New approaches for predicting protein retention time in hydrophobic interaction chromatography††

M. E. Lienqueo[1]*, A. Mahn[1†], G. Navarro[1], J. C. Salgado[1], T. Perez-Acle[2], I. Rapaport[3] and J. A. Asenjo[1]

[1]Department of Chemical and Biotechnology Engineering, Centre for Biochemical Engineering and Biotechnology, University of Chile, Beauchef 861, Santiago, Chile
[2]Centre for Genomics and Bioinformatics CGB, School of Biological Sciences, Catholic University of Chile, Santiago, Chile
[3] Centre for Mathematical Modelling, University of Chile, Santiago, Chile

Hydrophobic interaction chromatography (HIC) is an important technique for the purification of proteins. In this paper, we review three different approaches for predicting protein retention time in HIC, based either on a protein's structure or on its amino-acidic composition, and we have extended one of these approaches. The first approach correlates the protein retention time in HIC with the protein average surface hydrophobicity. This methodology is based on the protein three-dimensional structure data and considers the hydrophobic contribution of the exposed amino acid residues as a weighted average. The second approach, which we have extended, is based on the high correlation level between the average surface hydrophobicity of a protein's hydrophobic interacting zone and its retention time in HIC. Finally, a third approach carries out a prediction of the average surface hydrophobicity of a protein, using only its amino-acidic composition, without knowing its three-dimensional structure. These models would make it possible to test different operating conditions for the purification of a target protein by computer simulations, and thus make it easier to select the optimal conditions, contributing to the rational design and optimization of the process.

Keywords: **hydrophobic interaction chromatography; retention time prediction; average surface hydrophobicity; local hydrophobicity**

INTRODUCTION

Hydrophobic interaction chromatography (HIC) is a powerful technique for the separation and purification of proteins and other biological compounds. HIC has found widespread use for the purification of membrane proteins, serum proteins, nuclear proteins, and recombinant proteins (Roettger and Ladisch, 1989).

HIC combines both the non-denaturing characteristics of salt precipitation and the precision of chromatography to yield high resolution and activity recoveries. HIC is based on several non-specific affinity interaction biomolecules (e.g., van del Waals type interaction, weak ion exchange interaction, and the most important hydrophobic interaction) to a weakly hydrophobic surface at high salt concentrations, followed by elution with a descending salt gradient. HIC is an ideal 'next step' after precipitation with ammonium sulfate or other salt, or elution in high salt during ion exchange chromatography (Queiroz *et al.*, 2001). Moreover, careful manipulation of the conditions can enable it to be very sensitive, for example, HIC is capable of separating proteins that differ by as few as one amino acid residue, and separating native from incorrectly folded forms (Fexby and Bülow, 2004).

The process of protein binding to and elution from HIC adsorbents, an entropic process which is driven by the release of water molecules from the solute and stationary phase surface, has been studied to increase recovery and resolution (Jennissen, 2000; Machold *et al.*, 2002; Chen and Sun, 2003; Hahn *et al.*, 2003; Lienqueo and Mahn, 2005). The theoretical background of HIC is based on an extension of the solvophobic theory developed by Imre Molnar, Wayne Melander, and Csaba Horváth (Melander and Horváth, 1977). Jennissen (2000) proposed testing different operating conditions, using the critical hydrophobicity method. This method involves three basic steps: (i) selection of an appropriate alkyl chain length; (ii) determination of the critical surface concentration of alkyl residues (critical hydrophobicity of the adsorbent); and (iii) determination of the minimal salt concentration necessary for a complete adsorption of a protein. On the other hand, Chen and Sun (2003) proposed a model to describe salt effects on protein adsorption equilibrium on the hydrophobic media, and then select the optimal operating conditions.

The main characteristics of the system affecting protein retention in HIC include concentration and type of salt (Sofer and Hagel, 1998), properties of the stationary phase resin (ligand and backbone chemistry and ligand density) (Eriksson, 1998; Ladiwala *et al.*, 2005), temperature, buffer pH and additives in the buffer (Xia *et al.*, 2005), and the operation

*Correspondence to: M. E. Lienqueo, Department of Chemical and Biotechnology Engineering, Centre for Biochemical Engineering and Biotechnology, University of Chile, Beauchef 861, Santiago, Chile.
E-mail: mlienque@ing.uchile.cl
†A. Mahn's present address is Centre for Molecular Studies of the Cell, Institute for Biomedical Sciences, School of Medicine, University of Chile, Santiago, Chile.
††This paper is published as part of a special issue entitled 'Bioaffinity 2005, Uppsala, August 14–18, Sweden'.

mode (e.g., isocratic, gradient, displacement, etc.). The main physicochemical protein properties that determine chromatographic behavior in HIC are hydrophobicity (Queiroz *et al.*, 2001) and protein size (Fausnaugh *et al.*, 1984).

In this paper we review three approaches, based on different protein hydrophobicity parameters, for predicting protein retention time in HIC, and we have extended the second of them. The first one is based on the relationship between protein retention time in HIC and protein average surface hydrophobicity, which is calculated using three-dimensional structure data and an amino acid hydrophobicity scale. The second approach considers the surface hydrophobicity distribution of a protein, and correlates the average surface hydrophobicity of the interaction zone of a protein with protein retention time in HIC. Finally, the third approach performs the average surface hydrophobicity of a protein using only its amino-acidic composition, without considering its three-dimensional structure, and then uses this structure to predict retention time in HIC.

PROTEIN CHROMATOGRAPHIC BEHAVIOR IN HIC

The chromatographic behavior can be represented by different variables, for example, elution (retention) volume, V_r, or elution (retention) time, TR, of the protein.

In the case of isocratic elution, the chromatographic behavior is usually represented by the capacity (or retention) factor, k'

$$k' = \frac{V_r - V_o}{V_o} \tag{1}$$

where V_o is the void volume in the column.
In the case of salt gradient elution, the chromatographic behavior can be represented by the dimensionless retention time, DRT (Lienqueo *et al.*, 2002).

$$\mathrm{DRT} = \frac{\mathrm{TR} - t_o}{t_f - t_o} \tag{2}$$

where t_o is the time corresponding to the start of the elution gradient, and t_f is the time corresponding to the end of the salt gradient. If the hydrophobic resin does not retain a protein, the DRT for that protein is zero. In contrast, if a protein elutes only after the salt gradient has been completed, the DRT for that protein is one.

The ability to predict retention time could be useful for designing purification processes and thus reduce experimental task (Mahn and Asenjo, 2005). In this paper, we review three methodologies for predicting protein retention time in HIC, and we extended the second one for all kinds of proteins, which have hydrophobic amino acid patches evenly or unevenly distributed on their surface.

RESULTS

First approach: average surface hydrophobicity, $\Phi_{surface}$

The first approach correlates the protein retention time in HIC with protein average surface hydrophobicity. This methodology is based on the protein three-dimensional structure and considers the hydrophobic contribution of the exposed aminoacid as a weighted average. This model can be applicable to stable proteins with a relatively homogeneous surface hydrophobicity distribution.
This methodology had been proposed by Lienqueo *et al.*, 2002, and it has three steps:

1. First, it is necessary to know the three-dimensional structure of the protein, by means of a Protein Data Bank (PDB) file (Berman *et al.*, 2000), meeting the requirements of (i) to give the coordinates for the majority of the heavy atoms, (ii) to have at least a resolution of 2.0 Å, (iii) and determined experimentally (by crystallization). The PDB files were appropriately modified using molecular modeling tools, for example, Homology and/or Builder modules of Insight II 2000 (Accelrys Software, Inc.).
2. Then, it is necessary to calculate the 'average surface hydrophobicity' of the protein, $\Phi_{surface}$, considering that each amino acid has a relative contribution to surface properties, proportional to its solvent accessibility, as calculated in Equation (3) (Berggren *et al.*, 2002).

$$\Phi_{surface} = \frac{\Sigma(s_{aai} \cdot \phi_{aai})}{s_p} \tag{3}$$

where s_{aai} is the solvent accessible area of each residue, s_p is the total solvent accessible area of the protein, and ϕ_{aai} is the amino acid hydrophobicity given by the normalized scale reported by Miyazawa and Jernigan (1985) and Lienqueo *et al.*, (2002). The program Graphical Representation and Analysis of Structural Properties (GRASP) (Nicholls *et al.*, 1991) was used in order to render protein surfaces and calculate the solvent accessible area of single residues in a protein.
3. The last step is achieved by using a simple quadratic model, Equation 4, whose parameters have been determined empirically, in order to predict the chromatographic behavior of proteins in different HIC media, based on the average surface hydrophobicity. A diagram with the different steps is summarized in Figure 1.

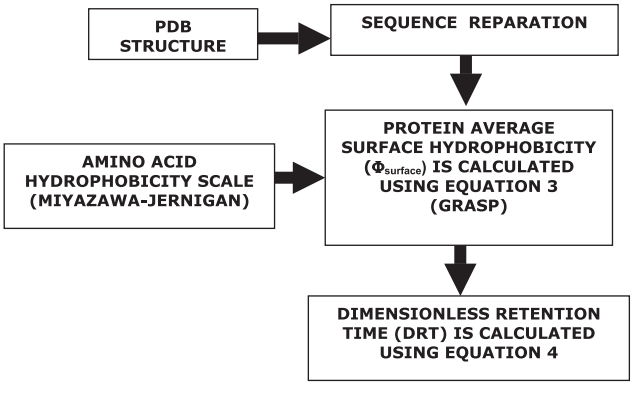

Figure 1. Block diagram of the first approach used in the calculation of protein average surface hydrophobicity and prediction of protein dimensionless retention time in HIC.

This approach was tested with nine monomeric proteins (conalbumin, cytochrome-c, ribonuclease A, α-chymotrypsinogen A, chicken lysozyme, α-lactoalbumin, myoglobin, α-chymotrypsin, thaumatin) under different operating conditions. The results of this approach, illustrated in Figure 2, exhibit a correlation between average surface hydrophobicity ($\Phi_{surface}$) and the experimental dimensionless retention time ($r^2 > 0.96$) (Lienqueo *et al.*, 2002). This relationship is shown in equation (4).

$$\begin{aligned} \text{DRT} &= 0 \qquad \text{If } \Phi_{surface} \leq 0.185 \\ \text{DRT} &= A^*\Phi^2_{surface} + B^*\Phi_{surface} + C \\ &\qquad\qquad \text{If } 0.185 < \Phi_{surface} < 0.345 \\ \text{DRT} &= 1 \qquad \text{If } \Phi_{surface} \geq 0.345 \end{aligned} \tag{4}$$

where $\Phi_{surface}$ is the average surface hydrophobicity value calculated using Equation (3). A, B, and C are the constants for each set of operating conditions (HIC media, salt type, and concentration). The values on constant A, B, and C of the models at different operating conditions are summarized in Table 1.

In addition, these correlations have been validated for several monomeric and multimeric proteins (ovalbumin and β-lactoglobulin) (Lienqueo *et al.*, 2002), and 'real' cell extracts (yeast producing human superoxide dismutase (Lienqueo *et al.*, 2003), and *E. coli* producing β-glucanase); results (see Figure 2) have always been satisfactory ($r^2 > 0.96$). However, when applying this methodology to predict retention time of proteins with a relatively heterogeneous surface hydrophobicity distribution, the results have not been as good as expected (Mahn *et al.*, 2004). Besides, a requirement for any model based on structural data is that the conformation of proteins remains the same as determined experimentally (PDB file) during the chromatographic process, that is, the protein should not suffer conformational changes during the chromatographic process.

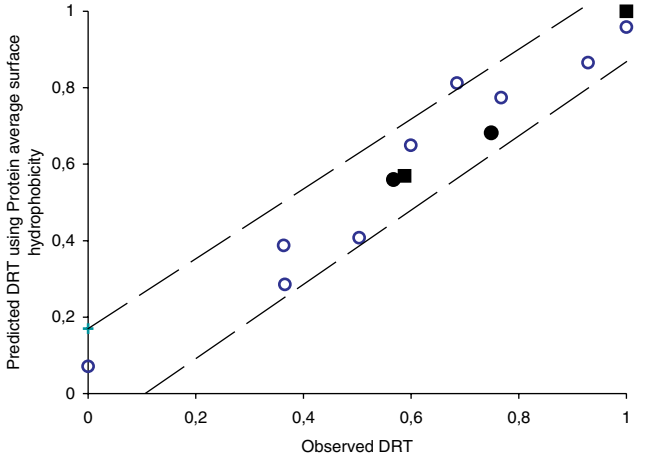

Figure 2. Comparison between predicted and observed dimensionless retention times on phenyl sepharose 6FF 2 M ammonium sulfate using the first approach based on protein average surface hydrophobicity ($\Phi_{surface}$) (○) standard proteins, (●) monomeric and multimeric proteins used for model validation, (■) 'real' cell extracts, (—) confidence intervals (95%).

In view of these results, this empirical model can be applicable to stable proteins with a relatively homogeneous surface hydrophobicity distribution. However, the main disadvantage of this methodology is that it does not consider the effect of the distribution of the surface hydrophobic patches on protein retention, and the prediction for proteins with a heterogeneous surface hydrophobicity distribution has been inadequate. Hence, a second approach considering the effect of surface hydrophobicity distribution on protein retention could be developed.

Second approach: local hydrophobicity (LH)

The idea of the second approach is to correlate the surface hydrophobicity distribution of a protein and the average surface hydrophobicity of the interaction zone of a protein with protein retention time in HIC. This is based on the results that showed that some proteins having very similar average surface hydrophobicity present different retention time in HIC (Mahn *et al.*, 2004). This variation in the chromatographic behavior has been attributed to differences in surface hydrophobicity distribution, for example, hydrophobic patches in globular proteins. Then, it was necessary to find a way to evaluate the surface hydrophobicity distribution.

Effect of surface hydrophobicity distribution on the retention factor and dimensionless retention time. Mahn *et al.* (2004) stated that surface hydrophobicity distribution could be represented by the parameter called 'Hydrophobic contact area' (HCA), which represents the contact area between the stationary phase and the protein when attached to the HIC resin.

Earlier reports by Melander *et al.* (1984) proposed that HCA could be estimated by a simplified thermodynamic model, which describes protein retention due to the combined effects of electrostatic and hydrophobic interactions. Then, the retention factor k' can be represented by the following equation:

$$\ln k' = \log\left(\frac{N_{AV} b^2 \delta_p}{1000e}\right) + \frac{Z_p}{Z_s} \cdot \log\left[\frac{1000e}{(N_{av} b^2 \delta_s m_s)(1 - Z_s \xi)}\right] + \frac{\Delta G^o_{aq}}{2.3RT} + \frac{\text{HCA} \cdot \sigma_s m_s}{2.3RT} + \log\alpha \tag{5}$$

where m_s is the molal salt concentration, N_{AV} the Avogadro's number, 'e' the base of the natural logarithm, 'b' the average spacing of fixed charges on the surface, δ_p the thickness of the condensation layer over the surface of the stationary phase where each fixed charge occupies an area of b^2, δ_s the layer thickness of salt counter ion, Z_p the characteristic charge of the protein, Z_s the valence of the salt counter ion, ξ a dimensionless structural parameter that characterizes the charged surface, and ΔG_{aq} is the reduction in free energy due to other effects different form hydrophobic interactions. R is the universal constant of gases, T the absolute temperature, α is the phase ratio (stationary phase/mobile phase) and σ_s is the surface tension increment of a solution due to the addition of a neutral salt.

Table 1. Prediction of protein retention times in HIC: parameters of quadratic model (Mahn and Asenjo, 2005; Lienqueo and Mahn, 2005)

HIC media	Salt type and concentration	Quadratic model coefficients		
		A	B	C
Phenyl sepharose	Ammonium sulfate (1 M)	11.79	−0.29	−0.35
	Ammonium sulfate (2 M)	−12.14	12.7	−1.74
	Sodium chloride (2 M)	−77.10	42.33	−5.13
	Sodium chloride (4 M)	−65.01	37.55	−4.71
Butyl sepharose	Ammonium sulfate (1 M)	36.76	−16.07	1.73
	Ammonium sulfate (2 M)	10.02	0.54	−0.38
	Sodium chloride (4 M)	−1.74	5.55	−1.01

The quadratic model is DRT $= A^*\phi^2 + B^*\phi + C$.

This equation could be simplified and expressed on the basis of salt molality as:

$$\ln k' = A - B \cdot \log m_s + C m_s \tag{6}$$

where A is a constant determined by all the system's characteristics, B the electrostatic interaction parameter, and $C\left(=\left[\frac{\mathrm{HCA}\cdot\sigma_s}{2.3RT}\right]\right)$ the hydrophobic interaction parameter.

At high salt concentrations, electrostatic interactions (parameter B) are negligible; then, parameters A and C can be obtained from isocratic retention data using different salt concentrations in the elution buffer. HCA can be calculated from the slope of the limiting plot of log k' versus salt molality. Unfortunately, the experimental methodology for determining the value of HCA needs a large number of tests.

Mahn *et al.* (2004) verified that the parameter HCA correlated very well with the 'dimensionless retention time', DRT, ($r^2 = 0.99$), in gradient elution with butyl sepharose 2 M ammonium sulfate, for three ribonucleases (RNAse T1, a variant of RNAse T1, and RNAse A) with similar average surface hydrophobicity, but different surface hydrophobicity distribution.

Characterization of surface hydrophobicity distribution using Local Hydrophobicity (LH). On the other hand, Mahn *et al.* (2005) defined a new parameter called 'local hydrophobicity' (LH). LH represents the average surface hydrophobicity of the interaction zone of the protein with the hydrophobic ligand. LH is calculated as follows:

$$\mathrm{LH} = \frac{\sum s_{i\mathrm{ZHI}}{}^*\phi_i}{s_{\mathrm{IZ}}} \tag{7}$$

Where $s_{i\mathrm{ZHI}}$ is the solvent accessible area of each residue in the interaction zone, s_{IZ} is the solvent accessible area of the interaction zone and ϕ_i is the amino acid hydrophobicity given by the normalized scale reported by Miyazawa and Jernigan (1985).

Then, to calculate LH it is necessary to locate the interaction zone of the protein (ZHI) with the hydrophobic ligand (see Figure 3). This localization was carried out with a conformational sampling procedure called 'molecular docking.' This procedure examined different protein–ligand conformations to find the correct one, based on energy minimization.

In addition, Mahn *et al.* (2005) demonstrated that the new LH parameter correlated extremely well with the 'dimensionless retention time' ($r^2 = 0.99$) for the same three ribonucleases in gradient elution with phenyl sepharose 2 M ammonium sulfate. The resulting relationship is shown in Equation (8).

$$\mathrm{DRT} = 0.77\mathrm{LH} + 0.21 \tag{8}$$

However, this study has a disadvantage in the sense that it was carried out only for a single HIC medium using a small set of homologous proteins. Then, it is necessary to extend this methodology to other proteins and different experimental conditions.

Prediction of protein retention time using local hydrophobicity calculated by molecular docking. In the present work, we have extended and automated the methodology proposed by Mahn *et al.* (2005) for predicting protein behavior in HIC, considering seven proteins (α-Lactalbumin, α-Chymotrypsinogen A, Lysozyme, RNAse T1, a variant of RNAse T1, RNAse A, and Concanavalin A) with heterogeneous and homogeneous surface hydrophobicity distribution and maintaining the same experimental conditions (phenyl sepharose 2 M ammonium sulfate).
The methodology has six steps:

(1) First, it is necessary to know the complete crystal structure of the proteins and that of the hydrophobic ligand (Phenyl).
   (i) In the case of the proteins, it could be necessary to review, repair, and complete the sequence atomic coordinates file using molecular modeling tools, for

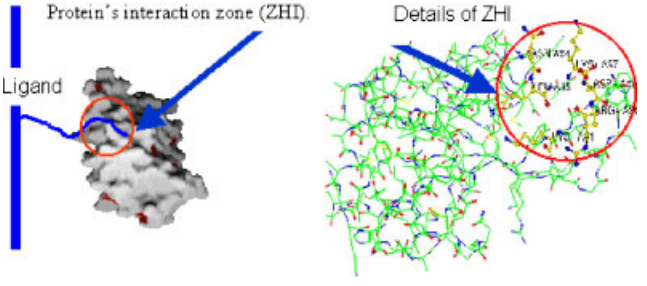

Figure 3. Amino acid in the protein's interaction zone (ZHI).

example, Homology and/or Builder modules of Insight II 2000 (Accelrys Software, Inc.).

(ii) In the case of hydrophobic ligand structure, the spatial coordinates can be estimated by molecular modeling tools, such as Builder and/or Biopolymer modules of Insight II 2000, using 1000 minimization steps.

(2) Then, using molecular docking tools (for instance, the Affinity module of Insight II), which provides automated docking by Grid method using consistent valence force field, 50 different conformations of the protein–ligand are found. Each conformation was inspected by the DeCipher module of Insight II.

(i) If the ligand is located in a protein's pocket or concave zone, this conformation is discarded.

(ii) Or else, the protein's interaction zone (ZHI) is identified and the conformation is clustered in families, considering the spatial localization of the ligand over the protein.

(3) After that, the different probable docked protein–ligand conformations were automatically scored using the module LUDI of Insight II, through the following scoring function (Böhm, 1992):

$$\text{Score} = -73.3\left[\frac{\text{mol}}{\text{kcal}}\right]\Delta G_{\text{binding}} \tag{9}$$

where $\Delta G_{\text{binding}}$, the free energy of binding at equilibrium, is expressed by the following empirical function (Böhm, 1994, 1998)

$$\begin{aligned}\Delta G_{\text{binding}} =& \Delta G_0 + \Delta G_{\text{hb}} \sum_{\text{h-bonds}} f(\Delta R) f(\Delta\alpha) \\ &+ \Delta G_{\text{ion}} \sum_{\text{ionic}} f(\Delta R) f(\Delta\alpha) \\ &+ \Delta G_{\text{lipo}} A_{\text{lipo}} + \Delta G_{\text{aro/aro}} \Delta N_{\text{aro/aro}} \\ &+ \Delta G_{\text{rot}} NR\end{aligned} \tag{10}$$

where the ΔG_0 term represents the contribution to the binding energy that does not directly depend on any specific interactions with the protein; the ΔG_{hb} and ΔG_{ion} terms represent the contributions from an ideal hydrogen bond and an undisturbed ionic interaction, respectively. The ΔG_{lipo} term represents the contribution from lipophilic interactions. The ΔG_{rot} term represents the contribution due to the freezing of internal degrees of freedom in the fragment. A_{lipo} is the area of hydrophobic contact between the ligand and protein; NR is the number of acyclic sp^3—sp^3 and sp^3—sp^2 bonds. The $\Delta G_{\text{aro/aro}}$ term represents the contribution due to aromatic–aromatic interactions, and $N_{\text{aro/aro}}$ is a count of aromatic–aromatic interactions. Rotations of terminal CH_3 or NH_3 groups and flexibility of cyclic portions of the ligand are not taken into account. Finally, $f(\Delta R)$ and $f(\Delta\alpha)$ represent the deviation from ideal values, 1.9 Å and 180°, respectively (Böhm, 1998).

(4) Next, the most probable conformation was selected considering three variables of each family:

(i) Proximity to the concave zone.

(ii) The number of cluster components.

(iii) The average score value of each cluster, <Score>.

(5) After that, the amino acid residues in the protein's interaction zone (ZHI) are identified by Subset for Insight II. The solvent accessible area of each residue ($s_{i\text{ZHI}}$) and the solvent accessible area of the interaction zone (s_{IZ}) were computed by the program Graphical Representation and Analysis of Structural Properties (GRASP). Then, the local hydrophobicity (LH) was calculated using Equation (7) below:

$$\text{LH} = \frac{\sum s_{i\text{ZHI}}{}^{*}\phi_i}{s_{\text{IZ}}} \tag{7}$$

(6) Finally, by using a linear equation, Equation (11), the parameters of which have been determined empirically, it is possible to predict the chromatographic behavior of proteins in HIC based on 'local hydrophobicity.' A diagram with the different steps is summarized in Figure 4.

The results of this approach (see Figure 5) show a suitable correlation level between 'local hydrophobicity' (LH) and the experimental dimensionless retention time ($r^2 > 0.87$). This relationship is shown in Equation (11).

$$\text{DRT} = 2.8^{*}\ \text{LH} - 0.53 \tag{11}$$

The relationship shown by Equation (11) has a similar tendency to that of the results shown by Mahn *et al.* (2005) in Equation (8).

Subsequently, we think that this methodology could be used to adequately represent the chromatographic behavior in HIC for proteins with a homogeneous and heterogeneous surface hydrophobicity distribution and without a large number of tedious experiments.

In view of these results, we think that the methodology proposed by Mahn has been validated for the experimental conditions used, and it could be used to represent the retention time in HIC for all kinds of proteins with well-known complete crystal structure, with a homogeneous and heterogeneous surface hydrophobicity distribution, and only using computational simulations and adequate score criteria.

Additionally, we propose to continue to test this method using different operating conditions (e.g., other HIC media and salt type and concentration), since it has been demonstrated that different HIC media (e.g., butyl, propyl, hexyl, octyl) can show differences in protein behavior in HIC (Machold *et al.*, 2002; Ladiwala *et al.*, 2005). Furthermore, it is important to consider that the phenyl media used in the present study could be enabled to interact for van der Waals interaction and hydrophobic interaction with the proteins; then, if different HIC is used, these interactions could be investigated separately.

Third approach: predicted average surface hydrophobicity (ASH)

Another different approach in our investigation is trying to predict the protein retention time in HIC only using its amino-acidic composition and computational simulations.

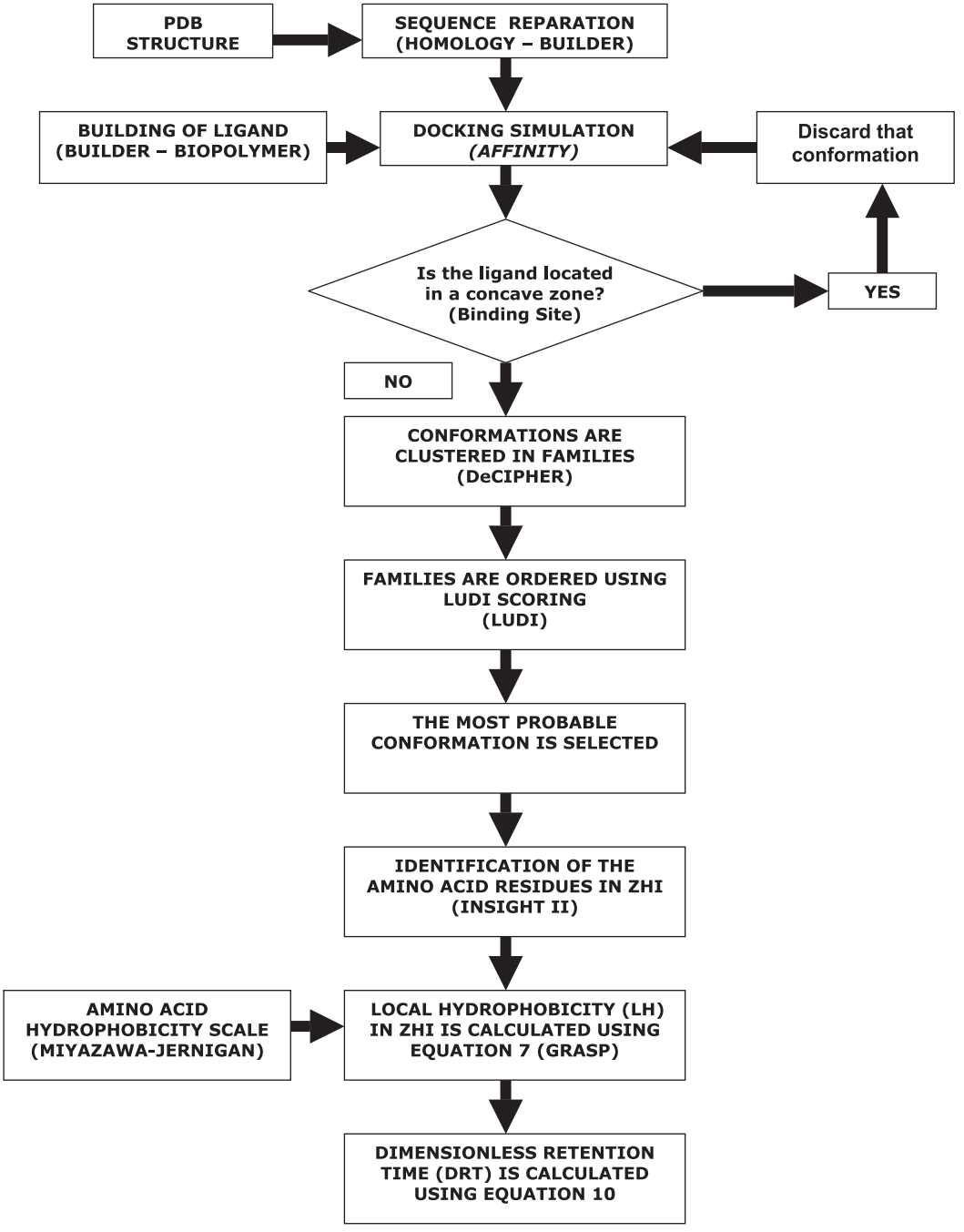

Figure 4. Block diagram of the second approach used in the selection of the most probable protein–ligand complex conformation, calculation of local hydrophobicity and prediction of protein dimensionless retention time in HIC.

Prediction of average surface hydrophobicity only based on amino acidic composition. In order to calculate the average surface hydrophobicity ($\Phi_{surface}$) of a protein, 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 through the methodology of comparative modeling, 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 the DRT could be determined by using low level information, as for instance, the amino-acidic composition.

Some features of proteins can be predicted based on their amino-acidic composition. For example, it has been reported that the prediction of the protein's secondary structural content (Pilizota *et al.*, 2004), and the protein's structural class (Luo *et al.*, 2002) can be carried out successfully from its amino-acidic composition only.

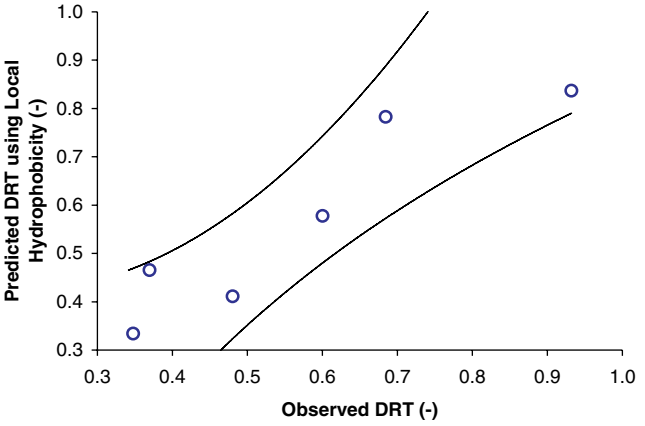

Figure 5. Comparison between predicted and observed dimensionless retention times on phenyl sepharose 6FF 2 M ammonium sulfate using the second approach based on protein local hydrophobicity (LH) (○) standard proteins (- - -) confidence intervals (95%).

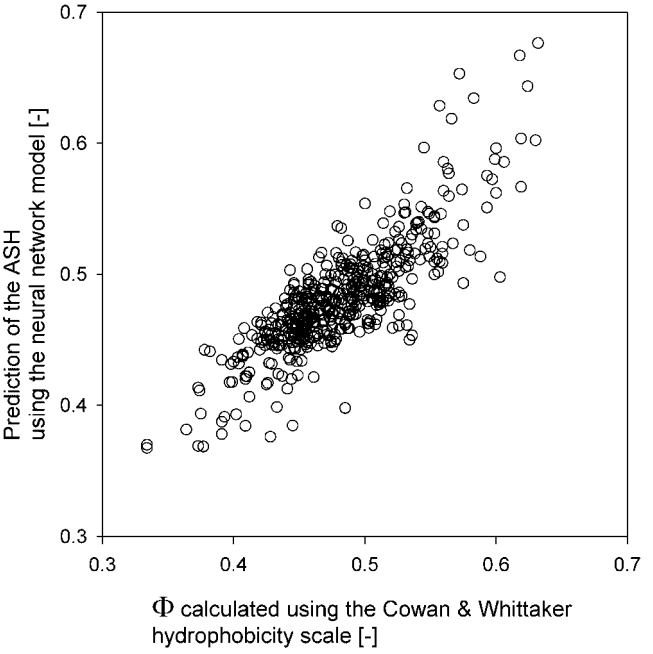

Figure 6. Scatter plot between the ASH calculated using the scale of Cowan–Whittaker and the prediction of the neural network in a test subset chosen randomly (Modified from Salgado JC, Rapaport I, Asenjo JA. 2005. Is it possible to predict the average surface hydrophobicity of a protein using only its amino acid composition? *J. Chromatogr. A* 1075 (1–2): 142, with permission from Elsevier).

Recently, we investigated the prediction of average surface hydrophobicity ($\Phi_{surface}$) calculated on the basis of the hydrophobicity scale of Cowan-Whittaker (Cowan and Whittaker, 1990) by using mathematical models based on the amino-acidic composition and measurements of the amino acids tendency to exposure. We proved that it is possible to predict the average surface hydrophobicity, called ASH, of a large set of proteins to an acceptable level for many practical applications (correlation coefficient > 0.8), using linear models and neural networks based on the protein's amino-acidic composition (Salgado *et al.*, 2005a).

The simpler ASH predictive model was a linear model based only on the amino-acidic protein composition. This model had 21 parameters and it was able to predict the ASH for a standalone test subset with a correlation coefficient of 0.769 for the case of the average surface hydrophobicity ($\Phi_{surface}$) calculated using the hydrophobicity scale of Cowan-Whittaker. In all cases, where evaluated, it showed low variability in its performance.

A model based on a neural network was also evaluated. It used the same inputs as the linear model. It was determined that the optimum configuration for the neural model was a network with a single hidden layer which had three neurons in it, and no pre-processing of the inputs. The neural network model had 67 parameters and improved the results shown by the linear model by a little more than 24%. The predictive performance of the neural network model was found satisfactory as shown by the scatter plot in Figure 6. The correlation coefficient obtained by this model was 0.831. The neural model was shown to be slightly more robust than the linear one, hence the diverse variability observed in the performance indices was not greater than 6.2% of the mean square error.

Prediction of chromatographic behavior only based on amino-acidic composition. Prediction of the DRTs of proteins by means of mathematical models based, essentially, only on the amino-acidic protein composition was the following stage. Our results show that such prediction is possible with a performance similar to that observed in models using much more sophisticated information as the three-dimensional structure of proteins (Salgado *et al.*, 2005b).

A DRT prediction model based on information concerning the three-dimensional structure of proteins was proposed by Lienqueo *et al.* (2002). They selected the Miyazawa and Jernigan hydrophobicity vector (Miyazawa and Jernigan, 1985) in the process of adjusting the parameters of their model. In that context, we showed that a model (called DRT 0), constructed using the Wertz and Scheraga vector (Wertz and Scheraga, 1978), is better, since the Jack Knife estimation of the prediction error was 38.2% smaller than that based on the Miyazawa and Jernigan vector (Miyazawa and Jernigan, 1985). We used the Jack Knife methodology because 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 that the arbitrary division of the data set in a training and test set (Chou and Zhang, 1995; Zhou, 1998; Zhou and Assa-Munt, 2001). The mathematical principle and a comprehensive discussion about this can be found in Mardia *et al.* (1979).

Our main contribution was the design of models that predict the DRT using the minimal information concerning a protein, its amino-acidic composition. We did not take into account the protein's amino acid sequence, nor its secondary structure or 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 giving best results was the

one based on a linear estimation of the amino-acidic surface composition (DRT III). The prediction error (MSE_{JK}) obtained by this model was almost 35% smaller than that obtained by the model assuming that all the amino acids are completely exposed (DRT I) and 40% smaller than that obtained by the model using a simple correction factor, and 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 amino-acidic properties, plus a set of 6912 vectors derived from these. The derived vectors were obtained using two mathematical tools: k-means (Seber, 1984) and self-organizing maps (SOM) (Kohonen, 1989) 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. In fact, the performance of DRT III with vector '$\boldsymbol{v}$' was 5% better than that observed in DRT 0 using the three-dimensional structure of proteins. This can be seen in Figure 7, which shows the scatter plots between experimental DRT and those estimated by means of the DRT 0 and DRT III models.

Finally, the plot in Figure 8 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 smallest 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 as in conalbumin (682 aa) were observed. This allows us to state that the DRT III model (as model DRT 0) is able to predict the retention times for a wide set of monomeric and multimeric proteins, and that the prediction error is not related to the length of these, nor to their average surface hydrophobicity.

CONCLUSIONS

In this paper, three different approaches for predicting protein retention time in HIC, based on a protein's structure or on its amino-acidic composition were reviewed; we have extended one of these approaches. The first approach is based on the protein's three-dimensional structure data and considers the hydrophobic contribution of the exposed amino acid residues as a weighted average, a parameter called 'protein average surface hydrophobicity.' The second approach, which we have extended, is based on the high correlation level between the average surface hydrophobicity of a protein's hydrophobic interacting zone (a parameter called 'local hydrophobicity'), and its retention time in HIC. Finally, a third approach carries out a prediction of the average surface hydrophobicity of a protein using only its amino-acidic composition, without knowing its three-dimensional structure. The results have shown that the use of these approaches could make it possible to predict protein chromatographic behavior in HIC for all proteins, with and without knowing their tertiary structure. Then, if a new relationship is developed for different operating conditions, it could be possible to test different operating conditions for the purification of a target protein, and finally select the best conditions *in silico* to contribute to the rational design and optimization of the process involving an HIC step. Currently, studies with different experimental conditions are being carried out in our laboratory to extend the second approach, based on 'local hydrophobicity,' to other HIC media (e.g., butyl and octyl).

EXPERIMENTAL MATERIALS

Materials

Seven proteins of known three-dimensional structure were used: α-Lactalbumin(1A4V), α-Chymotrypsinogen A(2CHA), Lysozyme(2LYM), RNAse T1(1RGC), a variant of RNAse T1(1TRP), RNAse A (1AFU), and Concanavalin

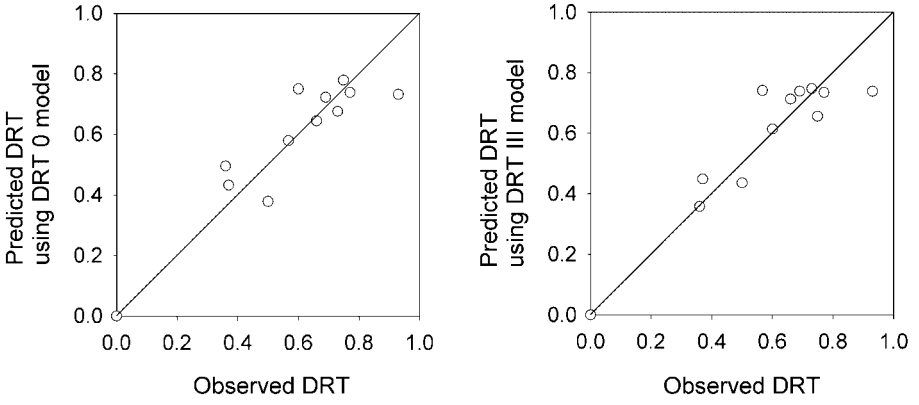

Figure 7. Scatter plots between the experimental dimensionless retention time (DRT) and DRT predicted by the DRT 0 model based on the three-dimensional structure of the proteins and DRT III model based on the amino-acidic composition of the proteins. (Modified from Salgado JC, Rapaport I, Asenjo JA. 2005. Prediction of retention times of proteins in hydrophobic interaction chromatography using only their amino acid composition. *J. Chromatogr. A* 1098 (1–2): 44–54, with permission from Elsevier).

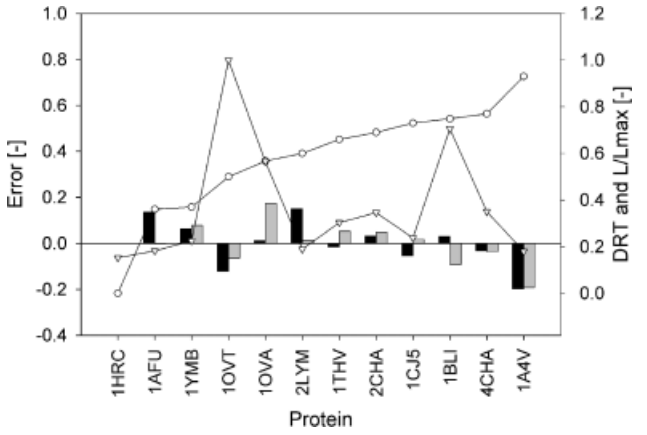

Figure 8. Plot of the residual error between the experimental dimensionless retention time (DRT) and DRT estimated by the DRT 0 model (:) and the DRT III model (:). The experimental DRT (●), and the dimensionless length (▲) are also shown (Modified from Salgado JC, Rapaport I, Asenjo JA. 2005. Prediction of retention times of proteins in hydrophobic interaction chromatography using only their amino acid composition. *J. Chromatogr. A* 1098 (1–2): 44–54, with permission from Elsevier).

A(1QDO). Water prepared with Milli-Q water cleaning system (Millipore, Bedford, MA, USA) and analytical-regent grade ammonium sulfate were used in the preparation of the eluent. Protein solutions were prepared to contain up to 2.0 mg/ml dissolved in the initial eluent. All protein solutions were filtered through 0.22-μm Millipore filters.

Equipment

The high-performance liquid chromatography system employed consisted of a fast protein liquid chromatography (FPLC) system (Pharmacia, Uppsala, Sweden) equipped with a 500-μl injection loop. The chromatographic columns were 1 ml Phenyl Sepharose (GE Health care; Uppsala, Sweden). The experiments were performed at room temperature, using a flow-rate equal to 0.75 ml/min and 10-column volumes (CVs). After that, retention times (TR) were recorded. Finally, the chromatographic behavior of proteins was characterized by the 'dimensionless retention time' (DRT) parameter, using Equation (2).

Operating conditions

Elution was obtained by a decreasing gradient of ammonium sulfate. The initial eluent was 20 mM Bis–Tris, pH 7.0 plus 2 M ammonium sulfate (solvent B). The final eluent was 20 mM Bis–Tris, pH 7.0 (solvent A). The gradient steepness used was 7.5% B/min. All buffers were filtered through 0.22-μm Millipore filters after preparation, and degassed with helium for 10 min.

Acknowledgement

This work was supported by the National Foundation of Science and Technology (Fondecyt) Project 1030668.

REFERENCES

Berggren K, Wolf A, Asenjo JA, Andrews BA, Tjerneld F. 2002. The surface exposed amino acid residues of monomeric proteins determine the partitioning in aqueous two-phase systems. *Biochim Biophys Acta.* **1596**: 253–268.

Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE. 2000. The protein data bank. *Nucleic Acids Res.* **28**: 235–242.

Böhm HJ. 1992. The computer program LUDI: a new method for the de novo design of enzyme inhibitors. *J. Comp. Aided Molec. Design.* **6**: 69–78.

Böhm HJ. 1994. The development of a simple empirical scoring function to estimate the binding constant for a protein-ligand complex of known three-dimensional structure. *J. Comp. Aided Molec. Design.* **8**: 243–256.

Böhm HJ. 1998. Prediction of binding constants of protein ligands: a fast method for the prioritization of hits obtained from the de novo design or 3D database search programs. *J. Comp. Aided Molec. Design.* **12**: 309–323.

Chen J, Sun Y. 2003. Modelling of the salt effects on hydrophobic adsorption equilibrium of proteins. *J. Chromatogr. A* **992**: 29–40.

Chou KC, Zhang CT. 1995. Prediction of protein structural classes. *Crit. Rev. Biochem. Mol. Biol.* **30**: 275–349.

Cowan R, Whittaker RG. 1990. Hydrophobicity indices for amino acid residues as determined by high-performance liquid chromatography. *Pept. Res.* **3**: 75–80.

Eriksson K. 1998. Hydrophobic interaction chromatography. In *Protein Purification: Principles, High-Resolution Methods, and Applications*, Janson J-C, Ryden L (eds). Wiley-Liss: New York; 283– 309.

Fausnaugh JL, Kennedy LA, Regnier FE. 1984. Comparison of hydrophobic interaction and reversed phase chromatography of proteins. *J. Chromatogr.* **317**: 141–155.

Fexby S, Bülow L. 2004. Hydrophobic peptide tags as tools in bioseparation. *Trends Biotechnol.* **22**: 511–516.

Hahn R, Deinhofer K, Machold C, Jungbauer A. 2003. Hydrophobic interaction chromatography of proteins II. Binding capacity, recovery and mass transfer properties. *J. Chromatogr. B.* **790**: 99–114.

Jennissen HP. 2000. Hydrophobic interaction chromatography. *Int J. Bio-Chromatogr.* **5**: 131–138.

Kohonen T. 1989. *Self-Organization and Associative Memory*, 3rd edn. Springer-Verlag: Berlin-Heidelberg-New York-Tokyo.

Ladiwala A, Xia F, Luo Q, Breneman CM, Cramer SM. 2005. Investigation of protein retention and selectivity in HIC systems using quantitative structure retention relationship models. *Biotechnol Bioeng.* DOI: 10.1002/bit.20771.

Lienqueo ME, Mahn AV, Asenjo JA. 2002. Mathematical correlations for predicting protein retention time in hydrophobic interaction chromatography. *J. Chromatogr. A.* **978**: 71–79.

Lienqueo ME, Mahn A, Vásquez L, Asenjo JA. 2003. Methodology for predicting the separation of proteins by hydrophobic interaction chromatography and its application to a cell extract. *J. Chromatogr. A.* **1009**: 189–196.

Lienqueo ME, Mahn A. 2005. Predicting protein retention time in hydrophobic interaction chromatography. *Chem. Eng. Technol.* **28**: 1326–1334.

Luo RY, Feng ZP, Liu JK. 2002. Prediction of protein structural class by amino acid and polypeptide composition. *Eur. J. Biochem.* **269**: 4219–4225.

Machold C, Deinhofer K, Hahn R, Jungbauer A. 2002. Hydrophobic interaction chromatography of proteins I. Comparison of selectivity. *J. Chromatogr. A.* **972**: 3–19.

Mahn A, Lienqueo ME, Asenjo JA. 2004. Effect of surface hydrophobicity distribution on protein retention in hydrophobic interaction chromatography. *J. Chromatogr. A.* **1043**: 47–55.

Mahn A, Zapata-Torres G, Asenjo JA. 2005. A theory of protein–resin interaction in hydrophobic interaction chromatography. *J. Chromatogr.* **1066**: 81–88.

Mahn A, Asenjo JA. 2005. Prediction of protein retention in hydrophobic interaction chromatography. *Biotechnol. Adv.* **23**: 359–368.

Mardia KV, Kent JT, Bibby JM. 1979. *Multivariate Analysis*. Academic Press: London.

Melander W, Horváth Cs. 1977. Salt effect on hydrophobic interactions in precipitation and chromatography of proteins: an interpretation of the lyotropic series. *Arch. Biochem. Biophys.* **183**: 200–215.

Melader W, Corradini D, Horváth Cs. 1984. Salt-mediated retention of proteins in hydrophobic-interaction chromatography. Application of solvophobic theory. *Chromatogr.* **317**: 67–85.

Miyazawa S, Jernigan R. 1985. Estimation of effective inter residue contact energies from protein crystal structures Quasi-chemical approximation. *Macromolecules.* **18**: 534–552.

Nicholls A, Sharp K, Honing B. 1991. Protein folding and association: insights from the interfacial and thermodynamic properties of hydrocarbons. *Proteins: Struct. Funct. Gen.* **11**: 281–296.

Pilizota T, Lucic B, Trinajstic N. 2004. Use of variable selection in modelling the secondary structural content of proteins from their composition of amino acid residues. *J. Chem. Inf. Comput. Sci.* **44**: 113–121.

Queiroz JA, Tomaz CT, Cabral JMS. 2001. Hydrophobic interaction chromatography of proteins. *J. Biotechnol.* **87**: 143–159.

Roettger BF, Ladisch MR. 1989. Hydrophobic interaction chromatography. *Biotechnol Adv.* **7**: 15–29.

Salgado JC, Rapaport I, Asenjo JA. 2005a. Is it possible to predict the average surface hydrophobicity of a protein using only its amino acid composition? *J. Chromatogr. A* **1075**: 133–143.

Salgado JC, Rapaport I, Asenjo JA. 2005b. Prediction of retention times of proteins in hydrophobic interaction chromatography using only their amino acid composition. *J. Chromatogr. A.* **1098**: 44–54.

Seber GAF. 1984. *Multivariate Observations*. Wiley: New York.

Sofer G, Hagel L. 1998. Handbook of process chromatography: a guide to optimization, scale-up, and validation. Academic Press: San Diego; 387.

Wertz DH, Scheraga HA. 1978. Influence of water on protein structure. An analysis of the preferences of amino acid residues for the inside or outside and for specific conformations in a protein molecule. *Macromolecules* **11**: 9–15.

Xia F, Nagrath D, Cramer SM. 2005. Effect of pH changes on water release values in hydrophobic interaction chromatographic systems. *J. Chromatogr. A.* **1079**: 229–235.

Zhou GP. 1998. An intriguing controversy over protein structural class prediction. *J. Protein Chem.* **17**: 729–738.

Zhou GP, Assa-Munt N. 2001. Some insights into protein structural class prediction. *Proteins.* **44**: 57–59.