European Symposium on Computer Aided Process Engineering – 15
L. Puigjaner and A. Espuña (Editors)

An MILP Model for Optimal Design of Purification Tags and Synthesis of Downstream Processing

Evangelos Simeonidis[a], Jose M. Pinto[b] and Lazaros G. Papageorgiou[a,*]

[a]Centre for Process Systems Engineering, Department of Chemical Engineering
UCL (University College London), Torrington Place, London WC1E 7JE, U.K.

[b]Department of Chemical and Biological Sciences and Engineering
Polytechnic University, Six Metrotech Center, Brooklyn NY 11201, U.S.A.

Abstract

Downstream protein processing in biochemical production plants can be improved significantly with the use of peptide purification tags: comparatively short sequences of amino acids fused onto the product protein, which modify the physical properties of the desired product in a way that enhances its separation from contaminants. A two-step MINLP framework that integrates the selection of optimal peptide tags with the synthesis of downstream processing has previously been developed by the authors. The objective of this work is to transform this framework to a simpler MILP model. The methodology is validated by an illustrative example based on experimental data.

Keywords: protein purification processes, peptide tags, mixed integer linear programming

1. Introduction

Recent advances in biotechnology have given immense impetus to the introduction of biopharmaceutical and biotechnological products. Downstream processing is typically among the most difficult and complex stages and the source of a large portion of the manufacturing and investment costs in a biochemical production plant. The quality of the product is predominantly determined at the purification level, which may therefore be regarded as the most important production stage. Early systematic methods for the synthesis of downstream protein processing made use of expert knowledge systems for selecting operations (Lienqueo *et al.*, 1996). Vasquez-Alvarez and Pinto (2004) presented a mixed integer linear programming (MILP) framework, in which mathematical models for each chromatographic technique rely on physicochemical data on the protein mixture that contains the desired product, and provide information on its potential purification.

Considerable improvement of downstream protein purification processes can be achieved with the use of peptide purification tags (Steffens *et al.*, 2000, Simeonidis *et al.*, 2004). Peptide tags are comparatively short sequences of amino acids, genetically fused on the protein product, in order to modify its physicochemical properties in a way that will enhance the separation, thus simplifying the purification flowsheet. The

* Author to whom correspondence should be addressed: l.papageorgiou@ucl.ac.uk

development of a framework for the optimal design of case-specific peptide tags that alter the properties of a particular protein product in the most beneficial way, and the concurrent synthesis of downstream protein processing has been previously presented by the authors (Simeonidis *et al*., 2004); a methodology based on a two-step, mixed integer non-linear programming (MINLP) framework has been developed.
In this work, the above model is reformulated as a mixed integer linear programming (MILP) model through piecewise linear approximations of the nonconvex, nonlinear functions. The new model utilises physicochemical property data to specify the amino acid composition of the shortest and most advantageous peptide tag configuration, and concurrently select operations among a set of candidate chromatographic techniques in order to achieve a specified purity level. The applicability of the model is demonstrated by an example that relies on experimental data.

2. Problem Statement

Overall, the problem of simultaneous optimal tag design and synthesis of downstream protein processing can be stated as follows:

Given:
- a mixture of proteins (p: 1,...,P) with known physicochemical properties;
- a set of available chromatographic techniques (i: 1,...,I) each performing a separation task by exploiting a specific physicochemical property;
- the properties of the twenty amino acids (k: 1,...,20); and
- a minimum purity level for the desired product (dp).

Determine:
- the amino acid composition of the shortest and most advantageous peptide tag;
- the physicochemical properties of the tagged protein (desired product + tag); and
- the flowsheet of the high-resolution purification process.

So as to optimise a suitable performance criterion.

3. Mathematical Formulation

Next, the main components of the proposed mathematical framework are briefly described. The resulting MILP representation, designed for the synthesis of purification bioprocesses, so as to consider the optimal design of purification tags, extends an earlier MINLP formulation (Simeonidis *et al*., 2004).

3.1 Physicochemical property constraints

The tagged protein's net charge (Q_{dp}) is predicted based on the methodology suggested by Mosher *et al*. (1993).

$$Q_{i,dp} = \hat{Q}_{i,dp} + \sum_{k \in \mathrm{BA}} \frac{n_k}{\dfrac{K_k}{[H^+]_i} + 1} - \sum_{k \in \mathrm{AA}} \frac{n_k}{\dfrac{[H^+]_i}{K_k} + 1} \tag{1}$$

where *BA* and *AA* are the acidic and basic amino acid groups respectively; K_k is the ionisation constant; n_k is the integer number of amino acids *k* in the tag and $\hat{Q}_{i,dp}$ is the initial product charge.

The tagged protein's hydrophobicity (H_{dp}) is estimated using the work by Lienqueo *et al*. (2002). The calculation is based on the relative contribution of each amino acid to the surface properties of the product protein and the knowledge of its 3D structure.

3.2 Dimensionless retention times

Dimensionless retention times KD_{ip} are defined as a function of net charge Q_{ip} or hydrophobicity H_p. For ion exchange chromatography, retention times for the tagged protein product are estimated based on approximations of the chromatograms by isosceles triangles and on physicochemical property data for the product and contaminants (Vasquez-Alvarez *et al*., 2001). The methodology presented by Lienqueo *et al*. (2002) is used to estimate the dimensionless retention times for hydrophobic interaction ($KD_{HI,p}$). Both relationships between retention time – physicochemical property are nonlinear; therefore piecewise linear approximations are used for their linearisation, as presented in Figure 1.

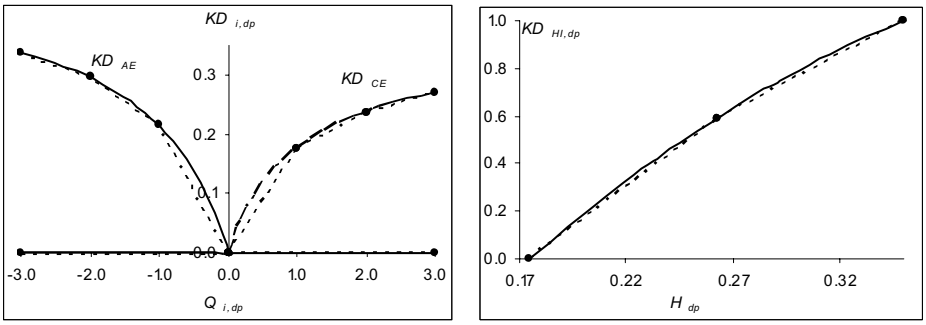

Figure 1. Piecewise linear approximations of retention times for ion exchange chromatography (AE: anion exchange; CE: cation exchange) and for hydrophobic interaction (HI)

3.3 Deviation factors

Deviation factors DF_{ip} indicate the distance between the protein product's chromatographic peak and a contaminant's chromatographic peak. They are defined as the difference between the dimensionless retention times of the product and each contaminant *p* for each particular chromatographic step *i*.

$$DF_{ip}^{+} - DF_{ip}^{-} = KD_{i,dp} - KD_{ip} \qquad \forall i, p \neq dp \qquad (2)$$

$$DF_{ip}^{+} \leq M \cdot x_{ip} \qquad \forall i, p \neq dp \qquad (3)$$

$$DF_{ip}^{-} \leq M \cdot (1 - x_{ip}) \qquad \forall i, p \neq dp \qquad (4)$$

$$DF_{ip} = DF_{ip}^{+} + DF_{ip}^{-} \qquad \forall i, p \neq dp \qquad (5)$$

where DF_{ip}^{+}, DF_{ip}^{-} are auxiliary positive continuous variables; and x_{ip} is a binary variable equal to 1 if DF_{ip} is positive and 0 otherwise.

3.4 Concentration factors

Deviation factors are used to calculate the concentration factors CF_{ip}, which represent the ratio of the mass of contaminant p after chromatographic step i to the mass of contaminant p before step i. The relationship between deviation factors and concentration factors is also nonlinear (Vasquez-Alvarez *et al.*, 2001), so another piecewise linear approximation is needed, as presented in Figure 2.

3.5 Purity constraint

The mass $m_{I,dp}$ of protein product dp after the last chromatographic step I must meet a specified purity level, SP. Since it is assumed that the separation is performed without product loss, the final product mass $m_{I,dp}$ is constant and equal to the initial mass $m_{0,dp}$.

$$m_{I,dp} \geq SP \cdot \sum_{p} m_{I,p} \Rightarrow (1-SP) \cdot m_{0,dp} \geq SP \cdot \sum_{p \neq dp} m_{I,p} \tag{6}$$

The remaining mass $m_{I,p}$ of each contaminant protein p after the final technique I is calculated from the initial mass $m_{0,p}$ by:

$$m_{I,p} = m_{0,p} \cdot \prod_{i=1}^{I} \overline{CF}_{ip} \qquad \forall\ p \neq dp \tag{7}$$

where $\overline{CF}_{ip} = CF_{ip}, \quad \text{if} \quad w_i = 1$
$\overline{CF}_{ip} = 1, \quad \text{if} \quad w_i = 0$ $\qquad \forall\ i, p \neq dp$

Binary variable w_i is used to indicate the selection of technique i. Variables $\overline{CF}_{ip}$ can be expressed as an exponential function of concentration factors CF_{ip} and decision variables w_i:

$$\overline{CF}_{ip} = e^{(\ln CF_{ip}) \cdot w_i} \qquad \forall\ i, p \neq dp \tag{8}$$

Therefore, using equations (7) and (8), purity constraint (6) can now be rewritten as:

$$(1-SP) \cdot m_{0,dp} \geq SP \cdot \sum_{p \neq dp} m_{0,p} \cdot e^{\sum_{i} \xi_{ip}} \tag{9}$$

where $\xi_{ip} \equiv (\ln CF_{ip}) \cdot w_i$. Constraint (9) incorporates the nonlinear factor $e^{\sum_{i} \xi_{ip}}$, which can also be linearised with a piecewise linear approximation (Figure 2).

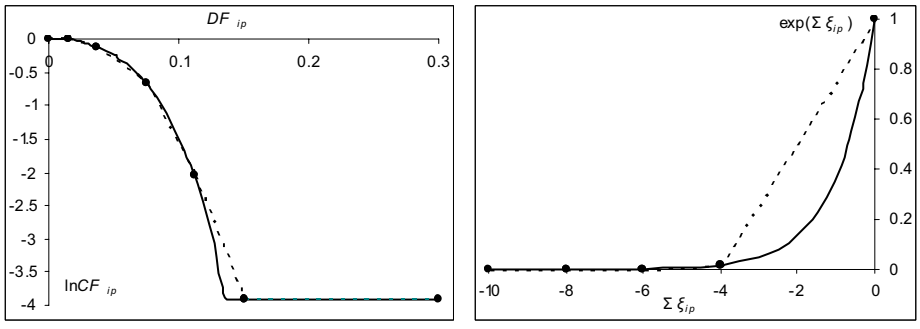

Figure 2. Piecewise linear approximations for concentration factors CF_{ip} and for ξ_{ip}

3.6 Solution approach

The overall problem is formulated as an MILP model, in order to identify the chromatographic techniques and the shortest amino acid sequence that can produce the optimal flowsheet of the purification process. The objective is to minimise the total number of selected chromatographic steps i in the purification process and, using a penalty parameter c, to force the model to select the minimum number of amino acids n_k in the tag.

$$minimise \sum_i w_i + c \cdot \sum_k n_k \tag{10}$$

First the MILP is solved without the use of a peptide tag for the purification of protein *dp*. Then the MILP is solved again with a tag fused to the product protein; but this time the candidate chromatographic steps i are chosen only among those selected in the first stage of the solution.

4. Computational Results

Solutions were obtained with the GAMS software (Brooke *et al*., 1998), using the CPLEX 6.5 solver. All computational experiments were performed on an IBM RS6000 workstation. The methodology was tested with a four-protein mixture: thaumatin (*dp*), conalbumin (*p*1), chymotripsinogen A (*p*2) and ovalbumin (*p*3). The physicochemical properties of the mixture are presented in Table 1.

Table 1. Physicochemical properties of protein mixture.

Protein	$m_{0,p}$	MW_p	H_p	$Q_{ip} \times 10^{-17}$ (*C/molecule*)				
	(mg/mL)	(Da)		pH 4.0	pH 5.0	pH6.0	pH7.0	pH8.0
Dp	2	22200	0.27	1.60	1.57	1.64	1.55	0.75
p1	2	77000	0.23	0.93	0.33	-0.12	-0.34	-0.50
p2	2	23600	0.31	2.15	1.46	1.17	0.78	0.38
p3	2	43800	0.28	1.16	-0.63	-1.36	-1.82	-1.95

A maximum number of 6 amino acids per tag is imposed on the number of amino acids that can be present in the peptide tag, so as to avoid interference with the tertiary structure of the protein product, as well as the possibility of formation of an alpha-helix or a beta-sheet from the tag itself. At the same time, hydrophobic amino acids should be balanced by polar residues so that the tag is soluble and does not bury itself within the protein. This possibility is avoided by imposing an upper bound to the number of hydrophobic residues that may be included in the peptide tag.

The purity level required for the desired product (*dp*) is 98%. There are 11 available chromatographic steps: anion exchange chromatography (AE) at pH 4, pH 5, pH 6, pH 7, pH 8, cation exchange chromatography (CE) at pH 4, pH 5, pH 6, pH 7, pH 8 and hydrophobic interaction (HI). From these, CE pH 6, CE pH 7, CE pH 8 and HI are needed for the purification without the use of a peptide tag fused to protein *dp*, which achieves a product purity of 98.1%. The solution is significantly improved with a tag of 3 lysine residues; a purity of 98.1% can be achieved with only three separation steps: CE pH 7, CE pH 8 and HI. The results are illustrated in Figure 3.

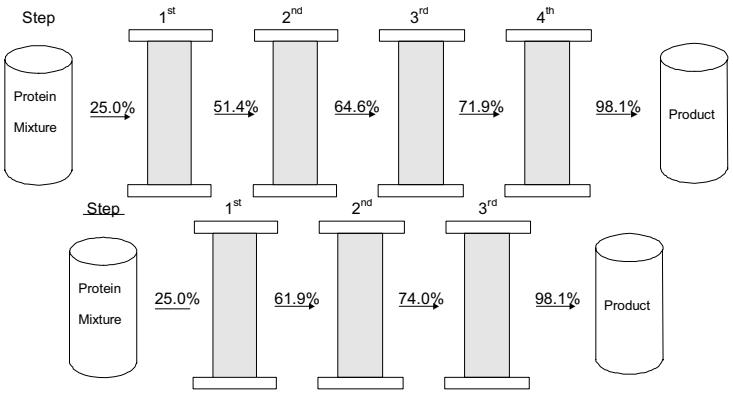

Figure 3. Optimal result for protein mixture with no tag and with a tag of 3 lysines

The MILP solution is almost identical to the one provided by the MINLP model presented in Simeonidis *et al*. (2004), which selected the same 3 chromatographic steps and a tag with lysines only. The selection of a peptide tag that only contains lysine amino acids implies that the increase of the product charge benefits the purification and that a hydrophobicity increase would be detrimental. Even though there are amino acids with a stronger effect on charge than lysine, they would increase hydrophobicity as well, which remains unchanged when lysine is used. Indeed, when the model is tested with a pre-fixed tag containing any amino acids that would increase hydrophobicity, a purity of 98% is not achievable.

5. Concluding remarks

An optimisation framework for the simultaneous selection of optimal peptide tags and the synthesis of chromatographic steps for the purification of protein mixtures in downstream protein processing has been presented. The framework was formulated as an MILP mathematical model, developed from a previous MINLP model (Simeonidis *et al.*, 2004) through piecewise linear approximations of nonlinear functions. The methodology was validated through its application on an example protein mixture involving 3 contaminants and a set of 11 candidate chromatographic steps. Results were indicative of the benefits of peptide tags in purification processes and provide a useful guideline for both downstream process synthesis and optimal tag design.

References

Brooke, A., D. Kendrick, A. Meeraus, and R. Raman, 1998, GAMS: A User's Guide. GAMS Development Corporation, Washington.

Lienqueo, M.E., E.W. Leser and J.A. Asenjo, 1996, Comput. Chem. Eng. 20, S189.

Lienqueo, M.E., A. Mahn and J.A. Asenjo, 2002, J. Chromatogr. A 978, 71.

Mosher, R.A., P. Gebauer and W. Thormann, 1993, J. Chromatogr. 638, 155.

Simeonidis, E. J.M. Pinto and L.G. Papageorgiou, 2004, Proc. ESCAPE-14, Portugal, 289.

Steffens, M.A., E.S. Fraga and I.D.L. Bogle, 2000, Comput. Chem. Eng. 24, 717.

Vasquez-Alvarez, E., M.E. Lienqueo and J.M. Pinto, 2001, Biotechnol. Progr. 17, 685.

Vasquez-Alvarez, E. and J.M. Pinto, 2004, J. Biotechnol. 110, 295.