ser	0.970	0.595
Thr	1.080	0.669
Trp	1.000	0.615
Tyr	0.830	0.500
Val	1.370	0.865

Fig. 6 shows the scatter plots between the experimental DRT and the predictions carried out by the models. In the case of the model based on the ASP of Browne [16], it is possible to notice that the difference is considerable at least in two proteins; in addition the greater magnitude errors are concentrated in the proteins with DRT < 0.6. However, in the case of the multivariable model, the error is distributed in a most uniform way throughout all the scale, explaining the inferior MSE_{JK} shown by this model and its best predictive characteristics.

The distribution of the residual error for DRT predictive models is shown in Fig. 7. In the case of the model ASP, the greater magnitude residuals are observed in the case of proteins cytochrome *c* (1HRC), myoglobin (1YMB) and RNAse S (1RBC). It is possible to notice that cytochrome *c* is a hard protein for both models; however, in the case of RNAse S, the problem is presented exclusively for the model ASP. This hindering in its modelling was reported previously by Mahn et al. [4] and attributed to its great flexibility. Nevertheless, the multivariable model does not present this problem. In fact, if we took in consideration only the four ribonucleases reported by Mahn et al., the correlation coefficient for the multivariable model was 0.906, this is 62.7 and 87.6% greater than the correlation coeffi-

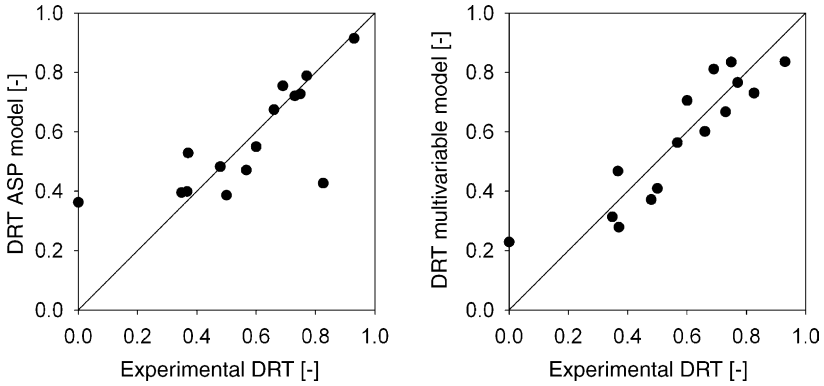

Fig. 6. Scatter plots between the experimental dimensionless retention time (DRT) and DRT predicted by the models based on the average surface properties (ASP) and the multivariable model.

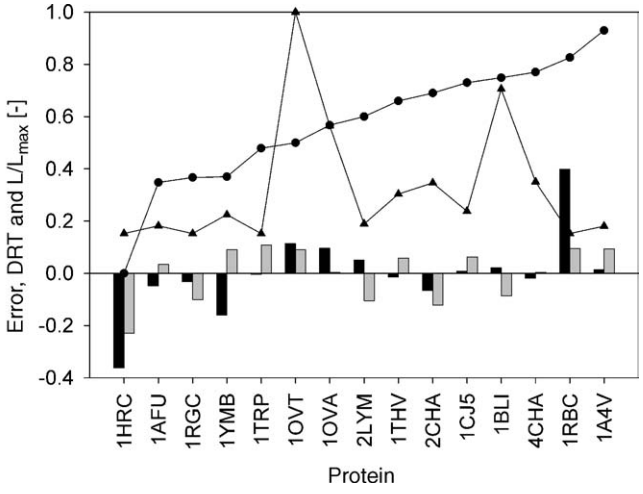

Fig. 7. Plot of the residual error between the experimental dimensionless retention time (DRT) and DRT predicted by the ASP model (■) and the multivariable model (□). The experimental DRT (●), and the dimensionless length (▲) are also shown. The proteins are arranged in ascending order with respect to their DRT.

cients of the models based on the LH and HCA, respectively. On the other hand, it was not observed a direct relation between the residual magnitude and the protein length or with the value of the DRT, in fact, the correlation coefficient between these magnitudes was inferior to 0.300 and 0.540, for the length and the DRT, respectively.

4. Conclusions

In this paper, the use of surface amino acid distribution to predict the behaviour of proteins in hydrophobic interaction chromatography (HIC) was investigated. The main contribution of this work was the hydrophobic imbalance (HI). This parameter, obtained from the characteristics of the protein surface, represents the displacement of the superficial geometric centre of the protein when the effect of the hydrophobicity of each amino acid is considered.

The HI was calculated for a set of four ribonucleases reported in [4] with similar ASP and very different DRTs and therefore with a DRT hard to predict using only the ASP. The HI calculations were carried out using simple hydrophobicity scales. These calculations showed that the HI obtained correlation coefficients remarkably better (al least 67%) than the models based on the local hydrophobicity (LH) and the hydrophobic contact area (HCA). The DRTs of 15 proteins and a predictive linear model based on the HI were correlated, we obtained a correlation coefficient of 0.784, slightly superior (5%) to those observed in models based on an average surface hydrophobicity (ASH). In addition, this model decreased the MSE_{JK} by 9.1% with respect to the model based on the ASH.

The linear combination of the HI, ASP and the $ASP_{HC(85^\circ),T}$ (ASP calculated in an hydrophobic cone (HC) at 85°) allowed the development of a multivariable model that decreased the MSE_{JK} by 24.9% with respect to the best model found previously (the linear model based on HI) and 31.8% with respect to the model based on the ASH. The correlation coefficient obtained for the multivariable model was 0.899. We observed that the coefficients associated to HI in all the models were negative and that the coefficients associated to the measures of the protein hydrophobicity were positive. The interpretation of the sign of the HI coefficient is not trivial because the amount of effects that take part in the calculation of its magnitude prevent a direct interpretation.

Although the best predictive model developed in this article was a multivariable model, this improvement did not justify, necessarily, the reduction of the degrees of freedom (from 13 to 11) and the increase in the complexity of the model with respect to the linear model based on HI.

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.E. Lienqueo, A. Mahn, L. Vásquez, J.A. Asenjo, J. Chromatogr. A 1009 (2003) 189.
[4] A. Mahn, M.E. Lienqueo, J.A. Asenjo, J. Chromatogr. A 1043 (2004) 47.
[5] A. Mahn, G. Zapata-Torres, J.A. Asenjo, J. Chromatogr. A 1066 (2005) 81.
[6] W. Melander, D. Corradini, Cs. Horváth, J. Chromatogr. 317 (1984) 67.
[7] D. Eisenberg, R.M. Weiss, T.C. Terwilliger, Nature 299 (1982) 371.
[8] B.D. Silverman, Proteins 53 (2003) 880.
[9] J.C. Salgado, I. Rapaport, J.A. Asenjo, J. Chromatogr. A. 1107 (2006) 120–129.
[10] A. Ladiwala, F. Xia, Q. Luo, C.M. Breneman, S.M. Cramer, Biotechnol. Bioeng., DOI: doi:10.1002/bit.20771.
[11] J.C. Salgado, I. Rapaport, J.A. Asenjo, J. Chromatogr. A 1075 (2005) 133.
[12] J.C. Salgado, I. Rapaport, J.A. Asenjo, J. Chromatogr. A 1098 (2005) 44.
[13] K. Berggren, A. Wolf, J.A. Asenjo, B.A. Andrews, F. Tjerneld, Biochim. Biophys. 1596 (2002) 253.
[14] D. Frishman, P. Argos, Proteins 23 (1995) 566.
[15] 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.
[16] C.A. Browne, H.P. Bennett, S. Solomon, Anal. Biochem. 124 (1982) 201.
[17] J.M. Zimmerman, N. Eliezer, R. Simha, J. Theor. Biol. 21 (1968) 170.
[18] K.J. Wilson, A. Honegger, R.P. Stotzel, G.J. Hughes, Biochem. J. 199 (1981) 31.
[19] G. Deleage, B. Roux, Protein Eng. 1 (1987) 289.
[20] S. Miyazawa, R.L. Jernigan, Macromolecules 18 (1985) 534.
[21] R. Cowan, R.G. Whittaker, Peptide Head Cattle 3 (1990) 75.
[22] T.P. Hopp, K.R. Woods, Proc. Natl. Acad. Sci. U.S.A. 78 (1981) 3824.
[23] J.K. Rao, P. Argos, Biochim. Biophys. Acta 869 (1986) 197.
[24] A.A. Aboderin, Int. J. Biochem. 2 (1971) 537.
[25] D.J. Abraham, A.J. Leo, Proteins 2 (1987) 130.
[26] A. Bairoch, Release you notice for Swiss-Prot release 41, February 2003.
[27] R. Bhaskaran, P.K. Ponnuswamy, Int. J. Pept. Protein. Head Cattle 32 (1988) 242.
[28] S.D. Black, D.R. Mould, Anal. Biochem. 193 (1991) 72.
[29] H.B. Bull, K. Breese, Arch. Biochem. Biophys. 161 (1974) 665.
[30] C.J. Chothia, Mol. Biol. 105 (1976) 1.
[31] P.Y. Chou, G.D. Fasman, Adv. Enzym. 47 (1978) 45.
[32] M.O. Dayhoff, R.M. Schwartz, B.C. Orcutt, in: Atlas of Protein Sequence and Structure, vol. 5, Suppl. 3, 1978.
[33] D. Eisenberg, E. Schwarz, M. Komarony, R. Wall, J. Mol. Biol. 179 (1984) 125.
[34] 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.
[35] J.L. Fauchere, V.E. Pliska, Eur. J. Med. Chem. 18 (1983) 369.
[36] S. Fraga, Dog. J. Chem. 60 (1982) 2606.
[37] R. Grantham, Science 185 (1974) 862.
[38] H.R. Guy, Biophys. J. 47 (1985) 61.
[39] S. Hellberg, M. Sjöström, B. Skaberger, S. Wold, J. Med. Chem. 30 (1987) 1126.
[40] J. Janin, Nature 277 (1979) 491.
[41] J.C. Jesior, J. Protein Chem. 19 (2000) 93.
[42] D.D. Jones, J. Theor. Biol. 50 (1975) 167.
[43] J. Jonsson, L. Eriksson, S. Hellberg, M. Sjöström, S. Wold, Quant. Struct. Act. Relat. 8 (1989) 204.
[44] J. Kyte, R.F. Doolittle, J. Mol. Biol. 157 (1982) 105.
[45] M. Levitt, Biochemistry 17 (1978) 4277.
[46] S. Lifson, C. Sander, Nature 282 (1979) 109.
[47] P. Manavalan, P.K. Ponnuswamy, Nature 275 (1978) 673.
[48] P. McCaldon, P. Argus, Proteins 4 (1988) 99.
[49] J.L. Meek, Proc. Natl. Acad. Sci. U.S.A. 77 (1980) 1632.
[50] J.M.R. Parker, D. Guo, R.S. Hodges, Biochemistry 25 (1986) 5425.
[51] G.D. Rose, A.R. Geselowitz, G.J. Lesser, R.H. Read, M.H. Zehfus, Science 229 (1985) 834.
[52] M.A. Roseman, J. Mol. Biol. 200 (1988) 513.
[53] M. Sandberg, L. Eriksson, J. Jonsson, M. Sjöström, S. Wold, J. Med. Chem. 41 (1998) 2481.
[54] R.M. Sweet, D. Eisenberg, J. Mol. Biol. 171 (1983) 479.
[55] G.W. Welling, W.J. Weijer, R. Van der Zee, S. Welling-Wester, FEBS Lett. 188 (1985) 215.
[56] D.H. Wertz, H.A. Scheraga, Macromolecules 11 (1978) 9.
[57] R.V. Wolfenden, L. Andersson, P.M. Cullis, C.C.F. Southgate, Biochemistry 20 (1981) 849.
[58] K.C. Chou, C.T. Zhang, Crit. Rev. Biochem. Mol. Biol. 30 (1995) 275.
[59] G.P. Zhou, J. Protein Chem. 17 (1998) 729.
[60] G.P. Zhou, N. Assa-Munt, Proteins 44 (2001) 57.
[61] K.V. Mardia, J.T. Kent, J.M. Bibby, Multivariate Analysis, Academic Press, London, 1979.