Efficient MILP Formulations for the Simultaneous Optimal Peptide Tag Design and Downstream Processing Synthesis

João M. Natali and José M. Pinto
Othmer-Jacobs Dept. of Chemical and Biological Engineering, Polytechnic University, 6 MetroTech Center, Brooklyn, NY 11201

Lazaros G. Papageorgiou
Centre for Process Systems Engineering, Dept. of Chemical Engineering, University College London, Torrington Place, London WC1E 7JE, U.K.

DOI 10.1002/aic.11913
Published online July 10, 2009 in Wiley InterScience (www.interscience.wiley.com).

Novel and efficient linear formulations are developed for the problem of simultaneously performing an optimal synthesis of chromatographic protein purification processes, and the concomitant selection of peptide purification tags, that result in a maximal process improvement. To this end, two formulations are developed for the solution of this problem: (1) a model that minimizes both the number of chromatographic steps in the final purification process flow sheet and the composition of the tag, by use of weighted objectives, while satisfying minimal purity requirements for the final product; and (2) a model that attempts to find the maximal attainable purity under constraints on the maximum number of separation techniques and tag size. Both models are linearized using a previously developed strategy for obtaining optimal piecewise linear approximations of nonlinear functions. Proposed are models to two case studies based on protein mixtures with different numbers of proteins. Results show that the models are capable of solving to optimality all the implemented cases with computational time requirements of under 1 s, on average. The results obtained are further compared with previous nonlinear and linear models attempting to solve the same problem, and, thus, show that the approach represents significant gains in robustness and efficiency.
 AIChE J, 55: 2303–2317, 2009
Keywords: bioprocess synthesis, protein purification, piecewise linear approximation, mixed integer linear optimization, chromatographic techniques, purification tags

Introduction

Preparative purification of protein products are major determinants of the fixed and operational costs associated with many bioprocesses.[1] This major component of downstream processing is also pivotal to the definition of the quality of the desired product, commonly determined by purity specifications. The crucial role performed by downstream processing in the modern biotechnology industry makes its efficient design a fundamental part of a successful process. Among many possible techniques employed for the purification of proteins in complex mixtures, liquid chromatography are of major interest in many industrial applications. These purification processes usually require several chromatographic steps to achieve a final product within

Correspondence concerning this article should be addressed to J. M. Pinto at his current affiliation, Praxair, 39 Old Ridgebury Road, Danbury, CT 06810; e-mail: Jose_M_Pinto@Praxair.com

defined specifications, resulting in complex process flow sheets.

For that reason, many authors have attempted to systematize efforts to efficiently design preparative purification processes. At the level of a single operating unit, Nagrath et al.[2] developed a system for balancing confronting chromatography design objectives. Steffens et al.[3] initially developed a synthesis approach based on physicochemical property data and managing multiple design objectives with a heuristic implicit enumeration algorithm. Later, the same group incorporated in their synthesis method the use of purification tags, small sequences of amino acids which are attached to a target protein product and facilitate the purification at subsequent stages.[4] Concomitantly, an expert system based on the division of separation processes into recovery and purification parts was developed.[5,6] These authors then relied on heuristic rules to obtain insights into the development of large-scale downstream bioprocesses.

Subsequently, Vasquez-Alvarez et al.[7] developed mathematical models based on mixed-integer linear programming (MILP) for the synthesis of protein purification processes, by the optimal selection and sequencing of purification steps. The models made use of physicochemical properties of all components in a protein mixture, and of a set of available chromatographic steps to minimize the number of purification stages for a specified purity level of the product, and to maximize product purity. Later, the same group improved on the previous formulation by generating MILP formulations that incorporated product losses along of the process in order to evaluate the trade-off between product purity and recovery.[8]

Recently, Simeonidis et al.[9] developed nonlinear models based on mixed-integer optimization that simultaneously perform the selection of peptide purification tags and the synthesis of protein purification processes. The resulting MINLP was successfully applied to a small-scale system composed of a mixture of four proteins, and to a larger-scale system with 13 proteins. However, the approach suffered from drawbacks inherent to nonlinear formulations such as the lack of possibility to guarantee the optimality of the solutions, and a general inefficiency of current nonlinear integer solvers when compared to their linear counterparts. For that reason, the developed models could only be solved by a suboptimal two-step procedure in which an untagged solution is initially generated and used to reduce the search space for the solution of the complete model. This group later converted the previous framework into a simpler MILP model through piecewise linear approximations of the nonconvex, nonlinear functions.[10] However, the developed new linearized models were composed by simple linearizations of distinct parts from the original MINLP models and still suffered from some of the previous limitations, which included the necessity for the two-stage solution approach and the incapacity to obtain a solution to the large-scale 13-protein system.

The objective of this work is to develop a novel linear formulation based on the MINLP approach developed by Simeonidis et al.[9] that presents significant gains in efficiency, and is capable of a de facto simultaneous solution of the problems of separation process synthesis and tag optimization. The developed models incorporate optimal piecewise linear interpolation and approximation strategies developed by Natali and Pinto[11] for the use of the closest possible linearized representations of nonlinear functions, which adds to the accuracy of the solutions. Finally, we expand the previous formulation to include models that perform the maximization of the target proteins final purity by limiting the total number of chromatographic techniques to be used in the separation process flow sheet and the total number of amino acids to be added to the tag.

The remainder of this article is structured as follows. The next section describes the problem for the simultaneous synthesis of protein purification processes and selection of tag composition. Then, we present the mathematical formulation of the linearized models developed, followed by a detailed account of the piecewise linear formulations employed. Then, we present the definition of the systems in which the developed models are applied. Next, numerical results are presented and analyzed, and the computational performance of the proposed frameworks is evaluated. Finally, the main conclusions of this work are discussed.

Problem Description

The problem considered in this work consists of defining an optimal set of chromatographic steps that are capable of separating a target protein from a complex mixture to a predefined purity requirement, while simultaneously determining a minimal set of amino acids in a tag that, when genetically added to the target protein, would result in a gain in efficiency to the final separation flow sheet. Such a task is accomplished with efficient mixed integer linear formulations that remedy many of the drawbacks presented by the nonlinear strategy previously proposed by Simeonidis et al.[9]

The description of the problem relies on the definitions of a set of proteins $P = \{1, \ldots, n_P\}$, whose physicochemical properties of interest are all known, and a set of chromatographic techniques $I = \{1, \ldots, n_I\}$, that are able to perform the separations by exploiting individual characteristics of the proteins in the mixture. One of the proteins in the mixture is defined as the target protein which is to be ultimately separated from the remaining proteins, and is named $dp \in P$. Furthermore, the physicochemical characteristics of the set of 20 standard proteinogenic amino acids $K = \{Ala, Arg, \ldots\}$, are used to account for the influence of the tag in the separation processes. Figure 1 is a representation of the superstructure of the purification process. An initial mixture containing the tagged protein and contaminants is passed through a selection of chromatographic techniques to yield the purified desired product. The problem is composed of the optimal selections of both the tag composition and the optimal sequence of purification steps.

To formulate the mathematical model used to solve the proposed problem, we rely on the same simplifying assumptions used by Simeonidis et al.[9] Namely, we assume initially that the tags are small in comparison to the target protein, which results in that the original properties of *dp* are only marginally influenced by the addition of the tag. The main consequence of this assumption is that it allows for the expression of the final physicochemical properties of the target protein as linear functions of the properties of the tag. These relationships are defined in the next section.

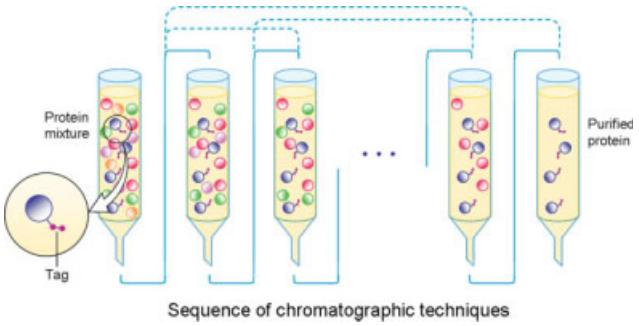

Figure 1. Superstructure of the purification process with tagged target protein.

Dotted lines represent the bypass of chromatographic steps not selected in the optimal sequence. [Color figure can be viewed in the online issue, which is available at www.interscience.wiley.com.]

Furthermore, the amino acids that comprise the tag are assumed to fold outwards in the protein and to have a fully exposed surface. The possibility of the tag to be folded into the core of the protein is avoided by a proper choice of the placement of the amino acids' code in the engineered gene sequence, which is assumed a priori in our model, and by imposing an upper bound to the number of hydrophobic residues that may be included in the tag.

We further assume that each chromatographic step is capable of recuperating the target protein in its entirety, resulting in no product loss for the overall process. Such assumption includes the inexistence of product loss in both the columns and in the necessary membrane steps for buffer exchange and protein concentration between chromatographic techniques. Note that due to this simplifying assumption, the order of the selected techniques in the model solution is immaterial. Moreover, it is assumed that the distributions of protein concentrations in the exit of chromatographic columns can be approximated by isosceles triangles.[7,12,13] It will be clear in the next section, however, that the choice of approximation strategy for the concentration peaks, although relevant to the quality of the results, is immaterial with respect to the structure of the presented models.

Mathematical Formulation

This section presents a description of the proposed MILP formulation that is based on and expands nonlinear models previously introduced in the literature.[9] The models are comprised of three parts that allow for the optimal definition of the minimal process purification flow sheet and tag composition. Initially, the model uses approximations to define relevant physicochemical properties of the target protein as functions of the number and type of amino acids in the selected tag. These properties are employed to the quantification of the efficiency of each chromatographic step for the separation of compounds in the mixture. Finally, a material balance is defined for the directed selection of the superstructure of all possible flow sheets that comply with desired specifications.

Two models are developed that differ in the objective to be optimized. Initially, we consider a model in which a minimum final purity for the target protein is defined a priori, and the objective of the optimization problem is to obtain a minimal set of techniques and tag composition that is able to attain the purity specification. Furthermore, we study the complementary problem of finding the maximal obtainable purity of the desired product when the available numbers of chromatographic techniques and amino acids that may be added to the tag are limited.

Peptide tags and chromatographic techniques

We initially define an integer decision variable y_k, $\forall k \in K$, which specifies the number of each of the 20 amino acids in the selected tag. The restrictions on the size of the tags are expressed as follows

$$\sum_{k \in K} y_k \leq Naa \tag{1}$$

Specifically, the restriction of the hydrophobic nature of the tag is imposed by the following constraint

$$\sum_{k \in HA} y_k \leq 0.5 \cdot \sum_{k \in K} y_k \tag{2}$$

where $HA \subset K$ is the subset of hydrophobic amino acids.

Equation 1 determines that the maximum number of amino acids in the tag may not exceed *Naa*. Equation 2 specifies that the number of amino acids with a predominantly hydrophobic character is limited to half of the total number of amino acids in the tag.

Furthermore, we define a binary decision variable w_i, $\forall i \in I$, that is equal to 1 if and only if technique i is selected as part of the separation flow sheet. Two types of chromatographic techniques are considered in this study, ion-exchange $(IE \subseteq I)$, and hydrophobic interaction $(HI \subseteq I)$ chromatography. It was previously determined experimentally that the former technique is mainly influenced by the overall charge of the proteins in the mixture, and the latter by the proteins' hydrophobicity.[14] Additionally, the set of ion exchange techniques can be further divided into anion exchange and cation exchange chromatography $(AE, CE \subseteq IE)$, depending on the charge of the filling in the column.

The optimal definition of the aforementioned decision variables constitutes the solution of the proposed problems, thus, it is further necessary to define the relationship between these variables and the concentration of the target protein and contaminants during the separation process, in order to satisfy purity constraints and define optimal objectives.

Physicochemical properties constraints

The introduction of tags into the target protein has the effect of altering its physicochemical properties with the goal of facilitating the separation process. In order to quantify the efficiency improvement in the chromatographic techniques due to the tag, it is initially necessary to account for the properties of the target protein that are manipulated.

Hence, it is necessary to correlate the target protein's overall charge and hydrophobicity to the composition of the tags employed. Mosher et al.[15] proposed a linear relationship to determine the net charge of a protein based on its amino

acid composition. This relationship has been modified[9] to account only for the influence of the added tag to the original protein, and it is defined as

$$Q_{i,dp} = \hat{Q}_{i,dp} + \sum_{k \in BA} \frac{y_k}{\frac{Ka_k}{[H^+]_i} + 1} - \sum_{k \in AA} \frac{y_k}{\frac{[H^+]_i}{Ka_k} + 1} \quad \forall i \in IE \quad (3)$$

where $Q_{i,dp}$ is the charge of the tagged protein, $\hat{Q}_{i,dp}$ is the original charge of the desired protein, $BA \subset K$ is the set of basic amino acids, $AA \subset K$ is the set of acidic amino acids, Ka_k is the ionization constant of amino acid k, and $[H^+]_i$ is the hydronium concentration in each ion-exchange technique.

Similarly, the tagged protein hydrophobicity can be calculated linearly from the untagged protein using the following relationship[16]

$$H_{dp} = \hat{H}_{dp} + \sum_{k \in K} h_k \cdot \frac{s_k \cdot y_k}{\hat{S}_{dp}} \quad (4)$$

In Eq. 4, H_{dp} denotes the hydrophobicity of the tagged protein, whereas $\hat{H}_{dp}$ is its original hydrophobicity; h_k is the hydrophobicity assigned to each amino acid k, s_k is the total exposed area of an amino acid, and $\hat{S}_{dp}$ and is the exposed surface area of the original protein. The linearized relationship in Eq. 4 is the one developed by Simeonidis et al.[10] and differs from previous formulations[9] in that the total exposed surface of the tagged protein is now assumed to be approximately equal to the original protein's surface, which is justified by the assumption that the tags are very small in comparison to the desired product.

Chromatographic separation model

The model for the chromatographic techniques used in this work is based on the works previously published in the literature.[16,7] Initially, we define the dimensionless retention time of protein i in chromatographic technique p, $KD_{i,p}$, which was previously experimentally determined,[16] for each type of chromatographic column. The authors observed that the dimensionless retention times for techniques in IE were only a function of the charge densities of the proteins $\left(Q_{i,p}/MW_p\right)$ for the conditions of operation considered. Two different correlations were obtained for techniques in sets AE and CE, defined as follows:

- Anion Exchange Chromatography

$$\left.\begin{array}{ll} KD_{i,p} = \frac{8826 \cdot \left|Q_{i,p}/MW_p\right|}{1+18845 \cdot \left|Q_{i,p}/MW_p\right|} & \text{if} \quad Q_{i,p} < 0 \\ KD_{i,p} = 0 & \text{if} \quad Q_{i,p} \geq 0 \end{array}\right\} \forall i \in AE, \forall p \in P \quad (5)$$

- Cation Exchange Chromatography

$$\left.\begin{array}{ll} KD_{i,p} = 0 & \text{if} \quad Q_{i,p} \leq 0 \\ KD_{i,p} = \frac{7424 \cdot \left|Q_{i,p}/MW_p\right|}{1+20231 \cdot \left|Q_{i,p}/MW_p\right|} & \text{if} \quad Q_{i,p} > 0 \end{array}\right\} \forall i \in CE, \forall p \in P \quad (6)$$

For hydrophobic interaction chromatography, the dimensionless retention time was experimentally fitted to a second-order polynomial to yield

$$KD_{HI,p} = -12.14 \cdot H_{dp}^2 + 12.07 \cdot H_{dp} - 1.74 \qquad \forall p \in P \quad (7)$$

The dimensionless retention time of a protein carries the information of how much longer, on average, it takes to pass through the column in comparison to the moving phase. To quantify the separation of different proteins in a mixture from the desired product, it is further necessary to account for both the shape of the peaks at the exit of the chromatographic column and the distance between these peaks.

The shape of the chromatographic peaks is approximated by isosceles triangles and the breadth of these peaks, defined as the width of the base of the triangles, was experimentally determined to be a function of the chromatographic column only.[16] This width is, thus, represented by σ_i, $i \in I$, and the obtained values are $\sigma_i = 0.15$, $i \in IE$, and $\sigma_{HI} = 0.22$. The distance between the peaks of every contaminant protein and the target protein are defined by the parameter $DF_{i,p}$, such that

$$DF_{i,p} = \left|KD_{i,dp} - KD_{i,p}\right| \quad \forall i \in I, \forall p \in P | p \neq dp \quad (8)$$

Finally the separation process can be quantified with the definition of the concentration factor $CF_{i,p}$, defined as the ratio between the mass of contaminant p before and after chromatographic technique i. The computation of only requires information contained in $DF_{i,p}$ and σ_i, and is based on the correlations defined by Vásquez-Alvarez et al.[13] to be

$$\left.\begin{array}{ll} CF_{i,p} = 1 & \text{if} \quad 0 \leq DF_{i,p} < \sigma_i/10 \\ CF_{i,p} = (1+\Delta)\frac{\sigma_i^2 - 2DF_{i,p}^2}{\sigma_i^2} & \text{if} \quad \sigma_i/10 \leq DF_{i,p} < \sigma_i/2 \\ CF_{i,p} = \max\left(\Delta,\ 2(1+\Delta)\frac{\left(\sigma_i - DF_{i,p}\right)^2}{\sigma_i^2}\right) & \text{if} \quad \sigma_i/2 \leq DF_{i,p} \end{array}\right\} \quad \forall i \in I, \forall p \in P | p \neq dp \quad (9)$$

where the parameter Δ accounts for a safety factor for the assumed approximations. Note that the aforementioned expression presents a small discontinuity in $DF_{i,p} = \sigma_i/10$. We later argue that this discontinuity is naturally treated by the linearization procedure described in the Appendix, and does not generate any impediment for the implementation of the models developed in this work.

With the definition of the concentration factors from proper choices of tag selection, it is possible to define the material balances around each contaminant protein for each chromatographic technique[7,10] in a convex hull formulation

$$\begin{array}{ll} m_{1,p} = CF_{1,p} m_{0,p} w_1 + m_{0,p}(1 - w_1) & \forall p \in P \\ m_{i,p} = CF_{i,p} m_{i-1,p}^1 + m_{i-1,p}^2 & \forall i \in I | i \geq 2, \forall p \in P \\ m_{i-1,p} = m_{i-1,p}^1 + m_{i-1,p}^2 & \forall i \in I | i \geq 2, \forall p \in P \\ 0 \leq m_{i-1,p}^1 \leq U_p w_i & \forall i \in I | i \geq 2, \forall p \in P \\ 0 \leq m_{i-1,p}^2 \leq U_p (1 - w_i) & \forall i \in I | i \geq 2, \forall p \in P \end{array} \quad (10)$$

where $m_{i,p}$ represents the mass of protein p after the chromatographic step i, $m_{0,p}$ denotes the initial mass of each protein in the mixture, $m_{i,p}^1$ and $m_{i,p}^2$ are disaggregated mass

variables corresponding to the situations in which chromatographic technique i is, respectively, selected and not selected, and U_p is a large constant.

Model Definition

In the beginning of this section, we propose the definition of two distinct models for the optimization of the tag composition and separation flow sheets in the studied problem. The first problem deals with the definition of minimal sets of chromatographic techniques and amino acids composing a tag that is capable of meeting a predefined minimal purity acceptable. The second model attempts to find the maximal purity that can be obtained with upper bounds in the number of techniques, and in the number of amino acids allowed to compose the tag. We shall refer to the first model as minimization of decision variables model (MDV), and to the second model as maximization of purity model (MP).

We start by considering the MDV model. The objective function of this model involves two independent goals: the minimization of the number of chromatographic techniques selected to compose the separation flow sheet; and the minimization of the number of amino acids selected to compose the tag added to the target protein. These two goals are weighted in the objective function, defined as

$$\min \; Z^{MDV} = \sum_{i \in I} w_i + c \cdot \sum_{k \in K} y_k \tag{11}$$

where c defines the weight balancing the two goals. Model MDV counterposes the minimization of the objective function defined by Eq. 11 with the requirement of a minimum purity attained by the optimal process. Using the definition of contaminant concentrations provided in the preceding section, the minimal purity constraint can be defined as

$$m_{n_I,dp} \geq SP_{dp} \cdot \sum_{p \in P} m_{n_I,p} \tag{12}$$

where SP_{dp} is the minimal purity requirement. Equation 12 can be further simplified with the use of the assumption that the mass of the target protein remains unchanged during the separation process, to yield

$$\left(1 - SP_{dp}\right) \cdot m_{0,dp} \geq SP_{dp} \cdot \sum_{\substack{p \in P \\ p \neq dp}} m_{n_I,p} \tag{13}$$

Model MP employs an objective function that seeks to maximize the purity of the desired product upon completion of the separation process. Such objective can be formulated using the same simplifications as in Eq. 13, as

$$\max \; Z'^{MP} = \frac{m_{0,dp}}{\sum\limits_{\substack{p \in P \\ p \neq dp}} m_{n_I,p} + m_{0,dp}} \tag{14}$$

To avoid the nonlinearity imposed by Eq.14, we reformulate the objective function by defining $Z^{MP} = 1/Z'^{MP}$, and converting the problem to a minimization

$$\min \; Z^{MP} = \frac{\sum\limits_{\substack{p \in P \\ p \neq dp}} m_{n_I,p} + m_{0,dp}}{m_{0,dp}} \tag{15}$$

It is further noted that the right-hand side of Eq. 14 has strictly positive values, thus, Eq. 15 does not present any singularity. In close analogy to model MDV, model MP counterposes the objective of maximization of the target protein's purity—or, precisely, the minimization of its inverse—with upper bounds on the number of chromatographic techniques that can be employed, and the number of amino acids that compose the tags. The maximum number of amino acids, and further constraints in their hydrophobic nature have already been defined in Eqs. 1 and 2, with the added consideration that model MP employs the parameter *Naa* as a tunable specification corresponding to different solution scenarios, whereas this parameter is considered a predefined physical limitation of the tag size in model MDV. Finally, the constraints regarding the maximum number of selected techniques is defined as

$$\sum_{i \in I} w_i \leq Ntech \tag{16}$$

where *Ntech* is a parameter that specifies the corresponding upper bound.

The following section completes the definition of the linear model by describing a procedure for the linearization for the material balances previously presented, followed by the definition of optimal piecewise linear approximations for the nonlinear terms that are still present in the model.

Linearization of Material Balances

The material balances formulation defined in the previous section makes use of nonlinear terms provided that the concentration factors are variable, and defined in accordance to the selection of tags. It has to be noted that the purpose of the material balances around contaminants is the calculation of the final purity of the desired product either as a lower bound constraint in the first model proposed or as the objective function of the second one. Thus, a previously proposed strategy[10] consists of a reformulation of the purity calculation. We will resort to a similar treatment of the purity calculation by considering that the assumption of no product loss renders the material balances around dp unnecessary. Furthermore, we can state that the final concentrations of all contaminants are given by the following relationships

$$m_{n_I,p} = m_{0,p} \cdot \prod_{i \in I} \overline{CF}_{i,p} \qquad \forall p \in P | p \neq dp \tag{17}$$

where

$$\begin{array}{lll} \overline{CF}_{i,p} = CF_{i,p} & \text{if} & w_i = 1 \\ \overline{CF}_{i,p} = 1 & \text{if} & w_i = 0 \end{array} \qquad \forall i \in I, \forall p \in P | p \neq dp \tag{18}$$

Variable $\overline{CF}_{i,p}$ corresponds to a variation of $CF_{i,p}$ that enforces the condition of no separation for the case in which a technique is not selected. $\overline{CF}_{i,p}$ can be expressed in exponential form as

$$\overline{CF}_{i,p} = e^{\left(\ln CF_{i,p}\right)\cdot w_i} \qquad \forall i \in I, \forall p \in P | p \neq dp \qquad (19)$$

Thus, the final concentrations of the contaminants after all chromatographic steps are defined as

$$m_{n_l,p} = m_{0,p} \cdot e^{\sum_{i \in I} \left(\ln CF_{i,p}\right)\cdot w_i} \qquad \forall p \in P | p \neq dp \qquad (20)$$

While this formulation does not avoid the use of nonlinear terms for the calculation of final masses of contaminants, it has the advantage that all nonlinearities are present in a single term in Eq. 20 that can be decomposed into simpler functions that can be linearly approximated. In the next section, we describe a procedure to represent these nonlinearities as piecewise linear functions and implement an approach previously derived by our group to optimally define these approximations.[11]

Piecewise Linear Approximations

In the previous section, it was specified that the only required set of variables for the definition of the overall separation obtained with a specific set of chromatographic steps, namely the final concentrations of contaminants after the flow sheet, can be calculated by Eq. 20, which depends on concentration factors and on the topology of the separation process. In this section, Eq. 20 is reformulated by piecewise linear approximations[17] that are capable of linearly relating the target protein's overall charge and hydrophobicity to the final mass of contaminants.

We initially note that Eqs. 5 to 9 nonlinearly relate variables $Q_{i,dp}$ and H_{dp}, which are linear functions of the selection of amino acids in the tag, to variables $CF_{i,p}$. More specifically, we define a new variable $\overline{\ln CF}_{i,p}$ such that

$$\left.\begin{array}{ll} \overline{\ln CF}_{i,p} = \left(\ln CF_{i,p}\right) \cdot w_i = f_{i,p}\left(Q_{i,dp}, w_i\right) & \forall i \in AE \cup CE \\ \overline{\ln CF}_{i,p} = \left(\ln CF_{i,p}\right) \cdot w_i = f_{i,p}\left(H_{dp}, w_i\right) & \forall i \in HI \end{array}\right\} \forall p \in P | p \neq dp \qquad (21)$$

where $f_{i,p}$ are nonlinear piecewise continuous and non-smooth functions that represent the composition of Eqs. 5 to 9 and the natural logarithm function, with the added discontinuity introduced by the multiplication by w_i.

The $f_{i,p}$ functions are further represented by piecewise linear expressions and modeled with the use of SOS2 type variables $\lambda_{i,p,j}$, where $j \in J_{i,p} = \left\{1, 2, \ldots, n^J_{i,p}\right\}$, the index on which the SOS2 constraints are imposed, is the set of knots in the piecewise approximations. Therefore, variable $\overline{\ln CF}_{i,p}$ can be linearly represented by the following set of constraints

$$\overline{\ln CF}_{i,p} = \sum_{j \in J} \alpha_{i,p,j} \lambda_{i,p,j} + sl_{i,p} \quad \forall i \in I, \forall p \in P | p \neq dp \qquad (22)$$

where $\alpha_{i,p,j}$ are parameters that correspond to the values of the approximating function in the set of knots. The value of $\overline{\ln CF}_{i,p}$ is further related to the charges and hydrophobicity of the tagged protein by the following equations

$$\left.\begin{array}{ll} Q_{i,dp} = \sum_{j \in J} \beta_{i,p,j} \lambda_{i,p,j} & \forall i \in AE \cup CE \\ H_{dp} = \sum_{j \in J} \beta_{i,p,j} \lambda_{i,p,j} & \forall i \in HI \end{array}\right\} \forall p \in P | p \neq dp \qquad (23)$$

where $\beta_{i,p,j}$ are the values of the abscissae in the set of knots. The piecewise linear approximations are completed with the imposition of the following constraint

$$\sum_{j \in J} \lambda_{i,p,j} = 1 \qquad \forall i \in I, \forall p \in P | p \neq dp \qquad (24)$$

Finally, $sl_{i,p}$ are slack variables that allow for the imposition that $\overline{\ln CF}_{i,p}$ is equal to zero—i.e., no separation occurs—when a specific technique is not selected, through the following constraints

$$\left.\begin{array}{l} 0 \leq sl_{i,p} \leq -\ln(\Delta) \cdot (1 - w_i) \\ \ln CF_{i,p} \geq \ln(\Delta) \cdot w_i \end{array}\right. \quad \forall i \in I, \forall p \in P | p \neq dp \qquad (25)$$

The calculation of the final concentrations of contaminants through Eq. 20 contains a further nonlinear term in the form of one simple exponentiation. This nonlinear term can be linearized in a similar way as the used above. We initially define a new variable $\xi_p = e^{\sum_{i \in I} \overline{\ln CF}_{i,p}}, \forall p \in P | p \neq dp$, and another SOS2 type variable $\rho_{p,j}, j \in J_q = \{1, 2, \ldots, n^{J_q}\}$. We can, thus, define the piecewise linearization of the exponential function as

$$\left.\begin{array}{l} \xi_p = \sum_{j \in Jq} \eta_j \rho_{p,j} \\ \sum_i \overline{\ln CF}_{i,p} = \sum_{j \in Jq} \gamma_j \rho_{p,j} \\ \sum_{j \in Jq} \rho_{p,j} = 1 \end{array}\right\} \quad \forall p \in P | p \neq dp \qquad (26)$$

where η_j and γ_j are the values of the ordinate and the abscissa values, respectively, of the exponential function in the knots.

Parameters α, β, η and γ define the piecewise linear approximations used. Therefore, the proper selection of these parameters dictates the proximity of the piecewise linear functions to the original nonlinear relationships. To guarantee the use of the best possible linear approximations with a limited number of segments, we have employed an approach developed which is able to find the optimal approximation to nonlinear functions using continuous piecewise linear approximations.[11] It is noted here that this optimal linearization procedure is capable of approximating discontinuous functions, such as the case for the computation of the concentration factors in Eq. 9, and still generate continuous piecewise linear functions. A summary of the procedure is provided in Appendix A.

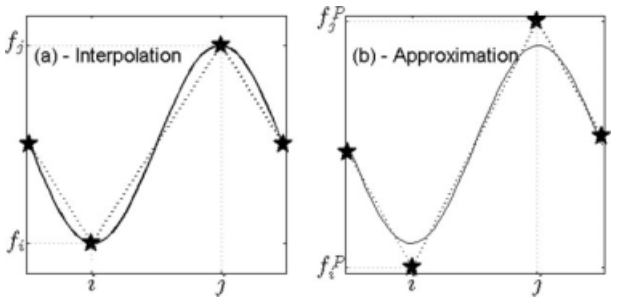

Figure 2. PWLI and PWLA approaches.

(a) Piecewise linear interpolation of a sine period, and (b) piecewise linear approximation of a sine period. Pentagrams correspond to the set of knots.

The optimal linearization procedure employed makes use of two distinct approaches for the definition of piecewise linear approximations. In the first approach, hereon named piecewise linear interpolation (PWLI), a set of points (the knots) within the graph of the nonlinear function is optimally chosen so that the resulting piecewise linear function is composed of all linear segments between these points. In the second approach, hereon named piecewise linear approximation (PWLA), the value of the linearized function does not need to coincide with the original nonlinear function in the knots, adding more degrees of freedom and allowing for a better approximation for the same number of knots utilized, at the expense of computational performance. Figure 2 exemplifies both approaches employed for the approximation of a sine period. In this figure, i and j represent internal knots for the interpolation and approximation procedures, whereas f_i and f_j correspond to the nonlinear function values in the knots and f_i^P and f_j^P correspond to the approximated function values in the knots. This article later presents an example of the PWLI and PWLA methods applied to the approximation of a function from the case studied developed in this work.

The next sections contain the definition of two case studies to which the procedure was applied and the results of the implementation. For both cases, results are generated using both PWLI and PWLA approximation strategies and their accuracy is compared.

System Definition

In the two previous sections, a general mathematical framework for the representation and solution of the problem of optimally separating a complex mixture of proteins, with concomitant selection of an optimal tag composition, has been fully defined. In this section, we initially summarize the models that compose our approach and we further define the relevant physical characteristics of the systems in which the proposed framework will be applied.

Model summary

Aiming clarity of representation of the developed approaches, we here summarize the models developed in the previous sections. We start by describing components that are common to both MDV and MP models.

The physicochemical properties of the target protein due to the addition of the tag are given by

$$Q_{i,dp} = \hat{Q}_{i,dp} + \sum_{k\in BA} \frac{y_k}{\frac{Ka_k}{[H^+]_i}+1} - \sum_{k\in AA} \frac{y_k}{\frac{[H^+]_i}{Ka_k}+1} \quad \forall i \in IE \tag{3}$$

$$H_{dp} = \hat{H}_{dp} + \sum_{k\in K} h_k \cdot \frac{s_k \cdot y_k}{\hat{S}_{dp}} \tag{4}$$

The set of variables $\overline{\ln CF}_{i,p}$ is defined by piecewise linear interpolations and approximations by the following set of constraints

$$\overline{\ln CF}_{i,p} = \sum_{j\in J} \alpha_{i,p,j}\lambda_{i,p,j} + sl_{i,p} \quad \forall i \in I, \forall p \in P | p \neq dp \tag{22}$$

$$\left.\begin{aligned} Q_{i,dp} &= \sum_{j\in J} \beta_{i,p,j}\lambda_{i,p,j} \quad \forall i \in AE \cup CE \\ H_{dp} &= \sum_{j\in J} \beta_{i,p,j}\lambda_{i,p,j} \quad \forall i \in HI \end{aligned}\right\} \forall p \in P | p \neq dp \tag{23}$$

$$\sum_{j\in J} \lambda_{i,p,j} = 1 \qquad \forall i \in I, \forall p \in P | p \neq dp \tag{24}$$

$$\left.\begin{aligned} 0 \leq sl_{i,p} &\leq -\ln(\Delta)\cdot(1-w_i) \\ \ln CF_{i,p} &\geq \ln(\Delta)\cdot w_i \end{aligned}\right\} \quad \forall i \in I, \forall p \in P | p \neq dp \tag{25}$$

The linearized exponentiation of the sum of the logarithms of the concentration factors is represented by

$$\left.\begin{aligned} \xi_p &= \sum_{j\in Jq} \eta_j \rho_{p,j} \\ \sum_i \overline{\ln CF}_{i,p} &= \sum_{j\in Jq} \gamma_j \rho_{p,j} \\ \sum_{j\in Jq} \rho_{p,j} &= 1 \end{aligned}\right\} \quad \forall p \in P | p \neq dp \tag{26}$$

The constraints in Eq. 26 allows for the calculation of the final mass of the contaminants

$$m_{n_I,p} = m_{0,p} \cdot \xi_p \qquad \forall p \in P | p \neq dp \tag{27}$$

Furthermore, the limits in the size and composition of the tags and the process flow sheets are determined by the following constraints

$$\sum_{k\in K} y_k \leq Naa \tag{1}$$

$$\sum_{k\in HA} y_k \leq 0.5 \cdot \sum_{k\in K} y_k \tag{2}$$

$$\sum_{i\in I} w_i \leq Ntech \tag{16}$$

It is important to note that the size constraints Eqs. 1 and 16 are employed as design limitations in model MDV, whereas in model MP they specify direct counterpoints to the purity maximization in the objective.

Table 1. Physicochemical Properties of Amino Acids

Amino Acid	h_k	s_k	pK	Properties
Ala	0.391	115	–	–
Arg	0.202	225	12.5	Basic
Asn	0.125	160	–	–
Asp	0.105	150	3.91	Acidic
Cys	0.819	135	8.30	Acidic, Hydrophobic
Gln	0.151	180	–	–
Glu	0.115	190	4.25	Acidic
Gly	0.252	75	–	–
His	0.354	195	6.50	Basic
Ile	0.967	175	–	Hydrophobic
Leu	0.908	170	–	Hydrophobic
Lys	0.000	200	10.79	Basic
Met	0.987	185	–	Hydrophobic
Phe	1.000	210	–	Hydrophobic
Pro	0.151	145	–	–
Ser	0.188	115	–	–
Thr	0.253	140	–	–
Trp	0.775	255	–	Hydrophobic
Tyr	0.484	230	10.95	Acidic
Val	0.770	155	–	Hydrophobic

Both models also share domain constraints that can be defined as

$$\begin{aligned} &w_i \in \{0,1\}, \forall i \in I; y_k \in \mathbb{Z}, y_k \geq 0, \forall k \in K \\ &Q_{i,dp} \geq 0, \forall i \in I;\ H_{dp} \geq 0; m_{n_l,p} \geq 0, \forall p \in P | p \neq dp \\ &\lambda_{i,p,j} \in SOS2^I, \forall i \in I, \forall p \in P | p \neq dp \\ &\rho_{p,j} \in SOS2^{Jq}, \forall p \in P | p \neq dp \end{aligned} \tag{28}$$

Finally, model MDV is completely defined with the specification of its objective function and the requirement of a minimal final purity

$$\min\ Z^{MDV} = \sum_{i \in I} w_i + c \cdot \sum_{k \in K} y_k \tag{11}$$

$$\left(1 - SP_{dp}\right) \cdot m_{0,dp} \geq SP_{dp} \cdot \sum_{\substack{p \in P \\ p \neq dp}} m_{n_l,p} \tag{13}$$

and the objective function of model MP attempting to minimize the final concentration of contaminants is given by

$$\min\ Z^{MP} = \frac{\sum_{\substack{p \in P \\ p \neq dp}} m_{n_l,p} + m_{0,dp}}{m_{0,dp}} \tag{15}$$

Note that this objective is equivalent to maximizing the final purity of the target protein.

Definition of case studies

It has been discussed that the set of chromatographic techniques considered in this work is composed of anion-exchange, cation-exchange and hydrophobic interaction chromatographers. Specifically, anion-exchange chromatography techniques are operated in integer values of pH ranging from 4 to 8. Therefore, the complete set of anion-exchange chromatographers is defined as $AE = \{\text{AE4}, \text{AE5}, \text{AE6}, \text{AE7}, AE8\}$, where the number in the element corresponds to the pH of operation. Analogously, the set of cation-exchange chromatography techniques is defined as $CE = \{\text{CE4}, \text{CE5}, CE6, CE7, CE8\}$. Only one type of hydrophobic interaction chromatography is used, and, thus, the set HI contains a single element and the complete set of techniques considered is $I = AE \cup CE \cup HI$. The properties of the 20 proteinogenic amino acids that are used to compose the tags are defined in Table 1.

We have applied the proposed framework to the two distinct cases studied by Simeonidis et al.[9] In the first case, the optimal flow sheets and tag compositions are obtained for the separation of a mixture of four proteins (a target protein and three contaminants), initially in the same concentration. The physicochemical properties of the mixture were obtained from the literature[16] and are presented in Table 2. This table contains values for the proteins hydrophobicity, exposed surface, and charge under different values of pH. The values for *dp* correspond to the properties of the untagged target protein.

The second case study consists of a mixture of 13 proteins. The data for this case were generated in a previous work[9] and are presented in Table 3. This case is employed here as an example of a higher dimension system that can be solved with the proposed approach. The exposed surface of the target protein was set to $\hat{S}_{dp} = 29287$.

Results and Discussion

In this section, we analyze the solutions of the proposed models applied to the two case studies defined in the previous section. We start by considering the four-component mixture and expand our analysis with results from the larger scale 13-protein problem. We have defined two different models for the optimal selection of separation flow sheets and tags, namely MDV which implements the weighted minimization of the number of techniques and amino acids in the tag, and MP, which attempts to find the maximal achievable purity with limitations in the number of techniques and tags. We have also specified two approaches for the piecewise linear approximation of the nonlinear function in the model — e.g., PWLI and PWLA. We hereon refer to the models applied to the 4-protein mixture problem as MDV-

Table 2. Physicochemical Properties of First Case Study—4 Proteins Mixture

					$Q_{i,p}$ (Charge/molecule) $\times 10^{-17}$				
Protein	$m_{0,p}$ (mg/mL)	MW_p (Da)	H_p	S_p	pH 4	pH 5	pH 6	pH 7	pH 8
dp	2	22200	0.27	9573.15	1.6	1.57	1.64	1.55	0.75
p1	2	77000	0.23	29287.6	0.93	0.33	−0.12	−0.34	−0.5
p2	2	23600	0.31	10910.8	2.15	1.46	1.17	0.78	0.38
p3	2	43800	0.28	15880.9	1.16	−0.63	−1.36	−1.82	−1.95

Table 3. Physicochemical Properties of Second Case Study—13 Proteins Mixture

Protein	$m_{0,p}$ (mg/mL)	MW_p (Da)	H_p	$Q_{i,p}$ (Charge/molecule) $\times 10^{-17}$				
				pH 4	pH 5	pH 6	pH 7	pH 8
dp	2	77000	0.28	2.04	1.06	−0.37	−0.81	−1.13
p1	2	22200	0.27	1.6	1.57	1.56	1.55	0.75
p2	2	23600	0.31	2.15	1.46	1.17	0.78	0.38
p3	2	13500	0.23	1.83	0.65	0.26	−0.2	−0.33
p4	2	43800	0.28	1.16	−0.63	−1.36	−1.82	−1.95
p5	2	15900	0.27	2.89	2.81	2.8	2.64	2.07
p6	2	14400	0.32	−0.46	−0.47	−0.63	−1.21	−1.25
p7	2	17500	0.21	0.45	−0.62	−0.79	−1.26	−1.7
p8	2	50000	0.27	−0.12	−0.32	−0.76	−0.91	−1.04
p9	2	12100	0.18	1.46	0.62	−1.02	−1.33	−1.52
p10	2	25500	0.3	1.01	−0.63	−1.27	−1.59	−1.76
p11	2	26000	0.28	2.96	1.26	0.92	0.54	0.01
P12	2	19900	0.25	0.93	0.33	−0.12	−0.34	−0.5

PWLI4 and MP-PWLI4, in case piecewise linear interpolation is applied, and MDV-PWLA4 and MP-PWLA4, for the model based in the piecewise linear approximation procedure. Similarly, the 13 proteins mixture problem is solved by models MDV-PWLI13, MP-PWLI13, MDV-PWLA13 and MP-PWLA13. We employ piecewise linear interpolations and approximations with a fixed number of 10 knots (eight internal knots $|J| = |J_q| = 10$). The safety factor parameter Δ (see Eq. 9) was set to 0.02 in all performed implementations. Figure 3 shows an example of both the PWLI and PWLA approaches applied to the approximation of $\overline{\ln CF}_{\text{AE6,p4}}$ and $\overline{\ln CF}_{\text{CE6,p4}}$ as functions of $Q_{\text{AE6},dp}$ and $Q_{\text{CE6},dp}$. The proteins and chromatographic techniques in this example are extracted from the 4-protein system. Two sets of approximations are provided, using ten and five knots. It is seen that the use of a larger number of knots clearly results in great improvements in approximation quality. However, there is a tradeoff between approximation quality and the size of the model in which they are implemented. The choice of using ten knots for all models in this work was based on the observation that it resulted in a very high-approximation quality, and still allowed for considerably low memory and computational time requirements. In systems containing a significantly larger size of proteins the balance may push for the use of a lower number of knots in the piecewise linear functions. As a final note, we observe that the piecewise linear functions for AE6 with five knots did not present considerable difference between the PWLI and PWLA approaches. The reason for this result lies on the physical limitation on the values of $\overline{\ln CF}_{i,p}$, which must be non-positive. The added constraint prevented the PWLA method to shift the knot to higher function values and further reduce the approximation error.

All models were implemented in the GAMS modeling language,[18] and solved to proven optimality (zero gap) with the ILOG CPLEX 11 solver.[19] All computational results were obtained in an Intel platform based on processor T7200 at 2 GHz with 1GB of RAM (only a single CPLEX thread was used). It is noted that all instances of the models applied to the system with four proteins were solved with less than 1 s of CPU time; thus, a computational requirements analysis is only performed for the 13-protein system.

Four-Protein Mixture

We start our analysis with the results obtained for the problem of purifying a target protein from a 4-protein mixture. As previously discussed, this problem can be solved using the approaches formulated in models MDV and MP. We initially consider model MDV, composed with piecewise linear interpolation and approximation approaches. The solution to these models depends on the specification of the minimal purity level desired after the purification process SP_{dp}, and the choice of weighting parameter c in the objective function. Figure 4 presents results for the solution of different instances of models MDV-PWLI4 and MDV-PWLA4, in which the values of c and Sp_{dp} were changed. Note that the

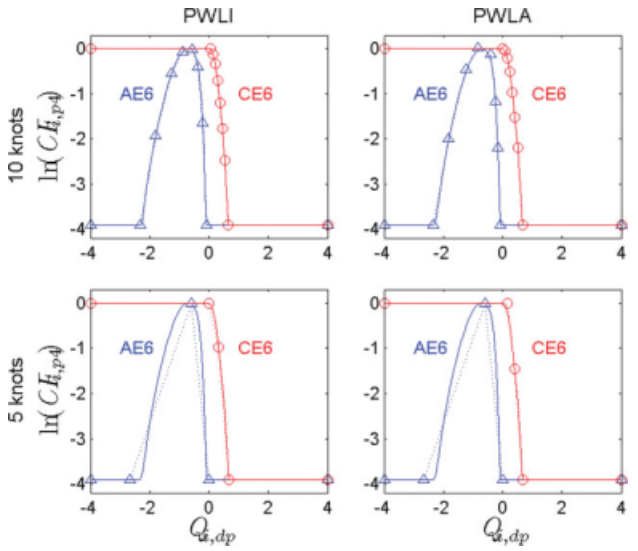

Figure 3. Examples of PWLI and PWLA applied to the approximation of $\overline{\ln CF}_{i,p}$.

Continuous lines represent the original nonlinear functions, dotted lines represent the piecewise linear interpolations and approximations. Triangles and circles are the sets of knots. [Color figure can be viewed in the online issue, which is available at www.interscience.wiley.com.]

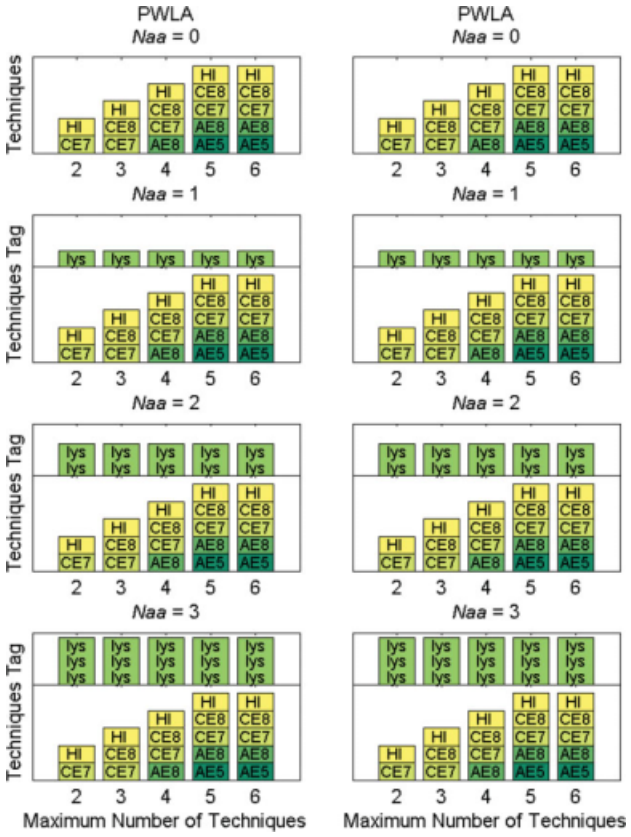

Figure 4. Results for models MDV-PWLI4 (left), and MDV-PWLA4 (right).

[Color figure can be viewed in the online issue, which is available at www.interscience.wiley.com.]

selected chromatographic techniques are presented in alphabetical order since the order of the steps is irrelevant to the proposed formulation. The results show that for purity requirements lower than 99.1%, no tag was selected, regardless of the weighting of the objective, for both linearization strategies. In contrast, the achievement of marginally higher levels of purity was only possible with the addition of lysines in the tag, while no other amino acid was selected in any of the solutions. With the upper bound of six amino acids in a tag, purities higher than 99.3% were unattainable. It can be seen for both approximation models that an increase in the value of c resulted in the selection of smaller tags at the expense of a larger number of chromatographic techniques.

It is also important to note that the chosen objective function composed of the simple minimization of numbers of amino acids and techniques results in the occurrence of alternate optimal solutions, as can be seen in the results in which no tag was selected. As an example, the solution of model MDV-PWLI4 with c = 0.1 and = 99% is naturally also an optimal solution for the same case but with $Sp_{dp} = 98\%$. The arbitrary dependence on weighting in the objective function, and the occurrence of alternate optima could be mitigated with the use of economical coefficients, appraising the variable costs of operating chromatographic columns against the fixed costs associated with genetically engineering tags in expressed proteins. However, such analysis is beyond the scope of this work.

It can be also observed that the results from models generated with the linearization approaches PWLI and PWLA show good agreement, but are not identical, which emphasizes the importance of the use of optimal linearization techniques. In particular, model MDV-PWLA4, solved for a purity lower bound of 99.2%, and with $c = 1$, obtained a solution requiring one less technique (AE7), and one more Lysine in the tag than model MDV-PWLI4 under the same conditions. The discrepancy in this instance of the models is analyzed in Figure 5, which shows the process flow sheets obtained with the PWLI and PWLA models. However, it is crucial to consider that when $c = 1$, the cost of adding one technique to the process, and one amino acid to the tag are the same, and, thus, both solutions in Figure 5 have the same objective function value and, hence, may be simply alternate optimal solutions. Nonetheless, it is observed that the use of an extra lysine in the tag by the MDV-PWLA4 model resulted in the requirement of one less chromatographic technique in order to perform the separation. In this case study, the remaining instances of the separation problem generated the same required numbers of techniques, albeit not the same types of chromatographic columns, neither amino acids in the tag.

We proceed to analyzing the results from the models MP, based on the maximization of the final purity of the target protein under constraints on the allowable number of techniques to be utilized and amino acids to compose the tag. The solution of these models depends on the value of parameters Ntech and Naa, which specify upper bounds to the number of used techniques and the number of amino acids in the tag, respectively. Figure 6 shows the results of the solution of models MP-PWLI4 and MP-PWLA4 for different values of these parameters. It can initially be seen that a good agreement with the solutions of model MDV is observed. The choice of chromatographic techniques performed by models MP is consistent with the techniques chosen by models MDV with high-purity requirements. The discrepancies related to the solutions of MDV for a 98% purity requirement are explained by the existence of alternate optima, as discussed previously. Moreover, lysine was the only amino acid selected for the tags, regardless of the chosen upper bounds in tags and techniques, in accordance with results from MDV. It must be noted that the choice of lysines as the sole component of the tags obtained in this system is a direct consequence of the physicochemical properties of the

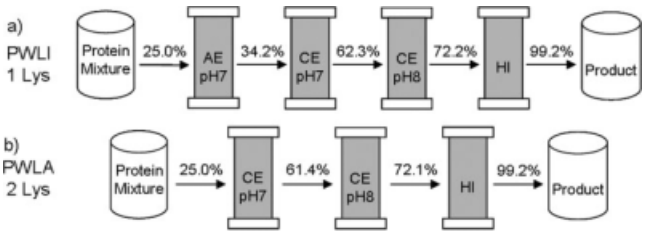

Figure 5. Flow sheets for the purification process obtained by (a) MDV-PWLI4 model, and (b) MDV-PWLA4 model, for $c = 1$ and a purity requirement of 99.2%.

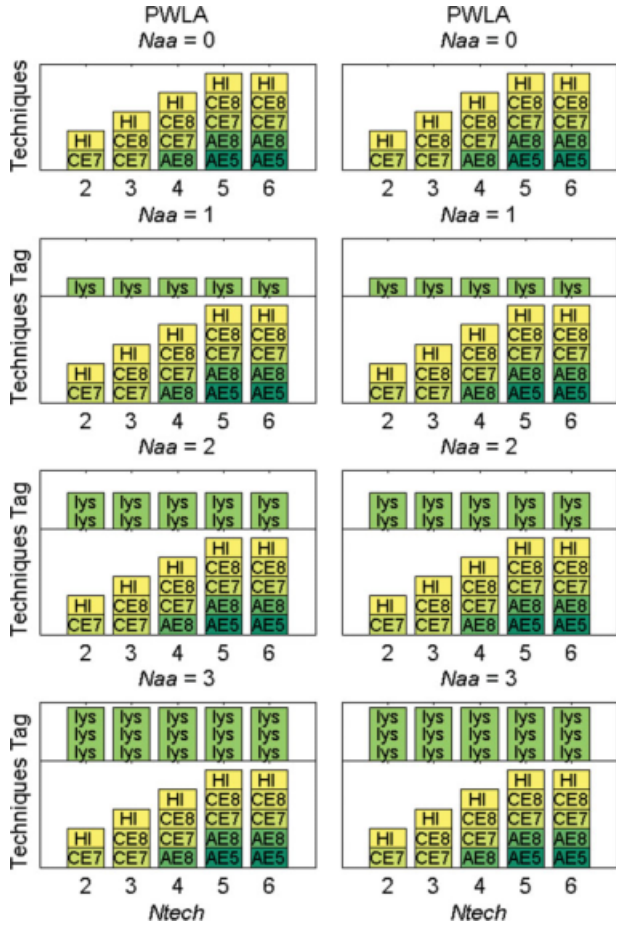

Figure 6. Results for models MP-PWLI4 (left) and MP-PWLA4 (right).

[Color figure can be viewed in the online issue, which is available at www.interscience.wiley.com.]

protein mixture. Table 2 shows that *dp* is more positively charged than the contaminant proteins at higher pH. For that reason, the addition of a polycationic tag composed of lysines aids the separation from the remaining proteins in the mixture by chromatographic techniques in high pH environments. Charged polycationic tails composed of lysines have been experimentally shown to improve the selectivity of ion-exchange chromatography.[20]

In addition, it is also seen that in this system the maximal obtainable purity is saturated for numbers of chromatographic techniques larger than 5, which is evidenced by the fact that no more than five techniques are chosen in any instance of the model in which *Ntech* is 6. This phenomenon has the same origin as the observed impossibility of obtaining a higher than 99.3% purity with model MDV, and it lies in the use of the safety factor, Δ, which has the consequence that no contaminant can be perfectly separated from the target protein.

Furthermore, the maximal purity—obtained from the OF values of model MP—for distinct upper bounds in tags and techniques is shown in Figure 7. The purity saturation previously discussed can be readily observed in this figure. It is seen that there is a large jump in the maximal purity attainable with two and three techniques, but the improvement with further addition of other columns is very small, becoming null after five techniques. It is also observed that in this system there is only a small influence of size of the tag in the quality of the separation.

In order to gauge the efficiency of the proposed models, we offer a comparison with the previous MINLP approach developed by Simeonidis et al.[9] and the linearized MILP model proposed by the same group.[10] Both models attempt to solve the problem of minimizing the separation process flow sheet and the tag composition in a similar way as this model MDV. However, it is important to note that neither of these previous models was able to generate solutions for the simultaneous minimization of both the number of techniques and amino acids in the tag, as is the case for the MDV model. Conversely, they required a two-stage solution approach that relied on the initial solution of an untagged problem (y_k set to zero), and the subsequent minimization of the tag in a second problem. The possibility of simultaneously solving both minimizations, although dependent on an arbitrary weighting of the objective function, is crucial for the guarantee of optimality in the final solution.

Furthermore, both previous methods were incapable of dealing with the discontinuity and general nonsmoothness introduced by the calculation of the concentration factors in Eq. 9, and, thus, relied on a sigmoidal approximation of these relationships. Such an approximation resulted in a lower precision in the obtained solutions, exemplified by the observation that models MDV-PWLI4 and MDV-PWLA4 in this study were capable of obtaining purities of up to 99.1% (Figure 4) using three chromatographic techniques and without the requirement of tags, whereas the solutions reported for the 4-protein system by Simeonidis et al.[9] and Simeonidis et al.[10] required at least a tag of two Lysines to accomplish a purity of 98% with three chromatographic steps.

Finally, the computational time requirements for the present approaches—all instances of the 4-protein system were solved to proven optimality in less than 1 s—were significantly lower than the requirements for the previous models. The MINLP approach required a total of 33 s to solve one instance of the problem with purity requirement of 98%, whereas the previous MILP model the time requirement for

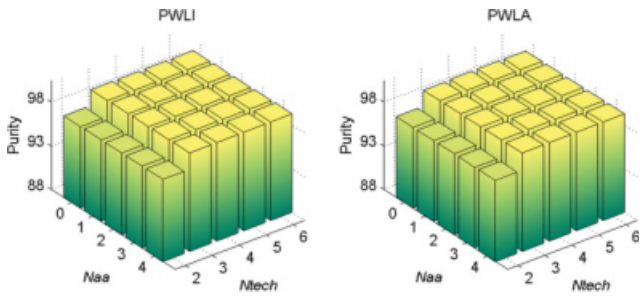

Figure 7. Maximal obtainable purity by models MP-PWLI4 (left), and MP-PWLA4 (right), for different values of the upper bounds on the number of chromatographic techniques and the number of amino acids in the tag.

[Color figure can be viewed in the online issue, which is available at www.interscience.wiley.com.]

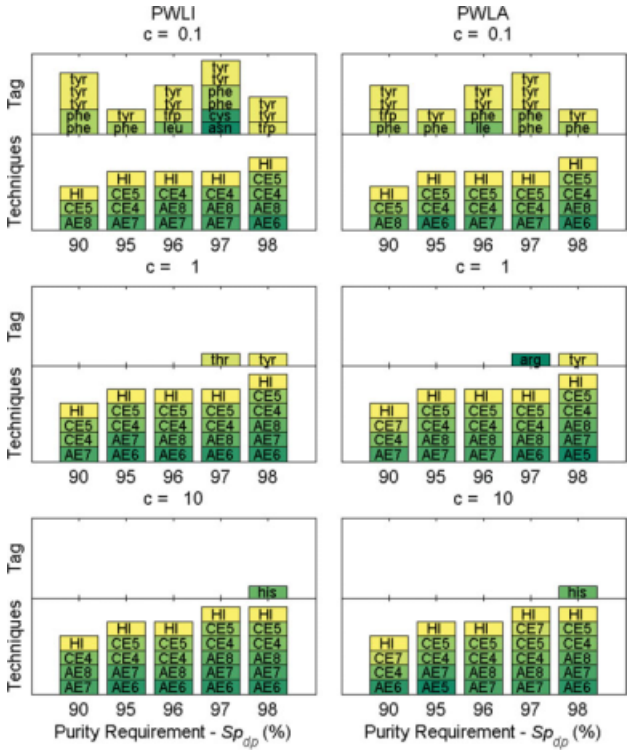

Figure 8. Results for models MDV-PWLI13 (left), and MDV-PWLA13 (right), with different values of the objective function weight and purity requirements.

[Color figure can be viewed in the online issue, which is available at www.interscience.wiley.com.]

the same problem was of 3 s. Nonetheless, a comparison of computational time requirements for a larger problem such as the 13-protein system, provided in the next section, is capable of better demonstrating the efficiency gap between the approaches.

Thirteen-protein mixture

We proceed our analysis of the computational results from the proposed models with the solutions to the 13-protein system, previously defined. Apart from numerical results, a computational requirement analysis is provided at the end of this section. Figure 8 contains the solutions from models MDV-PWLI13 and MDV-PWLA13 for different values of c and purity requirement. The larger size and complexity of this problem with respect to the 4-protein system is evidenced by the larger requirements for techniques and tags to attain lower levels of purity. Apart from that, we observe similar trends already seen in the solutions for the 4-protein system. Despite the larger size of the problem, alternate optimal solutions are still frequently obtained, as is evident from the analysis of the solutions in which no tags were selected. In these systems, all solutions based on models using the PWLI and PWLA approaches required the same number of techniques for the separations, albeit of different types. The numbers and types of amino acids that compose the tags varied considerably, especially when c was set to 0.1 and larger tags were selected. In this case, large tags were used in an attempt to minimize the number of techniques required. When the addition of tags was not sufficient to perform the separation, the number of techniques was increased and the size of the tag reduced. For higher values of c, the use of tags was reduced only for instances with high purity requirements.

The results from models MP-PWLI13 and MP-PWLA13 are presented in Figures 9 and 10. Figure 9 contains the composition of the separation flow sheets and the tags obtained in the solutions, while the number of techniques and amino acids were constrained by upper bounds specified in the figure. It is noted that the composition of the tags is much more homogeneous than the selections performed by model MDV. Only amino acids tyrosine and phenylalanine were selected to the tags, and the preference of inclusion constantly lies with the former. The predominance of these amino acids in the tags selected with model MDV is also apparent; however, this feature is by no means consistent.

It is also observed that the selection of chromatographic techniques is generally consistent between models composed with different approximation approaches, specially in instances with high upper bounds. The solutions to model MP have the added advantage of being able to sort the chromatographic techniques and the amino acids in the tag in order of effect in the overall separation. Naturally, the selections of

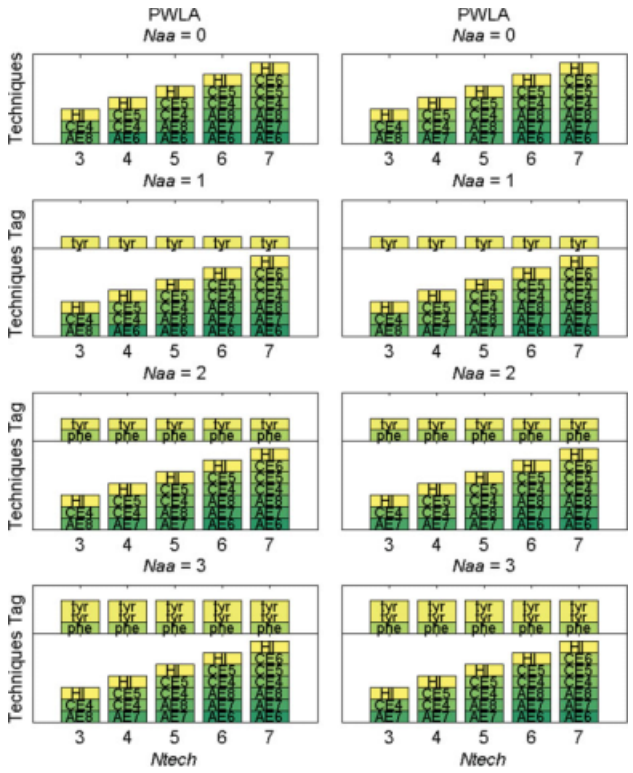

Figure 9. Results for models MP-PWLI13 (left) and MP-PWLA13 (right), with different values of the upper bounds on the number of chromatographic techniques and the number of amino acids in the tag.

[Color figure can be viewed in the online issue, which is available at www.interscience.wiley.com.]

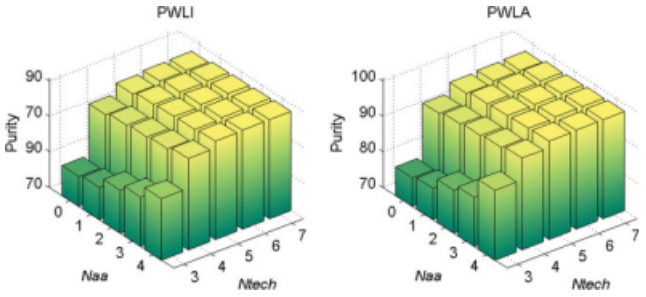

Figure 10. Maximal obtainable purity by models MP-PWLI13 (left) and MP-PWLA13 (right), for different values of the upper bounds on the number of chromatographic techniques and the number of amino acids in the tag.

[Color figure can be viewed in the online issue, which is available at www.interscience.wiley.com.]

flow sheets and tags made under stringent upper bounds will be composed by subsets of amino acids and techniques that individually contribute more significantly for the separation process. Few instances of this result contradict this general observation, e.g., the solutions for the model MP-PWLI13 with upper bounds on number of techniques of three and four. In all these instances, we see that the anion-exchange technique selected with an upper bound of three techniques differed from the anion-exchange technique selected when this upper bound was 4. Such phenomenon may indicate a poor capacity from the model of distinguishing between two chromatographic techniques with very similar physicochemical properties.

Finally, we can observe in Figure 10 the effects in the maximal purity of the upper bounds on the numbers of amino acids in the tag and techniques in the separation flowsheet. We see that the presence and size of tags exert a much higher influence in this problem, in comparison to the 4-protein system. Nonetheless, the effect of tags is more pronounced with the use of lower numbers of techniques. The same phenomenon of saturation of maximal purity with larger numbers of techniques can be observed.Next, we focus on the computational performance from the 13-protein system with the computational requirements for the solution of the models. Figure 11 contains the computational time used to solve all instances of models MDV-PWLI13, MDV-PWLA13, MP-PWLI13 and MP-PWLA13 to proven optimality. The first observation that can be made is that the maximum computational time required to solve one instance of the model is in the order of 2 s, which shows the high efficiency of the proposed linear formulation. Model MDV presented considerably higher computational time requirements for instances with $c = 0.1$ than for the remaining values of the OF weight. Note from the results in Figure 8 that the solutions for this instance made use of larger and more diverse tags, which may account for the higher time required to prove optimality. No clear trend between solution time and the values of purity requirements was observed.

The solution of model MP had a maximal computational time requirement of the order of 0.8 s, making it more computationally efficient than to model MDV. It is observed that the solution times increase with larger values for the upper bound in number of amino acids in the tag. Such behavior is credited to the higher complexity of solutions with larger tags. On the other hand, the computational time requirements show a more complex dependency on the upper bound on the number of chromatographic techniques employed. When no tag is allowed, the solution times increase with increments to *Ntech*. However, when tags of distinct sizes are allowed, the computational time requirements decrease with increments to the upper bound on techniques. In any case, it is safe to say that the computational time requirements for model MP presents a stronger dependence on the size of the tag than on the number of techniques employed.

A comparison between the results obtained for the 13-protein system by the approaches developed in this work and the previously published models by Simeonidis et al.[9] and Simeonidis et al.[10] is presented next. Initially, we note that the previous MILP formulation was unable to find any solution to the 13-protein problem, even with the use of the two-stage solution approach previously described. Such incapacity may have arisen from the suboptimal piecewise linear interpolations employed, or from the mentioned use of a sigmoidal approximation to the computation of the concentration factors.

The MINLP model yielded solutions for purity requirements of 95 and 98% with the two-stage suboptimal approach. The solutions obtained by Simeonidis et al.[9] for a purity requirement of 95% required a set of six chromatographic steps without the use of a tag, and four steps with a tag composed of four amino acids. In comparison, the solutions of model MDV presented here resulted in purities of up to 96% with five chromatographic steps, and up to 97% with six chromatographic steps without the requirement for tags, as can be seen in Figure 8 for the cases in which c is set to 10. Under purity requirements of 98%, the MINLP model required a total of five chromatographic steps with a tag composed of four amino acids. For the same conditions,

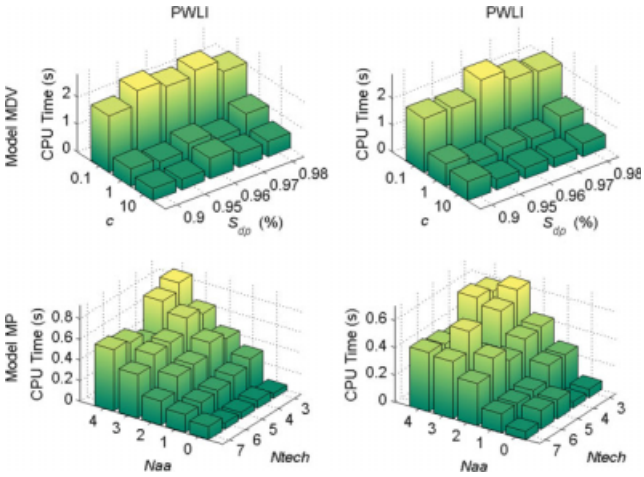

Figure 11. CPU time requirements for obtaining the optimal solutions to models MDV (top row) and MP (bottom row), with approximation strategies PWLI (left column) and PWLA (right column), for the 13-proteins system.

[Color figure can be viewed in the online issue, which is available at www.interscience.wiley.com.]

and with the objective function weight set to 0.1, model MDV-PWLI13 required five steps and a tag of three amino acids to attain 98% of purity of the target protein, whereas model MDV-PWLA13 also required five chromatographic steps, but with a tag composed of only two amino acids. As is the case for the 4-protein system, these observed discrepancies can be credited to imprecisions associated with the requirement for a sigmoidal approximation to the computation of concentration factors.

The computational time requirements for the formulations developed in this work in comparison to the previous formulation by Simeonidis et al.[9] show larger differences for the 13-protein system. The MINLP model required 245 s to solve the problem with a purity requirement of 95%, and 415 s for the problem with 98% purity required. In comparison, Figure 11 shows that the largest time required to solve any attempted instance of the 13-protein system with the models developed here was of the order of 2 s. It is further important to note that due to the nonlinear and nonconvex nature of the MINLP model, the obtained solutions can not be guaranteed to be optimal solutions by most available algorithms capable of solving such problems. In summary, the results from the larger 13-protein system are capable of further underlying the gain in efficiency obtained by the formulations developed in this work, when compared to previous published formulations.

Conclusions

This article developed a novel and efficient MILP formulation for the problem of simultaneously optimizing a process flow sheet composed of distinct chromatographers capable of purifying a target protein, and determining the optimal composition of an amino acid tag to be added to the desired protein to facilitate the separation process. Two models were developed: a minimization of decision variables (MDV) model, which attempts to find the minimal set of chromatographic techniques and amino acids in the tag that are capable of satisfying a lower bound in the final purity of the target protein; and a maximization of purity (MP) model, which is capable of defining what is the maximal attainable purity of the desired product, provided that the number of chromatographic techniques composing the process flow sheet and amino acids composing the tag are limited. The linearization of the model was achieved by using an approach to generate optimal piecewise linear interpolations and approximations.[11]

The results obtained show that the developed models were capable of obtaining optimal solutions with time requirements that varied from fractions of 1 s for a system containing an initial mixture of four proteins, to 2 s for the most challenging instances of a problem with a 13-protein mixture. The obtained solutions obtained identified in both systems studied minimal sets of chromatographic techniques and tags that resulted in maximal efficiency separation processes. It was verified that model MP is computationally better conditioned than model MDV, and is able to obtain optimal solutions with less time. Moreover, model MP has the added advantage of providing an ordering of importance for chromatographic techniques and amino acids, identifying which of these have the greatest effects in the overall process. Further comparisons with results from previously published models further underlined the efficiency of the proposed formulation, which was able to obtain optimal solutions for problems that are intractable by the other models, and the solutions were obtained with significantly less computational time requirement.

It was pointed that the proposed MDV model still suffers from the drawback of requiring an arbitrary weighting of the objective function, which balances the both the minimization of the number of chromatographic steps and of amino acids in the tag. Future work in the models will include the substitution of this arbitrary weighting for a more accurate economic objective function, balancing the fixed costs associated with engineering the heterologous gene responsible for coding the target protein in the host organism, and the operational costs related to the use of different chromatographic techniques.

Literature Cited

1. Datar R. Economics of primary separation steps in relation to fermentation and genetic engineering. *Process Biochem.* 1986;21:19–26.
2. Nagrath D, Bequette BW, Cramer SM, Messac A. Multiobjective optimization strategies for linear gradients chromatography. *AIChE J*. 2005;51:511–525.
3. Steffens MA, Fraga ES, Bogle IDL. Multicriteria process synthesis for generating sustainable and economic bioprocesses. *Comp Chem Eng*. 1999;23:1455–1467.
4. Steffens MA, Fraga ES, Bogle IDL. Synthesis of purification tags for optimal downstream processing. *Comp Chem Eng*. 2000;24:717–720.
5. Lienqueo ME, Salgado JC, Asenjo JA. An expert system for selection of protein purification processes: experimental validation. *J Chem Technol Biotechnol*. 1999;74:293–299.
6. Lienqueo ME, Asenjo JA. Use of expert systems for the synthesis of downstream protein processes. *Comp Chem Eng*. 2000;24:2339–2350.
7. Vasquez-Alvarez E, Lienqueo ME, Pinto JM. Optimal synthesis of protein purification processes. *Biotechnol Prog*. 2001;17:685–696.
8. Vasquez-Alvarez E, Pinto JM. A mixed integer linear programming model for the optimal synthesis of protein purification processes with product loss. *Chem Biochem Eng Q*. 2003;17:77–84.
9. Simeonidis E, Pinto JM, Lienqueo ME, Tsoka S, Papageorgiou LG. MINLP models for the synthesis of optimal peptide tags and downstream protein processing. *Biotechnol Prog*. 2005;21(3):875–884.
10. Simeonidis E, Pinto JM, Papageorgiou LG. An MILP model for optimal design of purification tags and synthesis of downstream processing. *Proc ESCAPE-15*. 2005;20B:1537–1542.
11. Natali JM, Pinto JM. Piecewise polynomial interpolations and approximations of one-dimensional functions through mixed integer linear programming. *Optim Method Softw*. 2009; DOI: 10.1080/10556780802614507.
12. Vasquez-Alvarez E, Pinto JM. MILP models for the synthesis of protein purification processes. *Proc ESCAPE-11*. 2001;579–584.
13. Vasquez-Alvarez E, Pinto JM. Efficient MILP formulations for the optimal synthesis of chromatographic protein purification processes. *J Biotechnol*. 2004;110:295–311.
14. Lienqueo ME, Leser EW, Asenjo JA. An expert system for the selection and synthesis of multistep protein separation processes. *Comput Chem Eng*. 1996;20S:S189–S194.
15. Mosher RA, Gebauer P, Thormann W. Computer-simulation and experimental validation of the electrophoretic behavior of proteins: iii. use of titration data predicted by the protein's amino acid composition. *J Chromat*. 1993;638:155–164.
16. Lienqueo ME, Mahn A, Asenjo JA. Mathematical correlations for predicting protein retention times in hydrophobic interaction chromatography. *J Chromat A*. 2002;978:71–79.
17. Williams HP. *Model Building in Mathematical Programming*. New York: John Wiley & Sons; 1985.
18. Brooke A, Kendrick D, Meeraus A, Raman R. *GAMS User Guide*. GAMS Development Corp., Washington, DC; 1997.

19. ILOG, Inc. ILOG CPLEX 11.0 User's Manual; 2008.
20. Kweon DH, Kim SG, Seo JH. Purification and refolding of cyclodextrin glycosyltransferase expressed from recombinant. *Escherichia Coli J Incl Phen Marc Chem*. 2004;50:37–41.

Appendix: PWLI and PWLA Formulations

Here we describe the procedures based on MILP models previously developed[11] for the optimal interpolation and approximation of single-dimensional nonlinear functions by piecewise linear functions. The procedures make use of a discrete representation of a nonlinear function described by the pairs $(x_i, f_i), i \in Q = \{1, 2, \ldots, n_Q\}$, where Q is the predefined set of sampling points. We further define a binary variable $W_{i,j}$ that is equal to 1 if $i \in Q$ and $j \in Q$ are two consecutive knots, and 0, otherwise. Furthermore, let N be the number of knots in a solution, given *a priori*.

Based on the aforementtioned definition, the feasible region of the binary variable $W_{i,j}$ can be defined by the following sufficient constraints.

At most one polynomial piece of the approximating function may begin and one piece may end in each of the points in Q

$$\begin{aligned} \sum_{j\in Q} W_{i,j} \le 1 \qquad & \forall i \in Q | i > 1 \\ \sum_{i\in Q} W_{i,j} \le 1 \qquad & \forall j \in Q | j < n_Q \end{aligned} \tag{A1}$$

The first and the last points of Q are necessarily knots

$$\begin{aligned} \sum_{j\in I} W_{i,j} = 1 \qquad & i = 1 \\ \sum_{i\in I} W_{i,j} = 1 \qquad & j = n_Q \end{aligned} \tag{A2}$$

Any knot, with exception of the first and the last ones, has to be both the start of a polynomial piece of the approximating function and the end of another

$$\sum_{i\in Q} W_{i,k} = \sum_{j\in Q} W_{k,j} \quad \forall k \in \{2, \ldots, n_Q - 1\} \tag{A3}$$

The approximating function is predefined to have N internal knots

$$\sum_{i\in Q}\sum_{j\in Q} W_{i,j} = N - 1 \tag{A4}$$

In the piecewise linear interpolation (PWLI) model, the values of the approximating function are defined by the following set of constraints

$$f_k^P = \sum_{\substack{i\in Q \\ i\le k}} \sum_{\substack{j\in Q \\ j\ge k}} \frac{\left[(x_k - x_i)\cdot f_j + (x_j - x_k)\cdot f_i\right]}{(x_j - x_i)} \cdot W_{i,j} \quad \forall k \in Q \tag{A5}$$

Alternatively, in the piecewise linear approximation (PWLA) model, in which the approximating function is not required to coincide with the value of the original nonlinear function in the knots, the approximation is defined by

$$f_k^P - \frac{\left[(x_k - x_i)\cdot f_j^P + (x_j - x_k)\cdot f_i^P\right]}{(x_j - x_i)} \le M(1 - W_{i,j}) \quad \forall (i,j,k) \in Q^3 |\ i < k < j \tag{A6}$$

$$f_k^P - \frac{\left[(x_k - x_i)\cdot f_j^P + (x_j - x_k)\cdot f_i^P\right]}{(x_j - x_i)} \ge -M(1 - W_{i,j}) \quad \forall (i,j,k) \in Q^3 |\ i < k < j \tag{A7}$$

Finally, the measure for the quality of the approximations is defined using the 1-norm of the distance between the vectors describing the original function and the piecewise linear interpolations and approximations. The objective function is to minimize this norm and is given by the following equations

$$Z = \sum_{i\in Q} z_i \tag{A8}$$

$$\left.\begin{aligned} z_i &\ge (f_i - f_i^P) \\ z_i &\ge -(f_i - f_i^P) \end{aligned}\right\} \forall i \in Q \tag{A9}$$

Manuscript received Aug. 10, 2008, and revision received Feb. 26, 2009.