Journal of Chromatography, 584 (1992) 43–57
Biomedical Applications
Elsevier Science Publishers B.V., Amsterdam

CHROMBIO. 6598

Review

Rational design of purification processes for recombinant proteins

E. W. Leser[☆] and J. A. Asenjo

Biochemical Engineering Laboratory, University of Reading, Reading RG6 2AP (UK)

ABSTRACT

This paper discusses the elements important for rational design of purification processes for recombinant proteins. Main issues involved in selection of operations and process design are reviewed with particular emphasis on the challenges posed by recombinant proteins. This includes thermodynamic characterization of target protein and main contaminants, use of correlations and of expert knowledge for the development of an expert system for optimization and design (selection) of separation and purification (chromatographic) processes. The main deficiency in accurate information for rational process selection is in that required for high-resolution chromatographic processes. The authors show that a database with detailed information on properties of the main contaminants present in the fermentation streams of usual recombinant protein sources can be integrated to an expert system with an open architecture. This will allow more precise selection of unit operations for the design of protein purification processes.

CONTENTS

1. Introduction 44
2. Rationalization of procedure 44
   2.1. Protein isolation and purification process 45
   2.2. Challenges with recombinant proteins 48
   2.3. Selection of operations: chromatography 49
3. Process design 50
   3.1. Use of mathematical models 51
   3.2. A prototype Expert System 51
   3.3. Use of protein properties 52
   3.4. Implementation and testing of prototype Expert systems 54
4. Conclusions 56
5. Acknowledgement 56
References 56

Correspondence to: J. A. Asenjo, Biochemical Engineering Laboratory, University of Reading, Reading RG6 2AP, UK.

[☆] On leave from Fundação Oswaldo Cruz, Fiocruz/MS, Brazil.

0378-4347/92/$15.00

1. INTRODUCTION

Discoveries and achievements in modern biology and recombinant DNA technology in the last few years have resulted in the development of several new therapeutic and diagnostic products and thus the possibility of their industrial large-scale production. This poses a tremendous challenge for the chemical and biochemical engineer in terms of downstream processing of these new proteins. Due to rigorous criteria in terms of quality control and the fact that some of them are intended for human use, the required levels of purity can be 99.5%, 99.9% or even higher (depending on the dosage) or, from the point of view of some contaminants, their presence should be reduced to the allowable limits.

A critical element of modern process biotechnology is the separation and purification of the target product from a fermentation broth or cell rupture supernatant. As it represents the major manufacturing cost, competitive advantage in production will depend not only on innovations in molecular biology and other areas of fundamental biological sciences but also on innovation and optimization of separation and downstream processes [1].

The design of an economic process to purify a protein, maintaining a high yield, yet obtaining a virtually pure product while minimizing the cost, demands three main considerations: (1) clearly defining the final product requirements, (2) characterizing the starting material, and with these two pieces of information in hand (3) defining possible separation steps and constraints regarding operations and conditions to be used [2].

The product characteristics are mainly determined by the final product utilization. The analysis of the starting material, with evaluation not only of common engineering data but determination of biochemical and thermodynamic properties of the major contaminants will provide the tools for the third step, when the expertise and knowledge will be the instruments to compare data, to judge the alternatives and to select the adequate sequence of operations to achieve the product as previously defined.

The following main heuristics or rules-of-thumb [2,3] provide a good basis for process selection.

Rule 1: "Choose separation process based on different physical, chemical or biochemical properties".
Rule 2: "Separate the most plentiful impurities first".
Rule 3: "Choose those processes that will exploit the differences in the physicochemical properties of the product and impurities in the most efficient manner".
Rule 4: "Use a high-resolution step as soon as possible".
Rule 5: "Do the most arduous step last".

In any event we have to keep a very open mind for any possible changes in the process and for keeping it as simple as possible. The main steps in a large-scale protein purification procedure are usually not more than four or five necessary ones and they normally consist of:

Recovery/isolation:
1. cell separation,
2. cell disruption and debris separation (for intracellular proteins only),
3. concentration;

Purification:
4. pretreatment or primary isolation,
5. high-resolution purification,
6. polishing of final product.

2. RATIONALIZATION OF PROCEDURE

An important point that needs consideration is that once the purification procedure is set and regulatory approval of the product is in progress the procedure cannot be changed. Only a particular product obtained by a specific procedure obtains regulatory approval, therefore once this is given, the purification method is fixed. This stresses the value of early rationalization of the purification process search. It also means that for a protein to be used for therapeutic applications or other human use for which, if successful, large

quantities of the product will be required, even in the very early stages of protein purification one should only use in the laboratory such techniques that can be realistically used in large scale, *i.e.* for which suitable large-scale equipment either exists or might be developed in the foreseeable future. Also, from early on, concepts related with maximization of yield in each step and in the whole separation sequence, minimizing the number of steps used and minimizing resources (economics) should be introduced [4].

It is important to consider the process of fermentation and downstream processing as a single system so that, for example, the effect of decisions about the fermentation conditions on subsequent purification stages is made clear. Product concentration and activity will partly depend on the binomial biological system–fermentation system employed. For instance, it is known that activity usually can change during the different phases of fermentation. The fermented broths leaving a stirred tank reactor or an air-lift system or a bioreactor are essentially distinct. The presence of proteases as well as bacterial contamination have to be minimized, which creates a need for rapid processing. Utilization of calf or foetal bovine sera will usually increase the number of purification stages required. Recombinant proteins sometimes are present in particles that need to be solubilized and refolded. It is thus important not only to discuss upstream processing in the light of all the protein purification stages but also to make the necessary decisions that will improve the recovery of the protein product early in the process development steps.

2.1. Protein isolation and purification process

Isolation comprises obtaining a cell-free solution with a total protein concentration around 60–70 g l^{-1} [5,6]. If cell separation is necessary the most frequently used methods, at the laboratory scale, are centrifugation and filtration. This operation, concerning solid–liquid separation, will pose difficulties at the large scale. The smaller the size of the particles the more difficult it is to isolate them. Large-scale centrifuges are complex

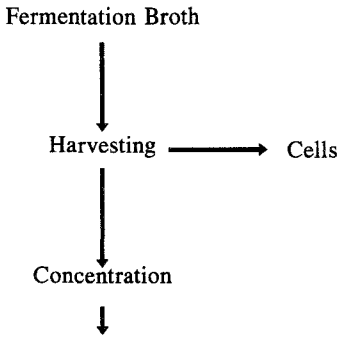

Fig. 1. Recovery subprocess of extracellular product (yeast, mammalian and bacterial) [19].

to operate and to maintain, specially when one must consider the operation with pathogenic sources or a contained process. On the other hand, cross flow filtration is a relatively new alternative but there are some technological barriers to overcome. If the product location is extracellular, then the liquid part is kept (Fig. 1); if the product is intracellular, the solid fraction of the operation is kept (Fig. 2). When a mammalian cell culture is used, the product is usually secreted by the cells. Production of monoclonal antibodies has been extensively performed using hollow-fibre reactors and thus, the fermentation broth is free of particles, with exception of a small amount of cell debris that can be eliminated by gel permeation [7].

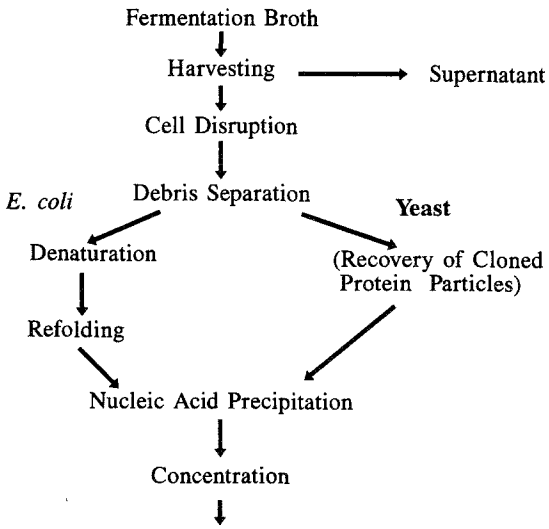

Fig. 2. Recovery subprocess of intracellular product (*E. coli*, yeast) [19].

Cell disruption is required when the product is intracellular. Methods and equipment are selected mainly based on the biological source and product. The choice of disruption technique determines the size of the resulting debris that in turn has an influence on subsequent operations. Typical methods used can be classified into four groups: non-mechanical, ultrasonic, high-pressure homogenization and mechanical grinding (bead milling) [8], but, for large-scale purposes, only the last two categories are important. Bead milling is used with gram-positive bacteria and specific yeast applications; pressure homogenization for most bacteria including *Escherichia coli* and yeast [9]. Mechanical disruption releases nucleic acids that need to be precipitated. The standard method is precipitation with polyethylenimine. Separation of cell debris has to be undertaken once the cells are disrupted and due to the small size of particles this brings extra difficulties for the large-scale process. The result of this step is a solution containing the product, cell metabolites and remaining components of the culture medium. At this point the addition of proteolytic inhibitors should be evaluated.

If the intracellular product is manufactured in *E. coli*, high expression of heterologous proteins will usually accumulate as insoluble inclusion bodies [10]. This makes necessary the processing of the inclusion bodies into the native protein by denaturing and refolding. If the intracellular product is manufactured in yeast, often the protein is present in homogeneously particulate form, typically 30–60 nm particles such as virus-like particles [11]. Although the processing of intracellular particulate recombinant proteins is an important aspect of downstream processing, not many satisfactory methods exist for large-scale separation, denaturation and refolding of the particulate proteins.

Concentration is usually required when the protein concentration of the harvested, disrupted and separated stream is below 60–70 g l^{-1}, the suitable range for chromatographic purification [4–6]. With some proteins it is very difficult to obtain higher concentrations without a serious increase in viscosity, which would then impose very poor transport characteristics on the system. If a membrane (ultrafiltration) is used for concentration, the resulting flux characteristics will decide the highest possible concentration that can be obtained from the operation. If at the point where flux has dropped below an acceptable limit the concentration is below 60 g l^{-1}, then the proteins can be precipitated (*e.g.*, with ammonium sulphate) to increase the final concentration to an adequate level.

At this point the broth will contain proteins and some other components such as lipids and/or wall or other polysaccharides, salts and water. Here product purification begins and there will be many alternative combinations of processes (Tables 1 and 2). For the recovery, resolution and purification of a single protein, ideally one would like one step to extract virtually 100% of the target protein from the mixture with no contaminants. As this is almost impossible, two or sometimes three or four stages will probably be needed to achieve the final purity required for the particular application (Fig. 3).

As most of the excess water has been extracted, a purification step of extremely high resolution should be used to minimize the number of stages and hence maximize yield. However, in many cases this may not be possible at this stage, as some contaminants still present may produce fouling of the affinity or high-resolution ion-exchange column and, consequently, shorten its life. Therefore, a first step in protein purification from other contaminants probably will be necessary. This would constitute a clean-up step of pretreatment or primary isolation. For this, a somewhat inexpensive treatment to clarify the stream from suspended materials and non-protein contaminants in addition to salts should be utilised. This operation will not give a very high purity but must result in a very high recovery yield in terms of the protein product retrieved. Typical operations for this step would include inexpensive or disposable adsorption devices like a Whatman DE52 ion-exchange cartridge, a hydrophobic interaction step, aqueous two-phase partitioning or precipitation of the proteins using salt. After this procedure, a higher res-

TABLE 1

SEPARATION AND PURIFICATION OPERATIONS FOR LARGE-SCALE RECOVERY AND PURIFICATION OF PROTEINS AND THE CORRESPONDING MAIN PHYSICOCHEMICAL PROPERTY THAT DRIVES THE OPERATION [3]

Operation	Physicochemical property
Centrifugation	Sedimentation velocity
Filtration	Particle size
Microfiltration	Particle size
Homogenization	Intracellular nature (pressure gradient)
Bead milling	Intracellular nature (liquid/solid shear)
Ultrafiltration	Molecular size
Two-phase extraction	Partition coefficient
Precipitation	Solubility (hydrophobic internation)
Adsorption	Van der Waals forces, H bonds, polarities, dipole moments
Ion exchange	Charge (titration curve)
Hydrophobic interaction	Surface hydrophobicity
Affinity chromatography	Biological affinity
Gel permeation	Molecular size
Reversed-phase liquid chromatography	Hydrophilic and hydrophobic interactions

olution ion exchange will most probably be used, giving a product of up to 99% (usually 95–98%) purity. Typical operations will include one or two high resolution ion-exchange chromatography steps or affinity chromatography. Although high resolution is the main concern in this stage, an adsorbent that also will give a high recovery should be chosen or designed.

TABLE 2

CHROMATOGRAPHIC OPERATIONS AND THE RELATED PHYSICOCHEMICAL DRIVING PROPERTIES, PROCESS CHARACTERISTICS AND APPLICATIONS FOR LARGE SCALE PURIFICATION OF PROTEINS

Adapted from ref. 4.

Physicochemical property	Operation	Characteristic	Application
Van der Waals forces, H bonds, polarities, dipole moments	Adsorption	Good to high resolution, good capacity, good to high speed	Sorption from crude feedstocks, fractionation
Charge (titration curve)	Ion exchange	High resolution, high speed, high capacity	Initial sorption, fractionation
Surface hydrophobicity	Hydrophobic interaction	Good resolution, speed and capacity can be high	Partial fractionation (when sample at high ionic strength)
Biological affinity	Affinity chromatography	Excellent resolution, high speed, high capacity	Fractionation, adsorption from feedstocks
Hydrophilic and hydrophobic interactions	Reversed-phase liquid chromatography	Excellent resolution, intermediate capacity, may denature proteins	Fractionation
Molecular size	Gel permeation	Moderate resolution, low capacity, excellent for desalting	Desalting, end polishing, solvent removal

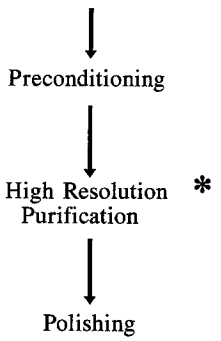

Fig. 3. Purification subprocess [19].

After the high resolution step a polishing step is frequently necessary to obtain ultra high purity. This will depend on the final use of the protein, and in some cases it is probably the most difficult task to perform. If another physicochemical property cannot be exploited, gel permeation will be used, which can separate dimers of the product (due to aggregation phenomena) or its hydrolysis products (due to action of proteases) solely based on their different molecular weights. High-performance liquid chromatography (HPLC) also can be used for polishing; however, this is an expensive technique for preparative purposes. It gives extremely high resolution but it may denature larger proteins.

2.2. Challenges with recombinant proteins

When the intracellular product is manufactured in *E. coli*, heterologous proteins will usually accumulate as large insoluble particles called inclusion bodies [10]. If the intracellular product is manufactured in yeast, in a number of cases the protein is present in particulate form, typically 30–60 nm particles such as virus-like particles [11].

Regarding cell breakage, some of the advantages of chemical and enzymatic permeabilization and lysis methods recently discussed [9,12–14] should be investigated in greater detail, particularly because mechanical disruption techniques have several drawbacks related to obtaining high product yield (micronized wall materials, nucleic acids, high viscosity, complex mixture of contaminants and partially damaged product) that are difficult to overcome. Release of recombinant intracellular proteins by chemical permeabilization and enzymatic lysis techniques as well as the release of recombinant protein from yeast [12] has been successfully achieved [9]. For this, however, greater availability of specialized reagents (*e.g.*, wall lytic enzymes) will be necessary as currently these are almost only available as laboratory reagents.

Inclusion bodies have to be solubilized, in many cases chemically or enzymatically modified, and correctly refolded, otherwise the process will produce large quantities of inactive product. This is usually the case when bacteria are used for the manufacture of human proteins. A very recent study of a process with *E. coli* [15] showed that denaturation and solubilization of inclusion bodies with, for example, guanidine · HCl are steps that account for most of the cost of downstream raw materials. Of all downstream processing costs related to raw materials 77% corresponds to guanidine · HCl and carboxypeptidase and 92% to the main four items, including formate and cyanogen bromide which are all specific to the denaturation and refolding of inclusion bodies into the active protein [15]. This clearly shows that currently there are no satisfactory methods for large scale denaturation and refolding of particulate proteins. Recent developments in the use of reverse micelles for protein refolding [16] and of two-phase aqueous systems for separation of virus-like particles from yeast homogenates [12] appear particularly attractive.

Separation of inclusion bodies from debris can be achieved on a large scale by the use of centrifuges even if the material is small (about 1.0 μm) mainly because of the relatively large density of inclusion bodies (*e.g.*, 1.3 g ml^{-1}) [15]. However, flow-rates have to be reduced several-fold compared to the separation of whole *E. coli* cells where flow-rates are already low. This results in large capital requirements.

Aqueous two-phase systems are a very attractive alternative for the separation of cell debris from target product protein [4,12,17]. The separation of recombinant particles from yeast has been demonstrated using this technique [12,18]. In the presence of debris and recombinant particles two stages were more appropriate, the first to separate the cell debris and the second to separate contaminant proteins [12].

2.3. Selection of operations: chromatography

Selection of operations required for recovery/isolation is relatively straightforward if the product is extracellular. When a product is intracellular, however, no satisfactory procedures are available for large scale processing of inclusion bodies into native proteins. Recent advances in cell permeabilization and differential product release as an alternative to disruption should show important developments in the next few years [13,14].

Selection of purification operations, on the other hand, is more cumbersome, and choosing those operations that will give the best results is not an easy task, particularly with the high resolution chromatographic operations that are carried out in one, two or even three steps. To be able to separate one protein from another, a difference in physicochemical properties between them is exploited. To design an optimal separation process is to exploit these differences in the most efficient manner to accomplish the desired separation. Individual separations will generally depend on more than one property difference for their overall performance, but one property will usually form the primary basis for separation.

Thermodynamic property information should be available for the target protein and also for the major contaminants. It is also useful to have some information on the fermentation supernatant from which the protein has to be separated and on some of the intermediate process streams such as those shown in Figs. 1–3, including thermodynamic equilibrium and transport properties such as density, total protein concentration, particle size distribution, when these are present, and viscosity. Main sources of recombinant proteins in modern biotechnology are few: *E. coli* (intracellular) and some bacilli, mammalian cells (extracellular), and yeast, normally *Saccharomyces cerevisiae* and *Picchia sp.* (extra- and intracellular). The characterization of product protein and major contaminants has to be carried out in terms of charge and titration curve of major proteins, molecular weight, hydrophobicity, p*I* and available biospecific interactions. Determination of this information can be performed on a case-by-case basis for the individual product proteins. General distribution of physicochemical properties of the host cells just mentioned (*E. coli*, yeast and mammalian cells) should be generated as it will allow selection of purification operations on a much more rational basis. This is shown in a later section in this paper on implementation of protein properties in a prototype expert system.

High resolution purification is usually carried out by chromatography. Selection of these purification operations is based on the efficiency of different chromatographic techniques to separate the target protein from the contaminating ones. Different techniques exploit different thermodynamic properties and some are much more efficient than others in exploiting the differences in these properties between two proteins. Ion-exchange chromatography will separate the proteins based on their difference in charge. The charge of a protein changes with the pH following the titration curve [19]. Hence, if carried out at considerably different pH values, at which the difference in charge of three or more proteins is significant, this technique can be used twice to purify a protein from different protein contaminants [19]. Ion exchange can use small differences in charge to give a very high resolution and hence is an extremely efficient operation to separate proteins.

Affinity chromatography can have a very high specificity for a particular protein or a small group of proteins; therefore, it can also have a very high resolution. The matrix can be expensive, but it can be reused many times. Ligand leakage into the product can be a problem. Regarding cost, affinity chromatography will usu-

ally be more expensive than ion-exchange chromatography [20,21]. Hydrophobic interaction chromatography has been proposed only as a pretreatment step or as a first high resolution purification step. The resolution is good but not always particularly high as the distribution of surface hydrophobicity in a protein can be random, thus giving only adequate resolution. Gel permeation for protein fractionation is normally not used as a high resolution operation in the large scale because of the low efficiency in exploiting differences in molecular mass.

3. PROCES DESIGN

Process design and selection of operations is a complex procedure where the design evolves from an initial stage to the final stage in a trial-and-error fashion, repeatedly revising and refining the initial assumptions and restrictions:

1. flowsheet generation (qualitative/semi-quantitative);
2. quantitative design of units;
3. revise flowsheet (1), then (2) etc. until some objective is reached.

An important aspect of process design involves the selection of operations and design of a process sequence (process synthesis). In the initial phases this process is usually done using heuristics: utilization of rules-of-thumb to arrive at a rapid (and reliable) specification of equipment type, size and, maybe, cost.

The problems that have to be solved in process synthesis and optimization of downstream protein separations are of two types: (i) choosing between alternative operations (*e.g.*, homogenizer *versus* bead mill or centrifugation *versus* cross flow microfiltration) and (ii) the design of an optimal chromatographic sequence with maximum yield and minimum number of steps (1, 2 or 3), a problem that is combinatorial in nature. The first type of problem can be adequately solved provided appropriate mathematical correlations and mathematical models that can be used as useful simulation tools are developed. The second type of problem has been partially tackled in classical chemical process engineering (*e.g.*, distillation *versus* extraction) by finding a rigorous solution using numerical methods like mathematical programming techniques (*e.g.*, resolution of "tree structure" [3]) or more novel techniques of Artificial Intelligence (AI) (*e.g.*, Expert Systems, ES). (ES) are intelligent computer programs that emulate human reasoning to solve problems. Fig. 4 shows a diagram of such a system with its two main parts: a *knowledge base* and an *inference engine*. To create an ES the domain or body of knowledge specific to the class of problem must be organized into the knowledge base. The organization consists of logical statements regarding hierarchic structures and relational rules. The inference engine will provide the reasoning simulation based on the search paths along the problem space and external inputs of information and data. The interface with databases and other numerical calculation programs is also shown. For the design of an optimal separation sequence the use of purely mathematical techniques has limited use in biotechnology due to a lack of useful design equations and databases. The ES approach appears more attractive since it allows the use of empirical knowledge that is not rigorous in nature and is typical of that used by experts in the field.

Computer-based ES are an important tool in the field of AI. Efforts have been made to develop

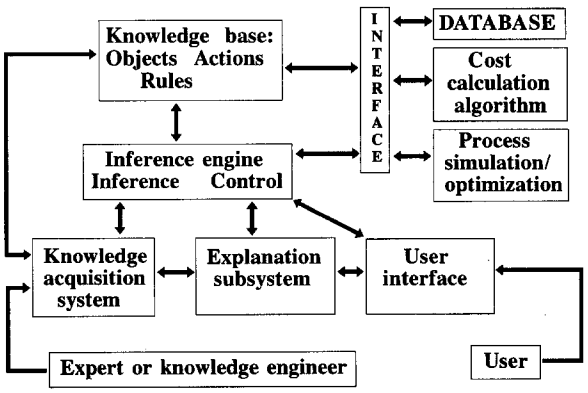

Fig. 4. Architecture of the Expert System and the links with external databases and algorithms.

ES for this purpose [22–24] or to adapt the existing software systems, commonly called "shells", both for the manipulation of heuristics, databases and simple algebraic design equations [5]. Conventional scientific and engineering scientific programs consist of a set of statements where the order of execution is predetermined, simple step-by-step procedures that guarantee that the right conclusion will be reached when the correct data have been entered. The programmer must, always, ensure complete specification of the problem and uniqueness of the solution; updating must be made in the precise place and sequence. ES alleviate this strictness by making a clear distinction between the knowledge base and the control strategy. This partition allows for incremental addition of knowledge without manipulating the complete program structure, and, by choosing strategies it is possible to provide the system with the tools to find the solution, and, if desired, to provide a number of ranked alternative solutions.

Process synthesis is one area where expert systems can be of important help, particularly in the selection of equipment and in the design of separation trains or sequences, especially in the field of biotechnology.

3.1. Use of mathematical models

Mathematical models and mathematical correlations of the operations will allow simulation of performance. They also may be used to scale up individual operations. Computer simulations are a useful tool to optimize individual separations [25]. Examples of useful downstream process simulations, and investigation of process conditions are microbial cell breakage and selective product release using enzymes [26,27], affinity and ion exchange chromatography of proteins [28,29] and continuous adsorption recycle extraction [30], in which case the model has been used for process optimization and search for appropriate control strategies.

3.2. A prototype Expèrt System

Today there are well developed expert software systems, called "shells", that help to develop an organised knowledge base from the domain knowledge and provide the inference engine. Asenjo *et al.* [5] found that for the creation of prototype systems there were adequate shells, particularly if they can evaluate uncertainties associated with the inference process. Expert knowledge was obtained mainly from industrial experts working on the large scale separation and purification of therapeutic, diagnostic and analytical proteins. Soon it became apparent that the true bottleneck in the development of expert systems for protein purification is not in its implementation but in the acquisition, clarification, formalization and structuring of the knowledge domain.

In the first basic prototype the knowledge was expressed in around 65 rules, some of which carried a degree of uncertainty [5]. The downstream process was divided into two distinct subprocesses, the first called 'Recovery/isolation" after which the total protein concentration is 60–70 g l^{-1} and the second called "Purification". Processing of recombinant proteins present in intracellular inclusion bodies or other particles was not considered in this prototype but was eventually included in the expanded version (next sections).

The proposed process consists of a sequence of operations to obtain the stated design objective. There might be several different sequences of operations that accomplish the same objective. In those cases, a quantitative degree of performance (given by a certainty factor) of each operation is assigned by the expert and carried by the system into the proposed design.

This prototype expert system (65 rules) did not have a database on the properties of main protein contaminants (physicochemical and thermodynamic) so the selection of high resolution purification operations was rather empirical and based on knowledge of the efficiency of the different techniques to separate a protein from its main contaminants [5].

In the development of the prototype expert system it was found that selection of operations in the recovery subprocess could be well structured. Concerning the purification subprocess, the structuring of the knowledge was more difficult. The main deficiency of available information was in the selection of high resolution chromatographic operations that should be based on the properties of the target protein and the contaminants in the solution. This informations is vital to select the right operations and in the best possible order according to their relative efficiencies [19,31,32] (next section). A considerable lack of information was also the case for the separation of minor contaminants present that are removed in the final polishing stage (also by chromatography).

For selection of operations, information generated at a very small scale in terms of "efficiency" of separation or, alternatively, information on thermodynamic properties (charge-titration curve, p*I*, surface hydrophobicity, molecular mass, bioaffinity) is necessary. Here, the deviation of the value for the product protein from those of the main contaminants should be used. A factor for efficiency of the operation in exploiting this difference also has to be included in this evaluation [19].

3.3. Use of protein properties

Selection of actual operations is based on information generated at a small scale to determine performance and efficiency of particular separations. Alternatively information on physical, chemical, biochemical and thermodynamic properties of product and contaminants can be used to predict such performance. Then the deviation (DF = deviation factor) of the value of the protein product from those of the main contaminants should be found. A factor for efficiency (η) of the separation operation in exploiting this difference or deviation (DF) of property has to be included in this evaluation. It is possible then to define a separation coefficient (SC) that can be used to characterize the ability of the separation operation to separate two or more proteins [19].

$$SC = f(DF,\eta)$$

When using chromatography there are differences in the cost of the matrices used (*e.g.*, protein A affinity chromatography uses a more expensive matrix than CM-Sepharose ion-exchange chromatography) although most of the cost in such a process is associated with the hardware (column, accessories, control system) as most matrices can be reused many times resulting in reduction of associated cost. Also different adsorption capacities and flow characteristics of the matrices will result in columns of different size. However, column hardware cost is only a fraction of the total cost hence the total hardware cost of a chromatographic step is relatively constant. Differences in the cost of a purification operation can be taken into account by using a cost factor (CF) giving a expression for the economic separations coefficient (ESC).

$$ESC = f(SC,CF)$$

The values of the parameters in these two expressions should range between $0 < \eta \leq 1$ and $0 \leq DF \leq 1$. As such values are relative, the maximum value for DF for individual properties has to be defined within this range and the value for η given to a particular operation also will depend on the range (or maximum possible value for the deviation of a specific protein property) that must be standardized for different operations. The value of the cost factor (CF) will be $CF < 1$ or $CF \geq 1$ and a standard operation (such as ion-exchange chromatography using CM-Sepharose) should be given a value of 1. We have made a first attempt to define both a separation coefficient, SC, and an economic separation coefficient, ESC. They are shown in Table 3. The inclusion of a term for concentration was suggested, as this will affect the selection criteria, since the contaminants in higher concentrations have to be removed first (heuristics: rule 2). But, as concentration does not appear to intrinsically influence the actual separation coefficient, the suggestion of using the term "Separation Selection Coefficient" (SSC), when including the concentration term θ has been preferred.

$$SSC = DF \cdot \eta \cdot \theta$$

TABLE 3

OVERVIEW OF SEPARATION COEFFICIENTS, THEIR DEFINITIONS AND "TENTATIVE" VALUES ASSOCIATED TO SOME OPERATIONS

Separations coefficients

$SC = DF \cdot \eta$

DF = Deviation factor for hydrophobicity, molecular weight, and pI

$$DF = \frac{\text{Protein value - contaminant value}}{\text{Max.[protein value, contaminant value]}}$$

DF = 1.0 for affinity chromatography

η = Efficiency =

- 1.00 for affinity chromatography
- 0.70 for ion exchange
- 0.35 for hydrophobic interaction chromatography
- 0.20 for gel permeation

$SSC = DF \cdot \eta \cdot \theta$

θ = concentration factor

$$\theta = \frac{\text{Concentration of contaminant protein}}{\text{Total concentration of contaminant proteins}}$$

$$ESC = \frac{SSC}{CF}$$

CF = Cost factor =

- 1.0 for affinity chromatography
- 0.6 for gel permeation
- 0.3 for ion exchange
- 0.3 for hydrophobic interaction chromatography

It should be noted that the two parameters η and the cost factor, CF, are thus far empirical and subjective. A more rigorous estimate is presently under study in our group. The cost factor is not based on a rigorous economic evaluation, such as has recently been carried out [20], but on an "approximate", preliminary evaluation of the cost involved using such an operation. There are many elements apart from the direct variable and capital costs that affect the choice of process and therefore the approximate evaluation of cost and thus CF (*e.g.*, availability of matrix in the pilot plant, reliability, robustness with variation in feedstock, speed of process implementation or quality control). This role of other elements is partly related to the fact that cost of production of a therapeutic or diagnostic protein is still only a small fraction of the final price. Hence the cost differences found in a rigorous economic evaluation are much more marked than those shown in the expression in Table 3. All values shown in Table 3 will be subjected to modifications as the rationale proposed is tested in real cases.

3.4. Implementation and testing of prototype Expert Systems

The rationale for selection of high-resolution purification operations has been implemented into the prototype ES. This was done by interfacing a program in PASCAL in which the main physicochemical properties of a target product protein were compared with those of the main protein contaminants and then used to select the most appropriate high resolution chromatographic operations [31]. The rationale discussed in the previous section and Table 3 was used.

The main sources used for the large scale production of recombinant proteins today are few. For the purpose of our prototype only three main production systems were chosen for the characterization of the main protein contaminants present in these sources. These are *E. coli* (intracellular proteins), *S. cerevisiae* (intra- and extracellular) and mammalian cells (extracellular proteins). Initial results of our present work on the characterization of the main proteins in these sources are shown in Tables 4–6 [33]. This approach appears conceptually valid for molecular weight and for hydrophobicity as exploited in hydrophobic interaction chromatography, but care has to be taken in the selection of ion-exchange chromatography as a suitable method. Values of the isoelectric point, p*I*, of proteins are useful for the selection of operating conditions when using an anion- or cation-exchange matrix but not for the selection of operations that will give better separation resolution between proteins. Data on the charge of the proteins as a function of pH [19] or on its adsorption properties and relative retention times on the different matrices under different conditions are necessary, as has been discussed in much more detail previously [32]. The isoelectric point is only directly relevant for the selection of chromatofocusing as a separation operation, but chromatofocusing is not a particularly feasible operation for large scale use, owing, not only to the high cost of polybuffer, but mainly because such buffers are unacceptable for use with therapeutic proteins. As more accurate and detailed information on the protein contaminants is made available, it will be appropriately implemented in the database and in the rationale for selection of operations.

TABLE 4

PROPERTIES OF MAIN PROTEIN CONTAMINANTS IN *E. COLI* LYSATE [33]

Cell lysate was prepared by bead milling.

Band No.	Molecular mass[a]	Hydrophobicity, Φ^b (M)	Isoelectric point[c]
1	90 000	0.02	4.8
2	145 000	1.12	4.8
3	80 000	0.13	4.9
4	200 000	1.02, 0.13	4.8
5	12 800	0.64	5.1
6	25 000	0.26	4.5
7	45 000	0.13	5.4
8	40 000	0.64	4.6
9	44 000	0.13	4.3
10	120 000	0.02	5.4
11	80 000	0.13	4.6

[a] Measured by gel permeation.

[b] Measured by hydrophobic interaction chromatography using a Phenyl-Superose gel in an FPLC system and a gradient elution from 2.0 M to 0.0 M $(NH_4)_2SO_4$ in 0.1 M KH_2PO_4. Units used are the concentration of $(NH_4)_2SO_4$ at which the protein eluted.

[c] Measured by isoelectric focusing using a Sephadex gel.

The "expanded" prototype with access to the databases and the more rational selection of high resolution separation operations discussed in this paper resulted in an ES with approximately 130 rules in addition to the PASCAL interface that implements the database and the rationale shown in Table 3. Presently were are using a much more advanced shell (NEXPERT OBJECT) that allows the development of an object-oriented ES. The feature of inheritance of properties amongst the hierarchical structure will be an important tool for our project. In order to make this concept more clear, a good example is to consider the class of chromatographic operations, with subclasses (*e.g.*, ion exchange, affinity, hydrophobic interaction, etc.) and objects belonging to this class (input stream, output stream or any other).

TABLE 5

PROPERTIES OF THE TEN MAIN PROTEIN BANDS PRESENT IN *S. CEREVISIAE* LYSATE [33]

Cell lysate was prepared by bead milling.

Band No.	Molecular mass[a]	Hydrophobicity, Φ^b (M)	Isoelectric point[c]
1	80 000	0.50	6.6
2	44 000	0.60, EtOH	6.4
3	22 000	0.25	5.6
4	80 000	EtOH	6.6, 8.8
5	49 000	ppt.	5.5
6	71 000	0.30	5.7
7	170 000	0.40	5.7, 6.9
8	12 000	ppt.	7.1
9	170 000	0.15	5.7
10	65 000	0.65	6.0, 7.7

[a] Measured by gel permeation.

[b] Measured by hydrophobic interaction chromatography using Octyl-Sepharose gel in an FPLC system and a gradient elution from 1.5 M to 0.0 M $(NH_4)_2SO_4$ to avoid protein precipitation. Some protein bands still precipitated (ppt. in table). EtOH means tightly bound band that needed to be eluted with 24% ethanol in deionized water.

[c] Measured by isoelectric focusing using a Sephadex gel.

TABLE 6

PROPERTIES OF THE TEN MAIN PROTEIN BANDS IN CHINESE HAMSTER OVARY CELLS CULTURE SUPERNATANT [33]

Band No.	Molecular mass[a]	Hydrophobicity, Φ^b (M)	Isoelectric point[c]
1	66 000	0.83	5.0
2	140 000-205 000	0.83, ppt.	5.4, 8.7
3	295 000	0.83	6.0
4	72 000	0.70	5.4
5	53 000	1.25	5.2
6	72 000	0.70	5.4
7	170 000	1.10	4.6
8	3000	1.25	5.4
9	6000	0.02	4.0
10	170 000	0.71	5.7

[a] Measured by gel permeation.

[b] Measured by hydrophobic interaction chromatography using Phenyl-Superose gel in an FPLC system and a gradient elution from 1.7 M to 0.0 M $(NH_4)_2SO_4$ to avoid protein precipitation. Some protein bands still precipitated (ppt. in table).

[c] Measured by isoelectric focusing using a Sephadex gel.

If we define a property for the class (*e.g.*, concentration of the product), it will be automatically transferred to the object input stream. To describe the knowledge base more accurately changes in the basic structure of the purification flowsheet are being investigated. The idea is to use the linear mode of this process and choose the precise reasoning paths through the problem space. The open architecture of this shell allows direct integration with databases, without the need of an interface even for calculation of separation coefficients and use of mathematical correlations with an extremelly fast and efficient two-way flow of information. It also grants the union of the expert system with an external algorithm. This will enable the integration of data and correlations with the expert system for process selection and design resulting in a sequence of operations with highest confidence factor and minimum number of high resolution chromatographic operations. The next best choices for individual operations are also given (Fig. 5). Such a "process synthesis" (selection) scheme can then be used for cost calculation analysis based on material and energy balances

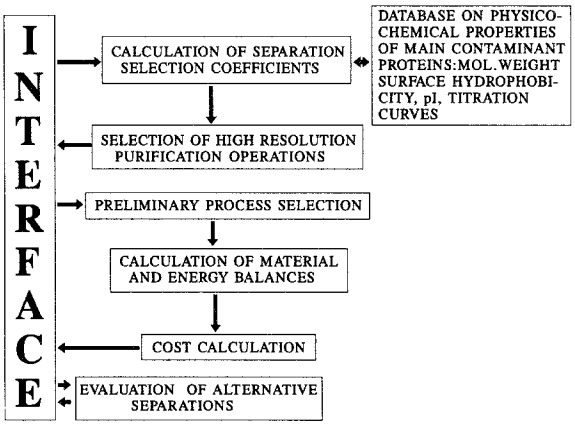

Fig. 5. Proposed scheme for the "tasks" and calculations to be performed by the Expert System, in addition to the heuristic "expert" knowledge (rules), for the rational design of protein purification processes.

using flow-sheeting techniques (Fig. 5). The process performance can be tested at the laboratory and pilot plant levels. Discrepancies between practical performance and the computer-selected process are analyzed and investigated in order to modify and validate the databases, selection rationale and/or mathematical correlations used.

4. CONCLUSIONS

The work described on rational design of protein purification processes clearly shows that properly developed ES can be a vital tool to assist with solving the knowledge-intensive and heuristic-based problem of selection of downstream processes, particularly the high resolution chromatographic steps. The overall separation process selection and synthesis problem in biotechnology does not have a strict combinatorial nature whereas the high resolution purification stages within the purification subprocess (one, two or even three chromatographic stages where several alternatives in different order combinations can be used) do.

It clearly appears that the limiting factor in the development of ES for protein purification is not the implementation of new AI programs but the acquisition, clarification, formalization and structuring of the domain of expert knowledge.

The use of an expert system shell with an open architecture which allows direct integration of a database of physicochemical and thermodynamic properties of main protein contaminants in main production streams used within recombinant production systems is an important improvement. This will allow the selection of chromatographic high resolution purification operations on a much more rational basis resulting in a very improved process selection and thus rational process design. In order to advance the further development of this field, there is an important need for generating more detailed databases for protein products, fermentation streams and contaminants. These include databases on surface hydrophobicity, molecular weight, isoelectric point and also, most importantly, titration curves for proteins.

5. ACKNOWLEDGEMENT

Financial assistance by Conselho Nacional para o Desenvolvimento Científico e Tecnológico (CNPq), Brazil, for support of E.W.L. is gratefully acknowledged.

REFERENCES

1 S. M. Wheelwright, *Bio/Technology*, 5 (1987) 789.
2 J. A. Asenjo, presented at the *32nd International IUPAC Congress, Stockholm Aug. 2–7, 1989.*
3 G. J. Prokopakis and J. A. Asenjo, in J. A. Asenjo (Editor), *Separation Processes in Biotechnology*, Marcel Dekker, New York, 1990, pp. 571–601.
4 J. A. Asenjo and I. Patrick, in E. L. V. Harris and S. Angal (Editors), *Protein Purification Applications: A Practical Approach*, IRL press, Oxford, 1990, pp. 1–28.
5 J. A. Asenjo, L. Herrera and B. Byrne, *J. Biotechnol.*, 11 (1989) 275–298.
6 Pharmacia, *Scale Up to Process Chromatography, Guide to Design,* Pharmacia, Uppsala, 1983.
7 S. Lee, in S. S. Seaver (Editor), *Commercial Production of Monoclonal Antibodies, A Guide for Scale-Up*, Marcel Dekker, New York, 1987, pp. 199–216.
8 S. M. Wheelwright, *Protein Purification, Design and Scale-Up of Downstream Processing*, Hanser, Munich, 1991.
9 J. A. Asenjo, in D. L. Pyle (Editor), *Separations for Biotechnology II*, Elsevier, New York, 1990, pp. 11–20.
10 J. F. Kane and D. L. Hartley, *Trends Biotechnol.*, 6 (1988) 95.
11 F. Muller, K. Bruhl, K. Freidel, K. V. Kowallik and M. Ciriacy, *Mol. Gen. Genet.*, 207 (1989) 421–429.
12 R. B. Huang, *Ph. D. Thesis*, University of Reading, 1990.
13 B. A. Andrews, R. B. Huang and J. A. Asenjo, in M. White, S. Reuveny and A. Shaffermann (Editors), *Biologicals from Recombinant Microorganisms and Animal Cells — Production and Recovery*, VCH Publishers, Weinheim, 1991, pp. 307–321.
14 T. J. Naglak, D. J. Hettwer and H. Y. Wang, in J. A. Asenjo (Editor), *Separation Processes in Biotechnology*, Marcel Dekker, New York, 1990, pp. 177–205.
15 R. Datar and C. G. Rosen in J. A. Asenjo (Editor), *Separation Processes in Biotechnology*, Marcel Dekker, New York, 1990, pp. 177–205.
16 A. J. Hagen, T. A. Hatton and D. I. C. Wang, *Biotechnol. Bioeng.*, 35 (1990) 955–965.
17 B. A. Andrews and J. A. Asenjo, in E. L. V. Harris and S. Angal (Editors), *Protein Purification Applications: A Practical Approach*, IRL Press, Oxford, 1990, pp. 161–174.
18 V. Riveros-Moreno and J. E. Beesley, in D. L. Pyle (Editor), *Separations for Biotechnology II*, Elsevier, London, 1990, pp. 227–237.
19 J. A. Asenjo, in J. A. Asenjo (Editor), *Separation Processes in Biotechnology*, Marcel Dekker, New York, 1990, pp. 3–16.

20 R. Kosti, *M. Sc. Thesis*, University of Reading, 1989.
21 S. A. Duffy, B. J. Moellering and C. R. Prior, presented at the *196th ACS National Meeting, MBTD Division, Los Angeles, CA, Sept. 25–30, 1988.*
22 C. A. Siletti and G. Stephanopolous, presented at the *192nd ACS National Meeting, Anaheim, CA, Sept., 1986.*
23 S. Wacks, *M. Sc. Thesis*, Columbia University, New York, 1987.
24 C. A. Siletti. *Ph. D. Thesis*, MIT, Cambridge, MA, 1989.
25 P. Hedman, J. C. Janson, B. Arve and J. G. Gustafsson, in G. Durand, L. Bobichon and J. Florent (Editors), *Proceedings of 8th International Biotechnology Symposium, Paris*, Vol. 1, Société Française de Microbiologie, Paris, 1988, pp. 623–642.
26 J. B. Hunter and J. A. Asenjo, *Biotechnol. Bioeng.*, 31 (1988) 929–943.
27 L. C. Liu, G. J. Prokopakis and J. A. Asenjo, *Biotechnol. Bioeng.*, 32 (1988) 1113–1127.
28 H. A. Chase, presented at the *Symposium on Antibodies for Purification, SCI, London, March 1988.*
29 B. Arve, presented at the *32nd International IUPAC Congress, Stockholm, Aug. 2–7, 1989.*
30 M. I. Rodrigues, C. A. Zaror, F. Maugeri and J. A. Asenjo, *Chem. Eng. Sci.*, 47 (1992) 263–269.
31 J. A. Asenjo and F. Maugeri, in P. Todd, S. Sikdar and M. Bier (Editors), *Frontiers in Biopressing* II, ACS books, Washington DC, 1992, pp. 359–379.
32 J. A. Asenjo, J. Parrado and B. A. Andrews, *Ann. N.Y. Acad. Sci.*, 646 (1989) 334–356.
33 A. T. Andrews, I. Noble, S. Keeratipibul and J. A. Asenjo, *Physico Chemical Properties of the Main Matrix Proteins of Three Important Culture Vehicles*, in preparation.