JOURNAL OF MOLECULAR RECOGNITION
J. Mol. Recognit. 2004; **17**: 236–247
DOI:10.1002/jmr.676

Is there a rational method to purify proteins? From expert systems to proteomics[†]

J. A. Asenjo* and B. A. Andrews

Centre for Biochemical Engineering and Biotechnology, Department of Chemical Engineering, Millenium Institute for Advanced Studies in Cell Biology and Biotechnology, University of Chile, Beauchef 861, Santiago, Chile

The purification of recombinant proteins for therapeutic or analytical applications requires the use of several chromatographic steps in order to achieve a high level of purity. A range of techniques is available such as anion and cation exchange chromatography, which can be carried out at different pHs, and hence used at different steps, hydrophobic interaction chromatography, gel filtration and affinity chromatography. Evidently when confronted with a complex mixture of partially unknown proteins or a clarified cell extract there are many different routes one can take in order to choose the minimum and most efficient number of purification steps to achieve a desired level of purity (e.g. 98, 99.5 or 99.9%). In this review we will show how an initial 'proteomic' characterization of the complex initial mixture of target protein and protein contaminants can be used to select the most efficient chromatographic separation steps in order to achieve a maximum level of purity with a minimum number of steps. The chosen methodology was implemented in a computer based expert system. The first algorithm developed was used to select the most efficient purification method to separate a protein from its contaminants based on the physicochemical properties of the protein product and the protein contaminants. The second algorithm developed was used to predict the number and concentration of contaminants after each separation as well as protein product purity. The successful application of the expert system approach, based on an initial proteomic characterization, to the practical cases of protein mixtures and clarified fermentation supernatant is presented and discussed. The purification strategy proposed was experimentally tested and validated with a mixture of four proteins and the experimental validation was also carried out with an 'unknown' supernatant of *Bacillus subtilis* producing a recombinant *β*-1,3-glucanase. The system was robust to errors <10% which is the range that can be found in the experimental determination of the properties in the database of product and contaminants. On the other hand, the system was sensitive both to larger variations (>20%) in the properties of the contaminant database and the protein product and to variations in one protein property (e.g. hydrophobicity). Copyright © 2004 John Wiley & Sons, Ltd.

Keywords: **rational method; protein purification; proteomics; expert systems**

Received 11 December 2003; revised 1 February 2004; accepted 4 February 2004

INTRODUCTION

Until now it has been virtually impossible to select separation and purification operations for proteins either for therapeutic or analytical application in a rational manner due to a lack of fundamental knowledge on the molecular properties of the materials to be separated and the lack of an efficient system to organize such information. A range of techniques is available, such as anion and cation exchange chromatography, which can be carried out at different pHs, and hence used at different steps, hydrophobic interaction chromatography, gel filtration and affinity chromatography in addition to HPLC and aqueous two-phase partitioning. Evidently when we are confronted with a complex mixture of partially unknown proteins or a clarified cell extract there are many different routes one can take in order to choose the minimum and most efficient number of purification steps to achieve a desired level of purity.

For selecting the sequence of operations for high resolution purification of proteins, there are a large number of options that can be chosen almost in any order, as shown in Fig. 1 (Leser and Asenjo, 1994). This figure does not show all six alternatives at all stages since, for instance, aqueous two-phase separation would only be used as a first step and HPLC would not be used as a first stage in a multistage process. This paper describes how to use the physico-chemical data of the product protein and other proteins present ('contaminants') to select an 'optimal' or suboptimal process sequence using the smallest number of operations.

SEPARATION COEFFICIENTS

The rationale for selection of high-resolution purification operations that has been developed characterizes the ability

*Correspondence to: J. A. Asenjo, Centre for Biochemical Engineering and Biotechnology, Department of Chemical Engineering, Millenium Institute for Advanced Studies in Cell Biology and Biotechnology, University of Chile, Beauchef 861, Santiago, Chile.
E-mail: juasenjo@ing.uchile.cl

Abbreviations used: DF, deviation factor; SBTI, soybean trypsin inhibitor; SC, separation coefficient; SSC, separation selection coefficient.

[†]This paper is published as part of a special issue entitled 'Affinity Interactions 2003, 15th International Conference held on 27th July–1st August 2003, St. John's College, Cambridge, UK'.

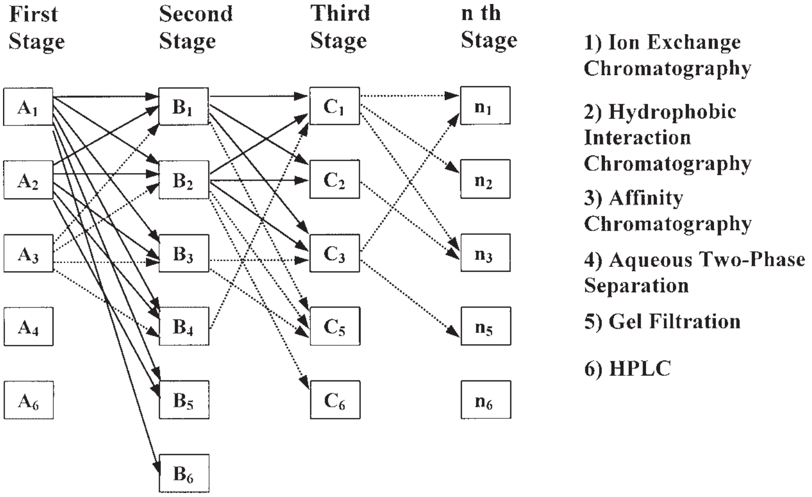
First Stage
Second Stage
Third Stage
n th Stage
A1
A2
A3
A4
A6
B1
B2
B3
B4
B5
B6
C1
C2
C3
C5
C6
n1
n2
n3
n5
n6
1) Ion Exchange Chromatography
2) Hydrophobic Interaction Chromatography
3) Affinity Chromatography
4) Aqueous Two-Phase Separation
5) Gel Filtration
6) HPLC

Figure 1. The combinatorial characteristic of choosing the sequence of operations for protein purification.

of the separation operation to separate one protein from another. The correlations used for this purpose are shown in Table 1. This rationale uses the theoretical concept of separation coefficients (Asenjo, 1990; Leser and Asenjo, 1992). It uses a relationship between the separation coefficient (SC) as shown in Table 1 and the variables that reflect the performance in a separation process: the deviation factor (DF) for differences among physicochemical properties and the efficiency (η) of the process, because some separations have high efficiency for exploiting differences in the deviation factor and some do not. The DF has been defined as the difference in a particular physicochemical property (such as molecular weight, charge or hydrophobicity) between two proteins, which correspond to the target protein and the particular contaminant protein being considered. The basic concept of DF has been defined in Table 1, as the relative difference in that particular physicochemical property between the product protein and each of the contaminant proteins. To include the rule-of-thumb that reflects the logic of first separating impurities present in higher concentrations, a relative contaminant protein concentration (θ), as defined by eq. (3) is included. A selection separation coefficient is defined, as shown in eq. (4), as the product of the separation coefficient and this relative concentration. In the expression shown here the exponent n can have values between 0 and 1 and the initial assumption has been taken as $n = 1$. Therefore, the system will base the choice of the best process on the comparison of the values of the selection separation coefficient, defined by eq. (4), calculated for the different alternative separations.

Table 1. List of expressions, variables and numerical values used for calculation of SC and SSC

$$\mathrm{SC} = \mathrm{DF} \cdot \eta \tag{1}$$

DF = deviation factor for molecular weight, charge and hydrophobicity

$$\mathrm{DF} = \frac{\text{protein value}^{\mathrm{a}} - \text{contaminant value}}{\max^{\mathrm{b}}\,[\text{protein value, contaminant value}]} \tag{2}$$

$$\eta = \text{efficiency} \begin{cases} 1.00 \text{ for ion exchange} \\ 0.86 \text{ for hydrophobic interaction chromatography (HIC)} \\ 0.66 \text{ for gel filtration(GF)} \end{cases}$$

$$\theta = \text{concentration factor}$$

$$\theta_i = \frac{\text{concentration of contaminant protein } i}{\text{total concentration of contaminant proteins}} \tag{3}$$

$$\mathrm{SSC} = \mathrm{SC} \cdot \theta_i^n;\ \mathrm{SSC} = \mathrm{DF} \cdot \eta \cdot \theta_i^n \tag{4}$$

$$n = 1$$

[a] Protein value or contaminant protein value corresponds to the value of the particular physicochemical property of the protein being considered (e.g. charge, hydrophobicity, molecular weight).

[b] Maximum value of the physicochemical property (this could be the value of the target protein or the contaminant protein).

RESOLUTION AND EFFICIENCY

In chromatography, resolution is a variable used to measure the column performance; and, although it does not necessarily predict elution profiles from fundamental properties (Leser *et al.*, 1996), it represents a means of interpreting column data and provides a basis for comparing results from different operating conditions. Considering two peaks in a chromatogram, chromatographic resolution is defined as the distance between the peak maxima divided by the mean peak width, as shown in Fig. 2, and the resolution is defined as expressed in eq. (5) in Table 2. As the separation coefficient is also a measure of the process performance, it can be assumed that it is proportional to the resolution [eq. (6), Table 2]. The relationship between SC, DF and η has been investigated (Watanabe *et al.*, 1994; Leser *et al.*, 1996). To avoid the possible influence of the concentration, the authors used the same protein concentration to determine relations between efficiency and resolution. When an equal concentration of all proteins in a mixture is present, it can be shown that the efficiency of the process can be defined by eq. (7).

This concept is not based on rate or equilibrium analysis but corresponds to a semi-empirical analysis of separation of two components. The first expert system developed in our group (Leser and Asenjo, 1992) used empirical values for the efficiency; however, experimental data have shown a relatively constant behavior of the efficiency for each particular separation process (Watanabe *et al.*, 1994; Leser *et al.*, 1996). In these publications the authors have used separation materials commonly used on the preparative scale. It should be noted that if, for combinations of different proteins, the efficiency (as defined here) shows a relatively constant value, then the relationship shown by eq. (7) is valid. Based on these results, the values of efficiency used in this paper have been calculated and are those shown in Table 1.

Table 2. List of expressions used to define efficiency (η) in terms of resolution

Resolution, R_S

$$R_S = \frac{V_2 - V_1}{1/2(W_1 + W_2)} \tag{5}$$

$$\mathrm{SC}\ \alpha\ R_S \tag{6}$$

$$\eta = \frac{\mathrm{SC}}{\mathrm{DF}}\ \alpha\ \frac{R_S}{\mathrm{DF}} \tag{7}$$

DEVIATION FACTORS AND SELECTION SEPARATION COEFFICIENT

A DF for each individual property such as charge, molecular weight and hydrophobicity of pairs of proteins has been defined. The efficiency (η) reflects the unequal ability of different separation processes (and/or different materials used) to exploit differences in the deviation factor to separate the proteins. This value is relatively constant for each type of separation and chromatographic material used and can be found experimentally using an expression for the resolution such as that used in chromatography making the resolution equivalent to SC as shown in the previous section (Leser *et al.*, 1996; Watanabe *et al.*, 1994). The property chosen for gel filtration was molecular weight (MW), for hydrophobicity it was either the concentration of $(NH_4)_2SO_4$ (M) at which the protein eluted or the chromatographic K_D value using a specific hydrophobic matrix and a decreasing gradient of $(NH_4)_2SO_4$ (Table 3). For ion exchange chromatography the property used to calculate the deviation factor was charge as a function of pH and charge density. Figure 3 shows the titration curves for four proteins as charge density (charge/volume or MW) as a function of pH. These values were compared with behaviour in ion-exchange chromatography both as charge (at a particular pH) and as charge density as a function of retention time. These results are shown in Fig. 4 and clearly charge density (charge/MW) gave a better correlation as a function of retention time in ion exchange chromatography.

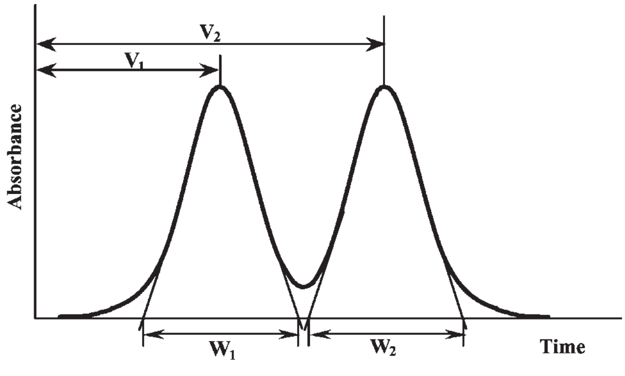

Figure 2. Determination of the resolution between two peaks.

A rule of thumb commonly used by experts is that the contaminants in higher concentrations should be separated first. For this reason the selection of operations was done using the separation selection coefficient (SSC), which includes the relative concentration θ as shown in Table 1.

To calculate the SSC the system will read a database containing the information on the properties of the main contaminant proteins present in the specific expression system used. Table 3 shows the data on the 13 predominant proteins present in a commercial strain of *E. coli* used for producing recombinant proteins that was kindly donated by an industrial source (Chiron) and is used in the expert system (Woolston, 1994). The proteins are identified by the value of the isoelectric point (pI). This database is then used to select the first high resolution purification step.

ELIMINATION OF PROTEIN CONTAMINANTS: A DYNAMIC DATABASE

After each high resolution step, the concentration of contaminant proteins decreases and the number of steps has to

Table 3. Concentration, molecular weight, hydrophobicity, and charge at different pHs for the main proteins ('contaminants' of the product) in *Escherichia coli*. Data from Woolston (1994)

					Charge (coulomb per molecule × IE25)									
					pH 4	pH 4.5	pH 5	pH 5.5	pH 6	pH 6.5	pH 7	pH 7.5	pH 8	pH 8.5
Contaminants	pI	(g/l)	MW (Da)	Hydrophobicity[a]	qA	qB	qC	qD	qE	qF	qG	qH	qI	qJ
Cont_1	4.67	11.29	18 370	0.71	1.94	0.25	−0.80	−1.41	−1.76	−1.97	−2.15	−2.33	−2.45	−2.67
Cont_2	4.72	7.06	85 570	0.48	2.35	0.29	−1.17	−2.17	−2.83	−3.24	−3.50	−3.63	−3.68	−3.64
Cont_3	4.85	4.63	53 660	0.76	1.83	0.67	0.04	−0.30	−0.49	−0.65	−0.85	−1.90	−1.34	−1.50
Cont_4	4.92	5.58	120 000	1.50	3.29	1.38	−0.03	−0.69	−1.07	−1.34	−1.73	−2.30	−2.85	−2.75
Cont_5	5.01	4.83	203 000	0.36	4.08	1.83	0.04	−1.17	−1.92	−2.46	−3.07	−3.90	−4.98	−5.65
Cont_6	5.16	2.48	69 380	0.36	5.22	3.17	1.02	−0.72	−1.90	−2.60	−3.05	−3.46	−3.90	−4.24
Cont_7	5.29	7.70	48 320	0.48	3.96	3.16	1.12	−0.58	−1.36	−1.34	−1.00	−0.95	−1.59	−2.84
Cont_8	5.57	6.80	93 380	0.93	10.90	5.81	2.78	0.77	−0.81	−2.18	−3.32	−4.12	−4.45	−4.31
Cont_9	5.65	7.53	69 380		1.09	0.55	0.26	0.10	−0.03	−0.12	−0.21	−0.28	−0.32	−0.32
Cont_10	6.02	6.05	114 450	0.63	10.40	5.94	3.15	1.51	0.56	−0.05	−0.53	−0.99	−1.43	−1.72
Cont_11	7.57	3.89	198 000	0.06	0.33	0.03	0.05	0.05	0.05	0.05	0.05	−0.69	−0.97	−1.57
Cont_12	8.29	1.48	30 400		5.17	4.22	3.20	2.25	1.46	0.87	0.50	0.30	0.20	0.08
Cont_13	8.83	0.83	94 670		11.70	7.94	5.39	3.73	2.66	1.97	1.50	1.13	0.80	0.51

[a] Expressed as the concentration (M) of ammonium sulphate at which the protein eluted (Lienqueo *et al.*, 1996).

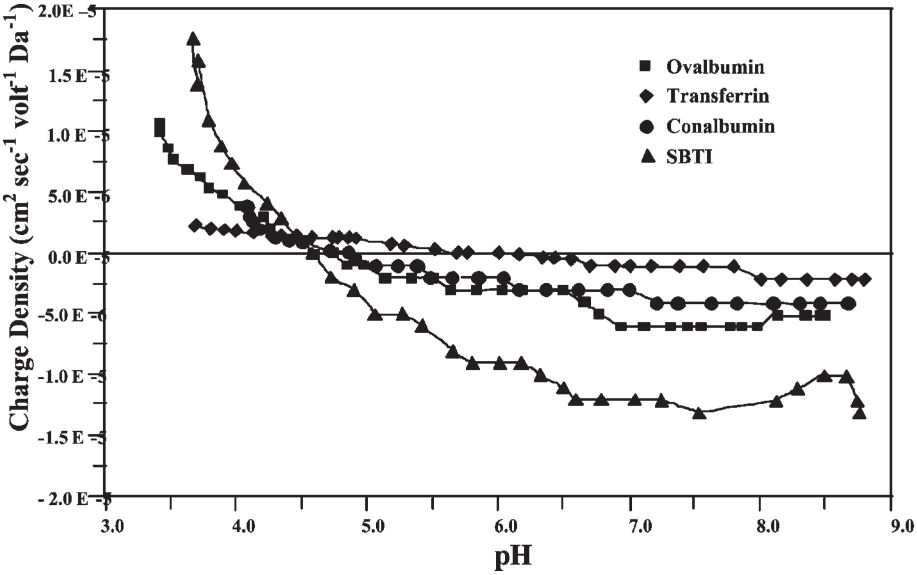

Figure 3. Charge density as a function of pH for the four proteins.

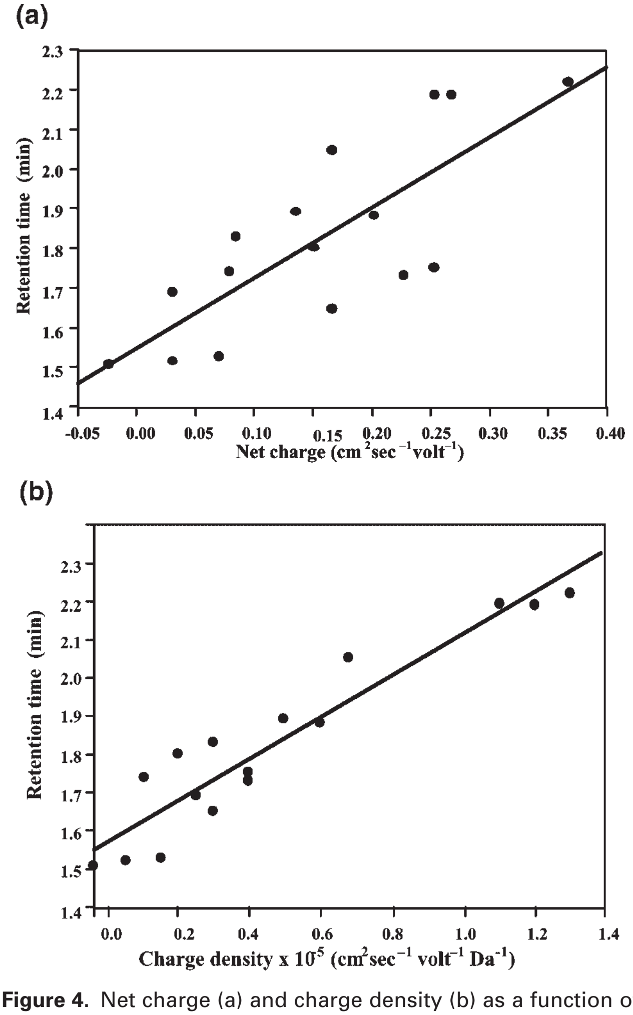

Figure 4. Net charge (a) and charge density (b) as a function of retention time for all pHs. Calculations were based on the results obtained for anion-exchange chromatography on HPLC.

be sufficient to eliminate contaminants until the product reaches the desired level of purity. In order to find the new concentration of all proteins in Table 3 after each separation step, a simple algorithm was developed based on the behaviour of a chromatographic separation, that gives an approximate value of what the concentration of each of the contaminants is after each separation step. It is clear that the second column in Table 3, namely the concentration of the protein contaminants, changes after each step.

The amount of a protein contaminant eliminated after a chromatographic step is graphically shown in Fig. 5 for three different situations. Figure 5(a) shows how the representation of the chromatographic peaks was simplified to a triangle. In Fig. 5(b) the protein product corresponds to the triangle on the left and the contaminant to the one on the right. The shaded area (S = ABC or ABCD) corresponds to the amount of contaminant left with the product in each case (Leser *et al.*, 1996). The variable Σ corresponds to the peak width and has been experimentally determined (Lienqueo *et al.*, 1996) and is shown in Table 3. It was found that under the conditions normally used for protein purification, which are well below saturation, the value of Σ is virtually independant of protein concentration. The concentration in grams per litre and the relative concentration (%) of the protein product (purity) and the main contaminants present in *E. coli* using a model protein (Leser, 1996), showing how these values evolve during a consultation with the expert system, are shown in Table 5. Since the question of 'peak cutting' was not investigated in this paper, yield of protein

Table 4. Values of Σ for the chromatographic processes used in the system

Chromatographic process	Σ
Size exclusion	0.46
Hydrophobic interaction	0.22
Ion-exchange	0.15

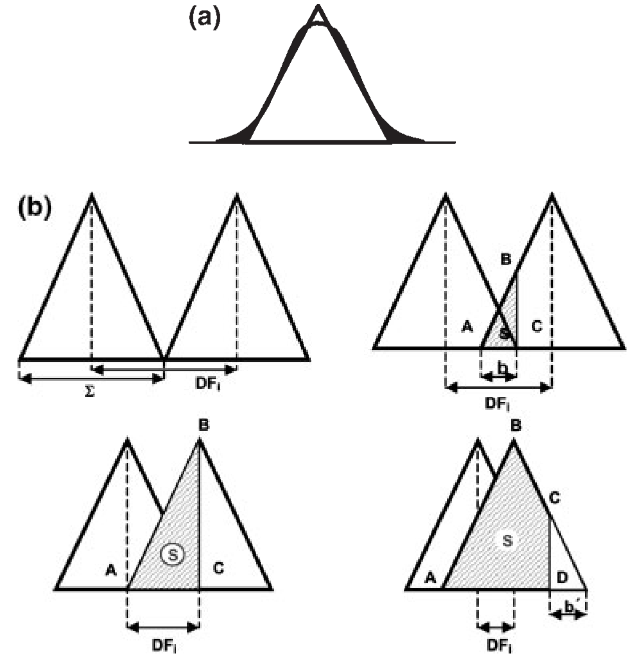

Figure 5. Representation of the peaks of a chromatogram as triangles. (a) Adjusting a triangle to a peak. (b) Variation in DF leads to different amounts of contaminant (triangle on the right) in the protein product (triangle on the left) ($\Sigma = 0.15$ for ion exchange, 0.22 for HIC and 0.46 for GF).

product was assumed to be as virtually 100%, as shown in Fig. 5(b). The questions of 'overlap', 'peak cutting' and their effect on purity and yield is only presently being studied in a systematic way in our laboratories.

In order to also consider affinity chromatography as a viable separation, as in many cases suitable affinity ligands for the protein product are well known, it was considered that if this technique is chosen by the user all contaminants will be reduced by a fixed percentage (e.g. 90%) in the affinity separation separation step (Leser, 1996). However, since affinity chromatography will have to be analysed on a case-by-case basis, given the nature of the different ligands that can be used (e.g. metal ions in IMAC, dye or other) it was not included in the present expert system.

ROBUSTNESS AND SENSITIVITY

A consultation was carried out using the expert system to find all the steps necessary to achieve the desired level of purity (e.g. 98%) for the purification of the protein somatotropin produced in *E. coli.* Once a process was found all of the values in the original databases of Tables 3 and 6 were randomly varied at the levels of 10 and 20% to see the effect on the process proposed, in terms of its robustness and sensitivity, of the system.

Somatotropin is produced in *E. coli* cells and forms inclusion bodies that represent 25% of total cellular protein. There are no published data for the hydrophobicity and the titration curve of somatotropin. Hence, they were estimated based on existing data. The Swiss Protein Database is an expert system that provides information on protein sequences and it can be accessed through internet (http://ExPaSy.hcuge.ch). Although mostly focused on molecular biology information, it was possible to obtain data such as molecular weight and pI, which, for somatotropin were 22 kDa and 7.86, respectively. The data of charge at different values of pH (titration curve) were estimated starting with zero at the pI and adopting average figures, taken from available titration curves, for the other values of pH. For hydrophobicity a value of 0.9 was used based on the data available in the protein database mentioned above (Lienqueo *et al.*, 1996). Other values were given by the user during consultation (e.g. product concentration 25 g/l, desired purity 98%) and were estimated based on process

Table 5. Concentration and relative concentration (%) of the main contaminants present in *Escherichia coli* and a model protein, showing how these values evolve during a consultation

	Loading		After the first step		After the second step		After the third step	
	Weight 0 (g/l)	Concentration 0 (%)	Weight 1 (g/l)	Concentration 1 (%)	Weight 2 (g/l)	Concentration 2 (%)	Weight 3 (g/l)	Concentration 3 (%)
Cont_1	11.24	14.97	0.22	1.62	0.00	0.00	0.00	0.00
Cont_2	7.06	9.40	0.06	0.44	0.00	0.00	0.00	0.00
Cont_3	4.63	6.17	0.24	1.76	0.01	0.19	0.01	0.20
Cont_4	5.58	7.43	0.11	0.81	0.00	0.00	0.00	0.00
Cont_5	4.83	6.43	0.09	0.66	0.00	0.00	0.00	0.00
Cont_6	2.48	3.30	0.04	0.29	0.00	0.00	0.00	0.00
Cont_7	7.70	10.25	0.02	0.11	0.02	0.39	0.00	0.00
Cont_8	6.80	9.05	0.13	0.95	0.00	0.00	0.00	0.00
Cont_9	7.53	10.03	7.56	55.51	0.15	2.89	0.00	0.00
Cont_10	6.05	8.06	0.12	0.88	0.01	0.19	0.00	0.00
Cont_11	3.89	5.18	0.00	0.00	0.00	0.00	0.00	0.00
Cont_12	1.48	1.97	0.02	0.15	0.00	0.00	0.00	0.00
Cont_13	0.83	1.11	0.01	0.07	0.00	0.00	0.00	0.00
Product	5.00	6.66	5.00	36.71	5.00	96.34	5.00	99.80

Table 6. Data for the protein product somatotropin used in the consultation for units and other information, see Table 3

				Charge									
Product	Weight	MW	Hydrophobicity	pH 4	pH 4.5	pH 5	pH 5.5	pH 6	pH 6.5	pH 7	pH 7.5	pH 8	pH 8.5
Somatropin	25	22000	0.93	4.77	3.81	2.42	1.50	1.50	0.67	0.12	0.07	−0.07	−0.50

Table 7. The downstream purification process of somatotropin (bovine growth hormone)

Published process (98% purity)	Prot_Ex suggestion (98.2% purity)
Centrifugation	Crossflow microfiltration
High-pressure homogenization	High-pressure homogenization
	Disk centrifugation
Pellet wash	
Solubilization	Solubilization
Renaturation	Renaturation
Microfiltration	
Concentration and diafiltration	Ultrafiltration
Anion exchange chromatography	Anion exchange chromatography at pH 7.5
Hydrophobic interaction chromatography	Hydrophobic interaction chromatography

conditions but are mostly not relevant to the work presented in this paper. The data on somatotropin used in the consultation is given in Table 6.

The suggested sequence for a purification process shown in Table 7 is very similar to one described as a ‘good’ industrial process that has been published (Wheelwright, 1991). That solution, however, corresponds to the overall process given by the expert system developed and described by Leser (Leser, 1996; Lienqueo *et al.*, 1996), which includes expert rules for all the initial downstream procces-ing steps. The two chromatographic steps (last two) are the ones relevant to the work presented in this paper. The differences with the published process are in stage l where centrifugation was used instead of microfiltration, and in stages 3, 4 and 5 solubilization and renaturation, which were done before separation of the cell debris. Although there is little difference between the two solid separation steps (1 and 3), for small cells like *E. coli* microfiltration usually makes more economic sense. Maybe microfiltration should be used in both steps. Regarding the two high-resolution chromatographic purification steps (7 and 8), they are the same as in the published industrial process. The purity obtained in the expert system was 98.2%, which is very similar to the one required, 98%. It is important to remember that, as exact data on somatotropin was not available, an estimated titration curve was used (from the pI, Table 3) and a possible variation of the value for hydropbobicity results in somewhat different solutions for the sequence of high resolution purification steps as shown in Table 8. If values of 0.5 or 1.3 are chosen, the sequence in both cases is hydrophobic interaction chromatography and size exclusion (gel filtration) for steps 7 and 8 and the final purity is 99.4 and 99.7% respectively. Interestingly enough, if the protein’s hydrophobicity is 0.7, three steps are needed instead of two, namely HIC, Anion exchange and size exclusion (final purity 99.7%). This clearly shows that the expert system is sensitive to the physicochemical parameters of the product protein to be purified.

The sensitivity of the proposed process to random changes in the values determined experimentally, shown in Tables 3 and 6, were investigated in order to assess the robustness of the system to either variations in the properties of the contaminant proteins present in the *E. coli* cells used (how universal is the data of Table 3) or in the experimental measurements. When only the *E. coli* data (Table 3) or both sets of data (Tables 3 and 6) were randomly varied at the level of 10% the sequence of operations was exactly the same as shown in Table 9. On the other hand, as can be seen

Table 8. Sequence of high-resolution purification operations and product purity for different values of hydrophobicity (of somatotropin)

	Cromatographic separation			
Hydrophobicity	First step	Second step	Third step	Purity (%)
0.2	Hydrophobic interaction	Size exclusion		99.7
0.5	Hydrophobic interaction	Size exclusion		99.4
0.7	Hydrophobic interaction	Anion exchange at pH 7.5	Size exclusion	99.7
0.9	Anion exchange at pH 7.5	Hydrophobic interaction		98.2
1.3	Hydrophobic interaction	Size exclusion		99.7

Table 9. Sequence of high resolution purification operations and product purity for 10% random variation of the databases (measured values) if only *E. coli* data varies or if both set of data (*E. coli* and somatotropin) vary

Only E. coli data varies	
First step	Anion exchange chromatography at pH 7.5
Second step	Hydrophobic interaction chromatography
Purity	98.23%
Both set of data vary	
First step	Anion exchange chromatography at pH 7.5
Second step	Hydrophobic interaction chromatography
Purity	98.04%

Table 10. Sequence of high-resolution purification operations and product purity for 20% random variation of the databases (measured values) if only *E. coli* data varies or if both set of data (*E. coli* and somatotropin) vary

Only E. coli data varies	
First step	Hydrophobic interaction chromatography
Second step	Size exclusion chromatography
Purity	99.68%
Both set of data vary	
First step	Hydrophobic interaction chromatography
Second step	Size exclusion chromatography
Third step	Anion exchange chromatography at pH 7.5
Purity	99.71%

in Table 10, when the data was varied at the level of 20%, the sequence changed. This clearly shows that the system has the necessary robustness to variations and possible errors in the experimental determination of the data shown in Tables 3 and 6 (<10%) but is sensitive enough to larger variations in protein properties (>20%).

EXPERIMENTAL TESTS

Purity criterion

Considering that the most important parameter after a separation step is the final product purity, and that an algorithm has been developed to calculate the purity after each step, this was also implemented as a possible selection criterion as an alternative to the SSC. This criterion compares the final purity level obtained after a particular chromatographic technique has been applied.
The purity concept is defined as:

$$\text{Purity} = \frac{\text{concentration of the target protein}}{\Sigma\text{concentration of all the proteins present}} \tag{8}$$

After determining which chromatographic technique gives the highest purity level, the system chooses this as the technique to use at this step. It then compares the purity with that required. A sequence of steps is chosen until the required level of purity is reached. Finally the system creates a list with the defined sequence of operations.

Two examples have been tested experimentally: a model protein mixture and a recombinant β-1,3-glucanase from *Bacillus subtilis* culture (Lienqueo *et al.*, 1999).

Purification of BSA

Assessments were done using both criteria (SSC and purity) implemented in Prot_Ex_Purification for purification of BSA from a mixture of four proteins [BSA, soybean trypsin inhibitor (SBTI), ovalbumin and thaumatin]. The data on this mixture used in the consultations are given in Table 11. The results obtained for a target of 94% purity are shown in Table 12.

The SSC criterion selects a purification sequence based on the elimination of the contaminant that gives the highest SSC value. Its contribution is described through the product of the concentration factor (θ), the efficiency factor (η) and the DF. In those cases where all contaminants have the same concentration (equal concentration factor) and the efficiency is constant, then DF is the variable that has the main contribution. The SSC criterion is based on the elimination of the contaminant that has properties the most different from those of the target protein [Fig. 6(a)]. In this example cation exchange chromatography at pH 6.0 is useful to eliminate the protein thaumatin. Nevertheless, the purity achieved after the purification is only 33%. However, if

Table 11. Physicochemical properties of protein mixture

Proteins	Initial concentration (mg/cm^3)	Molecular weight (Da)	Hydrophobicity [$(NH_4)_2SO_4$]	Charge (Coulomb/mol) 10^{-25}				
				pH 4.0	pH 5.0	pH 6.0	pH 7.0	pH 8.0
BSA	2	67 000	0.86	1.03	−0.14	−1.16	−1.68	−2.05
Ovalbumin	2	43 800	0.54	1.40	−0.76	−1.65	−2.20	−2.36
SBTI	2	24 500	0.90	1.22	−0.76	−1.54	−2.17	−2.13
Thaumatin	2	22 200	0.89	1.94	1.90	1.98	1.87	0.91

Table 12. Sequence suggested by the expert system to obtain a purity superior to 94% in the purification

SSC criterion chromatography steps	Purity	Purity criterion chromatography steps	Purity
Cation exchange at pH 6.0	33.1%	Anion exchange at pH 7.0	63.7%
Hydrophobic interaction	49.5%	Hydrophobic interaction	94.5%
Anion exchange at pH 7.0	97.0%		

anion exchange chromatography at pH 7.0 is used as suggested by the purity criterion it is possible to eliminate thaumatin and a part of SBTI, obtaining a purity of 64%. This has been confirmed experimentally, as shown in Figs 6 and 7 (Lienqueo *et al.*, 1999). This situation occurs because the purity criterion determines the optimum chromatographic step considering all the contaminants present. The SSC criterion considers only the contaminant that gives the highest SSC value. For this reason the chromatographic step chosen using the purity criterion was the optimum for that

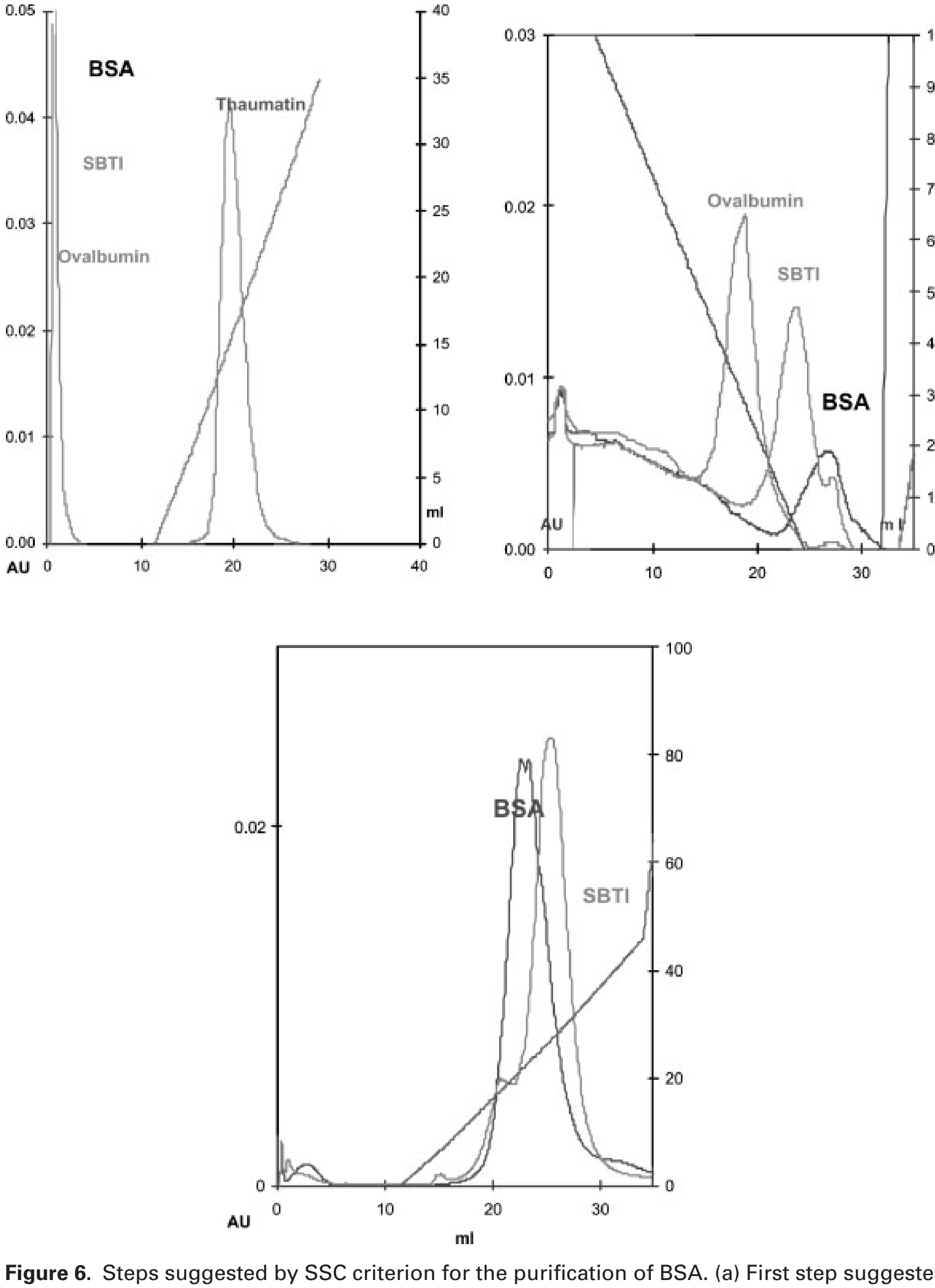

Figure 6. Steps suggested by SSC criterion for the purification of BSA. (a) First step suggested: cation exchange chromatography at pH 6.0. (b) Second step suggested: hydrophobic interaction chromatography. (c) Third step suggested: anion exchange chromatography at pH 7.0.

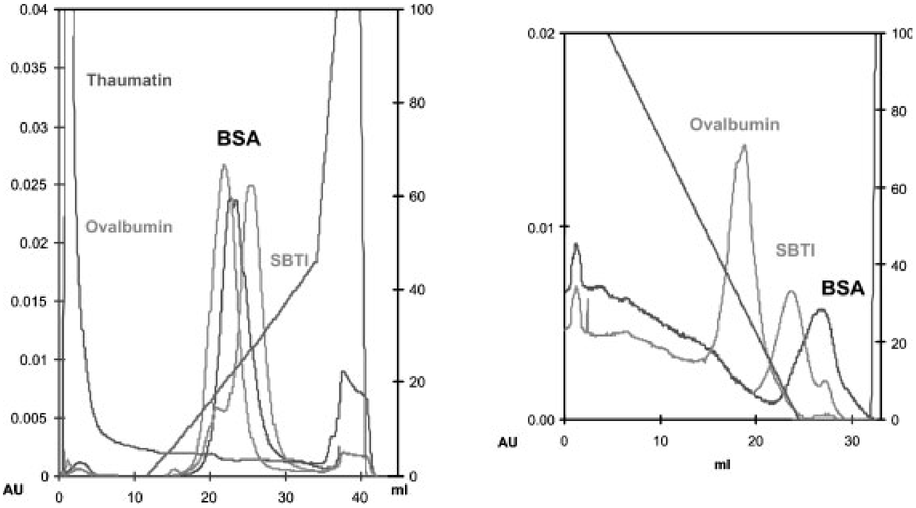

Figure 7. Steps suggested by purity criterion for the purification of BSA. (a) First step suggested: anion exchange chromatography at pH 7.0. (b) Second step suggested: hydrophobic interaction chromatography.

Table 13. Physicochemical properties and concentration for the main proteins in *Bacillus subtilus* ToC46(pFF1)

Proteins	Initial concentration (mg/cm^3)	Molecular weight (Da)	Hydrophobicity [$(NH_4)_2SO_4$]	Charge (Coulomb/mol) 10^{-25}				
				pH 4.0	pH 5.0	pH 6.0	pH 7.0	pH 8.0
β-1,3-Glucanase	0.60	31 000	0.00	1.46	−0.62	−1.02	−2.33	−2.52
Contaminants								
Low hydrophobic:								
Contaminant 1	2.74	41 000	1.50		0.26	−0.87	−1.65	−2.04
Contaminant 2	2.74	32 000	1.50		0.00	−2.70	−3.51	−3.51
Medium hydrophobic:								
Contaminant 3	0.25	35 500	0.20		−0.55	−0.22	−0.73	−1.82
High hydrophobic:								
Contaminant 4	0.42	62 500	0.00		−1.06	−1.17	−2.79	−3.32
Contaminant 5	0.25	40 600	0.00		−0.55	−0.22	−0.73	−1.82
Contaminant 6	0.25	69 600	0.00		−0.55	−0.22	−0.73	−1.82
Contaminant 7	0.09	40 600	0.00		1.46	−0.47	−1.06	−1.04
Contaminant 8	0.09	69 600	0.00		1.46	−0.47	−1.06	−1.04

Table 14. Sequence suggested by the expert system for both criteria

Both criteria chromatography steps	Purity	Experimental validation chromatography steps	Purity
Hydrophobic interaction	32.7%	Hydrophobic interaction	33–38%
Anion exchange at pH 6.5	70.3%	Anion exchange at pH 6.5	65–70%

stage and gave a higher purity than that obtained when the SSC criterion was used. For example, the second step suggested for both criteria was hydrophobic interaction chromatography which eliminates ovalbumin. Finally the SSC criterion considers an additional step (anion exchange chromatography at pH 7.0) to eliminate SBTI and to reach a final purity level of 97%. This was the first step suggested by the purity criterion.

Purification of β-1,3-glucanase and experimental investigation

Assessments were done using both the SSC criterion and the purity criterion implemented in the expert system for purification of a β-1,3-glucanase from *B. subtilis* ToC46 (pFFI) culture. In this case both criteria gave exactly the same sequence. The data on this system are given in Table 13. The

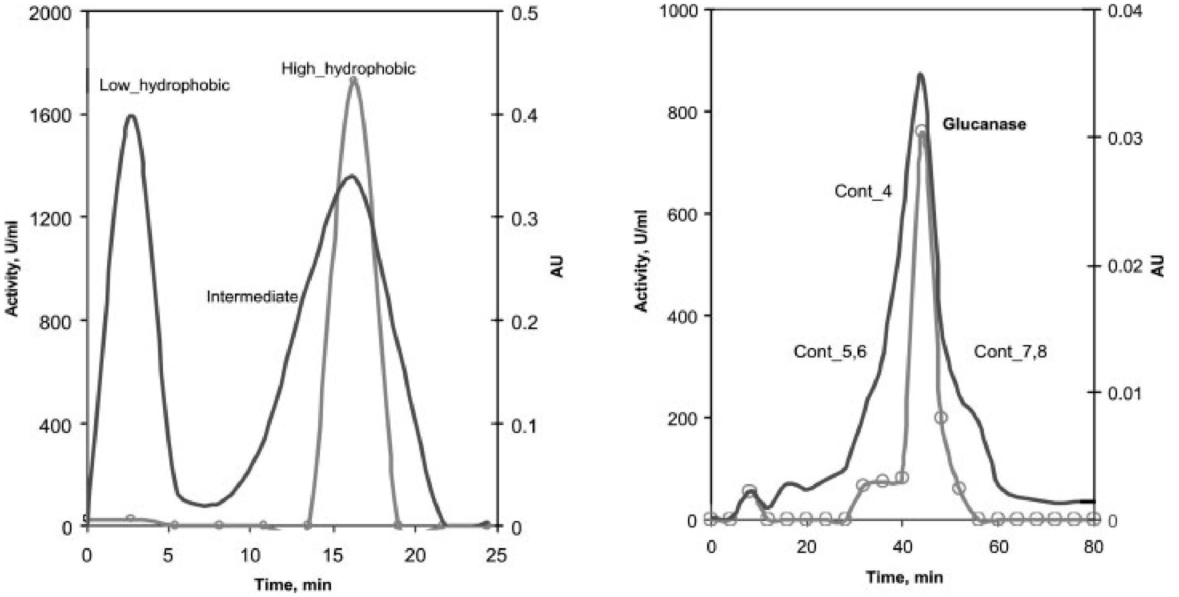

Figure 8. Steps suggested by both criteria for the purification of glucanase. (○) Glucanase activity. (a) First step suggested: hydrophobic interaction chromatography. (b) Second step suggested: anion exchange chromatography at pH 6.5.

results obtained for 70% purity are shown in Table 14 (Lienqueo *et al.*, 1999).

The chromatograms from the purification sequence are shown in Fig. 8(a) and 8(b) and Table 14. Figure 8(a) shows the separation of 'low hydrophobicity proteins' (contaminants 1 and 2) and part of the 'medium hydrophobicity proteins' (contaminant 3) from the β-1,3-glucanase. In this first step, the main contaminants were eliminated. Figure 8(b) shows the separation of contaminants 3–4, 5–6 and 7–8 from the β-1,3-glucanase. Figure 8 shows that the scheme for purification suggested by the expert system is valid for purification of this recombinant β-1,3-glucanase.

CONCLUSIONS

In this paper we have reviewed and discussed the proteomic approach to select a purification process for proteins based on physicochemical properties. The methodology described constitutes a rational proteomic procedure to separate the main contaminant proteins with a minimum number of steps.

An algorithm to calculate the SSC parameter used to select the actual purification at each step was developed, and the translation of physicochemical data of the proteins to chromatographic behaviour was also carried out for ion-exchange chromatography, hydrophobic interaction chromatography and gel filtration.

Another algorithm used to estimate concentration of each protein contaminant after a chromatographic process is performed, was also developed. The methodology described, which was handled by a computer-based expert system, was tested with recombinant proteins produced in *E. coli,* with a good database for the main protein contaminants, and purification of a recombinant protein product, with good results.

The system was robust to errors <10% which is the range that can be found in the experimental determination of the properties in the database of product and contaminants. On the other hand, the system was sensitive both to larger variations (>20%) in the properties of the contaminant database and the protein product and to variations in one protein property (e.g. hydrophobicity).

The purification strategy proposed was experimentally tested and validated with a mixture of four proteins and the experimental validation was also carried out with an 'unknown' supernatant of *Bacillus subtilis*, producing a recombinant β-1,3-glucanase.

In addition to SSC, final purity can also be used as a selection criteria given the fact that it is also calculated after each separation step is performed, to give the new protein contaminant concentrations in the database. Although both criteria SSC and purity will in most cases give similar results, purity may give fewer steps (and thus a better process) when concentrations of contaminant proteins are similar in the crude starting material.

REFERENCES

Asenjo JA. 1990. Selection of operations in separation processes. In *Separation Processes in Biotechnology*, Asenjo JA (ed.). Marcel Dekker: New York; 3–16.

Leser EW. 1996. Prot_Ex: an expert system for selecting the sequence of processes for the downstream purification of proteins. Ph.D. Thesis, University of Reading.

Leser EW, Asenjo JA. 1992. The rational design of purification processes for recombinant proteins. *J. Chromatogr*. **584**: 35–42.

Leser EW, Asenjo JA. 1994. The rational selection of purification processes for proteins: an expert system for downstream processing design. *Ann. NY Acad. Sci*. **721**: 337–347.

Leser EW, Lienqueo ME, Asenjo JA. 1996. Implementation in an expert system of selection rationale for purification processes for recombinant proteins. *Ann. NY Acad. Sci*. **782**: 441–455.

Lienqueo ME, Leser EW, Asenjo JA. 1996. An expert system for the selection and synthesis of multistep protein separation processes. *Comput. Chem. Eng*. **20**: S189–S194.

Lienqueo ME, Salgado JC, Asenjo JA. 1999. An expert system for selection of protein purification processes: experimental validation. *J. Chem. Technol. Biotechnol*. **74**: 293–299.

Watanabe E, Tsoka S, Asenjo JA. 1994. Selection of chromatographic protein purification operations based on physicochemical properties. *Ann. NY Acad. Sci*. **721**: 348–364.

Wheelwright SM. 1991. *Protein Purification Design and Scale-up of Downstream Processing*. Hanser: Munich.

Woolston PW. 1994. A physicochemical database for an expert system for the selection of recombinant protein purification processes. Ph.D. Thesis, University of Reading.