Control of Chromatography Column in Production Scale

Jonas Månsson

Department of Automatic Control
Lund Institute of Technology
May 1998
Title and subtitle
Control of Chromatography Column in Production Scale

Abstract
In the pharmaceutical industry there is a need for separating biochemical compounds, especially proteins. Ion exchange chromatography (IEC) is a widely used separation method. At the outlet of an IEC column the UV-absorption is measured to detect the concentration peaks from the separated proteins. The UV-absorption curve is called the chromatogram. The problem to find and collect the desired protein among other proteins is investigated in this thesis. This is called the pooling problem, and is normally done based on the chromatogram.

A simple model is developed that captures the main dynamics of an IEC process. The model considers two competing proteins, and the model is used to simulate the adsorption phase (protein binding) and the elution phase (protein desorption).

Process data from a real IEC process are used in an attempt to characterize the pooling problem. Multiple linear regression (MLR) models and principal component analysis (PCA) are used.

Proposals are given to provide tools, which could be used for a new approach to solve the pooling problem.

Key words
Ion exchange chromatography, Protein separation, Modeling, Simulation, Multiple linear regression, Principal component analysis

Classification system and/or index terms (if any)

Supplementary bibliographical information

ISSN and key title
0280–5316

Language
English
Number of pages
49

Security classification

The report may be ordered from the Department of Automatic Control or borrowed through:
University Library 2, Box 3, S-221 00 Lund, Sweden
Fax +46 46 222 44 22 E-mail ub2@ub2.lu.se
Contents

Acknowledgments ................................................. 2
1. Introduction .................................................. 3
2. Proteins ........................................................ 4
   2.1 Proteins are built from amino acids ..................... 4
   2.2 Proteins have unique amino-acid sequences ............. 5
   2.3 Proteins are charged macro molecules .................. 6
   2.4 Enzymes .................................................. 6
3. Ion Exchange Chromatography ............................... 7
   3.1 Equilibration ............................................ 7
   3.2 Sample application ..................................... 8
   3.3 Washing after sample application ....................... 8
   3.4 Elution .................................................. 8
   3.5 Regeneration ............................................ 10
   3.6 Quantities in Ion Exchange Chromatography ............ 10
4. Problem formulation ......................................... 14
5. Modeling ...................................................... 15
   5.1 Modeling ................................................ 15
6. Simulation of the model ...................................... 18
   6.1 The simulation setup ................................... 18
   6.2 An introducing simulation example ..................... 19
   6.3 Altering the constants of reaction velocity .......... 23
   6.4 Competition between proteins when altering the input concentrations .................................. 24
   6.5 Altering the length of the concentration pulse ........ 25
   6.6 Changing slope of the elution gradient ............... 26
   6.7 Non heterogeneous flow rate ............................ 28
   6.8 The number of sections ................................ 28
   6.9 The pooling problem: two simple approaches .......... 29
   6.10 Conclusions from the simulations ..................... 32
7. Data collection and data preparation ...................... 33
   7.1 Data collection ......................................... 33
   7.2 Data preparation ....................................... 33
8. Data analyzes ................................................. 34
   8.1 The age of the column .................................. 35
   8.2 A model for the eluate concentrations ................ 35
   8.3 Multiple Linear Regression models ..................... 38
   8.4 Conclusions from the data analyzes .................... 43
9. Proposals of data collection to make the pooling algo-
   rithm more robust ........................................ 46
   9.1 Automate the data collection ........................... 46
   9.2 Fraction measurements .................................. 46
10. References ................................................... 49
Acknowledgments

The work in this thesis has been performed as a collaboration between Pharmacia & Upjohn (PnU) and Department of Automatic Control at Lund Institute of Technology.

PnU kindly provided the process data needed, and I would like to thank all the staff involved at PnU for making this possible. Especially I would like to thank Jan-Peter Axelsson, who has been my supervisor at PnU.

I am also grateful to my supervisors at the Department, Per Hagander, and Mats Åkesson for all help and encouragement they have given me.
1. Introduction

In many industrial applications there is a need for separating chemical substances. Especially in the biochemical research and industry area, the separation and identification of biochemical compounds, such as proteins, is of great importance. In this Masters thesis the technique of protein separation or protein purification is studied. A worldwide method used for this purpose is the chromatography technique.

The inventor of chromatography was the Russian botanist Mikhail Tsvet. In a book written by Tsvet in 1910 he described a technique of how to separate different plant pigments. In a vertical column of glass Tsvet packed an adsorptive (adsorption = the ability to select and attach molecules onto ones surface) material. He then added the pigments to the top of the column and washed them through the column with an organic solvent. Due to relative rates of adsorption between the different pigments they separated into series of discrete coloured band in the column, divided by sections that were free of pigments. The adsorptive material was then pulled out of the column and the different coloured bands could be analyzed separately. Since Tsvet was working with coloured substances the technique was called chromatography, from the Greek words meaning colour writing.

The chromatography technique has then developed and today there exists a number of different chromatography methods, e.g., affinity, ion exchange, gel filtration chromatography etc. Although the methods do not depend on colour separation the technique is still called chromatography. The ion exchange chromatography will be treated in this thesis.
2. Proteins

When separating proteins using chromatography the characteristic of the proteins is an important factor. The size, net charge etc. of the proteins to be separated determines the output of the chromatography process and can also be used to choose the right type of method. An introduction to protein chemistry and physics is therefore needed. This section is based on [6].

The number of proteins in the biosphere is huge. The word protein was coined by Jöns Berzelius to associate the importance of this class of macromolecules. He derived it from the Greek word proteios which means "of the first rank". The proteins play crucial roles in almost all biological processes. Proteins are among other things involved in the following functions:

- the enzymes which catalyze chemical reactions in biological systems.
- transportation of specific molecules. For example, hemoglobin transports oxygen.
- immune protection. Antibodies are proteins that recognize and combine with for example viruses.
- generation and transmission of nerve pulses.
- control of growth and differentiation of cells.

2.1 Proteins are built from amino acids

Amino acids are the basic units to build up a protein. There are 20 different amino acids which build up the enormous number of different proteins. An amino acid consists of an amino group, a carboxy group, a hydrogen atom and an R-group and they all have the same structure.

Figure 1 shows the dipolar ion form of an amino acid. The difference between the amino acids is in the structure of the R-group. The different R-groups have different chemical and physical behaviour, e.g., solvent in water, the charge and the capability to form hydrogen bonds.

Proteins consist of long chains of amino acids. One amino acid is joined by another amino acid by a peptide bond or amide bond. The peptide bond is formed when the carboxy group of one amino acid reacts with the amino group of another amino acid. During the reaction, which is a condensation, a water molecule is lost, see Figure 2.

The carbon, hydrogen, and the nitrogen atoms are placed in one plane and mobility around the peptide is therefore limited. This property is an important factor determining what space structures possible for a peptide. Peptides

![Figure 1](image)

Figure 1 The structure of an amino acid.
containing more than 10 amino acids are called polypeptides. The amino acid containing the free amino group is called the amino-terminal residue (to the left in Figure 2) and the amino acid containing the free carboxyl group is called the carboxyl-terminal residue (to the right in Figure 2). By convention the polypeptide starts at the amino-terminal residue.

2.2 Proteins have unique amino-acid sequences

In 1953, Frederick Sanger showed for the first time that a protein has a precisely defined amino-acid sequence. Sanger determined the amino-acid sequence in insulin. Nowadays the amino-acid sequences of thousands of proteins are known.

What distinguishes one protein from another is the composition of amino acids and the sequence of them. Those differences give the proteins their different functions. To be more specific it is the different R-groups of the amino acids which give the properties of a protein.

The long chains of polypeptides only admit very special space structures depending on the surrounding environment. In order to describe the space structure of the protein it is divided into four levels of architecture.

The primary structure is the amino-acid sequence, i.e., what kind of amino-acids and in what order they are placed.

The secondary structure describes the spatial arrangement of amino acids in space. The α-helix and the β-structure are examples of secondary structures. The α-helix is very well defined with 3.6 amino acids each turn. The β-structure is also very well defined, and here the peptides are arranged side by side in a linear sequence.

The tertiary structure is close to the secondary structure, but one can say that it describes the total space structure of a protein.

Many proteins contain many chains of polypeptides. Each polypeptide chain is called a subunit. The quaternary structure refers to the spatial arrangement of subunits and the contact between them.
2.3 Proteins are charged macro molecules

Proteins in a water solution are related to a specific net charge. The net charge depends on the number of amino acids and the proportion between base and acid. The net charge is also dependent on the actual pH-value, see Figure 3. At a certain pH-value, specific for different proteins the net charge equals zero, and this pH-value is referred as the isoelectric point (pI). The space structure is also dependent on pH. If the space structure is changed a protein will lose its functionality in a process called denaturation. A protein is therefore only stable in a specific pH-range. Denaturation can also be caused by heating.

![Graph showing the net charge of a protein as a function of pH.](image)

Figure 3  The net charge of an arbitrary protein as a function of pH.

2.4 Enzymes

Enzymes are proteins which catalyze chemical reactions in biological systems. Like all catalyzators the enzymes speed up the reaction velocity without being consumed themselves. During enzyme reactions, the enzyme attracts substrate to its active center. In the active center the chemical reaction is then catalyzed by the surrounding R-groups of the amino acids. After the catalyzed reaction the product is formed and the product separates from the enzyme surface.

It is the space structure in the active center that determines what kind of substrates that fits into the center. It is like a lock and key function.
3. Ion Exchange Chromatography

The ion exchange chromatography technique, from now on IEC, origins from that proteins are charged molecules and that the net charge depends on the actual pH. This section is based on [2] and [6].

The principle for IEC is simple. In an IEC column there is a stationary phase called gel or binding matrix. The stationary phase is packed into the column and charge groups are bound to the gel. The binding type between the gel and the charged groups is covalent. The charged groups have the ability to bind or adsorb ions or other charged molecules of opposite charge, and therefore they are also known as ion exchangers.

The stationary phase is washed through with the mobile phase containing the proteins to be separated. Due to interaction between the mobile phase and the stationary phase, proteins with opposite charge to the charged groups bind to the gel. The higher opposite net charge of the protein the more strongly it will bind to the gel. Proteins with the same sign of the charge as the charged groups will not bind at all to the gel.

To separate the proteins from each other, after being bound to the gel, the ion concentration in the mobile phase is raised. The added ions with opposite charge to the charged groups tries to bind to the gel and thus some of proteins will be desorbed. The proteins that are more weakly bound to the gel will desorb first and then after increasing the ion concentration even more, the proteins that are more strongly bound to the gel will desorb. This process is called elution. At the outlet of the column the UV-absorption is measured and concentration peaks from the separated proteins are detected. Thus it is possible to find the wanted protein among the different peaks. This process is called pooling.

The IEC process can be divided into 5 steps [5]: the equilibration of the column, protein sample application, washing after sample application, elution and regeneration. The steps are discussed in more detail below.

3.1 Equilibration

In the equilibration step, the charged groups, or the ion exchangers of the stationary phase are brought to a starting state. The column is washed through with a starting buffer that gives a chemical environment suitable for the protein to be separated. The charge of the ion exchangers are, as for proteins, dependent on the pH-value. Therefore the ionic binding strength of the ion exchangers can be calibrated by altering the pH in the starting buffer. Since the wanted protein is only stable in a certain pH-range, see Figure 3, the starting buffer should give the right pH properties to the column.

There are two types of ion exchangers. If the gel is to bind positively charged molecules or ions then a negatively charged exchanger should be used, and the exchanger is called a cation exchanger. Vice versa, a positively charged exchanger is used to bind negatively molecules or ions and is called an anion exchanger. For the protein shown in Figure 3 an anion exchanger should be used since the protein is stable in a pH-range above the pI and the protein has a net charge that is negative in this range.

In the equilibration step the ion exchangers are associated with exchange-able counter ions that are to be replaced with sample molecules during the sample application. In Figure 4 the equilibration state is showed for an anion exchanger. The ion exchangers are modeled as free ions attached to the matrix by pins.
3.2 Sample application

In this stage the sample containing the wanted protein and other unwanted proteins is applied to the column by pumping. Proteins having the appropriate charge quantities adsorb to the gel by displacing the counter ions, see Figure 5. The flow rate is a crucial factor. The higher flow rate the smaller is the adsorption of proteins to the gel. Assuming a heterogeneous mixture between proteins with different net charge, the proteins which bind more strongly to the ion exchangers will be found closest to the top (input) of the column. Therefore it is sometimes possible to recognize different protein bands in the column [2].

3.3 Washing after sample application

After the sample application the column is washed with the starting buffer. Unbound substances and proteins that did not adsorb during the sample application will then be washed away from the gel.

3.4 Elution

At this stage the column contains substances bound to the gel and among these the wanted protein. To separate them, one has to affect their binding strength, so they desorb from the gel in an increasing order of ionic strength. This can be done by either changing the pH or by changing the ionic strength in the mobile phase. Here the latter technique is treated.

When the ionic strength in the mobile phase is low there is hardly no competition for the charged groups on the gel between the ions and the bound molecules. Thus, an increase in the ionic strength increases the competition for the charged groups, and the most weakly bound molecules will begin to lose their interaction to the gel, and they will leave the column first. Increasing the
Figure 5  The state in the column after adsorption.

Figure 6  The elution step.

Ionic strength further results in a elution of the bound molecules in an order determined by their relative charge and binding strength. The elution process is illustrated in Figure 6.

The change of the ionic strength can be either stepwise or a continuous gradient. The most common type is the continuous gradient, because it often leads to improved resolution between the eluted peaks and it is easy to use [5].
By mixing two buffers with differing ionic concentration the ionic concentration entering the column is changed linearly. A mixture of the starting buffer, used during equilibration, and a final buffer with higher ionic concentration is used. The ionic concentration in the buffers is equilibrated by adding salt, NaCl.

3.5 Regeneration

The final step is to regenerate the column. To desorb the molecules that were not eluted in the previous step, the column is washed through with the final buffer or with a buffer that have an even higher ion concentration. Sometimes lye is used to denature proteins which are still bound to the gel. The column can now be equilibrated for the next protein purification batch.

3.6 Quantities in Ion Exchange Chromatography

So far the principles of IEC has been discussed. To try to describe the IEC technique in terms of mathematical models some dynamic parameters are defined. The parameters defined below refer to the UV-absorption diagrams which form the output from IEC. The UV-absorption is correlated to the concentration of the different proteins and other substances involved. The diagrams are called chromatograms.

Retention. The retention volume \( V_R \) is defined for a specific substance and determines the elution volume that has passed through the column until a concentration peak of the substance appears in the chromatogram. The void volume \( V_0 \) is often defined as the volume of the mobile phase in the column. The parameters are defined in Figure 7.

To get a normalized retention parameter, the retention factor or capacity factor, \( k \), is defined:

\[
k = \frac{V_R}{V_0} - 1
\]  

The capacity factor thus determines how many void volumes it takes to elute a specific protein. It is important to remember that during gradient elution the capacity factor is dependent on the ionic concentration gradient.
Efficiency - the peak shape To give an adequate measurement of the concentration peak the efficiency factor \( n \) is introduced. An easily detected peak is narrow and high and gives a large value of the efficiency factor. The efficiency factor is defined by:

\[
n = 5.54 \left( \frac{V_R}{W_{1/2}} \right)^2
\]

(2)

The parameters are defined in Figure 8. The efficiency factor is also referred to as the number of theoretical plates. Another parameter is the height of theoretical plate \( HETP \), which is defined by:

\[
HETP = \frac{L}{n}
\]

(3)

\( L \) is the column length. The efficiency parameters defined above should give a measurement of the goodness of the column. It is often measured with substance that does not interact with the gel.

Equations (3) and (2) are derived from statistical theory [2] and they are only valid if the peak arises from a Gaussian probability curve. This only happens in simple cases. In practical applications the Gaussian function does not provide an accurate model for peak shapes. To get a better model of a peak shape the exponential modified Gaussian (EMG) is introduced [3]. The EMG can be used to describe an asymmetrical and tailed Gaussian curve, see Figure 9. The EMG is defined by:

\[
\begin{align*}
y(t) &= \frac{A}{\tau} \cdot e^{\left(\frac{1}{2}(\frac{t}{\tau})^2-(\frac{t-t_c}{\tau})\right)} \cdot \int_{-\infty}^{\frac{t}{\sqrt{2}}} e^{-x^2} dx \\
z(t) &= \frac{(t-t_c)}{\sigma} - \frac{\sigma}{\tau}
\end{align*}
\]

(4)

In equation (4) the function argument is time instead of volume. The relation between time and volume is linear if we assume a constant flow rate. The EMG results from the convolution of a Gaussian function with an exponential
decay function. The function is defined by four parameters, the peak area $A$, the retention time $t_r$, the standard deviation of the Gaussian component $\sigma$, and the time constant of the exponential time decay $\tau$. Using this new concept of the peak shape the efficiency factor is defined by:

$$n = \frac{41.7 \cdot (t_r/W_{10\%})^2}{(b/a) + 1.25}$$

(5)

This equation is described to be one of the best methods for efficiency calculations [3]. The parameters in equation 5 is defined in Figure 9. The peak width is defined by $W_{10\%} = a + b$ and the ratio $b/a$ is called peak asymmetry. $a$ and $b$ are measured at 10% of maximum peak height.

**Resolution** The Resolution, $R_s$, measures the relative separation between two peaks and is thus a very important parameter in protein separation. Resolution is the distance between two peaks divided by the mean peak width, see Figure 10, and thus:

$$R_s = \frac{V_{R2} - V_{R1}}{\frac{1}{2}(W_{b1} + W_{b2})}$$

(6)

An alternative definition of resolution is:

$$R_s = \frac{\alpha - 1}{\alpha + 1} \cdot \frac{\bar{k}}{k + 1} \cdot \frac{\sqrt{n}}{2}$$

(7)

Here, $\bar{k}$ is defined as the mean of the two $k$-values for the two peaks and $\alpha$ is the relative retention, $V_{R2}/V_{R1}$. In equation (7) one can see that the resolution consist of three factors. The first factor is called the selectivity factor, the second the retention and the third efficiency. To get good resolution these three factors have to be optimized.

If $R_s < 1$ the two peaks are not separated. If $R_s = 1$ the peaks just touch each other, and if $R_s > 1$ the peaks are separated.

If the peak shape is not pure Gaussian the resolution equations are not valid by the same reason mentioned in the previous section. Using equation (5) to calculate the efficiency factor $n$, equation (7) becomes more accurate.
Figure 10  The retention volumes and the mean peak width of two peaks.

**Capacity**  The capacity of a column is a measurement of the capability to bind counter ions and proteins. The *total ionic capacity* is the number of charged groups associated to the gel and the *available capacity* is the amount of protein that can be bound to the gel.

The capacity defined here should not be confused with the *capacity factor* defined above.
4. Problem formulation

The separation of the desired protein, from now on product, is done during the elution step. The UV-absorption is measured at the outlet of the column, and the problem to find the wanted concentration peak among others is referred as the pooling problem. Since there are a lot of unwanted substances being eluted at the same time as the product, the chromatogram shows a sum of concentration peaks. To pool "right" is to collect the product so that the product concentration is high and the concentration of unwanted substances in the pool is low, see Figure 11. By dividing a wide pool in fractions and then analyzing the concentrations in each fraction a schedule of when to start and to stop the collection of the product can be made.

The experience is that for some separation processes, especially in the first steps in a product line, there can be considerable variation in the chromatogram between batches and the pooling schedule is therefore hard to apply in a general case. The process variation is mainly due to variations from batch to batch in the substance concentration in the applied sample. There are also effects due to gradual degradation of the column. After a number of batches the gel in the column will be re-packed, and after a number of such cycles the gel will be exchanged. An obsolete or badly packed column has influence on the chromatogram. The purpose of this master thesis is to characterize these process variations and to use such knowledge to make the pooling algorithm more robust.

![Figure 11](image)

Figure 11  The pooling problem. The concentration peak in the middle is the one that is pooled. The two other substances are treated as unwanted and are not included in the pool. The individual peaks are dotted.
5. Modeling

To get some understanding of the process dynamics of an IEC column, a simple model is developed. The goal is to get a model that as good as possible describes the dynamics during the adsorption step and the elution step. The latter determines the actual chromatogram.

To validate that the model captures the main process dynamics, it has to be evaluated during simulation.

5.1 Modeling

Consider the problem when only one kind of protein binds to the gel in the column. The following equation of chemical equilibrium then applies [2]:

\[ P + ES \leftrightarrow PS + E \]  

(8)

Here, \( P \) represents a molecule of the protein of interest in the mobile phase, \( ES \) represents a binding site to which exchangeable ions are bound, \( PS \) represents the binding site when the protein molecule is bound, and \( E \) represent the solute ions in the mobile phase.

The net charge of the protein in principal determines the number of ions that are desorbed when the protein binds to the gel. \( E \) can then be viewed as the cluster of ions bound to a macro molecule needed to be desorbed when a protein binds to the gel. From now on we treat \( E \) as a substance with the same functions as a protein and not as free ions, and we refer to it as an ionic macro molecule. By not considering the total ionic capacity, but instead the available capacity (amount of protein that can be bind to the gel) then in an equilibrated column \( ES \) equals the available capacity. Thus, the model is simplified so that we just consider the number of charged groups that are available for protein bounds.

Divide the column into \( N \) sections and in each section, \( i \), describe the change of concentration in time for each substance in equation (8), and let a constant flow rate \( q \) transport the molecules in the mobile phase from section \((i-1)\) to section \(i\)

\[
\begin{align*}
\frac{dC^i_P}{dt} &= -k_1 C^i_P C^i_{ES} + k_2 C^i_E C^i_{PS} + \frac{q}{V_m} (C^{i-1}_P - C^i_P) \\
\frac{dC^i_{ES}}{dt} &= -k_1 C^i_P C^i_{ES} + k_2 C^i_E C^i_{PS} \\
\frac{dC^i_{PS}}{dt} &= k_1 C^i_P C^i_{ES} - k_3 C^i_E C^i_{PS} \\
\frac{dC^i_E}{dt} &= k_1 C^i_P C^i_{ES} - k_2 C^i_E C^i_{PS} + \frac{q}{V_m} (C^{i-1}_E - C^i_E)
\end{align*}
\]  

(9)

Here, \( C^i_P \) denotes the concentration of the protein in the mobile phase, \( C^i_{ES} \) denotes the concentration of ionic macro molecules which are bound to the gel, \( C^i_{PS} \) denotes the concentration of protein bound to the gel and \( C^i_E \) denotes the concentration of ionic macro molecules in the mobile phase. \( V_m \) denotes the volume of the mobile phase in each section. The constant of reaction velocity for the reaction from left to right in equation (8) is called \( k_1 \), and vice versa.
holds for $k_2$. All concentrations are measured in moles or equivalent gram per liter of solid or mobile phase. The flow rate is denoted by $q$.

To explain the nature of the equations derived, consider the equation for the concentration change of $C_P^i$, in equation (9). According to equation (8) $C_P^i$ will decrease the more $C_P^i$ and $C_{ES}^i$ there is in section $i$ in every point of time. Similarly, $C_P^i$ increases the more $C_{PS}^i$ and $C_{AS}^i$ there is in every point of time. The decreasing and increasing rate are determined by $k_1$ and $k_2$, respectively. The equations are balanced with the inflow of $C_P^{i-1}$ and the outflow of $C_P^i$. Naturally, it is only the equations concerning the mobile phase concentrations that should be balanced with inflow and outflow.

By including the competition for binding sites between protein $P$ and another protein, protein $A$, the model can be extended. The chemical reaction of equilibration is then described by [2]:

\[
\begin{align*}
    P + ES & \rightleftharpoons PS + E \\
    A + ES & \rightleftharpoons AS + E
\end{align*}
\]

(10)

Again, the equations for the change of concentration can be derived for protein $A$:

\[
\begin{align*}
    \frac{dC_A^i}{dt} &= -k_3 C_A^i C_{ES}^i + k_4 C_B^i C_{AS}^i + \frac{q}{V_m}(C_{A}^{i-1} - C_A^i) \\
    \frac{dC_{ES}^i}{dt} &= -k_3 C_A^i C_{ES}^i + k_4 C_B^i C_{AS}^i \\
    \frac{dC_{AS}^i}{dt} &= k_3 C_A^i C_{ES}^i - k_4 C_B^i C_{AS}^i \\
    \frac{dC_B^i}{dt} &= k_3 C_A^i C_{ES}^i - k_4 C_B^i C_{AS}^i + \frac{q}{V_m}(C_{B}^{i-1} - C_B^i)
\end{align*}
\]

(11)

The equations (9) and (11) differs only in the reaction velocities constants. Thus, by having different values for the reaction velocities constants, dynamics of two competitive proteins can be modeled by combining (9) and (11).

The available capacity is regarded to be time invariant in each section of column. The sites are always occupied with a protein, $PS$ or $AS$, or exchangeable ionic macro molecules $ES$. Thus:

\[
C_{PS}^i(t) + C_{AS}^i(t) + C_{ES}^i(t) = Q_{av}
\]

(12)

$Q_{av}$ denotes the available capacity. Equation (12) can be rewritten as:

\[
C_{ES}^i(t) = Q_{av} - C_{PS}^i(t) - C_{AS}^i(t)
\]

(13)

By combining equation (9), (11) and (13) we get a system of five differential equations that describes the competitive dynamic for binding sites between two proteins and exchangeable ionic macro molecules:
\[
\begin{align*}
\frac{dC_P^i}{dt} &= -k_1 C_P^i (Q_{av} - C_{PS}^i - C_{AS}^i) + k_2 C_E^i C_{PS}^i + \\
&\quad + \frac{q}{V_m} (C_P^{i-1} - C_P^i) \\
\frac{dC_A^i}{dt} &= -k_3 C_A^i (Q_{av} - C_{PS}^i - C_{AS}^i) + k_4 C_E^i C_{AS}^i + \\
&\quad + \frac{q}{V_m} (C_A^{i-1} - C_A^i) \\
\frac{dC_E^i}{dt} &= [(Q_{av} - C_{PS}^i - C_{AS}^i) (k_1 C_P^i + k_3 C_A^i)] - \\
&\quad - [C_E^i (k_2 C_P^i + k_4 C_{AS}^i)] + \frac{q}{V_m} (C_E^{i-1} - C_E^i) \\
\frac{dC_{PS}^i}{dt} &= k_1 C_P^i (Q_{av} - C_{PS}^i - C_{AS}^i) - k_2 C_E^i C_{PS}^i \\
\frac{dC_{AS}^i}{dt} &= k_3 C_A^i (Q_{av} - C_{PS}^i - C_{AS}^i) - k_4 C_E^i C_{AS}^i
\end{align*}
\]
6. Simulation of the model

The system in equation (14) will be used in this section to simulate the adsorption phase and the elution phase.

During adsorption we will treat the column as equilibrated before the sample is applied. That means that all binding sites are occupied with ionic macro molecules and therefore, $C_{ES}^i(t = 0) = Q_{av}$. During sample application a pulse of limited length of the mobile phase containing proteins $P$ and $A$, is applied to the first section of the column. The sample is washed through the column according to equation 14. A time after the input concentration pulse, the system becomes stationary and all the time derivatives equals zero. At this point the column is loaded with proteins $P$ and $A$. There also exist some ionic macro molecules $E$ that were not exchanged, or desorbed, during the adsorption phase. In the column $C_{P5}$, $C_{A5}$ and $C_{ES}$ vary between different sections. Therefore concentration gradients or profiles have developed in the column after the adsorption step.

The adsorbed proteins are to be eluted during the elution step. This is done by increasing the concentration of ions in the column, and in our model, increasing the concentration of ionic macro molecules $C_E$ at the inlet of the first section of the column. The proteins will desorb from the gel and then be transported in the mobile phase through the column. The concentration of $P$ and $A$ in the mobile phase from the very last section determines the chromatogram. The concentration profiles in the column before elution will of course influence the resulting chromatogram.

6.1 The simulation setup

The system is simulated using Simulink, and the simulation setup is shown in Figure 12.

The column is divided into eight sections. Four sections forms a simulation unit and thus the setup consists of two such units. Each unit has three inputs and three outputs: the mobile phase concentrations of protein $A$ and $P$, and the ionic macro molecules $E$. The outputs of the first simulation unit is connected to the inputs of the second simulation unit. This is also done between the four sections inside each unit. There are also outputs for concentration measures from the internal sections in each unit.

During the adsorption step there is a nonzero input of sample concentration of protein $P$ and $A$. The input of ionic macro molecules is zero and the sample input is formed as a pulse with limited length.

After the adsorption phase the sample concentration input is set to zero and an increasing concentration gradient of ionic macro molecules serves as a input to the system. The chromatogram is the sum of mobile phase concentrations of protein $P$ and $A$ from the last section.

The physical column parameters used in the simulations are shown in table 1. It is here assumed that the mobile phase volume is 30 % of the total volume. The column parameters are not taken from a real IEC process. Especially the flow rate $q$ is low compared to a real process [1]. The reader should keep in mind that the purpose of the simulations is to study the main dynamics of the process, and therefore the parameters are chosen arbitrary. The simulated model does not take the geometrical dimension of the column (in terms of length and width) into account. Instead it is the volume of each section that is considered. Thus, a column that is very long and narrow equals a column that is very short and wide (if the volumes are equal).
**Figure 12** The simulation setup with two simulation units. The column is divided into 8 sections with four sections in each simulation unit.

**Table 1** Column parameters used in simulation.

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total Column Volume [ml]</td>
<td>8.0</td>
</tr>
<tr>
<td>Number of Sections</td>
<td>8</td>
</tr>
<tr>
<td>$V_m$ [ml]</td>
<td>0.3</td>
</tr>
<tr>
<td>$Q_{av}$ [mg/ml gel]</td>
<td>3.0</td>
</tr>
<tr>
<td>$q$ [ml/s]</td>
<td>0.08</td>
</tr>
</tbody>
</table>

**6.2 An introducing simulation example**

In this section we will simulate the model using two different sets of constants of reaction velocity. We will compare the resulting chromatograms and explain the differences that occur. In latter sections we will in more detail investigate how the shape of the chromatogram are affected when changing the process parameters. In both simulations an input concentration of 1.0 mg/ml was used for both protein $P$ and $A$ during adsorption. The concentration pulse duration was 50 s. The elution gradient was raised from 0 to 8.0 mg/ml of ionic macro molecules $E$ in 300 s.

In the first simulation, the following values were used for the reaction velocity constants: $k_1 = 0.3$, $k_2 = 0.05$, $k_3 = 0.05$ and $k_4 = 0.3$. Thus, protein $P$ binds more easily to the gel than protein $A$. It can also be viewed as if protein $A$ is not good at desorbing ionic macro molecules. The concentration changes
during adsorption and the final concentration gradients after adsorption are shown in Figure 13. The final concentration gradients have occurred when the system becomes stationary, and this occurs about 100 s after the sample application.

Since protein $P$ binds easily to the gel, the concentration of bound $P$ is high at the beginning of the column. By opposite reason, concentration of protein $A$ is low near the inlet. When the concentration of bound $P$ increases, the concentration of $P$ in the mobile phase decreases. Thus, the concentration gradient for protein $P$ decreases through the column. There are less protein molecules $P$ in the mobile phase that can be adsorbed to the gel. When the concentration of protein $P$ in the mobile phase decreases, the competition between protein $P$ and $A$ for binding sites also decreases, and thus it is more easy for protein $A$ to bind to the gel. The concentration gradient for protein $A$ is therefore positive from the beginning of the column, and when most of protein $A$ is already bound, it becomes close to zero at the end of the column. One can also see that some protein $A$ is not at all bound to the column.

The concentration peaks from protein $P$ and $A$ from the outlet of the column and the resulting chromatogram during elution is shown in Figure 14. The elution gradient shown is the concentration of ionic macro molecules $E$ in the mobile phase in the last section in our model. Thus the shown gradient is not completely linear due to interaction with protein $P$ and $A$. The gradient is also scaled by a factor 0.05 so it can be presented with the chromatogram.

The concentration peak from protein $A$ is narrow and high while the concentration peak from protein $P$ wide and tailed. The reason is that $A$ desorbs more easily than $P$, when the concentration of ionic macro molecules $E$ is
raised. In the resulting chromatogram one can see that the two peaks are separated and the resolution is therefore rather good.

In the second simulation, the following values were used for the reaction velocity constants: $k_1 = 0.3$, $k_2 = 0.05$, $k_3 = 0.05$ and $k_4 = 0.3$. Thus, protein $P$ still binds more easily to the gel than protein $A$, but in this case protein $A$ is more inclined to bind to the gel than it was in the first simulation. The concentration changes during adsorption, and the final concentration gradients after adsorption for the second simulation are shown in Figure 15.

The main difference between the first and second simulation, regarding the concentration gradients, is that the concentration of bound protein $A$ is on average twice as high as in the first simulation. This is of course due to that protein $A$ binds more easily to gel in the second simulation. The concentration of bound protein $P$ differs slightly because of the competition for the sites. One can notice that some protein $P$ is not bound to the gel, and is therefore wasted.

The concentration peaks from protein $P$ and $A$ from the outlet of the column and the resulting chromatogram during elution is shown in Figure 16.

When studying the chromatogram in the second simulation it is hard to tell that there are two peaks forming the chromatogram. The resolution between the two peaks is in this case not satisfied. The reason is that the concentration peak from protein $A$ has been wider and more tailed, and that the peak maxima has been delayed compared to the first simulation. The peak shape from protein $P$ has just slightly been changed compared to the first simulation.
Figure 15  Adsorption simulation. Concentration changes during adsorption and final concentration gradients after adsorption. Reaction velocity constants: $k_1 = 0.3$, $k_2 = 0.05$, $k_3 = 0.05$ and $k_4 = 0.05$.

Figure 16  Elution simulation. Concentration peaks from protein A (dotted) and P (solid) and the resulting chromatogram during elution. Reaction velocity constants: $k_1 = 0.3$, $k_2 = 0.05$, $k_3 = 0.05$ and $k_4 = 0.05$. 

22
6.3 Altering the constants of reaction velocity

In this section we will consider the situation when there is only one protein involved in the IEC process, and investigate how the chromatogram is affected when changing the constants of reaction velocity. This is done by letting one of the two input protein concentrations be set to zero in the simulation setup. In this simulation protein $P$ serves as the only input to the system, and the constants of reaction velocity of protein $P$ ($k_1$ and $k_2$) are altered. The ratio $k_1/k_2$ was changed from 0.1 to 10 according to table 2. In the simulation an input concentration of 1.0 mg/ml was applied to the column inlet for 50 s. During elution, the elution gradient was raised from 0 to 8.0 mg/ml of ionic macro molecules $E$ in 300 s. The resulting chromatograms are shown in Figure 17.

<table>
<thead>
<tr>
<th>$k_1$</th>
<th>$k_2$</th>
<th>Ratio: $k_1/k_2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.05</td>
<td>0.5</td>
<td>0.1</td>
</tr>
<tr>
<td>0.05</td>
<td>0.25</td>
<td>0.2</td>
</tr>
<tr>
<td>0.05</td>
<td>0.05</td>
<td>1.0</td>
</tr>
<tr>
<td>0.25</td>
<td>0.05</td>
<td>5.0</td>
</tr>
<tr>
<td>0.5</td>
<td>0.05</td>
<td>10.0</td>
</tr>
</tbody>
</table>

Figure 17 Resulting five chromatograms when changing the constants of reaction velocity as in table 2. The chromatogram at the very left in the figure results from when $k_1/k_2 = 0.1$. The chromatogram to the very right in the figure results from when $k_1/k_2 = 10.0$.

The chromatogram at the very left in the figure results from when $k_1/k_2 = 0.1$. The chromatogram to the very right in the figure results from when $k_1/k_2 = 10.0$. When the protein binds easily to the gel, it will result in a peak shape that is wide and tailed. In this case the peak shape is best described by the EMG function introduced in section 3.6. When the protein
does not bind easily to the gel it will result in a peak shape that is narrow and high and thus well described by the Gaussian function.

6.4 Competition between proteins when altering the input concentrations

Consider a chromatogram that consists of more than one concentration peak. Under what circumstances is it possible to recreate the chromatogram by adding the chromatograms from individual batches? To clarify the question, consider the situation when there are two proteins to be separated with IEC technique. In the first case the column is loaded with the two proteins at the same time, and then eluted, giving a chromatogram with two concentration peak components. This is of course the normal case. In the second case the same column is first loaded with one protein and eluted, and then we do the same procedure for the other protein. The latter case gives us two chromatograms, one for each protein. Under what circumstances will the sum of the two chromatograms from the second case equal the chromatogram components from the first case? Under these circumstances superposition can be used to build up the chromatogram, and we do not have to consider the competition between the two proteins.

To answer the question, two simulations of the situation mentioned above were performed using our model. The difference between the two simulations is the input concentration of protein $P$ and $A$. The values of the constants of reaction velocity in both simulations were $k_1 = 0.3$, $k_2 = 0.05$, $k_3 = 0.65$ and $k_4 = 0.3$. Thus protein $P$ binds more easily to the gel than protein $A$. In the first simulation an input concentration of 1.0 mg/ml were used for both protein $P$ and $A$ during adsorption. The sample application time was again 50 s. In the second simulation the input concentration of each protein was 0.1 mg/ml. The elution gradient was raised from 0 to 8.0 mg/ml of ionic macro molecules $E$ in 300 s. The result is shown in Figure 18.

The concentration peaks from the individual batches are dotted and the concentration peaks from the common batch are solid in Figure 18. The concentration peaks to the very left in Figure 18 results from protein $A$ since it is loose bound to the gel.

When using an input concentration of 1.0 mg/ml one can see in Figure 18 that the use of superposition cannot be used. The peak shapes from protein $P$ are almost equal and the difference between them differs only slightly. It is the peak shape from protein $A$ that differs. The reason is that we get different concentration gradients of protein $A$ in the column during adsorption. When both the proteins are applied to the column at the same time, protein $A$ finds it hard to bind to the gel since protein $P$ is a strong binder. When it is only protein $A$ that is applied to the column, it does not have to compete with any other proteins for binding sites. This will lead to a higher final concentration gradient for protein $A$ after adsorption. In this simulation protein $P$ is the superior protein, and thus, the peak shape from protein $P$ is almost the same in both batches. The available capacity, $Q_{av} = 3.0$ mg/ml, is in the same numerical range as the input concentrations and therefore the competition for binding sites between the proteins is high.

When using an input concentration of 0.1 mg/ml one can see in Figure 18 that the use of superposition can be used. The peak shapes resulting from the individual batches and the common batch are almost identical for both proteins. For protein $P$ one cannot see any difference between the peak shapes. Due to the low input concentration there are a lot of sites available for both
Figure 18 The concentration peaks from the individual batches are dotted and the concentration peaks from the common batch are solid. Top: when the column is heavy loaded (input concentrations and available capacity in the same numerical range), the competition between protein $P$ and $A$ for binding sites is high, and the principle of superposition cannot be used. Bottom: when the column is not heavy loaded, the competition between protein $P$ and $A$ for binding sites is low, and the principle of superposition be used.

the proteins. The available capacity, $Q_{av} = 3.0$ mg/ml, is high compared to the input concentrations and therefore the competition for binding sites between the proteins is low.

The conclusion from the simulation is: when the input concentration of the applied proteins is low compared to the available capacity $Q_{av}$, then the competition for binding sites between the proteins is low. The principle of superposition can then be used.

6.5 Altering the length of the concentration pulse

In this section we will study the impact on the chromatogram when the length of the input concentration pulse is changed. We will see that the pulse time have a heavy impact on the resolution between two peaks. If the pulse time is too long, there is a risk that all binding sites are occupied. This will lead to that some amount of the wanted protein $P$, will not be bound to the gel, and therefore wasted. Thus, there is an upper limit concerning the amount of protein $P$ that is applied to the column, due to that you do not want to waste any amount of protein $P$. The amount of protein that is applied to the column depends on the length of the concentration pulse, and the concentration in the input sample.

Two different pulse times, $t_p = 100$ s and $t_p = 25$ s, were examined. When using $t_p = 100$ s the column is heavy loaded compared to $t_p = 25$ s. The values of the constants of reaction velocity in the simulations were $k_1 = 0.3$, $k_2 = 0.5$, and $k_3 = 0.2$. The concentration of protein $P$ in the input is $C_{inP} = 0.5$ mg/ml, and the concentration of protein $A$ in the input is $C_{inA} = 0.1$ mg/ml. The available capacity of the column is $Q_{av} = 3.0$ mg/ml, and the column is heavy loaded.
$k_2 = 0.05$, $k_3 = 0.05$ and $k_4 = 0.3$ (protein $P$ binds stronger to the gel). The elution gradient was raised from 0 to 8.0 mg/ml of ionic macro molecules $E$ in 300 s. The chromatograms are shown in Figure 19.

At the top of Figure 19 the pulse time was $t_p = 100$ s, and the input concentration for each protein was 1.0 mg/ml during adsorption. The resolution is poor, and one can hardly imagine that the chromatogram consists of two peaks. Due to the long pulse time the column is overloaded with protein $P$. After the adsorption step, the concentration gradient from protein $P$ is equally distributed throughout the column. The same yields for protein $A$ but the amount of protein $P$ is about five times more than the amount of protein $A$. During elution the peak shape from protein $P$ will be wider and more tailed than usual, due to the large amount of the protein that had to be eluted. This will result in a chromatogram with poor resolution.

The chromatogram in the middle of Figure 19 was simulated using $t_p = 25$ s and an input concentration of 1.0 mg/ml for each protein. This time there was no overloading of protein $P$. In the resulting chromatogram the two peaks are well separated, giving good resolution.

It is of course the total amount of protein applied to the column that is the crucial factor in these simulations. An overloaded column gives separation problems. By having a lower input concentration the pulse time can be made longer. The chromatogram at the bottom of Figure 19 is formed in this way. The input concentration was 0.1 mg/ml and $t_p = 100$ s. The peaks are well separated and can easily be detected.

![Figure 19](image)

**Figure 19** Impact on the chromatogram when changing the length of the concentration pulse. Top: the column is overloaded when $t_p = 100$ s, and the resolution is poor. Middle: when using $t_p = 25$ s the column is not overloaded, giving good resolution. Bottom: when using $t_p = 100$ s and an input concentration of 0.1 mg/ml the column is not overloaded, giving good resolution.

### 6.6 Changing slope of the elution gradient

The chromatogram is dependent on the slope of the elution gradient. A very steep gradient will force all substances that are loaded to the column, to be
desorbed approximately at the same time, leading to bad time resolution. On the other hand the total time of separation would be faster and the peak shapes sharper.

A long and shallow gradient will increase the resolution between the peaks. The total time of separation would be long and since it will take more time for each protein group to be completely desorbed and thus the peak shape will be wider. The discussion above concerning the gradient effects is intuitive and easy to understand.

In this section we will simulate our model to see if the arguments above hold. The column were loaded with protein $P$ and $A$ with an input concentration of 1.0 mg/ml for 50 s. The values of the constants of reaction velocity in the simulations were $k_1 = 0.3$, $k_2 = 0.05$, $k_3 = 0.05$ and $k_4 = 0.3$. During the elution step three different gradients were used to separate the proteins. The final value of the concentration of applied ionic macro molecules $E$ was 8.0 mg/ml and to get different gradient slopes the time to reach the final value were altered from 150 s to 450 s. The result is shown in Figure 20.

The elution gradients shown in Figure 20 are the measures of ionic macro molecules $E$ in the mobile phase in the last section in our model. Thus the shown gradients are not completely linear due to interaction with protein $P$ and $A$. The gradients are also scaled by a factor 0.05 so they can be presented with the chromatogram.

Figure 20 shows that:

- a short and steep gradient results in decreasing resolution (the difference between the peak maxima), sharper peaks shapes and shorter separation time.
- a long and shallow gradient results in increasing resolution, peak broadening and longer separation time.

Thus our model agree with the intuitive discussion above.

Figure 20  The elution step using three different gradients.
6.7 Non heterogeneous flow rate

The gel in the column is used for a large number of batches and to keep the performance of the column the gel has to be re-packed. This is often done periodically. In a new packed column the gel is heterogeneous distributed inside the column. After the column has been used for a while the gel might get unevenly distributed. This will lead to areas were cavity of gel is missing which in turn leads to that the flow rate differs from section to section.

In this section we will take a look at what happens to a single peak shape when we simulate our model using a non heterogeneous flow rate. To model this we let the actual column consist of a number of smaller columns and we simulate each one of the smaller columns using different flow rates. This can be seen as if the column is built up with tubes and that the flow rate in the tubes differs. We will compare a peak simulated with ”normal” flow rate with a peak simulated with an uneven flow rate. The total flow rate equals 0.08 ml/s in both cases.

![Flow Components Diagram]

**Figure 21** The total flow rate divided into five flow rates.

The column was divided into five tubes and the quantities of the different flow rates are shown in Figure 21. The flow rate used in previous sections was 0.08 ml/s and the sum of the individual flow rates equals 0.08 ml/s. The largest flow was 30% of total flow (one flow component), the second largest 20% of total flow (two flow components) and the smallest flow was 15% of total flow (two flow components). The flow components are shown in Figure 21. The reaction velocity constants were \( k_1 = 0.05 \) and \( k_2 = 0.05 \) (just one protein). The other parameters were the same as in the introducing example in Section 6.2. The resulting chromatogram is shown in Figure 22.

From Figure 22 one can that when using an uneven flow rate the peak shape get more wider than before. This is natural since the different flow rates tenders to make the peak shape more broader. The slow flow rates ”pulls” the peak to the left and the fast ones ”pulls” the peak to the right. There is also a difference in the peak area. This is due to that the amount of bound protein in the column after adsorption, differs in the two cases. When the ”normal” flow is used some amount of protein is not bound to the gel during adsorption, and thus some protein are wasted. When the non heterogeneous flow rate is used, the tubes associated with the smallest flow components, will almost bind all the amount of protein that passes in the tube. Therefore, the amount of bound protein is larger in the whole column when using the non heterogeneous flow, compared to when using the normal flow.

6.8 The number of sections

In all the simulations the column has been divided into 8 sections. We will here examine what happens if we increase the number of sections. If we double the
Figure 22  Peak shape when the flow rate is heterogeneous (solid) and the peak shape using an uneven flow rate (dashed). The peak shape is wider when using an uneven flow rate. The peak area differs, due to that the amount of bound protein is different in the two cases.

number of sections to 16 and compare the chromatogram when the number of sections is 8, and use the process parameters in the first introducing example, we can see a difference between these. This is shown in Figure 23.

The difference in Figure 23 is that the peak shapes when using 16 sections tenders to be more narrow than when using 8 sections. The different peak areas also differs but the main process dynamics still remains (the overall peak shape is the same).

6.9 The pooling problem: two simple approaches

We will in this section look at two different approaches of how to formulate the pooling strategy and briefly analyze them. To use our model we will investigate the problem of separating two different proteins. The quality, $Q$, is stated as the percentage fraction of the unwanted substance in the pool. Thus the quality number should be low. In the chromatogram there will be two concentration peaks and the wanted product is the concentration peak associated with the longest retention time. Therefore we just have to consider when to start to pool and we will refer to this point as $t_{start}$.

The pooling strategies (stated below) are assumed to have been developed in pilot scale where a lot of experiments have been done to optimize how to choose $t_{start}$. We further pretend that the input sample concentrations of the wanted and unwanted product were 1.0 mg/ml in all pilot experiments. We also assume that if the quality number exceeds a certain value, the pool sample cannot be used and will be wasted. Using this scenario an interesting question arises: what happens to the quality of the pool if the input sample concentration of the unwanted substance is not the same as in the pilot experiments?
The two pooling strategies are:

- The first pooling strategy is stated as a constant time delay taken from when the elution gradient is applied to the column input. In this case \( t_{\text{start}} = 100 \text{ s} \) (thus the time delay equals \( t_{\text{start}} \)).

- The second pooling strategy states: \( t_{\text{start}} \) is taken when the minimum between the two concentration peaks appears in the chromatogram.

The two strategies are assumed to be developed under the circumstance that the input sample concentration of the wanted protein \( P \) and the unwanted protein \( A \) are both 1.0 mg/ml. If the quality number exceeds 2.5% the pool sample will be wasted. Two simulations were performed to analyze the strategies. The constants of reaction velocity were \( k_1 = 0.3 \), \( k_2 = 0.05 \), \( k_3 = 0.05 \) and \( k_4 = 0.15 \) (again, protein \( P \) binds more easily to the gel than protein \( A \)). In the first simulation the input sample concentration was the same as in the pilot scale, and in the second simulation the input concentration of the unwanted protein \( A \) was 1.2 mg/ml. The interesting question is how the quality of the pool is affected when raising the input concentration of protein \( A \).

The result of the two simulations are shown in Figure 24 and the quality numbers and each substance amount are shown in table 3.

Each substance amount is calculated as the area under the individual concentration curve from \( t = t_{\text{start}} \) to \( t = 400 \text{ s} \). The second strategy gave \( t_{\text{start}} = 102 \text{ s} \) in the first simulation and \( t_{\text{start}} = 106 \text{ s} \) in the second. The reason why the amount of protein \( A \) is so small compared to the amount of \( P \) is that a lot of protein \( A \) does not bind to the gel during the adsorption phase.

By studying the quality numbers gives that the second strategy is the best. One should remember that the the quality number does not take the amount of
Figure 24  The two pooling strategies. Top: the resulting chromatogram and the individual concentration peaks using the same input sample concentration as in the pilot scale. Middle: the input sample concentration of protein A is not the same as in the pilot scale. Bottom: combination of the two chromatograms. \( t_{\text{start}} \) is marked with a solid line using the first strategy and dashed lines using the second. The pool is taken from \( t_{\text{start}} \) to \( t = 400 \) s and the wanted protein is the concentration peak to the right.

Table 3  The input sample concentrations, choice of strategy (and \( t_{\text{start}} \)), amount of protein P and A in the pool, and the quality number Q.

<table>
<thead>
<tr>
<th>( P_0 )</th>
<th>( A_0 )</th>
<th>Strategy</th>
<th>Amount P</th>
<th>Amount A</th>
<th>Q</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.0</td>
<td>1.0</td>
<td>I, ( t_{\text{start}} = 100 )</td>
<td>44.57</td>
<td>1.04</td>
<td>2.3%</td>
</tr>
<tr>
<td>1.0</td>
<td>1.2</td>
<td>I, ( t_{\text{start}} = 100 )</td>
<td>44.28</td>
<td>1.26</td>
<td>2.8%</td>
</tr>
<tr>
<td>1.0</td>
<td>1.0</td>
<td>II, ( t_{\text{start}} = 102 )</td>
<td>44.14</td>
<td>0.86</td>
<td>1.9%</td>
</tr>
<tr>
<td>1.0</td>
<td>1.2</td>
<td>II, ( t_{\text{start}} = 106 )</td>
<td>42.85</td>
<td>0.70</td>
<td>1.6%</td>
</tr>
</tbody>
</table>

wanted product that is wasted in consideration (product that is not included in the pool).

When using the first strategy, the quality number will be larger when the input sample concentration is larger than in the pilot scale. When the input concentration of protein A is 1.2 mg/ml the quality number is 2.8%. Since the quality number exceeds 2.5% the pool sample will be wasted. Thus the first strategy is in this case too static to handle process variations concerning the input sample concentration.

The second strategy is more adaptive than the first strategy, but it has a disadvantage. For some input sample concentrations there is a risk that the minimum between the concentration peaks does not appear. If the minimum does not appear it will be difficult to follow the pooling rule from the second strategy.
A better strategy than the ones discussed above, is a strategy that takes the value of \(Q\) into account when choosing \(t_{\text{start}}\). This strategy could be stated as: choose \(t_{\text{start}}\) in order to maximize the amount of product and limit the quality number \(Q\) to be less than 2.5\%. Using this strategy, no batches have to be wasted. If strategy is applied, it will give \(t_{\text{start}} = 98\) s (when the both input concentrations are 1.0 mg/ml), and \(t_{\text{start}} = 102\) s (when the input concentration of protein A is 1.2 mg/ml).

6.10 Conclusions from the simulations

In this chapter we have been simulating our model. The model that we are simulating is a simplification of reality and the purpose of the model is that it must capture the main dynamics of the IEC process. The most straightforward approach would be to treat the sample injection in the column as a diffusion problem. This would lead to differential equations with partial derivates in space and time. In our model we have reduced the partial derivates in space by chopping the column into a number of sections, and we just treat one diffusion component by letting the flow have just one direction. By reducing the partial derivates in space the model is more easy to understand. The simulation times will also be magnitudes of orders faster using the simple model without the partial derivates in space, which of course is preferable. Although the model is simplified it is still useful for our purposes.

Another complication is that the real IEC process has in principal two time constants: the time constant describing the flow mechanism in the column and the time constant describing the chemical binding between the stationary and the mobile phase. It is easy to understand that the chemical time constant are much smaller than the flow rate time constant. The simulation time for a system is in general determined by the smallest time constant, and a small time constant will lead to long simulation times. Thus, in our model it would be preferable not to use the very small chemical time constant to avoid long simulation times. However, in our model we do not treat the interaction between the stationary and mobile phase as a pure chemical reaction but instead as a slow process when a protein binds to or desorbs from the gel. The time constant for the chemical reaction are determined by the constant of reaction velocity and we have the freedom to choose it to be in the same numerical range as the flow rate time constant.

A question that always arise when modeling a physical process is: does the model describe the real process and can we validate it? The simulation results does in general agree with the experience from professional use of IEC in practical applications [1]. The use of the simulations of the model in this thesis is to learn more about IEC and to understand the factors that affect the process dynamics. The model developed does not in detail predict the output of an IEC process, but it captures the main process dynamics. No validation is made for the model with real data.

One approach to extend the model is to consider the backward diffusion and the diffusion perpendicular to the main flow. The backward diffusion can modeled by letting a fraction of the main flow have opposite direction in each section. The perpendicular flow can be modeled by dividing the column into sections and tubes (see section 6.7) and let a flow propagate between the tubes. The net flow component would equal the inflow to the column.
7. Data collection and data preparation

To be able to characterize the process variations and to develop a new pooling algorithm according to the problem formulation (see section 4), a number of chromatogram from an IEC process had to be analyzed. Pharmacia & Upjohn (PnU) kindly provided the process data needed, and the data came from a product line IEC process. Analyzes concerning the concentration of the wanted product and a number of unwanted substances, and the start and stop time for the pool (in terms of the elution gradient) were also provided. These analyzes came from both the input concentrations, from now on sample input, and the pooled sample, from now on eluate. The age of the column was stated in terms of cycles from previous gel re-packing. This section describes the data collection and the data preparation.

7.1 Data collection

The chromatograms provided were printed on paper and to analyze them with computer tools they first had to be transformed to function coordinates, and then saved electronically. The first step was to use a scanner to read the chromatograms and thus transforming the paper print to a image in computer format. Additional software (UN-SCAN-IT from Silk Scientific and licensed to Pharmacia and Upjohn) were then used to decode the function curve (the chromatogram) into pairs of function coordinates. The analysis data from the sample input and the eluate were already saved electronically by PnU. A total number of 51 chromatograms were collected and analyzed.

7.2 Data preparation

To analyze and to compare the chromatograms they need to have a common reference point in time, e.g., the time when the elution gradient is applied to the column. This was not the case with the chromatograms collected. Anyhow, at every start and stop time for the pool, the measurements of the elution gradient (measured in conductivity, ion density) were written on the chromatogram. Since the start and stop time were stated in terms of the elution gradient it is natural to use the elution gradient as a time scale. This is only correct if the elution gradient has the same slope in every single batch.

When the chromatogram curves were decoded into pairs of function coordinates, the samples were not taken equidistantly. To make the computer analyzes more easy the function coordinates were made equidistant (respecting the time scale) with cubic spline interpolation. The cubic spline interpolation fits a third order polynomial between two aligned coordinates, and the interpolation resulted in 27 function coordinates for each chromatogram. The part of the chromatogram that was used in the further analyzes, begins outside the start of the pool and ends outside the stop of the pool. Thus, some data points outside the pool were included in the data set for the chromatograms. The reason why not the whole chromatogram is included, is that the further analyzes consider the eluate concentrations, and the information about these lies in the pool, and in the boundaries of the pool.

All the chromatograms and analysis data were transformed to be used in the Matlab environment.
8. Data analyzes

In this section the data collected is analyzed. The objective is to characterize the process variations in order to make the pooling procedure more robust. The practical experience is that the process variations from batch to batch is mainly due to concentration variations in the applied sample. An old column, stated in terms of cycles from previous re-packing, will also cause process variations. In the analyzes we will consider the chromatograms, the sample input analyzes, the eluate analyzes, the start and stop time for the pool, and the age of the column.

In the eluate and sample input analyzes the concentration of the product and a number of unwanted substances is given. We will concentrate on the product concentration, an unwanted substance associated with the beginning of the pool and two unwanted substances associated with the end of the pool. From now on we refer to them as: \( P \) (product), \( A \) (unwanted substance in the beginning of the pool), \( B \) and \( C \) (unwanted substances at the end of the pool). The concentration of \( C \) is not included in the sample input analyzes. The concentration of \( C \) is considered to be the most nasty substance in the pool.

A chromatogram that illustrates the discussion above is shown in Figure 25. This is an artificial chromatogram and not taken from a real IEC process. The chromatogram shows were the product and the unwanted substances appears.

![Chromatogram Illustration](image)

**Figure 25** An illustrating chromatogram. The sum of the individual concentration peaks (product concentration peak (dashed), the impurity peaks (dotted)) gives the resulting chromatogram (solid). The start and the stop time of the pool are marked with solid lines.

Since the data collected is taken from an industrial process, they are treated as confidential information. Thus, neither chromatogram from the real process nor the data concerning the eluate and sample input are shown in this thesis.
The axes of data plots, concerning the process data, are therefore normalized.

8.1 The age of the column
We will here investigate how the age of the column effects the chromatogram and the eluate concentrations. A re-packed column is used for a specified number of batches before it is re-packed again. The re-packing of the column is thus done with a certain re-packing period. An interesting question then arises: is it possible that the column during this period will be degraded therefore decreasing the performance of the column? If the performance is decreased the product concentration $P$ should be lower and the concentration of the unwanted substances $A$, $B$ and $C$ should be higher. To answer the question a plot of how the eluate concentrations are changed during a re-packing period is shown in Figure 26. Figure 26 shows that the performance of the column does not seem to be decreased during such a period. The behaviour of the eluate concentration is stochastic.

![Figure 26](image)

Figure 26  The eluate concentrations during a re-packing period.

How is the chromatogram effected? From a visual inspection of the chromatograms during a period it is hard to find a pattern that is dependent on the age of column. Again, the shape of the chromatograms seems to be stochastic.

The conclusion is: the performance of the column is not decreased during a period. Therefore, the re-packing period is low enough to keep the performance of the column. The re-packing period could be longer but the data collected from the process cannot verify this.

8.2 A model for the eluate concentrations
The information about the quality of the pool is stated in the eluate concentrations. In a high quality pool sample the product concentration $P$ should be
as high as possible, and the impurity concentrations should not exceed certain maximum values. It would be preferable to have a model that predicts the eluate concentrations.

A straightforward approach to such a model is:

\[ y = F(x, u, z) \]  

(15)

Here the eluate concentrations are represented in \( y \), the chromatogram data points are represented in \( z \), the start and stop time of the pool is represented in \( u \), and the sample input concentrations are represented in \( z \). Equation (15) states that \( y \) is a function of \( z, u \) and \( z \).

Is the model valid? It makes sense to state that the eluate concentrations depend on the shape of the chromatogram \( (z) \). Naturally the eluate concentrations also depend on the pooling strategy used \( (u) \). Finally, since the sample input concentrations \( z \) serve as an input to the system the eluate concentration will also depend on them. It is reasonable to assume, that \( y \) is a function of \( z, u \) and \( z \), and the function is probably linear for small variations in \( u \) and \( z \).

Below we will analyze the dependence between the eluate concentrations and the start and stop time of the pool, and the sample input concentrations.

The only parameters which can be used to control the eluate concentrations \( y \) are the start and stop times of the pool \( u \). The other parameters, \( z \) and \( z \), are determined by the system itself, and they are not effected by the choice of \( u \). Thus, optimizing the eluate concentrations equals: find the values of the start and stop time of the pool which make the product concentration as high as possible, and which in the same time ensures that the impurity concentrations do not exceed certain values. The optimization problem is easy to state but it is hard to solve if the function \( F \) is not known.

Is it possible to find the function \( F \), or at least find an approximation of it? If that is the case we can use \( F \) to solve the optimization problem by choosing the right start and stop time of the pool. To answer the question we can investigate how the eluate concentrations depend on \( u \). This is shown in Figure 27. Figure 27 shows that the data collected does not show any, or at least no strong, correlation between \( y \) and \( u \). Instead they seem to appear in an unpredictable order.

The intuition is that the eluate concentrations should be correlated with the start and stop time of the pool. If the start time of the pool is small we would expect that the impurity concentration of \( A \) to be high, and vice versa for the stop time of the pool and the impurity concentrations of \( B \) and \( C \). Why can we not see this correlation in the data collected?

In the data collected the same pooling strategy has been used in all batches. Thus, the data is collected "under closed loop control". To get information from a system, there has to be enough excitation. Since the pooling strategy is the same in all batches there is a lack of excitation, and thus the information needed is not provided.

There is also an obvious reason why one would not like to excite the system regarding the start and stop time of the pool. Since it is an industrial process the price of changing the start and stop time back and forth, would be that some batches have to be wasted because of high impurity concentrations.

Conclusion: from the data collected it is not possible to find a model, that can be used to determine start and stop time of the pool, so that the eluate concentrations can be optimized.
Figure 27  The eluate concentration dependence on start and stop time of the pool. The start and stop time are measured in terms of conductivity.

Figure 28  The eluate concentration dependence on the sample input concentrations.

As stated above it assumed that the eluate concentrations depend on the sample input concentrations, since they serve as an input to the system. Figure 28 shows the correlation between the eluate concentrations and the sample input concentrations. Since there are no measurements of impurity C in the sample input analyzes, the impurity concentration C in the eluate is plotted
versus the impurity concentration \( B \) in the sample input. This is done because impurity \( B \) and \( C \) are both associated with the end of the pool.

Figure 28 shows that there is strong correlation between the sample input and the eluate concerning the product concentration \( P \) and impurity concentration \( B \). It seems as if both product \( P \) and impurity \( B \) lies well inside the pool, while impurity \( A \) and \( C \) lies in the boundary of the pool. If we refer to Figure 25, the concentration peak for impurity \( B \) is the impurity peak in the middle (concentration peak for impurity \( A \) to the left, and concentration peak for impurity \( C \) to the right).

Since the correlation between the sample input and eluate in general is strong, the sample input concentrations can be used to predict the eluate concentrations. We will see in next section how this is used to predict the eluate concentrations.

### 8.3 Multiple Linear Regression models

One approach to determine the eluate concentrations \( y \) is the use of multiple linear regression (MLR) models. We will investigate three different MLR models and see how well they determine the eluate concentration \( y \). The models differ in the way of choosing the regression variables. The regression will contain the sample input concentrations \( z \) and the chromatogram data points \( z \).

Following notation is used. The eluate concentrations are represented in a four dimension row vector \( y \),

\[
y = \begin{bmatrix} y_P & y_A & y_B & y_C \end{bmatrix} \tag{16}
\]

the sample input concentrations are represented in a three dimension row vector \( z \),

\[
z = \begin{bmatrix} z_P & z_A & z_B \end{bmatrix} \tag{17}
\]

and the chromatogram data points are represented in a \( k \) dimension row vector \( z \)

\[
z = \begin{bmatrix} z_1 & z_2 & \cdots & z_k \end{bmatrix}. \tag{18}
\]

\( k \) represents the number of data points in a chromatogram. An arbitrary MLR model is written as:

\[
y = \varphi \theta + b + e \tag{19}
\]

where \( y \) denotes the observed variable (the variable to predict), \( \varphi \) denotes the regression variables, and \( \theta \) and \( b \) are the parameters of the model to be determined. \( e \) is the residue or the error that the model cannot predict. Thus, the MLR model assumes that the observed variable \( y \) is linearly dependent on the regression variables \( \varphi \).

**Model I** Our first MLR model the regression variables contains only the sample input concentrations.

\[
\begin{cases}
y = \varphi_1 \theta_1 + b_1 + e \\
\varphi_1 = z
\end{cases} \tag{20}
\]

In this case \( \theta_1 \) is a three times four matrix, since the columns in \( y \) equals four and the columns in \( \varphi_1 \) equals three.
Model II  In the second MLR model we add some of the chromatogram data points to the regression variables. The reason is that we would like to see if the predictions of \( y \) are improved by adding more data, concerning the chromatogram, to the regression variables.

\[
\begin{align*}
\begin{cases}
y = \varphi_2 \theta_2 + b_2 + e \\
\varphi_2 = [x, x_{2/3}]
\end{cases}
\end{align*}
\] (21)

The notation \( x_{2/3} \) indicates that 67% of the first chromatogram data points are included in the regression variables. In this case \( \theta_2 \) is a three plus \( k \cdot \frac{2}{3} \) times four matrix. Thus there are a lot of parameters that determines the model. By the use of principal component analysis (PCA) [4], \( \varphi_2 \) can be replaced with another regression variable, with fewer column elements than in \( \varphi_2 \). The new regression variable is denoted \( t_2 \). Since there are fewer column elements in \( t_2 \), there are fewer parameters to be determined. The column elements in \( t_2 \) are linear combinations of the column elements in \( \varphi_2 \). Using PCA the second model can be written:

\[
\begin{align*}
\begin{cases}
y = t_2 Q_2 + b_2 + e \\
t_2 = \varphi_2 P_2 \\
\varphi_2 = [x, x_{2/3}]
\end{cases}
\end{align*}
\] (22)

The matrix \( P \) is determined using PCA.

Model III  In the third model the regression variables contain the sample input data \( z \) and all chromatogram data points. As for the second model we use PCA to have fewer parameters to determine.

\[
\begin{align*}
\begin{cases}
y = t_3 Q_3 + b_3 + e \\
t_3 = \varphi_3 P_3 \\
\varphi_3 = [x, z]
\end{cases}
\end{align*}
\] (23)

Estimation of the model parameters  To estimate the model parameters, 25 chromatograms where chosen to serve as the calibration set. The MLR models can then be written:

\[ Y = \Phi \theta + B + E \] (24)

Here \( Y \) consists of 25 rows of the eluate concentrations and \( \Phi \) consists of 25 rows of the regression variables. Using PCA \( \Phi \) has been replaced by \( T \) in the second and third model.

\[ Y = T Q + B + E \] (25)

When using PCA in the second and in the third model one has to choose the number of columns in \( T \). The following strategy was used to choose the
number of columns. Assume \( n \) columns in \( T \). If the prediction with \( n \) columns in \( T \) is as good as the prediction with \( n + 1 \) columns in \( T \), then use \( n \) columns. If not, add the column in \( T \) and repeat. The evaluation of "as good as" is done in the section below (Testing the models). This strategy gave \( n = 7 \) for the second model, and \( n = 9 \) in the third model. The dimension of \( Q \) in the second model is therefore seven times four and, nine times four in the third model. The same value for \( n \) was used for each concentration \( P, A, B \) and \( C \).

The parameters of the models where then estimated in the sense of least squares. Before making the least square estimates and the PCA, the calibration set was centered around its mean value.

**Testing the models** To evaluate the models, 10 new chromatograms were chosen to serve as a test set. The models were then used to predict the eluate concentrations. To compare how well the models predicted the eluate concentrations, the predicted value and the true value were compared for each concentration in each model. To measure this, the following number was calculated for each concentration in each model:

\[
R_x = \frac{1}{10} \sum_{i=1}^{10} \frac{|y_i^x - \hat{y}_i^x|}{y_i^x}, \quad x = P, A, B, C
\]  

(26)

Here \( y \) denotes the true value and \( \hat{y} \) denotes the predicted value. The value \( R \) should of course be as small as possible.

**Predicting the product concentration \( y_P \)** The result when predicting the product concentration \( y_P \) is shown in Figure 29 and in table 4. The plots to the left in Figure 29 show the predicted values (marked with +) and the true values (marked with o). The plots to the right show the relative residues, that is \( (y - \hat{y})/y \).

<table>
<thead>
<tr>
<th>Model</th>
<th>( R_P )</th>
<th>( R_A )</th>
<th>( R_B )</th>
<th>( R_C )</th>
</tr>
</thead>
<tbody>
<tr>
<td>I</td>
<td>0.0148</td>
<td>0.2714</td>
<td>0.0272</td>
<td>0.5252</td>
</tr>
<tr>
<td>II</td>
<td>0.0142</td>
<td>0.2205</td>
<td>0.0427</td>
<td>0.5839</td>
</tr>
<tr>
<td>III</td>
<td>0.0133</td>
<td>0.3232</td>
<td>0.0440</td>
<td>0.6158</td>
</tr>
</tbody>
</table>

From table 4 one can see that on average the prediction gets better as we put more information in the regression variables, although the difference is small. This is also shown in Figure 29. Almost all predictions in the models hit the true value with a difference that is less than 5% and even 3% using model three (does not hold for prediction number four).

**Predicting the impurity concentration \( y_A \)** The result when predicting the impurity concentration \( y_A \) is shown in Figure 30 and in table 4.

Figure 30 shows that the overall prediction is no good. The relative residues are too high to be acceptable. The difference between the true value and the predicted is sometimes more than 50%. Thus, none of the three models can be used to predict an acceptable value for the impurity concentration \( A \).
Figure 29  Plots to the left: the predicted values + and the true values o. The solid line is the mean value of the product concentration in the calibration set, and the dotted lines are the mean values +/− 10%. Plots to the right: the relative residues.

Figure 30  Plots to the left: the predicted values + and the true values o. The solid line is the mean value of the impurity concentration A in the calibration set, and the dotted lines are the mean values +/− 10%. Plots to the right: the relative residues.

Predicting the impurity concentration $y_B$  The result when predicting the impurity concentration $y_B$ is shown in Figure 31 and in table 4.

The first model predicts the impurity concentration $B$ best. This is due to the strong correlation between the impurity concentration $B$ in the sample input and the eluate analyzes. When adding the first part of the chromatogram
the prediction is not improved. This is not strange since the information about impurity concentration $B$ is expected to be in the end of the chromatogram. When adding all the chromatogram data points the prediction is again not improved. Thus it is hard to tell how the new information in the chromatogram is correlated to the impurity concentration $B$.

The overall prediction is fairly good. Most of the predictions lies in a 5% range from the true value. Thus, the models can be used to predict the impurity concentration $B$.

**Predicting the impurity concentration $y_C$**  The result when predicting the impurity concentration $y_C$ is shown in Figure 32 and in table 4.

As with impurity concentration $A$ the models cannot be used to estimate the impurity concentration $C$. The predictions are poor and the relative residues are too high.

Adding the chromatogram data points does not improve the predictions.

**Extension of the regression variables**  In our MLR models we have not included the information of the start and stop time of the pool. This is not done since the correlation between the start and stop time of the pool and the eluate concentration is very weak. What happens if we extend the regression variables to include this information? To test this we extended the regression variables in all three models with this information, but the predictions were not improved. This again states that from our data, it is hard to develop a model of how the eluate concentrations depends on the start and stop time of the pool.
8.4 Conclusions from the data analyzes

In this chapter we have analyzed the data collected. To investigate what effects the eluate concentrations, MLR models has been developed.

The effects from the age of the column has also been investigated briefly. The result showed that it cannot be stated that the performance of the column is decreased during a re-packing period. Therefore the re-packing period is low enough to ensure that the column keeps up with performance.

The approach to find a model of how the eluate concentrations depend on the start and stop of the pool, did not succeed. The reason for this was the lack excitation of the system, due to that the same pooling strategy has been used in all batches.

The MLR models showed that the product concentration $P$, and the impurity concentration $B$ in the eluate, can be fairly well predicted. This does not hold for impurity concentrations $A$ and $C$. It is unfortunate that we cannot predict the impurity concentration $C$, since this is the most nasty impurity. If the system could provide information about impurity concentration $C$ in advance, the information could be used to adjust the stop time of the pool.

One thing is confusing using the MLR models. It is intuitive that when the chromatogram data points are added to the regression variable, more information about the system is gathered, but the prediction does not improve. We expect that information about the eluate concentration can be found in the chromatograms. The MLR models state that this information does not improve the predictions. Below, a discussion follows of why we cannot use the information in the chromatograms to make better predictions.

When comparing chromatograms to each other it is necessary that they have the same time reference. This is of course a crucial factor when the regression variable contains the chromatogram data points. As mentioned in Chapter 7 the elution gradient serves as a time reference for the chromatograms in our
analyzes. This time reference is only valid if the elution gradient is assumed to be constant in every batch. The elution gradient is measured in conductivity. When the operator starts the pool, the measure of the elution gradient is written onto the chromatogram and the operator also marks where the pool is started on the chromatogram. The procedure is repeated when the operator stops the pool. Thus, there are two values of the elution gradient in every chromatogram and these values have been used as the time reference. If the distance between the start and stop mark on the chromatogram is measured, we can estimate the slope of the gradient. The slope is calculated as the difference between the stop and start conductivity ($\Delta C$) divided by the difference between the start and stop time ($\Delta t$). The time difference was measured with a ruler on the chromatograms. If the slopes calculated are constant, the elution gradient is suitable as time reference, and if they are not constant, the chromatograms cannot be compared with respect to time.

The slopes were calculated, and the result showed that the slopes were spread in a range of $+/\sim 10\%$ from the mean value.

The result of that the slopes are spread, is that the chromatograms to be compared are shifted back and forth along the time axis. Anyhow, this was the only time reference that could be used. This is one explanation why the predictions are not improved when adding the chromatogram data points to the regression variable.

The slope variation is due to the insecurity when the start and stop are marked on the chromatograms, and the insecurity when writing down the conductivity of the elution gradient. This means that the real elution gradient can be inaccurate (constant), but the calculated slopes (using $\Delta C/\Delta t$) differ. Thus, the accuracy of the time reference depends critically on how accurately the operators indicate the start and stop mark. If this process is automated the time reference problem is eliminated.

To further state this, a MLR model was tested by just using the chromatogram data points as regression variables. The result is shown in Figure 33. The number of columns in the $T$ matrix (using PCA, using the same strategy as before) were $n = 8$. Figure 33 shows that when just using the chromatogram data points in the regression variables, the predictions are poor. The $R$ values are shown in Table 5.

<table>
<thead>
<tr>
<th>$R_P$</th>
<th>$R_A$</th>
<th>$R_B$</th>
<th>$R_C$</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.0534</td>
<td>0.3237</td>
<td>0.2366</td>
<td>0.5407</td>
</tr>
</tbody>
</table>

Comparing the $R$ values with the ones using Model I, Model II and Model III (see Table 4), one can see that they are indeed worse. Especially this holds for the prediction of the product concentration $P$ and the impurity concentration $B$. 

44
Figure 33  Prediction of the eluate concentrations using only the chromatogram data points in the regression variables. The predicted values are marked with +, and the true values are marked with o. The solid lines are the mean values of the eluate concentrations in the calibration set, and the dotted lines are the mean values +/- 10%.
9. Proposals of data collection to make the pooling algorithm more robust

In the previous section, the attempt to find a model for how the eluate concentrations depend on the start and stop time of the pool, failed. The reason was that there was no information about how the start and stop time of the pool was correlated with the eluate concentrations. If the modeling attempt had succeeded, the model could have been used to determine of how to choose the start and stop time of the pool, to make the eluate concentrations optimized.

In this section proposals of how to collect data, and proposals of how to use this information to develop a model, that makes the pooling process more robust, are given.

9.1 Automate the data collection

To make the data analyzes more easy, the data collection concerning the chromatogram and the elution gradient should be automated. This is easy to do with an $A/D$ converter and a computer, connected to the UV-recorder and the conductivity recorder. Then the scanning and unscanning procedure, mentioned in Section 7, does not have to be done (it is the authors point of view that this is a really hard and time consuming procedure!). The time reference problem, stated in section 8.3, disappears and makes the comparison between the chromatogram fair. A natural reference point for the time measurements, is the time when the elution gradient is started.

9.2 Fraction measurements

The chromatogram shows the UV-absorption of a sample, and the UV-absorption is correlated with the protein concentration. Thus, the chromatogram measures the actual protein concentration in every point of time. In an IEC process every protein, product or impurity, is associated with a specific concentration peak or equivalent an UV-absorption peak. A chromatogram then measures the sum of all individual UV-absorption peaks, which arise from the proteins in the sample passing the UV-recorder. If the chromatogram could be decomposed into all its concentration peaks, the information is captured of where, and how, a single protein contributes to the whole chromatogram. To have this information, especially at the boundary of the start and stop time of the pool, is of course very useful.

To collect this important information, the following can be done. In equidistant points of time, samples or fractions, are taken from the fluid passing the UV-recorder. Each fraction is then analyzed, and the different protein concentrations are determined in the fraction. The analyzes should concern those proteins which are assumed to form the whole chromatogram. Each protein can then be associated with its own UV-absorption peak.

The a priori information that exists before the elution of the column starts, is the sample input analyzes and the age of the column, again, stated in terms of cycles since last re-packing. During the elution, fractions are taken, and after analyzing them each protein is associated with a UV-absorption curve. This curve can be described as a function of time, and it makes sense to assume that there is a strong correlation between the shape of the curve and the sample input concentration concerning the protein.

If the fraction measurements are made for a number of batches, we can then estimate the relation between the sample input concentration for a pro-
tein, with its UV-absorption curve. If the time function describing the curve can be parameterized with respect to the sample input concentration, we can use regression models to estimate that specific parameter, and other parameters describing the function. The difficulty here lies in finding a function that describes the curve shape, and that can be parameterized with respect to the sample input concentration.

Assume that there are $N$ different proteins involved in the process. Protein $k$ is associated with the sample input concentration $C_k$, and with the function describing the UV-absorption curve $f_k(\theta_k, t)$. $\theta_k$ is the parameters of the function, and $t$ is the time argument. Further assume that $C_k$ is one of the parameters in $\theta_k$. From the fraction measurements and the sample input concentration, estimation of how the curve shape depends on the sample input concentration can be made.

This model is then used to predict the curve shape using the sample input concentration. Thus, for every sample input concentration $C_k$, we predict every UV-absorption curve $f_k(t)$. Denote the chromatogram $z(t)$, and predict it using $\hat{z}(t) = \sum_{k=1}^{N} f_k(t)$.

When estimating the relation between the sample input concentration and the individual curve shapes, a number of batches have to be analyzed with fraction measurements. The fraction measurements are messy, since there are a lot of analyses concerned with each fraction sample. Its therefore preferable to limit the number of batches to be analyzed. In the batches to be analyzed, the excitation should be high enough, to ensure that the relation between the sample input concentration and the UV-absorption curve is stated as well as possible. Obviously, there is a problem of choosing which batches to be analyzed. One approach would be to try to classify the sample input concentrations in certain groups. The fraction analyzes should then be done with batches associated with these groups.

An attempt to do the classification for the sample input, was done using the collected analysis from the sample input concentrations. The attempt failed, because it was hard to find a pattern of how to classify the sample input concentrations.

The following suggestion is therefore given: do a number of fraction analyzes for batches during a re-packing period. Say that for some sample input concentrations, the fractions are measured for that batch, then the next batch to be analyzed should have sample input concentrations that clearly differs from the first one. Since the fractions are taken during a re-packing period, the age of the column might be modeled in the function describing the UV-absorption curve shape. The UV-absorption curve shapes can then be predicted using the sample input concentrations and the age of the column.

The described model is shown in Figure 34.

The predicted chromatogram can be compared to the real chromatogram, that develops during elution. If they differ to much, something has happen to the system that the model cannot predict. Thus, if the model is accurate, it can be used to indicate the performance of the system.

The predicted information can also be used to choose the start and stop time of the pool. All individual concentration peaks are predicted, and from this the quality of the pool, in terms of impurity concentrations and product concentration, can be calculated using a set of start and stop time. Applying the pooling strategy, the best choice of the start and stop time of the pool are determined. The system can then provide guidelines for how to choose the
Figure 34  Model description for predicting individual concentration peaks, and from that predicting the chromatogram.

start and stop time of the pool.
10. References


