OPTIMIZATION-BASED SYNTHESIS OF PURIFICATION STEPS IN PROTEIN PRODUCTION PROCESSES

Elsa Vasquez-Alvarez and Jose M. Pinto*
Department of Chemical Engineering, University of Sao Paulo
Av. Prof. Luciano Gualberto t. 3 n. 380, Sao Paulo, SP, 05508-900 Brazil
Othmer Department of Chemical and Biological Sciences and Engineering,
Polytechnic University, Six Metrotech Center, Brooklyn, NY, 11201 USA

Abstract

The objective of this work is to develop optimization models for the synthesis of protein purification chromatographic processes that incorporate economic information and product losses. Mathematical models for each chromatographic technique rely on physicochemical data on the protein mixture and represent the chromatographic peaks by a normal distribution function. In terms of the synthesis model, formulations that are based on a convex hull representation are proposed to calculate several objective functions. The methodology is validated in examples with experimental data and compared to simpler MILP models as well as to expert systems. Results are shown to provide an important guideline for synthesizing purification processes.

Keywords

Bioprocess synthesis, protein purification, mixed-integer optimization.

Introduction

Preparative bioseparations are of principal interest in the biotechnology industry. Biomolecules such as proteins, peptides, and nucleic acids frequently need to be purified from complex mixtures composed of many similar molecules, making purification a difficult and expensive process Moreover, in many cases purification represents a major parcel of the manufacturing cost.

The most important techniques for recovery and purification of protein mixtures include liquid chromatography. One of the main challenges in the synthesis of downstream purification stages is the appropriate selection and sequencing of chromatographic steps.

Steffens et al. (1999) developed a synthesis approach that is based on physicochemical property data. Units whose relevant component properties are significantly different are selected. The methodology was implemented within an implicit enumeration algorithm. Steffens et al. (2000) developed an algorithm that considers purification tags, which are attached to a specific product and help the purification at subsequent stages. The work is integrated in a synthesis algorithm. The particular application also depends on the physicochemical properties (product + contaminants).

Lienqueo and Asenjo (2000) developed an expert system that uses heuristic rules applied to large-scale downstream processes. The study was divided in two parts: the recovery and the purification process. The authors used real examples for the validation of their methodology.

Vasquez-Alvarez et al. (2001) developed optimization approaches for synthesis of protein purification processes. They developed mathematical models based on mixed-integer linear programming (MILP) for the optimal selection and sequencing of purification steps. The objectives of these works were to minimize the number of purification stages for a specified purity level of the product, and to maximize product purity. The study

*E-mail: jpinto@poly.edu

assumed the total recovery of the product. Later, Vasquez-Alvarez and Pinto (2003) developed another MILP that incorporates product loss along of the process in order to evaluate the trade-off between product quality (given by the purity) and quantity.

The objective of this work is to extend the previous works by incorporating economic information. Models for each chromatographic technique represent the chromatographic peaks with a normal distribution function.

Problem Description

Consider a complex protein mixture in which one of them has to reach a high purity degree using high resolution chromatographic techniques. Information on physicochemical properties is used for the target and contaminant proteins and each technique is able to perform the separation of the mixture by exploiting a specific physicochemical property, such as surface charge as a function of pH, surface hydrophobicity etc.

Generally, several steps are necessary to purify a protein mixture. Losses are considered in the target protein along the purification process, in order to evaluate the trade-off between product by purity and quantity. In other words, the higher the purity achieved within each step, the smaller the product yield. In this sense, decisions involve the selection of techniques and their order as well as the percentage of product recovered.

Mathematical Model of Chromatographic Techniques

The chromatographic peaks are approximated by a normal distribution function (Gaussian curve). Moreover, the mass reduction of the contaminants is a function of the area formed by the intersection of the curves of the contaminant and the desired protein. Retention times and the characteristic width of the chromatographic peak are defined for each technique.

Each chromatographic technique i performs separation based on the physicochemical properties of the proteins ($P_{a,p}$) that are relevant to it (set A_i). These are used for the calculation of the dimensionless retention time ($Kd_{i,p}$), as in Equation 1 (Lienqueo et al., 1996):

$$Kd_{i,p} = f_i\,(P_{a,p} |\ a \in A_i) \qquad \forall\ i, p \tag{1}$$

Another parameter is the deviation factor ($DF_{i,p}$), that indicates the separation distance of the target protein chromatographic peak from the one of contaminant p for the technique i, as shown in Equation 2.

$$DF_{i,p} = |Kd_{i,dp} - Kd_{i,p}| \qquad \forall\ i, p \tag{2}$$

Another important parameter used in the model is the characteristic width that depends on chromatographic technique i. This dimensionless parameter was originally determined by Lienqueo (1999) for the triangular approximation of the chromatograms. In the proposed model, parameter the characteristic width is calculated assuming that the width of the triangle (triangular approach) and the width of the cumulative normal distribution are equal. Therefore, the width of each chromatographic technique i of the proposed model is equal the six times the standard deviation ($6\cdot\rho_i$), whose value is equivalent to 99.87% of the area of the Gaussian curve. Standard deviation values are shown in Table 1.

Table 1. Standard deviation values.

Chromatographic technique	standard deviation (ρ_i) (Lienqueo, 1999)
Ion exchange	0.0250
Hydrophobic interaction	0.0366
Gel filtration	0.0766

The proposed model is developed for the case of target protein loss along the process. Parameters $n_{i,dp,l}$ and $B_{i,p,l}$ are defined to determine the discrete levels of product recovery and contaminant proteins, respectively. Values of discrete values ($n_{i,dp,l}$) for the product recovery are presented in the example.

Figure 1 shows representations of chromatographic peaks of the Gaussian form for the model that considers losses of the product along the chromatographic process. We admit that the peaks have constant form and that the peak on the left refers to the product and the other to the contaminant. The area of the figure formed by the intersection of the two peaks (shaded area) represents the quantity of the contaminant p that remains in the mixture (with the product) after chromatographic technique i.

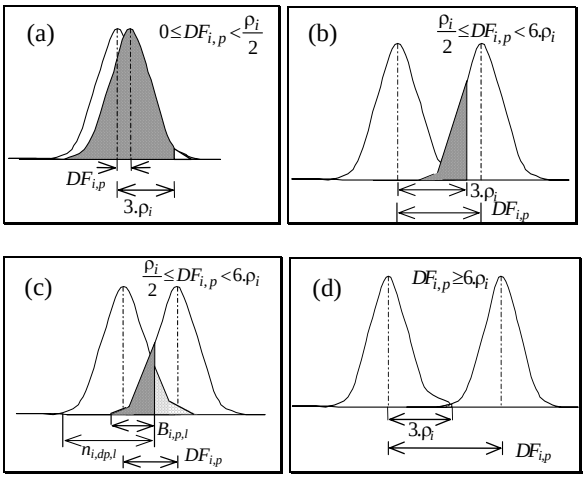

Figure 1. Gaussian peak representation

In Figure 1, parameter $B_{i,p,l}$ indicates the recovery levels imposed to the model (set of discrete values) for the case of contaminants. Four situations can happen depending on the peaks relative position, after application

of one chromatographic technique. Figures 1(a) and 1(d) indicate no purification and complete purification, respectively. Besides these extreme situations, Figure (1b) shows the presence of contaminant in the mixture (dark area) after chromatographic technique *i* is applied. Figure 1(c) indicates the amount of contaminant that remains in the mixture (darker area) and the quantity of product that is eliminated. The amounts of contaminant and product that remain in the mixture are determined from the concentration factors ($CF_{i,p,l}$) for any protein *p* in chromatographic step *i* and level *l*. Note that for total product recovery, (*L*=1) the base (width) of the corresponding peak of the product is $6\cdot\rho_i$ and for the contaminants is ($6\cdot\rho_i$ - $DF_{i,p}$).

If there are product losses in the process, the basis of the product curve takes values in accordance with the levels of product loss, whereas the basis of the contaminant peaks takes ($B_{i,p,l}$) values. Table 2 shows the relations for $CF_{i,p,l}$. It is important to note that the upper limit of the integral ($X_{i,p,l}$) represents the standardized value of the base ($B_{i,p,l}$), $x_{i,p,l} = \left(B_{i,p,l} - 3.\rho_i\right)\big/\rho_i$.

The mathematical model determines the mass ratio of each one of the components that remains in the mixture. The generated information is used in the synthesis model shown in the next section.

Table 2. Relations for the concentration factors

Base $B_{i,p,l}$		Conc. Factor ($CF_{i,p,l}$)
$B_{i,p,l} = n_{i,p,l}$	$\forall p=dp$	$CF_{i,p,l} = 1$
$B_{i,p,l} = n_{i,dp,l} - DF_{i,p}$	$\forall p\neq dp$	Fig. (1a)
$B_{i,p,l} = n_{i,p,l}$	$\forall p=dp$	$CF_{i,p,l} = \int_{-\infty}^{x_{i,p,l}} \frac{1}{2.\pi}.e^{\frac{x_{i,p,l}^2}{2}}.dx_{i,p,l}$
$0 \le B_{i,p,l} \le n_{i,dp,l} - DF_{i,p}$	$\forall p\neq dp$	Fig. (1b) and (1c)
$B_{i,p,l} = 0$		$CF_{i,p,l} = 0.001$ Fig. (1d)

Synthesis Model

The constraints of the synthesis model that minimizes the total chromatographic techniques for a given purity levels are presented in detail in Vásquez-Alvarez and Pinto (2003). This model considers a convex hull representation that is derived from the linear disjunction given in (3) that is defined for each order *k* in the sequence (*k* = 1,…*K*).

$$\begin{bmatrix} \bigvee_{i=1}^{I} \bigvee_{l=1}^{L} \begin{bmatrix} \Lambda_{i,k,l} \\ m_{p,k+1} = CF_{i,p,l}.m_{p,k} \ \forall p \end{bmatrix} \end{bmatrix} \vee \begin{bmatrix} \mathrm{A}_k \\ m_{p,k+1} = 0 \ \forall p \end{bmatrix} \qquad (3)$$

Disjunction (3) contains *I.L*+1 elements for each order *k*. The first *I.L* terms model the selection of step *i* in order *k* at level *l* (represented by Boolean variables $\Lambda_{i,k,l}$), whereas the last term models no step selection (represented by Boolean variable A_k). In each term, the mass of contaminant protein *p* at step *k* ($m_{p,k}$) is related to the mass at the previous step.

The constraints from the proposed MILP model that are based on a convex hull relaxation of disjunction rely on binary variables $\lambda_{i,k,l}$ that correspond to the Boolean variables in disjunction (3). Assignment, ordering, contaminant and specification constraints are defined in the MILP.

Objective function (4) selects a sequence with minimum number of steps for given purity as well as yield specifications, and is defined as follows:

$$Min\,\phi = \sum_i \sum_k \sum_l \lambda_{i,k,l} = \sum_k k.Z_k \qquad (4)$$

In (4), Z_k denotes the last step in the sequence.

Economical evaluation of the purification process.

With the purpose to evaluate the protein purification processes in economic terms, a new objective function is proposed. The objective function is the operational profit (*Pr*) maximization, composed by the product revenues and the operational costs for resins of the chromatographic columns. It can be written as follows:

$$\mathbf{Pr} = P_{dp} m_{dp,K+1} - \sum_i \frac{S_i}{Nc} \left(\sum_p \sum_{k\geq 1} \sum_l Cre_i m^1_{i,p,k,l} + \sum_p \sum_l Cre_i m_{p,0} \lambda_{i,1,l} \right) \qquad (5)$$

The first term of (5) accounts for the revenues of the desired product. The second parcel is constituted by operational costs that were calculated as function of the resin costs (Cre_i), life time (frequency of substitution or number of cycles, given by *Nc*), protein mass ($m^i_{i,p,k,l}$) along the process and the size factor (S_i), which represents the necessary volume per unit mass in step *i*.

It is important to note that the variables that denote the mass of proteins in (5) result from the convex hull representation of disjunction (3). Details are shown in Vasquez-Alvarez and Pinto (2003).

Computational Performance

The models are tested and implemented in GAMS/CPLEX 7.0 (Brooke et al., 1998) to generate their solution in several examples. Data (physicochemical properties as well as the initial concentration of the mixture) are shown in Vasquez-Alvarez et al. (2001).

An example for the purification of a recombinant mixture that contains nine components and considers 22 chromatographic techniques is shown next. The target protein is β-1,3 glucanase (*p*1) whose initial composition is 8.3 %, 98%purity level and 7% maximum product loss (*fr* in Table 4). The discrete recovery levels (*L*) are 4, as follows: 99.8%, 99.4%, 97.7%, and 93.3% obtained from standard deviation values. Figure 2 shows that three steps

are necessary to reach the required purity and recovery. Triangular approximations provide the same solution.

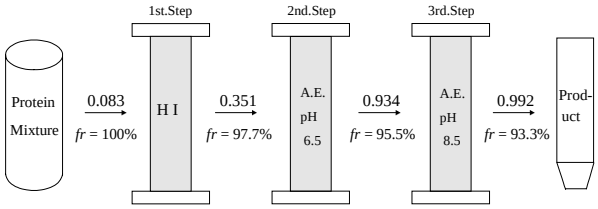

Figure 2. Solution for the minimization of steps

If no product loss is considered (only one recovery level – L=1) and the specified purity level is 94%, the optimal solution contains four stages (results not shown); note that it is not possible to achieve 98% without allowing for product loss. The result obtained by an Expert System (Lienqueo and Asenjo, 2000) is limited to 70% with two steps (HI and AE pH 6.5). The same stages are obtained in the first steps of the proposed model, reaching 78% of purity. The experimental validation process for this process gives a purity range of 60 -70%.

The same system was optimized with the economical objective function. The price of $p1$ is \$1460/g and cost data for the resins are given in Table 3. Interestingly, the solution provides a larger number of steps (see Figure 3) with \$798 maximum profit; the solution shown in Figure 2 presents a \$778 profit. Statistical data are given in Table 4.

Table 3. Cost data for the resins

Resin (SIGMA-ALDRICH, 2002)	Cost(\$/L)
Anion exchange – Q sepharose Fast Flow (FF)	700
Cation exchange – S sepharose FF	700
Hydrophobic interaction–phenyl sepharose FF	1519
Gel filtration sephacryl 200HR	450

Table 4. Statistical results for profit maximization

L	fr	Integer variables	Continuous variables	Nodes	CPU*(s)
1	100	288	2379	300	31
4	93	1080	8913	1690	440**

*Pentium II 384 MB, **4.9% relative gap

Conclusions

This paper presented an MILP model for the synthesis of purification process that considers Gaussian approximation of the chromatograms with product losses. An economic objective function is proposed that includes revenues from protein sales as well as resin operating costs. The methodology is validated in an example with experimental data and compared to more simplified MILP models as well as to an Expert System. Results provide an important guideline for synthesizing purification processes.

Acknowledgments

We acknowledge financial support from PADCT /CNPq under grant 62.0239/97 – QEQ, and from VITAE under grant B-11487/10B006.

References

Brooke, A., Kendrick, D., Meeraus, A, Ramesh, R. (1998). GAMS- A User's Guide (Release 2.25). *The Scientific Press*. San Francisco, CA.

Lienqueo, M. E., Leser, E. W., Asenjo, J. A. (1996). An expert system for the selection and synthesis of multistep protein separation processes. *Comp. Chem. Eng., 20*, S189.

Lienqueo, M. E. (1999). *Development of an expert system for the rational selection of protein purification processes.* Ph.D Thesis (in Spanish), University of Chile, Santiago.

Lienqueo, M. E., Asenjo, J. A. (2000) Use of expert systems for the synthesis of downstream protein processes. *Comp. Chem. Eng*., *24*, 2339.

SIGMA ALDRICH. (2002). Liquid chromatography. http://www.sigmaaldrich.com.

Steffens, M.A.; Fraga, E.S., Bogle, I. D. L. (1999) Multicriteria process synthesis for generating sustainable and economic bioprocesses. *Comp. Chem. Eng*., *23*, 1455.

Steffens, M.A., Fraga, E.S., Bogle, I.D.L. (2000). Synthesis of purification tags for optimal downstream processing. *Comp. Chem. Eng., 24*, 717.

Vasquez-Alvarez, E., Lienqueo, M.E., Pinto, J.M. (2001). Optimal synthesis of protein purification processes. *Biotechol. Prog., 17*, 685.

Vasquez-Alvarez, E., Pinto, J. M. (2003). A mixed integer linear programming model for the optimal synthesis of protein purification process with product loss. *Chem. Biochem. Eng. Q., 17*, 27.

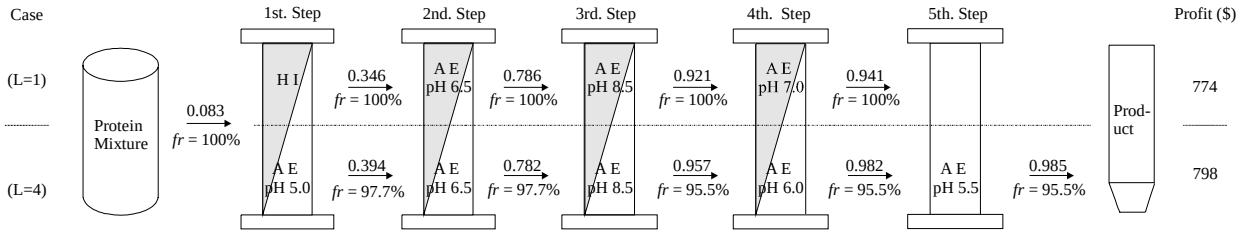

Figure 3. Solution for profit maximization.