

# Optimal Peptide Tag Design and Synthesis of Downstream Protein Processing

Evangelos Simeonidis<sup>1</sup>, Jose M. Pinto<sup>2</sup> and Lazaros G. Papageorgiou<sup>1,\*</sup>

<sup>1</sup> Centre for Process Systems Engineering, Department of Chemical Engineering, UCL (University College London), Torrington Place, London WC1E 7JE, U.K.

<sup>2</sup> Department of Chemical and Biological Sciences and Engineering, Polytechnic University, Six Metrotech Center, Brooklyn NY 11201, U.S.A.

## Abstract

In biochemical production plants, downstream protein processing can significantly be improved with the use of peptide purification tags; comparatively short sequences of amino acids fused onto the product protein, capable of simplifying the purification flowsheet. The objective of this work is to develop a framework that integrates the selection of optimal peptide tags with the synthesis of downstream protein processes. The methodology is validated by an illustrative example based on experimental data.

**Keywords:** protein purification processes, peptide tags, mixed integer optimisation

## 1. Introduction

Protein purification is regarded as the most complex and expensive stage of a biochemical plant, thus necessitating the development of systematic methods for the synthesis of downstream protein processing. Early approaches made use of expert knowledge systems for selecting operations in downstream protein processing (Lienqueo et al., 1996). Recently, methodologies based on optimisation techniques have been presented. Steffens et al. (2000a) incorporated heuristics based on physical property information in an implicit enumeration algorithm to solve the synthesis problem, thus reducing the search space. Vasquez-Alvarez et al. (2001) 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. The latter formulation was further improved exploiting the advantages of convex hull representations (Vasquez-Alvarez and Pinto, 2001).

The aim of this paper is to present a framework based on mixed integer non-linear programming (MINLP), which considers simultaneously the design of optimal peptide tags for each particular product and the synthesis of downstream protein processing. A purification tag is a peptide that can be fused genetically onto the product protein in order to modify its physical properties, in a way that will eventually enhance its separation from other contaminant proteins, thereby reducing the number of required

---

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

purification steps. It has recently been demonstrated that considerable improvements in yields and costs of downstream purification processes can be achieved with the use of such tags (Steffens et al., 2000b).

Although previous research has investigated the use of tags in protein purification (Sassenfeld, 1990), it was mainly focused on specific tags, which have advantages in certain situations, but are not necessarily optimal. The framework presented herein, utilises physicochemical properties data to specify the amino acid composition of the shortest and most advantageous tag, and concurrently select operations among a set of candidate chromatographic techniques that must achieve a specified purity level, while optimising a suitable performance criterion. Finally, the proposed methodology is validated with an illustrative example.

## 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 ( $p: 1, \dots, P$ ) with known physicochemical properties;
- a set of available chromatographic techniques ( $i: 1, \dots, I$ ) each performing a separation by exploiting a specific physicochemical property (charge or hydrophobicity);
- the properties of the twenty known amino acids ( $k: 1, \dots, 20$ ); and
- a specification for the desired product ( $dp$ ), in terms of a minimum purity level.

*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 MINLP representation extends an earlier MILP formulation (Vasquez-Alvarez and Pinto, 2001) designed for the synthesis of purification bioprocesses, so as to consider the optimal design of purification tags.

*Process synthesis constraints*

The convex hull representation applied for contaminant separation is based on previously developed MILP formulation (Vasquez-Alvarez and Pinto, 2001) for the selection of appropriate chromatographic steps and the indication of the remaining amount of protein after each step. Also, the mass of the product protein after the last chromatographic step is forced to meet a specified purity level.

#### *Physicochemical property constraints*

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

$$Q_{dp} = \hat{Q}_{dp} + \sum_{k \in G_b} \frac{n_k}{\frac{K_b}{[H^+]} + 1} - \sum_{k \in G_a} \frac{n_k}{\frac{K_a}{[H^+]} + 1} \quad (1)$$

where  $G_a$ ,  $G_b$  are the acidic and basic amino acid groups respectively;  $K_a$ ,  $K_b$  are the acidic and basic ionisation constants respectively;  $n_k$  is the integer number of amino acids  $k$  in the tag and  $\hat{Q}_{dp}$  is the initial product charge. It should be added that the method is not always accurate, especially for large molecules with complex structures, but is still reliable as an indication of how much and in what way the addition of a peptide tag will modify the net charge of the desired product.

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

#### *Dimensionless retention time constraints*

The retention time ( $KD_{ip}$ ) is defined as a function of a physicochemical property ( $P_p$ ).

$$KD_{ip} = f(P_p) \quad \forall i, p \quad (2)$$

For ion exchange chromatography, retention times for the tagged protein are estimated based on approximations of the chromatographs by isosceles triangles and on physicochemical property data for the product and contaminants (Lienqueo, 1999).

The methodology presented by Lienqueo et al. (2002) is used to estimate the dimensionless retention times for hydrophobic interaction ( $KD_{HL,p}$ ). As shown with equation 2, retention time is a function of hydrophobicity; the function in this case is a quadratic equation.

#### *Concentration factor constraints*

Concentration factors for the various chromatographic steps ( $CF_{ip}$ ) are calculated with a sigmoid function, which provides an accurate approximation of the previously used relationships describing the chromatographic peaks of the desired product and contaminants (Vasquez-Alvarez et al., 2001).

$$CF_{ip} = \frac{\alpha_i}{\beta_i + \gamma_i \cdot e^{(\delta_i \cdot DF_p + \varepsilon_i)}} + \zeta_i \quad \forall i, p \quad (3)$$

where,  $\alpha_i$ ,  $\beta_i$ ,  $\gamma_i$ ,  $\delta_i$ ,  $\varepsilon_i$ ,  $\zeta_i$  are suitable parameters.

The concentration factor ( $CF_{ip}$ ) is a function of the deviation factor ( $DF_{ip}$ ), which indicates the distance between the desired product's chromatographic peak and a contaminant's chromatographic peak (Vasquez-Alvarez et al., 2001). Deviation factors are defined as the difference between the retention times of product and contaminant for each particular chromatographic step.

$$DF_{ip} = |KD_{i,dp} - KD_{ip}| \quad \forall i, p \quad (4)$$

#### *Logical constraints*

An upper bound, typically between 5 and 10, is imposed on the number of amino acids in each tag. It should be noted that a smaller tag is amenable to molecular synthesis methods and has several practical advantages, including minimal effect on the protein structure, easier cleavage and simpler fusion onto the product protein.

A second constraint imposed on the amino acid composition of the tag is that only half of the amino acids are permitted to have a hydrophobic nature. Hydrophobic amino acids should be balanced by polar residues so that the tag is soluble and will not bury itself within the product protein.

#### *Solution approach*

Typical performance criteria to be optimised are the number of purification steps and/or the size of the tag. The overall problem is formulated as an MINLP model and a two-stage solution procedure is proposed, in order to identify the shortest amino acid sequence that can produce the optimal flowsheet of the purification process.

Stage 1: Designing flowsheets with fewer units can significantly reduce costs. The overall objective is to minimise the total number of selected chromatographic steps in the purification process subject to the described constraints (problem P1). A tag that will modify the properties of the product protein in the most beneficial way is selected.

Stage 2: The objective here (problem P2) is to minimise the number of amino acids in the tag subject to the same constraints, plus an additional constraint that fixes the number of steps as identified in stage 1. The smallest tag that can produce the optimal flowsheet is finally determined.

## **4. Computational Results**

Solutions were obtained with the network-enabled problem-solving environment NEOS Server 4.0 (Czyzyk et al., 1998) using the SBB solver for the solution of the MINLP models. The methodology was tested with a mixture of four proteins: thaumatin ( $dp$ ), conalbumin ( $p1$ ), chymotrypsinogen A ( $p2$ ) and ovalbumin ( $p3$ ). The physicochemical properties of the mixture are presented in table 1. The purity level requirement for the desired product ( $dp$ ) is 98%. Overall, there are 11 candidate 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).

Table 1: Physicochemical properties of protein mixture

protein	$m_{0,p}$ (mg/mL)	$MW_p$ (Da)	$H_p$	$Q_{ip} \times 10^{-17}$ (C/molecule)				
				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
$p_1$	2	77000	0.23	0.93	0.33	-0.12	-0.34	-0.50
$p_2$	2	23600	0.31	2.15	1.46	1.17	0.78	0.38
$p_3$	2	43800	0.28	1.16	-0.63	-1.36	-1.82	-1.95

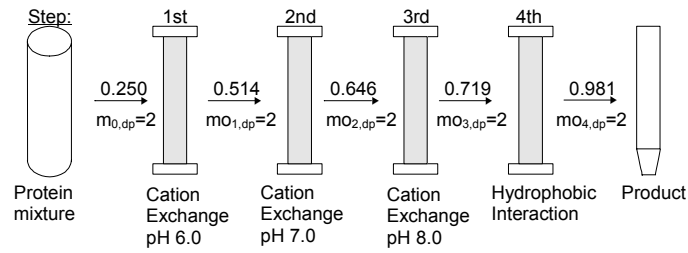


Figure 1: Optimal result for protein mixture with no tag

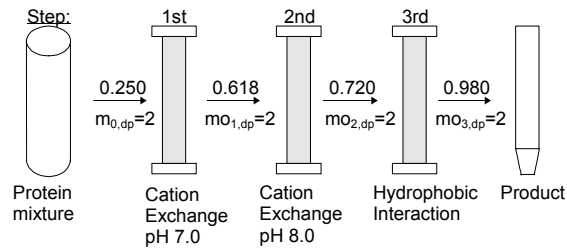


Figure 2: Optimal result for protein mixture with a tag of 2 lysines

In order to acquire a point of reference, the example was first solved without a tag fused to the product protein, *i.e.* using only the formulation of problem P1, with an upper bound of zero imposed on the number of amino acids in the tag. The optimal solution is presented in figure 1. The model was able to identify a solution that achieves a purity of 98.1% for the desired product. Four chromatographic steps are needed: CE pH 6, CE pH 7, CE pH 8 and HI.

An improved solution is suggested when using a peptide tag. The minimum number of steps is identified with an upper bound of 6 amino acids per tag (stage 1), then the number of purification steps is fixed and the model is solved again using the formulation of problem P2 (stage 2). With a tag of 2 lysine residues, a purity of 98.0% can be achieved and only three separation steps are needed: CE pH 7, CE pH 8 and HI. The results are presented in figure 2.

The selection of a tag that only contains lysines suggests that the recommended strategy is to increase the product charge. Other amino acids would have a stronger effect on charge than lysine, but they would also increase hydrophobicity, which remains unchanged when lysine is used. This strongly implies that a hydrophobicity increase is not beneficial in this case; it might affect the purity of the final product or raise the number of steps. This hypothesis can be tested by running the model with a pre-fixed tag, *e.g.* a tag containing phenylalanines, which would considerably increase hydrophobicity. Indeed, such experiments show that a purity of 98% is not achievable with a tag that includes hydrophobic residues.

## 5. Conclusions

An MINLP model for the simultaneous selection of optimal peptide tags and the synthesis of downstream purification for protein mixtures has been presented. The results are indicative of the benefits of peptide tags in purification processes and provide a useful guideline for both downstream process synthesis and optimal tag design.

Current work focuses on testing the mathematical framework with larger examples and investigating alternative solution strategies. Different approaches for the modelling of the purification are also considered, including sequencing of purification steps, inclusion of product loss, and application of alternative performance criteria (*e.g.* economical).

## Acknowledgments

The authors thank María E. Lienqueo for kindly supplying necessary data and for fruitful discussion; also Paul A. Dalby and Sophia Tsoka for many useful discussions. The authors acknowledge financial support from VITAE (Grant B-11487/10B006) and the Royal Academy of Engineering (Grant IJB/LB/ITG 03-330). E.S. was financially supported by the Engineering and Physical Sciences Research Council (Award No. 00319001) and the Centre for Process Systems Engineering.

## References

- Czyzyk, J., Mesnier, M.P. and Moré J.J., 1998, IEEE Comput. Sci. Eng. 5, 68.
- Lienqueo, M.E., 1999, PhD Thesis, University of Chile, Santiago, Chile (in Spanish).
- Lienqueo, M.E., Leser, E.W. and Asenjo, J.A., 1996, Comput. Chem. Eng. 20, S189.
- Lienqueo, M.E., Mahn, A. and Asenjo, J.A., 2002, J. Chromatogr. A 978, 71.
- Mosher, R.A., Gebauer, P. and Thormann, W., 1993, J. Chromatogr. 638, 155.
- Sassenfeld, H.M., 1990, Trends Biotechnol. 8, 88.
- Steffens, M.A., Fraga, E.S. and Bogle, I.D.L., 2000a, Biotechnol. Bioeng. 68, 218.
- Steffens, M.A., Fraga, E.S. and Bogle, I.D.L., 2000b, Comput. Chem. Eng. 24, 717.
- Vasquez-Alvarez, E., Lienqueo, M.E. and Pinto, J.M., 2001, Biotechnol. Progr. 17, 685.
- Vasquez-Alvarez, E. and Pinto, J.M., 2001, Proc. ESCAPE-11, Denmark, 579.