Available online at www.sciencedirect.com
SCIENCE DIRECT
Journal of Chromatography A, 1098 (2005) 44–54
JOURNAL OF CHROMATOGRAPHY A
www.elsevier.com/locate/chroma

Prediction of retention times of proteins in hydrophobic interaction chromatography using only their amino acid composition

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*

Received 1 March 2005; received in revised form 11 August 2005; accepted 15 August 2005
Available online 8 September 2005

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, essentially, only on aminoacidic composition. The results show that such prediction is indeed possible. Our main contribution was the design of models that predict the DRT using the minimal information concerning a protein: its aminoacidic composition. The performance is similar to that observed in models that use much more sophisticated information such as the three-dimensional structure of proteins. Three models that, in addition to the amino acid composition, use different assumptions about the amino acids tendency to be exposed to the solvent, were evaluated in 12 proteins with known experimental DRT. In all the cases analyzed, the model that obtained the best results was the one based on a linear estimation of the aminoacidic surface composition. The models were adjusted using a collection of 74 vectors of aminoacidic properties plus a set of 6388 vectors derived from these using two mathematical tools: *k*-means and self-organizing maps (SOM) algorithms. The best vector was generated by the SOM algorithm and was interpreted as a hydrophobicity scale based partly on the tendency of the amino acids to be hidden in proteins. The prediction error (MSE_{JK}) obtained by this model was almost 35% smaller than that obtained by the model that supposes that all the amino acids are completely exposed and 40% smaller than that obtained by the model that uses a simple correction factor considering the general tendency of each amino acid to be exposed to the solvent. In fact, the performance of the best model based on the aminoacidic composition was 5% better than that observed in the model based on the three-dimensional structure of proteins.

Keywords: Mathematical modelling; Hydrophobic interaction chromatography; Hydrophobicity; Retention time prediction; Proteins; Amino acid composition

1. Introduction

Hydrophobic interaction chromatography (HIC) is a technique used for the purification of proteins, which is based on the hydrophobic properties of the molecular surface and their interaction with a stationary matrix of non-polar molecules. Therefore, the separation of a protein mixture takes place when differences occur in the degree of interaction between proteins and the stationary matrix. The magnitude of these differences will affect the resolution and purification levels achieved by this technique [1].

The HIC principles show high sensitivity to changes in the tertiary structure of proteins as, for instance, the exposition of non-polar groups on the surface as the result of a polypeptide chain incorrectly misfolded or damaged. Also, these characteristics allow the identification of analogous proteins or degradation products, in addition to other contaminants. In the same way, this sensitivity allows its use as an analytical tool to detect conformational changes in proteins [2–6].

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 within the protein purification process, that follows ion exchange chromatography. It has been shown that the rational design of industrial protein purification processes normally requires an HIC stage [7].

On the other hand, hydrophobicity plays a key role in the definition of a protein's behaviour and this property is considered as one of the fundamental forces that govern protein folding [8]. Moreover, the hydrophobic characteristics of a protein perform a fundamental role defining its behaviour in solution and its

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

0021-9673/$ – see front matter © 2005 Elsevier B.V. All rights reserved.
doi:10.1016/j.chroma.2005.08.039

interaction with other biomolecules. The hydrophobicity value of a protein can be assigned by many different methodologies either experimental or theoretical. A method for establishing the hydrophobicity of a protein consists in considering the relative contribution of each one of the amino acids present on the surface, defining an average surface hydrophobicity (ASH) [9]. Using this definition, Lienqueo et al. found that the ASH can be correlated satisfactorily with retention times in hydrophobic interaction chromatography [10]. In this case, the aminoacidic hydrophobicity scales that best modelled the behaviour were those of Miyazawa-Jernigan [11] and Cowan-Whittaker [12].

In order to calculate the ASH, it is necessary to have the three-dimensional protein structure. Frequently this data does not exist, and the only information available is the amino acid sequence. In these cases, to estimate the surface composition of the protein, it is necessary to start with the construction of three-dimensional models, usually using the methodology of comparative modelling [13], or in some cases through the development of Ab initio models. As these methodologies are complex and time consuming, it would be desirable to investigate a methodology by which retention time could be determined when only the protein aminoacidic composition is available.

Some features of proteins can be predicted based on their aminoacidic composition. For example, it has been reported that the prediction of the protein's secondary structural content [14], and the protein structural class [15] can be carried out successfully from its aminoacidic composition only. In a previous paper, we investigated the prediction of ASH calculated with the hydrophobicity scales of Berggren and Cowan-Whittaker using mathematical models based on the aminoacidic composition and measurements of the amino acids tendency to exposition. We found that it is possible to predict the ASH of a protein in an acceptable degree starting from its aminoacidic composition, obtaining in the best case a correlation coefficient of 0.836 [16].

Therefore, the main objective of this paper is to investigate if it is possible to predict the retention time of a protein in HIC only from its aminoacidic composition using appropriate mathematical models.

2. Materials and methods

2.1. Materials

Twelve proteins with known dimensionless retention time (DRT) and three-dimensional structure were used: Cytochrome C (1HRC), Ribonuclease A (1AFU), 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). DRT values correspond to those used by Lienqueo et al. in [10].

Briefly, DRT corresponds to the dimensionless protein retention time observed in a hydrophobic interaction column, calculated according to:

$$\mathrm{DRT} = \frac{t_\mathrm{R} - t_0}{t_\mathrm{f} - t_0} \tag{1}$$

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.

DRT values used in this work were obtained in a 1 ml Phenyl-Sepharose Fast Flow column using 2 M ammonium sulfate as the eluent.

2.2. Mathematical models

2.2.1. DRT 0 model—average surface properties

The DRT 0 model is similar to that proposed by Lienqueo et al. [10]; however, our model considers a wider set of aminoacidic properties and not only hydrophobicity scales. We modelled the dimensionless retention time (DRT) using average surface properties (ASP) of proteins by the following equation:

$$\mathrm{DRT} = b_0 + b_1\Gamma + b_2\Gamma^2 \tag{2}$$

where Γ corresponds to the ASP of the protein and b_i to the adjustable coefficients of the quadratic model obtained by the least square procedure. The Γ of a protein was computed assuming that each amino acid on the protein surface contributes, proportionally to its abundance, to the properties associated to the protein surface [9]. According to the previous hypothesis, Γ can be calculated by the following equation:

$$\Gamma = \sum_{i \in A} \hat{r}_i \varphi_i \tag{3}$$

where A is the collection of the 20 possible amino acids and φ_i is the ith component of an aminoacidic property vector (APV). Lienqueo et al. used φ equal to a hydrophobicity scale ϕ for the modelling and prediction of DRT. They found that amongst a diversity of hydrophobicity scales, the best scales for the prediction of DRT were those of Miyazawa-Jernigan and Cowan-Whittaker. Here, however, additional properties were considered. Therefore, the expressions represented by Eqs. (2) and (3) correspond to a generalization of the equations used by Lienqueo et al. Finally, the variable, $\hat{r}_i$ represents the fraction of surface area occupied by the amino acids of class i and it is given by:

$$\hat{r}_i = \frac{S_i}{\sum_{j \in A} S_j} \tag{4}$$

where S_i is the sum of the accessible surface area (ASA) for all the amino acids of class i. The ASA was calculated using the software STRIDE from the three-dimensional structure of the proteins [17].

2.2.2. DRT I model—aminoacidic composition

In the case of the models based on the aminoacidic composition of the protein, Γ is estimated without using the three-dimensional structure of the protein. Three linear models described and analyzed in a previous paper were used [16].

DRT I model supposes that all the amino acids are completely exposed, so Γ is estimated by the following equation:

$$\Gamma^{\mathrm{I}} = c_0 + \sum_{i=1}^{20} c_i \hat{a}_i^{\mathrm{I}} + c_{21}\hat{l} \tag{5}$$

where c_i (i from 0 to 21) corresponds to the parameters of the linear model obtained by the least squares procedure, $\hat{l}$ is the ratio between the length of the protein sequence and the maximum length observed in the working database. The value $\hat{a}_i^{\mathrm{I}}$ corresponds to the fraction of the maximum accessible surface of the amino acids of class i when they are totally exposed, defined by:

$$\hat{a}_i^{\mathrm{I}} = \frac{n_i S_{\max,i}}{\sum_{j \in A} n_j S_{\max,j}} \tag{6}$$

where n_i is the number of amino acids of class i in the protein and $S_{\max,i}$ is the maximum possible value of ASA, obtained when arranging the amino acids of class i in a extended conformation tripeptide G-X-G [18]. The values of $S_{\max}$ in Å^2 are 113 (Ala), 241 (Arg), 158 (Asn), 151 (Asp), 140 (Cys), 189 (Gln), 183 (Glu), 85 (Gly), 194 (His), 182 (Ile), 180 (Leu), 211 (Lys), 204 (Met), 218 (Phe), 143 (Pro), 122 (Ser), 146 (Thr), 259 (Trp), 229 (Tyr), 160 (Val).

2.2.3. DRT II model—aminoacidic composition and exposition factor

The DRT II model incorporates a correction factor that considers the general tendency of each amino acid to be exposed to the solvent. In previous work, we found that the best results were obtained using a correction factor α equal to an estimation of the probability that an amino acid of class i had a RASA superior to a threshold $\mu = 0.6$ [16]. The RASA of an amino acid k in a protein is defined as the ratio between their ASA (s_k) and their maximum ASA ($S_{\max,k}$). Then, in the DRT II model, Γ is estimated by Eq. (5) where $\hat{a}^{\mathrm{I}}$ is replaced by $\hat{a}^{\mathrm{II}}$ given by:

$$\hat{a}_i^{\mathrm{II}} = \frac{n_i S_{\max,i} \alpha_i}{\sum_{j \in A} n_j S_{\max,j} \alpha_j} \tag{7}$$

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

2.2.4. DRT III model—aminoacidic composition and linear estimation of the surface

Finally, in the case of the DRT III model, we establish a linear relationship amongst the ASA S_i for all the amino acids of class i and the maximum possible ASA defined for $n_i S_{\max,i}$. Then, in the DRT III model, Γ is estimated by Eq. (5) when $\hat{a}^{\mathrm{I}}$ is replaced by $\hat{a}^{\mathrm{III}}$ described by:

$$\hat{a}_i^{\mathrm{III}} = \frac{n_i S_{\max,i} \beta_i + \eta_i}{\sum_{j \in A} (n_j S_{\max,j} \beta_j + \eta_j)} \tag{8}$$

where β_i and η_i are the coefficients of the linear model between S_i and $n_i S_{\max,i}$ calculated for all the amino acids of class i present in the working database using the least squares procedure [16].

By definition, the sum of coefficients $\hat{a}_i$ is one, so these coefficients conform a linear depending system. Therefore, the models analyzed in this work do not consider the data provided by the amino acid with φ_i equal to zero.

The determination of c_i coefficients of Eq. (5) was carried out by means of a least square adjustment on a set of 1982 proteins (working database) with known three-dimensional structure [16]. This set was derived from the non-redundant protein selection (identity cut-off 25%) published by Hobohm and collaborators in December of 2003 [19]. This subset was constructed eliminating the membrane proteins. The three-dimensional structures were obtained from the PDB database [20]. The c_i coefficients were determined as the average observed on 100 repetitions using different randomly generated subsets.

2.3. Collection of aminoacidic property vectors (APV)

A collection of 74 aminoacidic property vectors (APV) 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. [11–12,21–60]. All members in the APV collection were mathematically scaled at 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 to hydrophobic scales assigning 0 to the most hydrophilic amino acid and 1 to the most hydrophobic, the value for the rest of the amino acids was determined linearly. Other vectors non-associated to hydrophobicity scales were not modified.

2.4. Collection of derived APV

Additionally to the APV collection obtained from the literature, a set of vectors derived from these was used. This new set was constructed using algorithms that allow analysis of the underlying topology in high dimensionality data sets. The algorithms used in this work were k-means [61] and self-organizing maps (SOM) [62].

2.4.1. k-Means algorithm

The k-means algorithm can be described as a method to split a data set in k groups. Each group is represented by a prototype vector, which corresponds to the centroid of the vectors that belong to it. Given a fixed number of k prototype vectors at first located randomly in the space, the objective of this algorithm is to move the prototype vectors, so, for instance, the sum of the distance between each prototype and the set of vectors that it represents is minimized. If the number of partitions selected is suitable, it is possible to suppose that each prototype vector synthesizes in some way the characteristics of its group.

2.4.2. Self-organizing maps (SOM) algorithm

In the case of the SOM algorithm, the basic idea is similar, although in this case, the prototype vectors are, in addition, mapped to an ordered structure usually bi-dimensional called Kohonen map. The algorithm objective, however, is more ambitious than in the case of the k-means algorithm, since the bi-

dimensional structure must maintain the topological relations observed in the multidimensional space. In this way, if two prototypes are centroids of two groups of similar data, the map would have to maintain this relation locating these prototypes in the same zone or neighborhood in its structure.

Both algorithms can be trapped in local minima, so their performance will depend to a great extent on the initial location of the prototype vectors. For this reason, in the case of the *k*-means algorithm, each execution was repeated 100 times with different initial vectors. Since the SOM algorithm is very intensive in computational time the repetitions, in this case, only took place 20 times. On the other hand, the determination of the optimal *k*-value, in the case of the *k*-means algorithm, is not trivial. Therefore, a systematic test of all the *k*-values between 2 and 73 was carried out. In the same way, the a priori determination of the optimal dimensions of the Kohonen map is not easy either, so 16 maps with dimensions in the following sequence were evaluated: 1×2, 2×2, 2×3, 3×3, 3×4, ..., 8×9, 9×9. Finally, in each case, the best vector, in terms of its performance as APV in the models described in the previous section, was conserved.

2.5. Measurement of the model performance

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:

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

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

where DRT_k is the experimental DRT of protein k, $\widehat{\mathrm{DRT}_k}$ is the prediction of the DRT for protein k and $N = 12$ is the number of proteins with experimentally known DRT used.

The MSE_{JK} was used to estimate the prediction error of the models when using proteins not considered in the training data set. In this case, the size of the data set is modest, hence other techniques of re-sampling like *k*-folding cross-validation, boostrap or the use of an independent test set cannot be used. The Jack Knife re-sampling method (leave-one-out) is a widely known methodology [63]. Actually, it is regarded as the most objective and effective tool for the evaluation of predictor models [64,65]. The mathematical principle and a comprehensive discussion about this can be found in [66]. Briefly, this method consists in 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 *k*th adjustment, the *k*th element of the data is not considered. The model determined by means of the *k*th 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 *k*th element has been left out. So, the MSE_{JK} is obtained calculating the average on the collection of *N* proteins as indicated in the following equation:

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

3. Results and discussion

3.1. Analysis of the aminoacidic property vector (APV) collection

The aminoacidic property vector (APV) collection represents the distribution of physical, chemical and biological properties on the set of 20 amino acids. These 74 vectors are distributed in a vector space defined by their 20 components. In order to study the characteristics of this distribution, a principal component analysis (PCA) considering all these vectors was carried out. This was done setting the vectors as observations and their components as variables.

Fig. 1 shows a pareto graph which details the relative contribution of each principal component in the total variance observed in the collection. This contribution was related to the magnitude of the eigen value associated to each principal component. This graph indicates that 77% of the variability present in the APV collection is captured by the first four principal components. By means of this, it is possible to reduce the dimensionality of the APV collection from 20 to 4 (80%) with less than a 23% information loss.

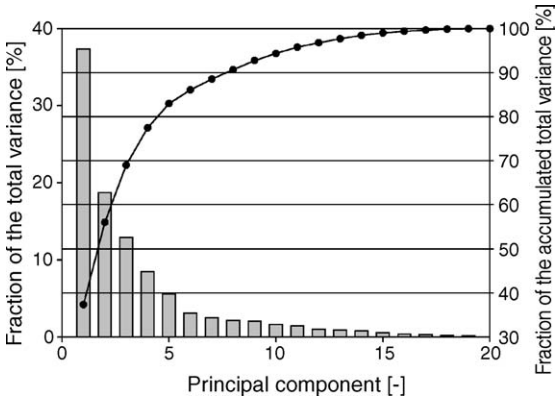

Fig. 1. Pareto plot of the variance contributed by each principal component obtained from the principal component analysis of the collection of aminoacidic property vectors (AVP). The variance was obtained from the eigen value associated to each principal component.

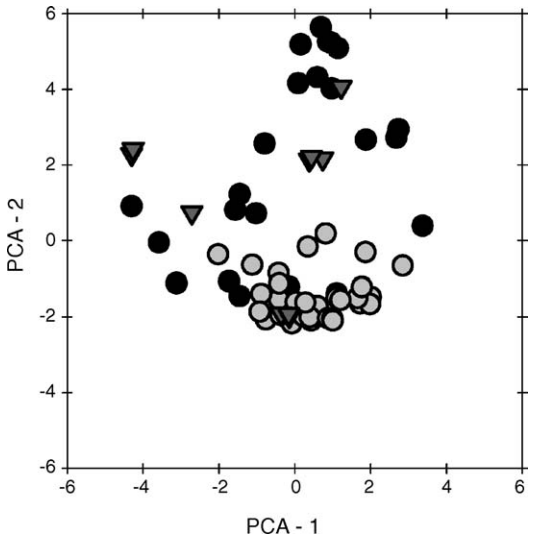

Fig. 2. Scatter plot between the two first components of the collection of aminoacidic property vectors when it has been projected on its two principal components. These vectors were separated in three qualitative categories: all vectors based on hydrophobicity scales (●); vectors constructed from a statistical analysis of different properties (▼); all the rest (●).

The reduction on the dimensionality of the APV collection allowed us to construct the scatter plot shown in Fig. 2. This figure shows the distribution of the members of the APV collection when they have been projected in the two first principal directions of the collection (PCA-1 and PCA-2). Additionally, the vectors were separated in three categories defined qualitatively. The first category contains all APV based on hydrophobicity scales (37); the second, all APV constructed from a statistical analysis of different properties (11); and the third, all the rest (26). The scatter plot of Fig. 2 suggests that the APV collection is distributed in the space of characteristics in a non-uniform way. It is observed that the APV are grouped in clusters that are relatively visible at first. Also, it is possible to notice that most of the members of the hydrophobicity scales category are separated from the rest, in a relatively defined cluster. That cluster also contains three members of category two; coincidently these three vectors were identified in the literature as associated to hydrophobic properties of amino acids [38,43,54].

This analysis allowed us to observe that although the APV associated to hydrophobicity scales was a relatively defined cluster, there is considerable diversity amongst them. This diversity may allow the construction of models with different properties and also will serve as a good starting point to derive new mixtures of vectors from the original ones.

3.2. Statistical Jack Knife evaluation of the DRT 0 model

Before analyzing the models based on the aminoacidic composition, a line of reference was established. This reference was established through the study of the DRT 0 model that uses the three-dimensional structure of the protein, not only using hydrophobicity scales, considering one by one the vectors belonging to the APV collection. This model was evaluated for the set of 12 proteins with know experimental DRT and the results of this evaluation are shown in Tables 1 and 2.

Table 1 shows the results obtained for the five best APV in ascending order with respect to the MSE. These results are consistent with those reported by Lienqueo et al. [10] and they show that, in general, the APV that gave the smaller MSE values correspond to vectors associated to hydrophobicity scales obtained through different methodologies. However, the MSE_{JK} values

Table 1
Effect of the aminoacidic property vectors (APV) on the performance indices of the DRT 0 model (based on the three-dimensional structure) in the prediction of the experimental DRT of the 12 proteins

No.	APV	Description	MSE × 10^3	Pearson	MSE_{JK} × 10^3
1	Miyazawa and Jerningan [11]	Hydrophobicity scale (contact energy derived from 3D data)	**5.812**	0.946	21.016
2	Cowan and Whittaker [12]	Hydrophobicity indices at pH 3.4 determined by HPLC	**6.902**	0.936	21.041
3	Deleage and Roux [31]	Conformational parameter for β-sheet	**7.572**	0.929	19.465
4	Browne [26]	Retention coefficient in heptafluorobutyric acid (HFBA)	**7.850**	0.926	16.869
5	Willson [58]	Hydrophobic constants derived from HPLC peptide retention times	**8.237**	0.923	19.298

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

Table 2
Effect of the aminoacidic property vectors (APV) on the performance indices of the DRT 0 model (based on the three-dimensional structure) in the prediction of the experimental DRT of the 12 proteins

No.	APV	Description	MSE × 10^3	Pearson	MSE_{JK} × 10^3
1	Wertz and Scheraga [57]	Fraction of buried amino acid in 20 proteins	8.754	0.917	**12.988**
2	Grantham [36]	Atomic weight ratio of hetero (non carbon) elements in end groups or rings to carbons in the side chain	8.657	0.918	**14.210**
3	Browne [26]	Retention coefficient in heptafluorobutyric acid (HFBA)	7.850	0.926	**16.869**
4	Willson [58]	Hydrophobic constants derived from HPLC peptide retention times	8.237	0.923	**19.298**
5	Deleage and Roux [31]	Conformational parameter for β-sheet	7.572	0.929	**19.465**

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.

in Table 1 do not follow the ascending order that those of MSE. This is reasonable, since it is known that the MSE corresponds to an optimistic estimation of the prediction error, produced by a loss in the predictive capacity of the model consequence of an over fitting to the training data and therefore Table 2 was constructed. This table shows the five best APV now ordered based on their MSE_{JK}.

Table 2 shows that, based on the analysis of the MSE_{JK}, the best APV in terms to assure a good predictive performance was the Wertz and Scheraga vector. This vector was constructed based on a measurement of the amino acid tendency to be hidden in proteins [57]. The MSE obtained by the Wertz and Scheraga vector was 8.754×10^{-3} which meant an increase of a 50% in relation to the one obtained by the Miyazawa–Jernigan vector. Nevertheless, the MSE_{JK} indicates that next to the increase in the MSE, a diminishment of 38.2% in the MSE_{JK} took place, allowing a substantial improvement in the predictive capacity of the model. This corresponds to an improvement with respect to the methodology proposed by Lienqueo et al., where the evaluation of the predictive performance of the model was carried out in only one protein set [10]. Therefore, the methodology proposed by Lienqueo et al. can be considered highly biased, since it depends strongly on the criterion used to construct the evaluation set. On the contrary, the methodology proposed in this paper estimates the prediction error of the model through the determination of the impact of the removal of each one of the elements in the data set in the model performance and hence it is more robust.

The second place in Table 2 corresponded to the Grantham vector. This vector represents an index of atomic composition defined as the atomic weight ratio of hetero (non carbon) elements in end groups or rings to carbons in the side chain [36]. The rest of the vectors are associated, in their majority, to hydrophobicity or exposition/hidden scales that also represents, in some way, hydrophobic characteristics.

3.3. Modelling and prediction of the dimensionless retention time (DRT) of a protein based on its aminoacidic composition

Having defined a line of reference, the results obtained with the models based on composition follow. Three models based on different assumptions about the amino acids tendency to be exposed to the solvent were evaluated: the first one supposes all the amino acids completely exposed (DRT I), the next one uses a simple correction factor considering the general tendency of each amino acid to be exposed (DRT II) and the last one is based on a linear estimation of the aminoacidic surface composition (DRT III).

The DRT I model reached a minimum MSE_{JK} equal to 22.749×10^{-3} when it was constructed using the Grantham vector and therefore was 1.8 times the minimum value obtained by the DRT 0 model (12.988×10^{-3}). The results obtained using the DRT II model were worse. The lowest MSE_{JK} (24.839×10^{-3}) was obtained when using the relative mutability vector proposed by Dayhoff [30]. The aminoacidic property represented by the vector of Dayhoff allows us to affirm that it corresponds to a model artefact and not to behaviour defined by the physical, chemical or biological nature of the vector.

With respect to DRT III model, Table 3 shows its performance when it was constructed using the same APV shown in Table 2. It is possible to note that the MSE_{JK} values in this table are not ordered in an ascending way, as, in this case, the DRT III model was evaluated like an estimator of the DRT 0 model and, of course, the quality of this approach depends strongly on the APV used. For the case of the Wertz and Scheraga vector, the difference in the MSE_{JK} between both models is considerable: DRT III model obtained a MSE_{JK} of 25.262×10^{-3}, this is more of the double of the value obtained by DRT 0 model using the same APV. On the other hand, in the fourth position in Table 3 is the Willson APV. Using this vector in the DRT III model, a MSE_{JK} equal to 13.501×10^{-3} was obtained, this is 3.9% greater than the value obtained by the DRT 0 model in its best case. This observation suggests the construction of Table 4, where the five better APV used to construct DRT III model are listed.

According to Table 4, the minimum MSE_{JK} for the DRT III model was obtained when using the Willson APV. This vector represents a hydrophobicity scale based on retention times in HPLC [58]. This vector obtained an MSE_{JK} equal to 13.501×10^{-3}. This MSE_{JK} was considered quite acceptable, since it is only 3.9% greater than the value obtained by the DRT 0 model. As in the case of the DRT 0 model, the APV present in Table 4 correspond, in their majority, to hydrophobicity scales. The appearance of the Sandberg and Jonson APVs is interesting, since both were derived by means of a statistical analysis on a diverse set of aminoacidic properties. In particular, the Sandberg vector is located in the third position and origi-

Table 3
Effect of the aminoacidic property vectors (APV) on the performance indices of the DRT III model (the best fit using only amino acid composition of the protein) in the prediction of the experimental DRT of the 12 proteins

No.	APV	Description	$MSE \times 10^3$	Pearson	$MSE_{JK} \times 10^3$
1	Wertz and Scheraga [57]	Fraction of buried amino acid in 20 proteins	18.168	0.820	**25.262**
2	Grantham [36]	Atomic weight ratio of hetero (non carbon) elements in end groups or rings to carbons in the side chain	18.980	0.811	**41.441**
3	Browne [26]	Retention coefficient in heptafluorobutyric acid (HFBA)	34.606	0.612	**102.666**
4	Willson [58]	Hydrophobic constants derived from HPLC peptide retention times	7.377	0.931	**13.501**
5	Deleage and Roux [31]	Conformational parameter for β-sheet	27.346	0.711	**58.317**

This table maintains the APV and the order defined by Table 2, so it can be compared directly. The mean square error (MSE), the correlation coefficient (Pearson) and the Jack Knife cross-validation mean square error (MSE_{JK}) are shown.

Table 4
Effect of the aminoacidic property vectors (APV) on the performance indices of the DRT III model (the best fit using only amino acid composition of the protein) in the prediction of the experimental DRT of the 12 proteins

No.	APV	Description	MSE × 10^3	Pearson	MSE_{JK} × 10^3
1	Willson [58]	Hydrophobic constants derived from HPLC peptide retention times	7.377	0.931	**13.501**
2	Cowan and Whittaker [12]	Hydrophobicity indices at pH 3.4 determined by HPLC	9.225	0.913	**15.508**
3	Sandberg [54]	Statistical analysis of aminoacidic properties, *z*3	17.310	0.829	**22.963**
4	Abraham and Leo [22]	Hydrophobicity scale	8.760	0.917	**23.023**
5	Jonson [43]	Statistical analysis of aminoacidic properties, *z*1	10.068	0.904	**24.016**

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.

nally it was described by its authors as a measurement of the amino acids electronic characteristics (for example, pK_a and p*I*) [54].

3.4. Modelling and prediction of the dimensionless retention time (DRT) of a protein based on its aminoacidic composition and derived aminoacidic property vectors (APV) using k-means and SOM algorithms

In this section, the results obtained by evaluating models DRT I, II and III on the basis of a new set of APV are described. This set was constructed with the APVs derived from the original APVs using the *k*-means and self-organizing maps (SOM) algorithms. In total, 6388 vectors were generated, 5340 from the *k*-means algorithm and 1572 using SOM algorithm.

3.4.1. APV derived using k-means algorithm

The performance of the best vectors found using the *k*-means algorithm is detailed in Table 5. This table shows that, in the case of the models DRT 0 and DRT I, the *k*-means algorithm was unable to locate prototypes that were able to improve the results shown previously. For these models, the algorithm located the best prototype in the same position as the best vectors found previously: Wertz and Scheraga and Grantham, in the case of DRT 0 and DRT I models, respectively.

In the case of models DRT II and III, this algorithm found two vectors that allowed an improvement in the MSE_{JK} in both cases. With respect to the DRT II model, a vector was found that was able to decrease the value of MSE_{JK} by almost 17%. The vector was very near to the *z*3 APV of Hellberg, and therefore, can be interpreted as a variation of it. The *z*3 vector of Hellberg is analogous to the *z*3 vector of Sandberg and it is described by its authors as a measurement of the amino acids electronic characteristics [38,54].

The best vector obtained for the DRT III model was found when a uniform distribution was used to initialize the centroides and it was obtained when considering $k = 31$. By means of this vector, the MSE_{JK} decreases a little more than 4% with respect to the value obtained using the Willson vector. This vector turned out to be a centroid of the vectors of Fauchere, Willson and Hopp [34,58,39] ordered on the basis of its distance to their centroid. All of them represent hydrophobicity scales, and therefore this vector can be interpreted as a consensus amongst them.

3.4.2. APV derived using the SOM algorithm

Table 6 shows that, as in the case of the *k*-means algorithm, the SOM algorithm was unable to find a vector that allowed an improvement in the results shown by the DRT 0 model when using the Wertz and Scheraga vector. The best prototype located by means of the SOM algorithm was found when using a 7×7 map built on the set of standardized APV. This vector presented an MSE_{JK} 11.7% greater than that obtained by the Wertz and Scheraga vector and corresponds to the prototype located in the proximities of the Wertz and Scheraga vector.

The DRT I model improved its performance by decreasing the MSE_{JK} in 16%. In this case, the closest original vector was the hydrophobicity scale of Erikkson, based on the change of the free energy in the transference of the amino acids from ethanol to water. The DRT II model did not present an improvement.

With respect to the DRT III model, the SOM algorithm found a prototype that improved the results obtained by the Willson vector by 7.3%. This vector was found when processing the not standardized original APVs by means of a 7×7 SOM map. When analyzing the characteristics of the map generated, it was found that this vector corresponds to the centroid of Cowan and

Table 5
Performance indices for the models on the prediction of DRT of the 12 proteins for the best APV obtained by the *k*-means algorithm

Model	MSE × 10^3	Pearson	MSE_{JK} × 10^3	*k*	*D*	*P*	Closest APV	
							APV	Description
DRT 0	8.754	0.917	**12.988**	50	0	1	Wertz and Scheraga [57]	Fraction of buried amino acid in 20 proteins
DRT I	10.401	0.901	**22.749**	54	0	1	Grantham [36]	At. weight ratio of hetero elements in end groups or rings to carbons in the side chain
DRT II	15.331	0.850	**20.553**	59	0.16	0.99	Hellberg [38]	Statistical analysis of aminoacidic properties, *z*3
DRT III	8.150	0.923	**12.914**	31	0.377	0.917	Fauchere [34]	Hydrophobicity scale (pi-r)

The mean square error (MSE), the correlation coefficient (Pearson) and the Jack Knife cross-validation mean square error (MSE_{JK}) are shown. In addition, distance (*D*), Pearson (*P*), *k* parameter and description of the closest APV are shown.

Table 6
Performance indices for the models on the prediction of DRT of the 12 proteins for the best APV obtained by the SOM algorithm

Model	MSE × 10^3	Pearson	MSE_{JK} × 10^3	Size	*D*	*P*	Closest APV	
							APV	Description
DRT 0	9.566	0.909	**14.508**	7 × 7	0.614	0.938	Wertz and Scheraga [57]	Fraction of buried amino acid in 20 proteins
DRT I	11.108	0.894	**19.112**	6 × 6	1.281	0.518	Erikkson [33]	ΔG in the transference of the amino acids from ethanol to water
DRT II	14.835	0.855	**24.658**	3 × 3	0.938	0.771	Chou and Fasman [29]	Conformational parameter for β-turn calculated on 29 proteins
DRT III	7.739	0.927	**12.332**	7 × 7	0.358	0.969	Abraham and Leo [22]	Hydrophobicity scale

The mean square error (MSE), the correlation coefficient (Pearson) and the Jack Knife cross-validation mean square error (MSE_{JK}) are shown. In addition, distance (*D*), Pearson (*P*), map size and description of the closest APV are shown.

Whittaker [12] and Abraham and Leo [22] vectors, both corresponding to hydrophobicity scales. This suggested that this vector corresponds to a consensus between these two hydrophobicity scales.

3.5. Final discussion

Table 7 shows the best APVs found in this work for each one of the DRT models considered. As has been mentioned, in the case of the vectors obtained by means of the *k*-means and SOM algorithms, these can be interpreted, in a certain sense, as mixtures of the original APV found in the literature. In particular, most of these were related to APVs associated to hydrophobicity scales. This fact is very important, as it is concordant with the application in which they are being used. An exception is the vector associated to the DRT II model, which was interpreted as an expression of the electronic properties of the amino acids.

The best models were those of the DRT 0 using the Wertz and Scheraga vector and DRT III using vector SOM 7 × 7. The coefficient of correlation between these vectors was of 0.763, which indicates that both vectors present differences. The analysis of these vectors showed that both agree in the allocation of the hydrophobic amino acids: phenylalanine, leucine, valine and isoleucine. Also, both vectors tend to privilege the non-polar group of amphipatic amino acids, i.e. that they contain, simultaneously, polar and non-polar groups, as for example, tryptophan, methionine and tyrosine. The greater discrepancy is in the case of the hydrophobic amino acid proline. The Wertz and Scheraga vector assigns an almost null hydrophobicity to it, whereas vector SOM 7 × 7, a medium-high hydrophobicity. In the case of the hydrophilic amino acids, both vectors agree, locating lysine as the most hydrophilic amino acid.

Also, important discrepancies in the other hydrophilic amino acids exist. The most important differences concern arginine and histidine. The vector SOM 7 × 7 tended assigning values lower than those found in the Wertz and Scheraga vector. The most remarkable case corresponds to cysteine, since both vectors assign a great hydrophobic character to it. This is explained by the fact that, although cysteine is usually classified as a hydrophilic amino acid, it tends to be inside of proteins forming disulfide bridges. On the basis of these observations, it is possible to conclude that the vector SOM 7 × 7 represents a synthesis between the hydrophobic character of the amino acids and their tendency to be located inside of proteins.

It was observed that the c_i coefficients from Eq. (5) presented, in most cases, variabilities smaller than 5% for the models DRT I, II and III. These variabilities were obtained from standard deviation in all repetitions. The highest variability was found in the coefficient associated to arginine, which in the case of the DRT III model displayed a variability of 110%. On the other hand, Table 8 shows that the b_i coefficients had, in general, confidence intervals at 95% of considerable size. Coefficients

Table 7
Improved amino acid property vectors (APV) found in this work for each one of the DRT models considered

AA	DRT 0 Wertz and Scheraga [57]	DRT I SOM 6 × 6	DRT II *k*-means, *k* = 59	DRT III SOM 7 × 7
ALA	0.375	0.013	0.517	0.519
ARG	0.321	0.633	0.000	0.026
ASN	0.196	0.214	0.598	0.254
ASP	0.107	0.082	0.760	0.333
CYS	0.929	0.158	1.000	0.732
GLN	0.071	0.522	0.295	0.283
GLU	0.125	0.066	0.458	0.345
GLY	0.179	0.044	0.515	0.492
HIS	0.696	0.351	0.561	0.222
ILE	0.857	0.729	0.283	0.979
LEU	0.821	0.235	0.302	0.941
LYS	0.000	0.459	0.090	0.000
MET	0.804	0.297	0.475	0.781
PHE	1.000	0.409	0.573	1.000
PRO	0.071	0.000	0.745	0.706
SER	0.321	0.101	0.587	0.358
THR	0.125	1.000	0.300	0.427
TRP	0.982	0.429	0.567	0.966
TYR	0.589	0.687	0.500	0.733
VAL	0.732	0.683	0.278	0.825

Table 8
Coefficients b_i from Eq. (2) for the models based on the aminoacidic composition

b_i	DRT I	DRT II	DRT III
b_0	−53.46 ± 21.97	−8.93 ± 22.04	−14.98 ± 10.29
b_1	314.80 ± 126.70	39.50 ± 113.30	75.32 ± 53.78
b_2	−456.40 ± 182.00	−39.55 ± 145.05	−90.15 ± 69.75
Γ_{min}	0.317	0.345	0.327
Γ_{max}	0.387	0.430	0.430

The confidence interval at 95% and the rank of the variable Γ in which they were obtained are included.

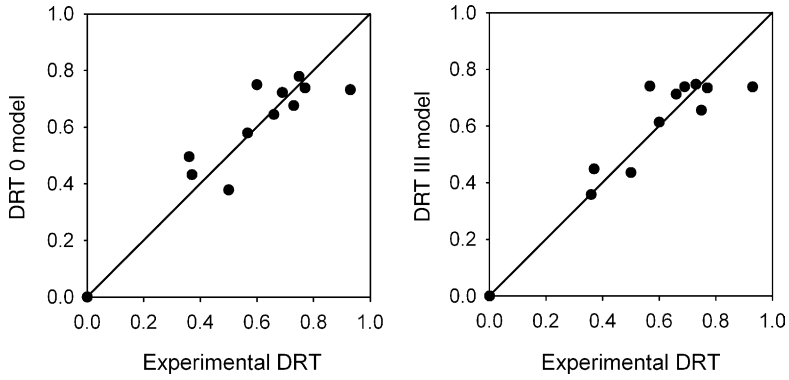

Fig. 3. Scatter plots between the experimental dimensionless retention time (DRT) and DRT predicted by the DRT 0 model based on the three-dimensional structure of the proteins and DRT III model based on the aminoacidic composition of the proteins.

determined with greater uncertainty were associated to DRT I model.

Fig. 3 shows the scatter plots between experimental DRT and those estimated by means of DRT 0 and DRT III models. The plots in Fig. 3 do not suggest a pattern between the dispersions and the experimental DRT. However, these plots show that DRT 0 and DRT III share the fact that one of the greater errors is in the zone where the most hydrophobic proteins are located. This observation can also be observed in Fig. 4, where a plot of the residual error of the models for each protein is shown. Additionally, the experimental DRT and the predictions carried out by the DRT III model are in Table 9. The plot in Fig. 4 shows that the biggest error was located in the protein α-lactalbumin (1A4V), followed by lysozyme (2LYM) in the case of the DRT 0 model and ovalbumin (1OVA) in the case of the DRT III model. On the contrary, in the case of the DRT III model, the smaller errors were found in cytochrome C (1HRC), ribonuclease A (1AFU), lysozyme (2LYM) and α-chymotrypsin (4CHA). A relation between the magnitude of the error and the length of the protein sequence was not observed, low residual errors in small proteins such as cytochrome C (104 aa) or of greater size like conalbumin (682 aa) were observed. This allows us to state that the DRT III model (like the model DRT 0) is able to predict the retention times for a wide set of proteins, monomeric and multimeric and that the prediction error is not related to the length of these, nor to their average surface hydrophobicity.

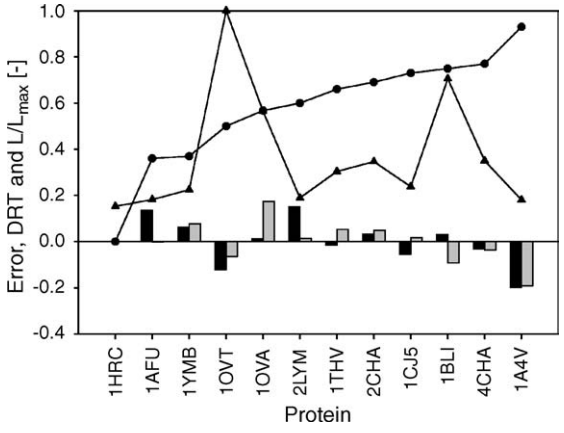

Fig. 4. Plot of the residual error between the experimental dimensionless retention time (DRT) and DRT estimated by the DRT 0 model (■) and the DRT III model (■). The experimental DRT (●), and the dimensionless length (▲) are also shown.

Table 9
The experimental dimensionless retention time (DRT) and DRT predicted by the DRT III model based on the aminoacidic composition of the proteins

Protein PDB ID	Experimental DRT	DRT III model prediction
1HRC	0.000	0.000
1AFU	0.360	0.357
1YMB	0.370	0.448
1OVT	0.500	0.436
1OVA	0.567	0.741
2LYM	0.600	0.613
1THV	0.660	0.712
2CHA	0.690	0.739
1CJ5	0.730	0.747
1BLI	0.749	0.656
4CHA	0.770	0.734
1A4V	0.930	0.738

4. Conclusions

In this paper, the prediction of the dimensionless retention time of proteins (DRT) in hydrophobic interaction chromatography (HIC) by means of mathematical models based, essentially, only on the aminoacidic composition was investigated. The results presented in this work show that such prediction was indeed possible, with a performance similar to that observed in models that use much more sophisticated information like the three-dimensional structure of proteins.

A DRT prediction model based on information concerning the three-dimensional structure of proteins was proposed by Lienqueo et al. [10]. They selected the Miyazawa and Jernigan hydrophobicity vector in the process of adjusting the parameters of their model. In that context, we showed that a model (called DRT 0), constructed using the Wetz and Scheraga vector, is better, since the Jack Knife estimation of the prediction error was 38.2% smaller than the one based on the Miyazawa and Jernigan vector. We used the Jack Knife methodology due it estimates the prediction error of the model through the determination of the impact of the removal of each one of the elements in the data

set in the model performance. In this case, the size of the data set is modest and therefore this approach is more robust than the arbitrary division of the data set in a training and test set.

Our main contribution was the design of models that predict the DRT using the minimal information concerning a protein: its aminoacidic composition. We did not take into account the protein amino acid sequence, nor its secondary structure nor its three-dimensional structure. Three models based on different assumptions about the amino acids tendency to be exposed to the solvent were evaluated. In all the cases analyzed, the model that gave best results was the one based on a linear estimation of the aminoacidic surface composition (DRT III). The prediction error (MSE_{JK}) obtained by this model was almost 35% smaller than that obtained by the model that supposes that all the amino acids are completely exposed (DRT I) and 40% smaller than that obtained by the model that uses a simple correction factor considering the general tendency of each amino acid to be exposed to the solvent (DRT II).

The models were adjusted using a collection of 74 vectors of aminoacidic properties, plus a set of 6388 vectors derived from these. The derived vectors were obtained using two mathematical tools: *k*-means and self-organizing maps (SOM) algorithms. The best results were observed in the DRT III model with a vector $\boldsymbol{v}$ generated by the SOM algorithm. This vector was interpreted as a hydrophobicity scale based partly on the tendency of the amino acids to be inside of proteins. The performance of DRT III with vector $\boldsymbol{v}$ was 5% better than that observed in DRT 0 which uses the three-dimensional structure of proteins. In fact, the MSE_{JK} decreased in DRT III. However, the models DRT I and DRT II obtained a MSE_{JK} 1.7 times bigger.

The best aminoacidic property vectors (APV) for models based on the aminoacidic composition were obtained using the *k*-means and SOM algorithms. These vectors were obtained from the mathematical synthesis of vectors that represent real properties of the amino acids and not as the result of a mere mathematical optimization. Therefore, their use and physical interpretation are possible. The determination of an optimal APV through, for instance, MSE_{JK} minimization, is not trivial. In fact, the large number of variables in comparison with the reduced number of proteins with known experimental DRT produces a system with an excessive number of degrees of freedom.

Finally, the relation between the quality of the prediction by DRT III and some protein properties like the length of the sequence, the retention time and the multimeric characteristics was studied. We did not identify any relation, since we observed low residual errors in proteins with different sizes and multimeric characteristics.

Acknowledgements

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

References

[1] J.A. Queiroz, C.T. Tomaz, J.M. Cabral, J. Biotechnol. 87 (2001) 143.
[2] S.L. Wu, K. Benedek, B.L. Karger, J. Chromatogr. 359 (1986) 3.
[3] S.L. Wu, A. Figueroa, B.L. Karger, J. Chromatogr. 371 (1986) 3.
[4] J. Withka, P. Moncuse, A. Baziotis, R. Maskiewicz, J. Chromatogr. 398 (1987) 175.
[5] R.E. Shansky, S.L. Wu, A. Figueroa, B.L. Karger, Chromatogr. Sci. 51 (1990) 95.
[6] E. Bramanti, F. Ferri, Ch. Sortino, M. Onor, G. Raspi, M. Venturini, Biopolymers 69 (2003) 293.
[7] J.A. Asenjo, B.A. Andrews, J. Mol. Recognit. 17 (2004) 236.
[8] W. Kauzmann, Adv. Protein Chem. 14 (1959) 1.
[9] K. Berggren, A. Wolf, J.A. Asenjo, B.A. Andrews, F. Tjerneld, Biochim. Biophys. 1596 (2002) 253.
[10] M.E. Lienqueo, A. Mahn, J.A. Asenjo, J. Chromatog. A 978 (2002) 71.
[11] S. Miyazawa, R.L. Jernigan, Macromolecules 18 (1985) 534.
[12] R. Cowan, R.G. Whittaker, Peptide Res. 3 (1990) 75.
[13] M.E. Lienqueo, A. Mahn, A. Olivera, J. Chromatogr. A, submitted for publication.
[14] T. Piližota, B. Lučić, N. Trinajstić, J. Chem. Inf. Comput. Sci. 44 (2004) 113.
[15] R.Y. Luo, Z.P. Feng, J.K. Liu, Eur. J. Biochem. 269 (2002) 4219.
[16] J.C. Salgado, I. Rapaport, J.A. Asenjo, J. Chromatogr. A 1075 (2005) 133.
[17] D. Frishman, P. Argos, Proteins 23 (1995) 566.
[18] S. Miller, J. Janin, A.M. Lesk, C. Chothia, J. Mol. Biol. 196 (1987) 641.
[19] U. Hobohm, M. Scharf, R. Schneider, C. Sander, Protein Sci. 1 (1992) 409.
[20] 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.
[21] A.A. Aboderin, Int. J. Biochem. 2 (1971) 537.
[22] D.J. Abraham, A.J. Leo, Proteins 2 (1987) 130.
[23] A. Bairoch, Release notes for Swiss-Prot release 41, February 2003.
[24] R. Bhaskaran, P.K. Ponnuswamy, Int. J. Pept. Protein Res. 32 (1988) 242.
[25] S.D. Black, D.R. Mould, Anal. Biochem. 193 (1991) 72.
[26] C.A. Browne, H.P. Bennett, S. Solomon, Anal. Biochem. 124 (1982) 201.
[27] H.B. Bull, K. Breese, Arch. Biochem. Biophys. 161 (1974) 665.
[28] C.J. Chothia, Mol. Biol. 105 (1976) 1.
[29] P.Y. Chou, G.D. Fasman, Adv. Enzym. 47 (1978) 45.
[30] M.O. Dayhoff, R.M. Schwartz, B.C. Orcutt, Atlas of Protein Sequence and Structure, vol. 5 (Suppl. 3), 1978.
[31] G. Deleage, B. Roux, Protein Eng. 1 (1987) 289.
[32] D. Eisenberg, E. Schwarz, M. Komarony, R. Wall, J. Mol. Biol. 179 (1984) 125.
[33] 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.
[34] J.L. Fauchere, V.E. Pliska, Eur. J. Med. Chem. 18 (1983) 369.
[35] S. Fraga, Can. J. Chem. 60 (1982) 2606.
[36] R. Grantham, Science 185 (1974) 862.
[37] H.R. Guy, Biophys. J. 47 (1985) 61.
[38] S. Hellberg, M. Sjöström, B. Skaberger, S. Wold, J. Med. Chem. 30 (1987) 1126.
[39] T.P. Hopp, K.R. Woods, Proc. Natl. Acad. Sci. U.S.A. 78 (1981) 3824.
[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. Argos, 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] M.J.K. Rao, P. Argos, Biochim. Biophys. Acta 869 (1986) 197.
[52] G.D. Rose, A.R. Geselowitz, G.J. Lesser, R.H. Lee, M.H. Zehfus, Science 229 (1985) 834.
[53] M.A. Roseman, J. Mol. Biol. 200 (1988) 513.
[54] M. Sandberg, L. Eriksson, J. Jonsson, M. Sjöström, S. Wold, J. Med. Chem. 41 (1998) 2481.
[55] R.M. Sweet, D. Eisenberg, J. Mol. Biol. 171 (1983) 479.
[56] G.W. Welling, W.J. Weijer, R. Van der Zee, S. Welling-Wester, FEBS Lett. 188 (1985) 215.
[57] D.H. Wertz, H.A. Scheraga, Macromolecules 11 (1978) 9.
[58] K.J. Wilson, A. Honegger, R.P. Stotzel, G.J. Hughes, Biochem. J. 199 (1981) 31.
[59] R.V. Wolfenden, L. Andersson, P.M. Cullis, C.C.F. Southgate, Biochemistry 20 (1981) 849.
[60] J.M. Zimmerman, N. Eliezer, R. Simha, J. Theor. Biol. 21 (1968) 170.
[61] G.A.F. Seber, Multivariate Observations, Wiley, New York, 1984.
[62] T. Kohonen, Self Organization and Associative Memory, third ed., Springer-Verlag, Berlin, Heidelberg, New York, Tokio, 1989.
[63] K.C. Chou, C.T. Zhang, Crit. Rev. Biochem. Mol. Biol. 30 (1995) 275.
[64] G.P. Zhou, J. Protein Chem. 17 (1998) 729.
[65] G.P. Zhou, N. Assa-Munt, Proteins 44 (2001) 57.
[66] K.V. Mardia, J.T. Kent, J.M. Bibby, Multivariate Analysis, Academic Press, London, 1979.