A COMPUTATIONAL FRAMEWORK FOR HARNESSING DATA AND KNOWLEDGE FOR BIOPROCESS DESIGN

A thesis submitted to University College London in fulfilment of the requirements for the degree of

Doctor of Philosophy (PhD)

By

Jun Zhang

Department of Biochemical Engineering
University College London
Torrington Place
London
WC1E 7JE
UK

August 2012

The author confirms that the work presented in this thesis is the author's own. All of the information derived from other sources has been indicated in the thesis.

Dedicate this work to my parents for their constant love and support

ABSTRACT

Bioprocess design requires substantial resources for the required experimental investigation of the options for each bioprocess step. With the aim of reducing the amount of experimentation needed for bioprocess development, a new computational framework called Bioprocess Data and Knowledge Framework (BDKF) has been developed to explore the data and knowledge systematically.

In BDKF, the representation of four types of data and knowledge i.e. experimental data, ontologies, theoretical knowledge and empirical knowledge, have been established. The experimental data is the data that comes from previous experiments. The ontologies are the systematic description of the bioprocess terminologies used in the experimental data and knowledge. It can organize the terminologies of a domain as a hierarchy that allows the experimental data to be searched. The theoretical knowledge is the knowledge represented by formal definitions in the bioprocess, such as fundamental equations. The empirical knowledge is the knowledge obtained from practical studies, e.g. the relationships between different scales established through ultra scale-down experimentation.

Three reasoning functionalities, search, prediction and suggestion, have been established to imitate human reasoning on using data and knowledge. The search functionality finds relevant experimental data to the bioprocess design problems. With this data, the prediction functionality analyses the data and estimates the possible performance of the bioprocess step. The suggestion functionality produces solutions for further experiments that either confirm the solutions or narrow down the design space.

A prototype applying the BDKF approach to illustrate how to capture data and knowledge and how reasoning functionalities work for the operating conditions identification was developed for a case study on centrifugation. Design queries that represented relevant process material information and separation requirements were generated to initiate the BDKF approach. The prototype demonstrated that data from strain variants and data from different scales can be utilized through ontologies, theoretical knowledge and empirical knowledge.

A more complicated prototype was developed for the chromatography case study. The prototype introduced a hierarchical heuristic approach to solve the chromatographic process design problems, such as column selection, buffer composition identification and operating conditions determination. This prototype demonstrated that BDKF can be used for both screening and optimisation to propose several potential bioprocess solutions. Evaluation results of each prototype showed that the BDKF approach can make good performance predictions and suggestions for further experiments. It is very promising as an early stage process development tool.

Finally, a method for finding a design solution for a giving sequence by using mass balance calculations has been developed. A case study including centrifugation, filtration and chromatography has been examined. This demonstrated that BDKE method had the potential to allow all of the data and knowledge to be used for the whole bioprocess design. Therefore, the BDKF approach can provide a systematic way to harness bioprocess data and knowledge to enhance the efficiency of bioprocess development.

ACKNOWLEDGEMENT

This thesis would not be finished without the guidance and help of my supervisors and colleagues. First, I would like to express my deepest gratitude to my supervisor, Dr. Yuhong Zhou, for her excellent explanation, caring and patience that allow me to finish this challenging research topic. I also would like to deeply thank Prof. Anthony Hunter, who let me experience the beauty of logic techniques and further develop my computational skills. Both of the two great supervisors give me substantial help on research and writing.

I would like to thank Dr. Andrew Tait, Dr. Jean Aucamp, Dr. Balasundaram Bangaru and Dr. Sunil Chhatre for their valuable experimental data. They also help me to develop my knowledge about various bioprocess development techniques.

Finally, I would like to thank the funding bodies, Overseas Research Scholarship (ORS), Engineering and Physical Sciences Research Council (EPSRC), University College London Graduate School and Institution of Chemical Engineers (IChemE). Their financial support allow me to finish my research and also have opportunities to present my research results on international conferences in order to communicate with experts from different fields.

Contents

List of Tables

List of Figures

ABBREVIATIONS

BDKF Bioprocess Data and Knowledge Framework

CBR Case Based Reasoning

KBES Knowledge Based Expert System

ERM Entity-relationship Model

CHO Chinese Hamster Ovary

USD Ultra Scale Down

OD Optical Density

CE Clarification Efficiency

DW Dewatering level

ERR Error of Retrieved Rate

IEX Ion Exchange Chromatography

AC Affinity Chromatography

HIC Hydrophobic Interaction Chromatography

DBC Dynamic Binding Capacity

HHA Hierarchical Heuristic Approach

Chapter 1

SCOPE AND INTRODUCTION

This Chapter describes the scope and motivation about this thesis, then the contents of the thesis are introduced.

1.1 INTRODUCTION

There is a constant battle for mankind to fight diverse diseases. However, as medical experience develops and diseases are being better understood, novel biological drugs are being developed to fight life threatening diseases such as cancer, HIV, and cardiac problems. Huge investment worldwide has been put into appropriate drug discovery, and the development of potent biopharmaceuticals to meet the demand of novel therapies has become important.

The number of companies focused on direct biological products grows rapidly. For instance, in 2007, the global biopharmaceutical market increased to $94 billion, which represented the fastest growing segment in the pharmaceutical market, and by 2010 it had grown to $300 billion, which is nearly 50% of the pharmaceutical market (Walsh, 2010). There are some 240 monoclonal antibody products currently in clinical trials, along with an additional 120 recombinant proteins (Kling, 2011). Furthermore,when patents lapse over the next few years biosimilars are likely to come to the fore which are being driven by the rapidly growing markets of particularly China and India (Roger, 2010). The demand for novel therapeutics e.g. age-related diseases is expected to grow as the ageing population worldwide increases.

Nowadays recombinant technology is of growing importance particularly for manufacturing these novel drugs. However, due to the complexity of protein molecules and the limitation of our understanding of cellular processes, it will take much investigation to achieve a higher product titre from a given starting material, and finding the right conditions to ensure the stability of the molecules in a variety of harsh conditions during the recovery and purification stages of any industrial scale production. The drugs also have to be produced in the required (large) quantities of the relevant therapeutic proteins at high purity and quality with low cost and reasonable time scales. Currently it takes over 7 years to bring a new biopharmaceutical candidate to the market with an expenditure of perhaps $800 million per candidate (DiMasi et al., 2003). In order to reduce the time and cost requirements, effective bioprocess design is needed, amongst other important factors such as molecule identification.

Currently, the approach of bioprocess design for the production of a specific biomolecule uses the input of a large number of experiments to screen the potential steps for process sequence, then to optimise the possible combinations of operating conditions, and finally to validate the process. This leads to a need for a large volume of processing materials, much data analysis, long development times and high costs so making the whole development process very expensive.

Various experiments are involved in the bioprocess design, e.g. experiments for characterizing the feed material, experiments for screening the operating conditions to separate the specific feed material and etc. One side-effect of these experiments is the rapid growth in the data and knowledge being generated by bioprocess design. Yet, this data and knowledge have not been adequately explored because the data is not stored for reuse, and insufficient consideration has been given to the ways that it could be harnessed for bioprocess design.

However, if these data and knowledge were systematically reused, they may hopefully enable the reduction of the number of experiments, so that the biprocess design could become more rapid and less expensive. This would be particularly useful at the initial investigation

and evaluation stages with the identification of potential bioprocess solutions within the vast feasible space.

Design tools developed in the chemical processing industry (including the biochemical processing industry) are currently based on the simulation and optimisation which require establishing the whole process models. However, in the bioprocess, there are limited models available and any proposed model will need to be re-validated when the biological product or micro-organism has been modified so as to ensure that no detrimental consequence has occurred. As it is at the early stage of the bioprocess development, there will be (too) many variables that influence the bioprocess performance to be considered. Again, it will be very costly. Thus such design tools are more suitable for process design at late stage when the process sequence has been defined and process characterization has been completed. A more generalized selection approach would be very beneficial for the earlier stages of bioprocess design.

1.2 AIMS AND ORGANIZATION OF THESIS

This thesis focuses on developing a novel computational framework to harness the bioprocess data and knowledge for bioprocess design which can be used at both early and late stage. This computational framework is expected to reduce the bioprocess development cost and bioprocess equipment cost.

The development of the computational framework is proposed to be achieved by the following steps. First, after the examination of the current computer-aided techniques used for the process design, and how the data and knowledge can be harnessed, the methodologies for development of the computational framework are suggested. Then, a typical bioprocess step, centrifugation, will be used to illustrate how this computational framework works. Later, this computer framework will be further developed to make it a useful tool for bioprocess design. Finally, the methodology of how to find solutions for a bioprocess sequence will be addressed. However, due to the time limitations, the thesis will focus on downstream

processing steps, but new directions for the other bioprocess steps, e.g. fermentation, are suggested.

The other chapters in this thesis are arranged as follows.

Chapter 2 introduces the background information about current methodologies of biopharmaceutical needs and design as well as the applicable computer-aided techniques used for the process design.

In Chapter 3, the methodologies of the computational framework are presented. The concepts of this framework are defined first, such as what types of data would be used and how to represent them, what kinds of knowledge could be captured and how to represent them. Then, the reasoning functionalities to harness these data and knowledge for bioprocess design problem are defined and explained. An explanation is given to demonstrate how to represent the design problem as the design query that can be understood by this computational framework. Then, the flowchart of the computational framework is used to illustrate how the bioprocess data and knowledge can be systematically harnessed for the bioprocess design problem and how the user can interact with this framework.

In order to investigate this computational framework, in Chapter 4, the centrifugation case study is considered. The background knowledge for centrifugation will be given to explain what is the role of centrifugation in a bioprocess and how it works. Following the methodologies discussed in Chapter 3, the representation of centrifugation data and knowledge is discussed and how they can be harnessed for design will be explained. The evaluation of the computational framework for the centrifugation case study is presented and discussed.

To further investigate this computational framework, a more complex case study concerning chromatography is considered in Chapter 5. This involves domain of the more data and knowledge requirement. The hierarchical heuristic approach is introduced to solve the chromatographic process design problem which is much more complicated than the cen-

trifugation design problem. The results generated by hierarchical heuristic approach demonstrates how the computational framework can be a promising approach for bioprocess design.

The feasibility of implementing the computational framework on bioprocess sequence is discussed in Chapter 6. A typical sequence consisting of centrifugation, filtration and chromatography is considered. The methodology of how to use the computational framework to conjoin these bioprocess steps to form a sequence will be explained with at practical purification case study.

In Chapter 7, conclusions of the main contribution of the work are provided and suggestions for the future work are presented.

Chapter 2

RESEARCH PROBLEMS AND TECHNIQUES REVIEW

This chapter provides an overview about bioprocess design and issues, and the reviews about the current computer-aided techniques that have been used for process design. For this, section 2.1 introduces the typical procedure of bioprocess design and the challenges; section 2.2 explains the issues of accumulated bioprocess data; section 2.3 reviews the computational techniques for the process design; section 2.4 discusses the limitations of these computational techniques and section 2.5 summarizes the conclusions of the whole chapter.

2.1 BIOPROCESS DESIGN AND CHALLENGES

2.1.1 Characteristics of bioprocess design

For the biopharmaceutical industry, the bioprocess is making use of microbial, animal and plant cells and components of cells, e.g. enzymes, to produce the bioproducts that are used as a specific therapy (Liu, 2012; Potvin et al., 2012). Bioprocess design is an creative process that needs to identify requirements to an end bioproduct, equipment and a bioprocess sequence that satisfies the purity and yield targets that are defined by economic considerations. The bioproduct could be an antibody, specific cell or others, the equipment

refers to the devices used for producing or purifying bioproducts, such as the bioreactor, centrifuge, chromatography column, the bioprocess sequence indicates how the equipment work. Therefore, bioprocess design is exploring a large set of feasible solutions constrained by specific performance and economic requirements. It is a intricate procedure determined by the characteristics of the *bioproduct* and *bioprocessing*.

- *Bioproduct*: The bioproducts are usually complex molecules, such as the protein, hormone. They are generated by the host cells. The generation procedure has not been understood yet. Unlike the chemical molecule production that can be described by the specific chemical reaction formula, the bioproduct generation is not robust that may be impacted by various factors, e.g. temperature, chemical concentration. Their physical properties are not well characterized, e.g. the density, molecule weight, viscosity.

- *Bioprocessing*: The common situation in bioprocessing is that the bioproduct is found in low concentration in a complex aqueous chemical environment, this makes recovery and purification more complicated than in traditional chemical compounds. The bioprocessing usually consists of a set of bioprocess steps. Some units are specific to the bioprocess only, e.g. chromatography, and some units have complex operation procedures that may need to identify multiple internal related variables.

2.1.2 Introduction of bioprocess design

The bioprocess design is a complicated procedure involving the synthesis task and the design task, e.g. equipment selection, operating condition determination. These two tasks are usually completed by experimentation and data analysis. In this section, a general description about bioprocess design is introduced, and then the experimental work in bioprocess design is given followed by the introduction of bioprocess modelling.

The product biomolecule is characterized for its stability and physio-chemical properties. Based on this information, the feasible sequences could be identified based on engineers experience and previous recipes. These sequences involve a set of bioprocess steps, e.g. fermentation, primary recovery, purification and formulation (Alford, 2006; Knapp et al., 2012).

Figure 2.1 gives a typical conceptual sequence employed by biopharmaceutical manufacturing. Based on the preliminary sequences, experiments are implemented to determine the operating conditions of each bioprocess step (design task). The robust and optimal sequence is achieved by comparison studies.

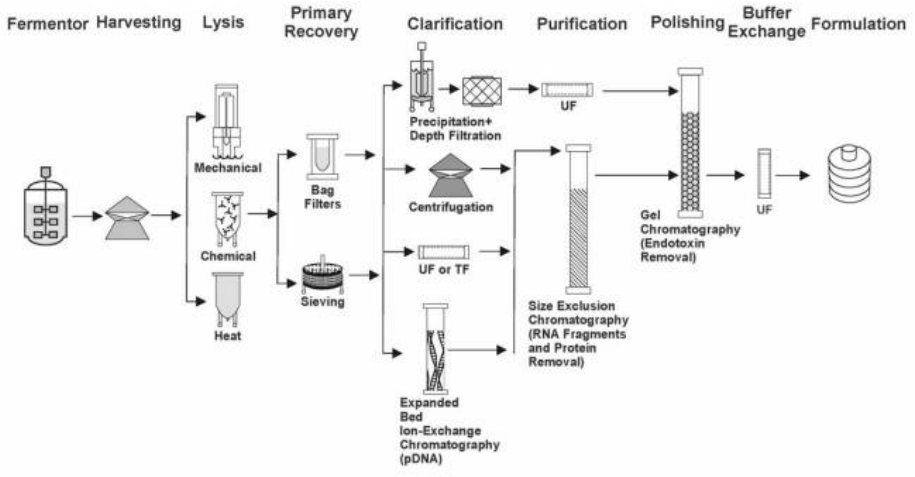


Figure 2.1: Illustration of bioprocess steps involved in the bioprocess sequence, source: Alford 2006. Permission to reproduce this figure has been granted by Computers and Chemical Engineering

2.1.2.1 Introduction of experimental work in bioprocess design

The amount of experiments required by design tasks depends on the number of the variables to be determined. The more variables to be studied, the more experiments are needed. These experiments are implemented at different scales, e.g. laboratory scale, pilot scale and manufacturing scale. The small scale experiments are usually used to screen the feasible design space at the early stage to save the time and money. The large scale experiments are necessary to verify the conditions identified by small scale experiments because some adverse factors maybe not show in the small scale equipment. In order to use the small scale experiments to simulate the performance of large scale experiments, a novel experimental approach, called ultra scale down (USD) approach, has been established. By using this approach, the desired large scale operation can be simulated in parallel with less processing materials and a faster speed. The review of USD achievements and implementations can be found elsewhere (Kumar et al., 2004; Titchener-Hooker et al., 2008; Chhatre and Titchener-

Hooker, 2009; Li et al., 2012).

The bioprocess steps involved in a sequence are referred as two parts, namely upstream processing and downstream processing. The upstream processing aims at producing the target biomolecule by optimising cell lines. The experiments of upstream processing are used to determine the operating conditions with respect to the specific titre requirement, e.g. which cell should be selected, what is the temperature or pH or dissolved-oxygen concentration or stirrer speed? what is the operation mode of fermentation, batch or semi-batch or continuous batch? In addition to the operating conditions, the bioreactor and the auxiliary service facilities also need to be considered, e.g. air supply, sterilization equipment, steam generator, pipe lines. The downstream processing consisting of other bioprocess steps performs recovery and purification of the target biomolecule, e.g. filtration, centrifugation, homogenization, membrane filtration, chromatography, crystallization and drying. The downstream processing design is complicated. Each individual step needs a series of experiments to determine the operating conditions, and the interactions between these steps also require substantial experiments to examine. Furthermore, additional experiments are needed to validate the upstream and downstream processing in order to ensure the purity and yield of the whole bioprocessing.

2.1.2.2 Introduction of bioprocess modelling in bioprocess design

Bioprocess modelling employs a series of mathematical equations to predict the performance of the bioprocess step or sequence, for example, economic projection, scheduling, cost of goods analysis. The process model is used to simulate the process operation by giving the information of inlet and outlet. It may help to understand how the process works, how the processing material components change (Rouf et al., 2001). The business model describes the relationships between the process operating conditions and the economic variables. It can determine which process scenario should be adopted for manufacturing by estimating the economic performance (Zott et al., 2011).

Bioprocess models are also used to further understand the bioprocess step. For example,

the computational fluid dynamics (CFD) describes the motion of fluid based on the Navier-Stokes equations (Aslam Bhutta et al., 2012), and it is used to analysis the aqueous environment in a specific equipment which would contribute to performance prediction (Boychyn et al., 2001).

Since the computers can perform complicated calculation fast, various tools or algorithms have been established to implement the modelling work. The reviews about these computer-aided techniques are given in section 2.3 and their limitations are explained in section 2.4.

2.1.3 Bioprocess design challenges and solutions

The obvious issue of bioprocess design is the requirement of substantial experiments. Since ‘one-step or ‘automation technology is not available yet for the bioprocess design, assessing any new idea or adjustment of operation still relies on experiments. These experiments not only burden the time of profiting, but also increase the cost of the goods. Therefore, the prior challenge is to reduce the number of required experiments in order to speed up the bioprocess design.

Although the USD approach can allow the experiments to be done with less time and processing material, the amount of experiments still cannot be minimized. In addition, one side-effect of these experiments is the continuous accumulation of experimental data and knowledge which conduct a data-rich world to the pharmaceutical industry, e.g. the amount of data generated during the drug devilment doubles every five years (Venkatasubramanian, 2009). Exploring this data and knowledge would be a promising way to narrow down the design space to be explored in order to reduce the amount of required experiments. For example, using historic data to predict the formulation of fermentation without experimentation (Jewaratnam et al., 2012). The benefits of using accumulated data for bioprocess design have been discussed elsewhere (Charaniya et al., 2008, 2010; Schaub et al., 2012), which provide the solid theoretical support.

However, this accumulated data has not been well used yet due to the characteristics of data and less attention has been paid on the way of how to use it. Therefore, before discussing the solutions to bioprocess data and knowledge utilization, the issues of bioprocess data should be described which may help reader to understand why the data has not been fully utilized yet, and then the review of computer-aided techniques is followed to explain why the current techniques are not suitable for this task.

2.2 BIOPROCESS DATA ISSUE

In 2005, Morris did a survey about the pharmaceutical development data utilization (Morris et al., 2005). In this survey, questionnaires were sent to the people in the development department of different pharmaceutical companies, the responds were used to benchmark the current state of data during the pharmaceutical and biopharmaceutical process development. Results showed that less than 10% of the data was utilized, which indicates that the benefits of data utilization have not been realized.

The inefficient data utilization could be caused by the improper data storage. The bioprocess data generated from various analytical instruments is in the form of images, spectra, tables, and they may be kept in the different carriers, e.g. laboratory desktop, personal laboratory notebook, personal desktop.

The data recorded in various format would make the data difficult to be harnessed. An typical example is data searching, for instance, if the information of bioproduct properties was kept as a picture, a table or a figure, it is difficult to search these information systematically. This problem has been pointed out by 'traceability' in the survey. More than 70% of respondents believed that the requested data can not be searched or only can be partially searched, and all of the respondents thought that the data can not be fully traced (searched by purpose) (Morris et al., 2005).

The data kept in different carriers would not be good for data sharing, e.g. the data kept

in the personal laboratory notebook would be difficult for other people to use. This situation was also reflected as 'visibility' in the survey, all of the respondents thought that they did not know where the required data was and how to access the data (Morris et al., 2005).

The bad data storage may cause two consequences. First, people has to spend considerable time on data searching. The survey indicated that two-thirds of the participants spent approximately four hours per week which accounted for 10% of the total working time per week. Second, people have to repeat the previous experiments to reproduce the data that can not be located. The repeated work was also indicated by the survey, almost 90% of respondents thought that they have to spend at least 10% of their working time to repeat the previous experiments. These two negative consequences may delay the progress of bioprocess development, e.g. spending long time to locate the requested data.

2.3 COMPUTER-AIDED PROCESS TECHNIQUES REVIEWS

2.3.1 Introduction

The other issue of using data for bioprocess design is lacking for suitable computational techniques. Although several statistical algorithms have been established for mining information from the accumulated data, e.g. the clustering algorithm that assigns the data into groups to extract the common characteristics (Jain et al., 1999; Kamimura et al., 2000), artificial neural network that can model the relations of data by the specific pattern recognition (Liu et al., 2012). These techniques are preferred to mining the information behind the data to answer the questions like what is the relation between these two variables? This information is helpful for modelling work to understand the specific process, but it may not easy to answer design questions directly, e.g. what is the equipment? what is the pH?

Since the computers are more powerful and more portable, people have developed various software that integrates different algorithms and models to provide solutions to the design

problem. However, these techniques have a certain degree of limits to harness the bioprocess data and knowledge for bioprocess design tasks. In the following, the review of computer-aided techniques is presented to help reader to understand the reason.

The computer techniques used for the process design can be classified as three types, i.e. flowsheet simulation, mathematical programming and knowledge based system. Since the literature about using computer-aided process techniques for bioprocess design are quite limited, the review in the following expands along the chemical engineering subject.

2.3.2 Flowsheet simulation

2.3.2.1 History of flowsheet simulation

Flowsheet simulation technologies started from 1960. At that time, research focused on the analytic aspects of design and resulted in the development of flowsheet simulator for the steady-state processes that performs a mass and energy balance of a stationary process (a process in an equilibrium state), e.g. the FLOWRTAN developed at Monsanto (Rosen and Pauls, 1977) which can calculate the steady state operation of the defined process and generate the information for each unit (e.g. inlet and outlet temperatures) and stream (e.g. chemical compositions).

In order to enhance the flexibility of steady state simulation, dynamic process simulation has been proposed which is an extension of steady-state process simulation whereby time-dependence is built into the models, e.g. accumulation of mass and energy. The first prototype is SPEED-UP developed at Imperial College London (Perkins et al., 1982; Pingaud et al., 1989). In this system, the speed of variations of the variables of the process with respect to changes were obtained, which can help for a better understanding of the process under different considerations.

2.3.2.2 Introduction of simulator

The flowsheet simulation approach requires well defined information to produce the results by harnessing the integrated mathematics models (Zhao et al., 2012), and it employs a set of mathematical equations to describe the interactions between steps involved in the processing, e.g. mass balancing, costing. Each flowsheet simulation approach consists of a simulator, and its general composition is given in Figure 2.2.

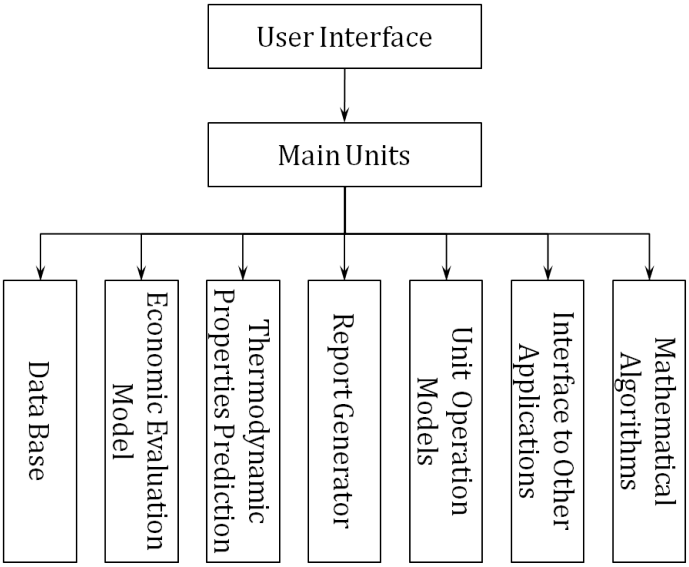


Figure 2.2: Main modules consisted in simulators

Overall of these units, the key part is the models that play as agents to describe the work of different units (Astrom, 2011). A large amount of work has been done over the past few decades to develop rigorous models of unit operation of chemical process (Marquardt, 1996). These models represent the complicated equilibrium process, reaction systems and energy balances encountered in the chemical processing. Because the chemical processes are generally operated continuously at the steady-state, the equations which represent the material and energy balances are non-linear and algebraic naturally.

2.3.2.3 Approaches used in flowsheet simulation

Three approaches are mainly used in the development of chemical process simulators, i.e. *sequential modular, equation-oriented and simultaneous modular* (Zhang et al., 2008,

2010). Their introductions are presented as follows.

Sequential Modular: It is the most commonly used approach in chemical process simulators (Liang et al., 2009; Marin, 2011). In this approach, the models are organized into unit operation modules which can calculate unit output stream values by giving information of unit input stream and equipment operation.

This approach is easy to be understood by users because the information flow follows the flow of materials in the steams. Thus it is easier to debug the errors if it is necessary. But the approach is not efficient in handling complicated processes which have a large number of nested recycle loops, and it dose not perform well for optimization problems since the entire flowsheet may take much calculation time before the optimum solution can be generated.

Equation-Oriented Approach: In this approach, all of equations, e.g. unit operation model equations, mass balance equations, are treated as the non-linear equations to be solved simultaneously (Ishii and Otto, 2008). Since these equations are processed simultaneously, solving recycle loops is not required. The Equation-Oriented Approach is reasonable for optimization problem which is usually formulated as the non-linear problem.

However, there are still lots of problems that restrict the approach from being adopted by industry. First, it demands more mathematically work for the large system of non-linear equations. Second, it does not perform well on reliability since it is complicated to convert these equations for real industry problems, e.g. it is difficult to fit the industrial problem in the specific liner equation. Finally, the flexibility of the equation-oriented approach allows the user to develop inconsistent specifications about process which makes the error checking difficult, e.g. the same information can be represented by different forms or different symbols.

Simultaneous Modular Approach: In this approach, the unit operation modules are represented in the same way as the sequential modular approach. Each module is used to generate

the output stream results by giving the information of input stream and equipment operation. But a fundamental difference exits, i.e. the following module for each unit operation should be written to relate each output value approximately to a linear combination of all input values (Chen and Stadtherr, 1985).

These linear equations are combined to form a large set of equations in order to describe the whole flowsheet. These equations can be calculated simultaneously as they are treated in the equation-oriented approach. Therefore, this approach combines the advantages and disadvantages of the previous two techniques. Similarly, the fundamental problems of the simultaneous modular approach include the representation of flowsheet level equations which requires non-linear equation solver and the reliability of results.

2.3.2.4 Commercial simulators

During the past decades, various simulators have been established. In the following, two commercial simulators used for steady state and dynamic state flowsheet simulations are introduced.

ASPEN PLUS (AspenTech, US) is a steady state flowsheet simulator, it consists a library of chemical unit operation models and uses the sequential modular approach to describe and process the chemical units in a stepwise way.

gPROMS (PSE, UK) is a dynamic state flowsheet simulator, whose central part is the gPROMS Modelbuilder. In this part, the process is described in terms of both the physic-chemical behavior of the unit and the external actions imposed on it. It employs the simultaneous modular approach that allows the calculation involved in each model to be calculated simultaneously.

Due to the time limitation, it is not possible to introduce every commercial simulator, however readers who are interested in the commercial simulators can find related information in Table 2.1, where the names of major simulator and companies are presented.

Table 2.1: Common commercial simulators with maintenance website

Simulator name	**Website information**
steady state flowsheet simulators	
Pro/II	www.quorum.simsci.com
ProSim Plus	www.prosim.fr
HYSLM	www.hyprotech.com
METSLM	www..ozemail.com.au
CHEMCAD	www.chemstations.net
dynamic state flowsheet simulators	
CHEMCADE REACS	www.chemstations.net
ProSim BatchColumn	www.prosim.fr
Aspen Dynamics	www.aspentech.com
Batch Plus	www.aspentech.com
Femlab	www.femlab.com

2.3.2.5 Use of simulators for bioprocess design

A set of simulators serving the biochemical process simulation have been developed (or improved based on the chemical process simulators because differences exist between the chemical process engineering and biochemical process engineering (see section 2.1.1)). In the following, three packages are considered, namely *BioProcess Simulator, SuperPro Designer and BioSolve*. These three packages were considered because their users cover various biopharmaceutical companies.

BioProcess Simulator (BPS): BPS is an extension of ASPEN (AspenTech,US) with a chemical process simulator (Petrides et al., 1989). Given a specific flowsheet, BPS carries out material and energy balances, estimates the size and cost of equipment and presents the economic evaluation. It is a steady-state simulator but can handle time-dependent processes with limitations. For instance, the equations used to describe fermentation in a batch process are integrated in the fermentation time. When the fermentation process is finished, the broth

is transferred to a holding tank before it starts being processed by the dowstream units. The economic evaluation in BPS is derived from the chemical process simulator AspenPlus, and the calculation mode is based on the practical project data. Thus, the results are close to the real situation, but it requires data to be updated in time to keep the economic calculation reliable.

Rouf used Bioprocess simulator (BPS) to assess the cost of tissue plasminogen activator (t-PA) generated from Chinese hamster ovary (CHO) cells (Rouf et al., 2001). The process of Ethanol production as well as the sensitivity of cost of production factors, e.g. material cost, annual production rate were determined (Fan and Lynd, 2007). Lee and Wankat used Aspen Chromatography, an extension of BPS, to simulate the isolation of intermediate solutes in ternary mixtures (Lee and Wankat, 2009).

SuperPro Designer (SPD): The functionality of SPD (intelligen, US) performance is similar to BPS. It produces the simulation results for a given flowsheet (Rouf et al., 2001). The information that is used to describe the flowsheet is different from BPS and SPD. In BPS, most required information is available in the literature or can be estimated by the built-in correlations, e.g. unit inlet stream, unit operating conditions. But in SPD, the information required is more specific. For instance, in column operation, the binding capacity and yield estimation should be defined for SPD simulation, and this information is needed to be confirmed by experiments. Unlike the BPS, the economic simulator in SPD are specially developed for the bioprocess design. Thus, it can do more bioprocess related economic analysis, such as equipment cost, labour cost, total cost, fixed cost as well as the sensitivity analysis.

SPD has been used to optimize the large scale biopharmaceutical facilities, e.g. bioreactor (Toumi et al., 2010), to simulate the production of small molecule active pharmaceutical ingredients (APIs) (Petrides et al., 2002a). The scale, titre and product changeover frequency of single-use system in biopharmaceutical processing was assessed by SPD (Papavasileiou et al., 2008). The process of therapeutic monoclonal antibody (Mab) was simulated by SPD to illustrate how to solve the problems of debottlenecking and cycle time reduction (Petrides

et al., 2002b).

BioSolve: BioSolve (Biopharm Services, UK) is a newly developed package that is famous for economic analysis (Sinclair and Monge, 2010). This software is based on Excel tool (MicroSoft, US), and the each model is developed in Visual Basic. The flowsheet is described as tables, e.g. each column represents one specific process unit and each row indicates one operating condition from the predefined database or specification given by users, such as the resin cost, binding capacity, life cycles. This software can perform economic analysis based on various aspects, e.g. time cost, labour cost, buffer cost, and it also can be used for sensitive analysis, such as the total cost breaking down. Since the economic calculations rely on the information kept in the database, keeping the database updated is necessary for the reliable results.

BioSolve has been used to evaluate the three cell culture techniques based on economic considerations (Lim et al., 2011). The impacts of manufacturing cost on process development (Sinclair and Monge, 2010) and the impacts of geography on cost of bio-manufacturing (Sinclair, 2010) have been assessed.

In addition to the commercial simulators, other academic applications were also reported. A model was established to describe the generation of enzyme alcohol dehydrogenase (ADH) from Saccharomyces cerevisiae, and its evaluation results showed the effectiveness of simulation on process design (Bulmer et al., 1996). A conceptual framework integrated with SIMBIOPHARMA was developed to model the biopharmaceutical manufacturing plant, the case study illustrated the importance and the effectiveness of simulation technique on making economic and operational decisions by simulation tool (Farid et al., 2007). A dynamic simulation framework was built up to determine the optimal column size for current and future fermentation titres which could facilitate the downstream processing design (Stonier et al., 2009).

2.3.3 Mathematical programming

The process design problem can be formulated as a mathematical optimisation problem and then solved by numerical methods, which is a sub discipline of process system engineering (PSE) (Rippin, 1993; Winston et al., 2003). In chemical engineering, the chemical process synthesis has been formulated as into Mixed Integer Nonlinear Programming (MINLP) and hierarchical decomposition (Grossmann et al., 1999).

2.3.3.1 Introduction of MINLP and hierarchical approaches

MINLP aims at finding solutions to the problems that have integer and real variables as well as both of linear and nonlinear constraints (Bussieck and Pruessner, 2003). A series of reviews on MINLP have been given by Grossmann (Grossmann, 1985; Papoulias and Grossmann, 1983; Karuppiah and Grossmann, 2006; Duran and Grossmann, 1986).

The fundamental idea about hierarchical approach is from the analytic hierarchy process (AHP) proposed by Saaty (Saaty, 1980). Since then, this technique has been extensively refined and studied. It is implemented widely for group decision making and used for variety of decision situations in management and industry process (Saaty, 2008).

For the chemical engineering, the Hierarchical approach breaks down the chemical design problem into a number of decision levels and solved it in a hierarchical order. The chemical process prototype was initially described by Douglas (Douglas, 1985), in which the chemical process design problem was decomposed into five hierarchical levels: Level 1 Batch vs. Continuous; Level 2 Input-Output structure of the flowsheet; Level 3 Recycle structure of the flowsheet and reactor considerations; Level 4 Separation system specifications and Level 5 Heat exchanger network. A decision is made at each decision level, but additional structures would be refined at the later decision levels when they are processed. Heuristics are used at all levels of decision to fix the structure of the flowsheet, impose design constraints and substantially reduced the number of alternatives to be considered. Each decision level terminates in the specific economic analysis, and it would be helpful to dis-

card the alternatives with poor economic performance which would not be realized in reality. This may be helpful to reduce the search space for next level, similarly the later level would be guided by the economic analysis of the early level decisions. The hierarchical approach provides a way to decompose the whole process design problem into reasonable sub-design problems. Each sub-design problem has the clear criteria for decision making, e.g. economic constraints. This is a 'logic' process that the decisions making is the same type and in the same order as an engineer would make naturally.

2.3.3.2 Use of mathematical programming for chemical and biochemical process design

The MINLP and hierarchical approach have been applied together for the chemical process synthesis for more than a decade (Daichendt and Grossmann, 1998). They were used to solve the synthesis problem of heat-induced separation networks (El-Halwagi et al., 1995), utility optimization (Savola et al., 2007) and decision making of the supply chain design under uncertainty (You and Grossmann, 2010), mass and energy of membrane-based gas separation (Gassner and Marchal, 2010), and the synthesis of mass and energy network in a separation process (Li et al., 2011).

Implementation of MINLP and hierarchical approach on biochemical process development was proposed by a new framework, called DYNSIM, which was introduced to represent the bioprocess steps and decompose the whole sequence as a set of inter-related equations for solutions (Wai et al., 1996). These two techniques were also used to optimized the size and time of the batch plants simultaneously (Asenjo et al., 2000; Pinto et al., 2001; Canovas et al., 2002).

2.3.4 Knowledge based system

2.3.4.1 Introduction of knowledge based system

The knowledge based system is a branch of applied artificial intelligence (AI) and developed since 1960s (Stephanopoulos and Han, 1996). The basic idea of knowledge based system is transforming the human knowledge into the rules to serve the generation of solutions to the queries. It behaves as an advisor to give suggestions or further explanation that indicates the logic is usually involved (Alavi and Leidner, 2001).

All of knowledge based systems involve the commercial expert system shells which provide an geographic interface between the system and users. These shells can interprets and executes the production rules that users may have, and they also provide the environment to develop the knowledge base, e.g. object oriented programming environment that can provide better organization of declarative knowledge. In this approach, the knowledge is organized by classes, and each class consists of different subclasses, such as the specific class, generic classes. For instance, the separator class can be represented as the class of unit operations; and in turn, the filter class can be represented as the separator class.

Four main components of the knowledge based system are usually an interface, a knowledge base, a knowledge engineering tool and specific user interface (Dhaliwal and Benbasat, 1996). The performance level that can be achieved by a knowledge based system depends on the quality of knowledge that has been captured. The system growth is supported by the size and quality of knowledge base which is a very intensive and expensive proposition.

The knowledge based system has drawn lots of academic attention over past 25 years, e.g. there were 10439 knowledge based system related articles from 1995 to 2004 (Liao, 2005). In the following, three approaches, namely the rule based approach, neural network approach and case based reasoning approach, were concerned, since they were applied for the process design frequently.

2.3.4.2 Rule based approach and applications

The rule based approach is a typical knowledge reasoning approach. It offers a natural way to capture and document knowledge. The knowledge is stored in the form of different rules, which are represented as the typical syntax 'IF...THEN'. A simple example is:

IF the pH of the Buffer is less than 7,
THEN the buffer material is an acid.

These rules can be used by reasoning programmes for the appropriate conclusions. In order to process the uncertainty involved in the problem, the fuzzy logic is used to simulate the human reasoning by allowing the computer to perform less logically reasoning than the conventional computer. Recent applications of rule based approach include management of supply chain performance (Olugu and Wong, 2012), weather prediction (Awan and Awais, 2011) and clinical risk assessment (Kong et al., 2011).

2.3.4.3 Neural network approach and applications

The neural network approach simulates the biologic neural network to process the information in massive parallel. The artificial neural network consists of a set of artificial neurons which is a specific programming construct that simulate the properties of biological neurons. Each artificial neuron receives an input and then generates an output for the next artificial neuron. This procedure is analogous to the electrochemical impulse passing through the biological neurons. Artificial neural networks may solve the artificial intelligence problem without necessarily creating a model of a real biological system. The recent applications of neural network approach includes: prediction of metabolic syndrome by using historic clinical data (Hirose et al., 2011), fault diagnosis (Jayaswal et al., 2011).

2.3.4.4 Case based reasoning approach and applications

The CBR proposes that a new problem is solved by noticing its similarity to previous solved problems and by adapting their known solutions instead of generating a solution from

scratch (Watson and Marir, 1994). The case in the CBR represents the past problem situation that includes specific information used to describe the situation, e.g. the chemical process recipe. The case representation allows the previous information can be organized in a consistent structure for reusing. Then the scientists proposed a logical way to reuse these represented cases for new problem solving which is decomposed into the following steps. Retrieving relevant cases from the case-base/case memory, selecting a set of best cases, deriving a solution, evaluating the solution and storing the newly solved case in the case-base/case memory.

This procedure above intimates the human behavior on resuing the previous experience for new problem solutions. Aamodt (Aamodt and Plaza, 1994) summarized the procedure above as '4R' rules, i.e. *Retrieve, Reuse, Revise* and *Retain*. This process can be illustrated by Figure2.3, where the initial description of a problem is defined as the targeted problem.

- *Retrieve* is to find the one or various previous cases that are similar to the target problem. The similarity between two cases is measured by function that compares features between the target problem and case in the case base. Similarity calculation depends on type of features, e.g. number.

- *Reuse* is used to propose a solution to the targeted problem, the solution could be derived from the retrieved cases. This would be the starting point to form a specific solution to the target problem. Reusing previous case to solve targeted problem would be trivial, because a gap exists between the previous case and targeted problem at most of time. Thus, adjustment would be needed for modifying the previous case to fit in the problem. The procedure would be complicated because it requires extra knowledge and rules to achieve this target.

- *Revise* proposes to test the solution to verify its adequacy. Many approaches may be adapted, e.g. simulation approach, experimental validation and etc.

- *Retain* makes the CBR system learn new information from the solution, the new case and its associated solution would be formalized as a new case that can be stored in the case base.

Figure 2.3: Cycle of Case Based Reasoning. Source: Aamodt and Plaza, 1994. This figure restricts access and has been removed.

Give the new problem, CBR would retrieve similar cases, reuse the and revise the case for the targeted problem and then retrain the new solutions for next cycle. The '4R' rules imitate the human typical behavior of problem solving by reusing previous experience or knowledge. It demonstrates that reasoning by reusing past cases is a powerful and frequently applied way to solve problems for humans. This claim is also supported by results of cognitive psychological research (Kolodner, 1997). Studies have also given the empirical evidence for the dominating role of specific, previously experienced situations in human problem solving (Ross, 1989). Anderson (Anderson, 1996) has explained that people use past cases as models when learning to solve problems, especially at the early stage of learning. Morris (Morris and Rouse, 1985) indicates that the use of past cases is a predominant problem solving method among experts as well. Studies of problem solving by analogy also shows the frequent use of past experience in solving new and different problems (Gentner, 1983). All of this research demonstrates that the previous cases can contribute to new problem solving and it is a general way adapted by people for new problem solving.

The CBR approach has various commercial implementations, e.g. ReCall by ISoft S.A. (France) which is a generic tool that has been applied to develop applications on fault diagnosis, bank loan analysis, teaching, risk analysis, control and supervision.

2.3.4.5 Use of knowledge based system for chemical and biochemical process design

The knowledge based system has not been widely applied for chemical or biochemical process development, and most of the applications were initiated by academia. Some relevant research is given in the following.

Most applications of rule based approach focused on the rule representation. A language was created to represent the chemical information as the rules for process design tasks (Weininger, 1988). The representation of the information that was extracted from the previous data as rules for the design of styrene-butadiene latex production has been defined (Nomikos and MacGregor, 1994). The interactions between biomolecules as a set of rules were represented in order to generate the reaction formulations for the target biomolecule (Faeder et al., 2005). Biochemical Abstract Machine (BIOCHAM) provided a precise semantic environment to represent the relations between biomolecules, and allowed the properties to be queried to help predict the unknown properties of new biomolecule (Chabrier-Rivier et al., 2005). React(C), an expressive programming language, was developed to represent the biochemical reactions with constraints (John et al., 2011). A rule based system involving 49 production rules and a fuzzy logic was used to control the penicillin production (Cosenza and Galluzzo, 2012).

Most applications of neural network approach focused on the process control, e.g. growth of Saccharomyces cerevisiae (Bulsari and Saxen, 1991). A Central Nervous System (CNS) was developed to select the desired molecule to optimize the drug discovery (Ghose et al., 2011). A hybrid neural network has been demonstrated to predict the biomass concentration of the bioreactor (Zhang et al., 2012). A model based on the neural network was used to predict the level of ethylene dichloride in a industrial fixed-bed catalytic ethylene oxide

reactor (Rahimpour et al., 2011). The accurate results demonstrated that it was a promising approach for process control.

The applications of case based reasoning approach concentrated on the process synthesis. Pajula et al. used the case based reasoning for separation synthesis, and the distillation column served as the criterion to determine the optimized solution (Pajula et al., 2001). Stephane used the case based reasoning to select the recipe for the new chemical reaction (Stephane and Marc, 2008). In this research, the previous chemical process recipes were represented as the combination of process features, e.g. compounds, pressure, temperature. Given the design task, a set of relevant source cases were searched by similarity measurement, and the desired case was retrieved based on the economic constraint. The small difference between the real solution and predicted solution in distillation design demonstrated that the CBR can produce accurate solution for chemical process design.

2.4 LIMITATIONS OF CURRENT COMPUTER-AIDED TECHNIQUES FOR BIOPROCESS DESIGN

Although the publication showed the great potential of computer-aided techniques on bioprocess design task, limitations still exist. In the following, the limitations of three computer-aided techniques referred above are presented.

2.4.1 Limitations of simulator

Based on the introductions of simulation techniques, it is known that the simulator requires complete input information, e.g. input stream of unit, output stream of unit, operating conditions, to call available models embedded in the system to generate the results. The results would not be generated if the required information is missing or not defined. However, this complete information is usually available at the late stage when the bioprocess sequence has been finalized. Therefore, the simulation tools are preferred to be used to solve the optimization of the bioprocessing candidates, e.g. cost of good, production time, annual yield.

The simulator cannot generate requested information at the early stage of bioprocess design, e.g. the specific operating conditions. Finding appropriate process data to verify the specific models may take considerable time and effort, which is unlikely to take place in early phase.

2.4.2 Limitations of mathematical modelling

The mathematical modelling needs to describe the process behaviors into a set of linear or nonlinear equations. However, it is very difficult to formulate the biological reactions or the bioprocess steps since there are several variables involved and the relationships between the variables and process performance are not fully understood. For example, few theories can explain the relations between the environmental factors and the protein generation because the protein behavior in the microorganisms is yet to be discovered. There is limited number of bioprocess models available and verified by pilot plant scale operations, hence more bioprocess models need to be developed.

2.4.3 Limitations of knowledge based system

The design problem is usually not well supported by computer-aided simulation tools and mathematical modelling unless it can be formulated as a set of objective functions and constraint. However, the knowledge based system can apply the knowledge represented by the designer's expertise to constrain the design problem without any formulation, e.g. excluding the inaccurate design space automatically. Hence,using knowledge based system for process design is promising. However, limitations still present when using the knowledge based system for bioprocess design.

The performance of rule based approach depends on the captured rules. These rules may come from persons experience or expertise. Hence, maintenance of these rules would be a challenge. For example, the new developer would not understand the semantics of captured rules unless he is fully trained. Furthermore, different experts may have different opinions

about the bioprocess design, hence the captured rules may be inconsistent leading to produce the conflict conclusions. Dealing with the conflict conclusions may require far more complicated techniques and would also impact the accuracy and the reliability of the results. So, the rule based approach may not be strong enough to process the bioprocess design task which may include the inconsistent knowledge.

The neural network approach requires the specific knowledge and well defined information to develop the neural with respect to the specific task. This is also the reason why most implementations of neural network approach were about process synthesis or optimization, because the economic variables and relations have been understood well so that the neural network can be constructed easily. For the bioprocess design, especially at the early stage, most information is unknown, e.g. properties of processing material and target biomolecule, and much process knowledge is still being learned, e.g. chromatography, hence building up the neural network would be very difficult. Based on this, the neural network approach would not be suitable for the bioprocess design, especially when the information is incomplete.

Although CBR approach is a promising approach to reuse data and knowledge for new problem, three limitations would appear when using the CBR approach to solve the bioprocess design problems. First, CBR is an easy operation, but requires precise strategies for retrieve or revise step, and these are crucial to elaborate a good solution. The complicated units and processing materials employed by bioprocess would make the development of a suitable retrieving strategy difficult. The retrieving strategy determines the quality of retrieved data that impacts the accuracy of the solution. Second, the similarity calculation adapted by CBR approach concerns on the numerical specifications, and it is not suitable to determine the similarity of the terminological specifications which are the common syntax used in the bioprocess information. Third, the CBR approach is not flexible enough for knowledge utilization. Studies demonstrated that it may have ability to harness the simple equations in revise step, but the knowledge required by bioprocess design task is far more complicated than that, e.g. mathematical models, background information of equipment.

2.4.4 Summary

Using simulator and mathematical modelling for bioprocess design needs to formulate the specific task into a set of mathematical equations. Since the bioprocess is complicated, the formulation is not easy to be realized. The knowledge based system provides a way to solve the bioprocess design task without mathematical formulation, however the knowledge referred in the knowledge based approach is too specific to make it flexible to deal with the bioprocess knowledge that has different formats and expressions. These limitations indicate that the three computer-aided techniques are not suitable to address the bioprocess design challenge, i.e. narrow down the design space to be explored.

Furthermore, all of the three techniques require well-defined, complete and consistent input information which indicate that they are not good choices to harness the incomplete and inconsistent bioprocess data and knowledge. Therefore, a new computational approach is required to fill this gap, and this approach not only is able to harness the various data and knowledge, but also can generate the solutions to reduce the number of experimentation.

2.5 CONCLUSIONS

This chapter introduces a general way of bioprocess design, i.e. experimental work and modelling work. A description is given to illustrate how to use the experiments to do the design tasks, followed by the introductions about the modelling work which mainly serves the bioprocess synthesis tasks. Since the substantial experimental work require long time and great expense, the bioprocess design challenge is summarized as reducing experimental work to speed up the bioprocess development.

One side-effect of experimental work is the accumulation of experimental data and knowledge, systematically exploring this data and knowledge may help engineers to narrow down the design space of the new bioprocess design problem. Usually, mining information is realized by computer-aided techniques which have been established since two decades ago. Systematical reviews of flowsheet simulation, mathematical modelling and knowledge based

system explain how they work on the process development. Then, the limitation analysis of three techniques shows that they are not appropriate for the bioprocess design challenge. Furthermore, the specific input requirements also indicate that these three techniques are not able to harness the incomplete and inconsistent bioprocess data and knowledge for the bioprocess design.

Based on these explanations, a specific achievement to be addressed by this thesis can be concluded as that a new computational approach needs to be developed to harness the bioprocess data and knowledge to narrow down the bioprocess design space to be explored. Since the biological reaction is complicated and the bioprocess has not been fully understood, the mathematical modelling would not be a efficient method for data and knowledge utilization. Instead, due to the advantages of knowledge based system, especially the case based reasoning that demonstrates the strong capability of reusing the previous information for the new problem, a logic based approach should be developed to realize gaining information from the accumulated data and knowledge to facilitate the bioprocess design.

Chapter 3

BIOPROCESS DATA AND KNOWLEDGE FRAMEWORK

3.1 INTRODUCTION

The Biprocess Data and Knowledge Framework (BDKF) approach is proposed to harness the bioprocess data and knowledge for bioprocess design. In this framework, four kinds of data and knowledge are considered, i.e. experimental data, ontology, theoretical knowledge and empirical knowledge, and the three reasoning functionalities are developed to use using bioprocess data and knowledge in a logic way, i.e. search, prediction and suggestion. For a specific bioprocess step, the implementation of BDKF approach is called a *BDKF system*.

The aim of this chapter is explaining the methodologies of BDKF approach. The sections of this chapter are arranged as follows, Section 3.2 describes the experimental data and the representation of it; section 3.3 describes the ontology and the representation of it; section 3.4 discusses the available knowledge that can contribute to data utilization; section 3.5 describes the reasoning functionalities and how they work; and section 3.6 describes the flowchart that how the BDKF approach harness data and knowledge for the new bioprocess design problem.

3.2 EXPERIMENTAL DATA AND REPRESENTATION

3.2.1 Definition of experimental data

The experimental data is the data generated from the previous bioproecss experimental studies, e.g. process and product characterization studies, specific modeling studies. It is a type of fact about previous experimental observations, e.g. what is the flowrate of a disk stack centrifuge to harvest mammalian cells? The experimental data is usually kept in different places, e.g. different laboratory notebooks, different computers, and its formats are inconsistent due to the different experimental purpose, e.g. the gel band generated for molecule qualitative analysis, the diagram generated to illustrate the relationship between the operating conditions and the performance. Since this experimental data inherits the characteristics explained in section 2.2, using them directly may be difficult. Thus, a systematic representation is required to make the experimental data be formalized and stored structurally.

3.2.2 Structure of experimental data

For each bioprocess step, the feed stream is the inlet stream, the reagents and the operating conditions are considered as the engineering specifications to the bioprocess step, e.g. flowrate, temperature and etc., then the feed stream is separated into two streams, namely product stream and waste stream. The waste stream includes the contaminant that is proposed to be separated by this bioprocess step. The product stream includes the target bioproduct and the contaminant that can not be removed by this step. These two streams are the outlet stream of the bioprocess step. Figure 3.1 gives the demonstration of input and output streams of a bioprocess step.

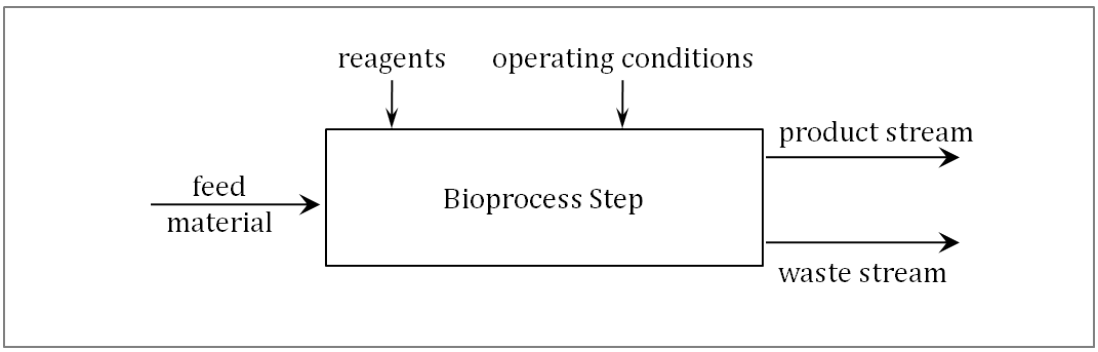


Figure 3.1: Illustration of bioprocess step input, step and output

The whole bioprocess sequence consists of a set of bioprocess steps. For the whole sequence, the processing material is the inlet stream of the first bioprocess step, then the product stream generated from this step would be the inlet stream for the next bioprocess step. This procedure is repeated until the product and waste streams are generated by the last bioprocess step. The contaminant included in the products steam after each bioprocess step would be reduced, and the pure target bioproduct is expected to be obtained from the product stream of the last bioprocess step. Each bioprocess step included in this bioprocess sequence has the specific operating conditions. Figure 3.2 gives a demonstration of input and output streams of three-step sequence.

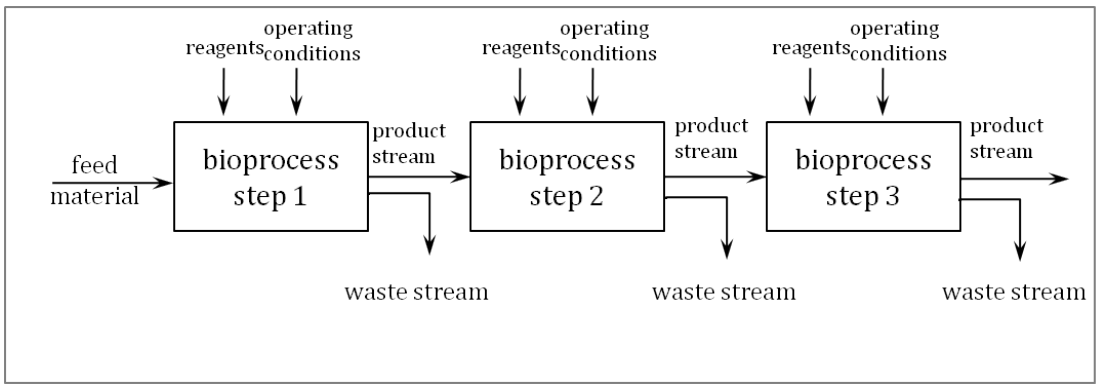


Figure 3.2: Illustration of bioprocess sequence input, step and output

Based on the Figure 3.1, the information included in the bioprocess step can be organized by three parts, i.e. *input*, *step* and *output*.

- *Input* is the information about the processing material, e.g. bioproduct properties,

contaminates properties.

- *Step* is the information about the unit operating conditions, e.g. equipment, flowrate.
- *Output* is the information about the performance of the unit, e.g. yield, purity.

According to the Figure 3.2, the three parts are also available to organize the information of the bioprocess sequence, where the input is the information about the processing material to each bioprocess step, the step is the operating conditions of bioprocess steps included in the sequence and the output is the information of the performance of each bioprocess step.

3.2.3 Representation of experimental data

The information included in each experimental data is proposed to be organized as the three parts. Each item of information about input, step or output is proposed to be given as a *specification* that consists of a specific parameter with an associated value. For instance, the input information 'bioproduct is IgG antibody' can be represented as a specification, 'product(IgG)', where the 'product' is the parameter and the 'IgG' is the associated value. Using parameter with value to represent the corresponding information is a common representation technique adopted by numerous computer-aided process design approaches (Stephanopoulos and Han, 1996; Stephane and Marc, 2008). The parameters used to represent the experimental data of different bioprocess steps are generally different. The Chapter 4 and 5 serve as examples to illustrate what specific parameters were used for representing the experimental data of centrifugation and chromatography.

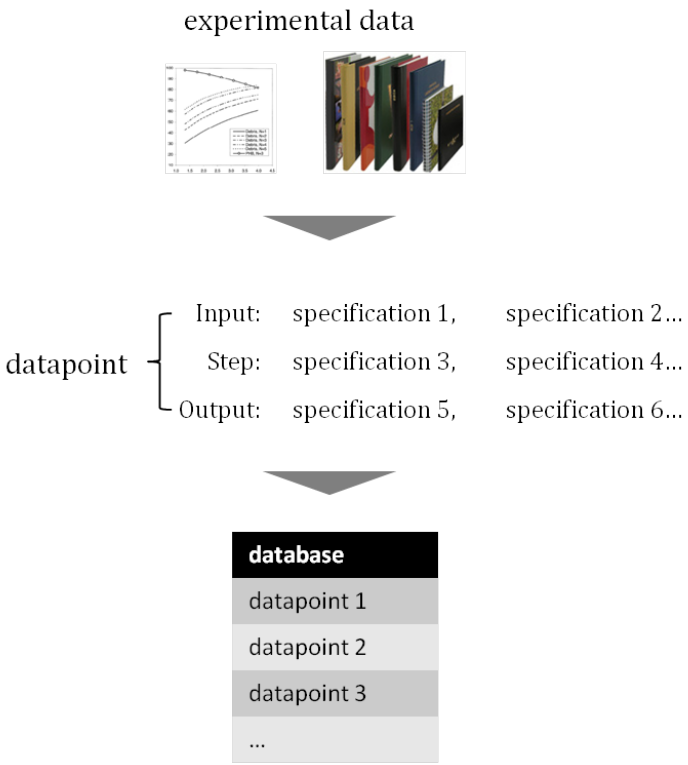


Figure 3.3: Illustration of experimental data representation

Figure 3.3 illustrates the procedure of experimental data representation. The experimental data in different formats or different carriers is represented as *datapoints*, these datapoints form a specific *database*. Each bioprocess step would have a database to store and manage its experimental data.

3.2.4 Capture of experimental data

The experimental data is usually kept in personal laboratory books. The initial effort was try to collect the experimental data from the previous laboratory books. However, the handwriting was difficult to recognize and the notes were not well organized, hence capturing experimental data from previous laboratory books was not considered.

The alternative is to capture the experimental data from the researchers who are working on bioprocessing development. It would be an efficient way since these researchers can provide accurate experimental data. In addition, the experimental data published on journal

papers which is clearly defined and explicitly described is also considered. Therefore, two types of experimental data, i.e. current researchers data and publications data, were considered as the raw data for this project. The workflow of capturing the experimental data is described in Figure 3.4.

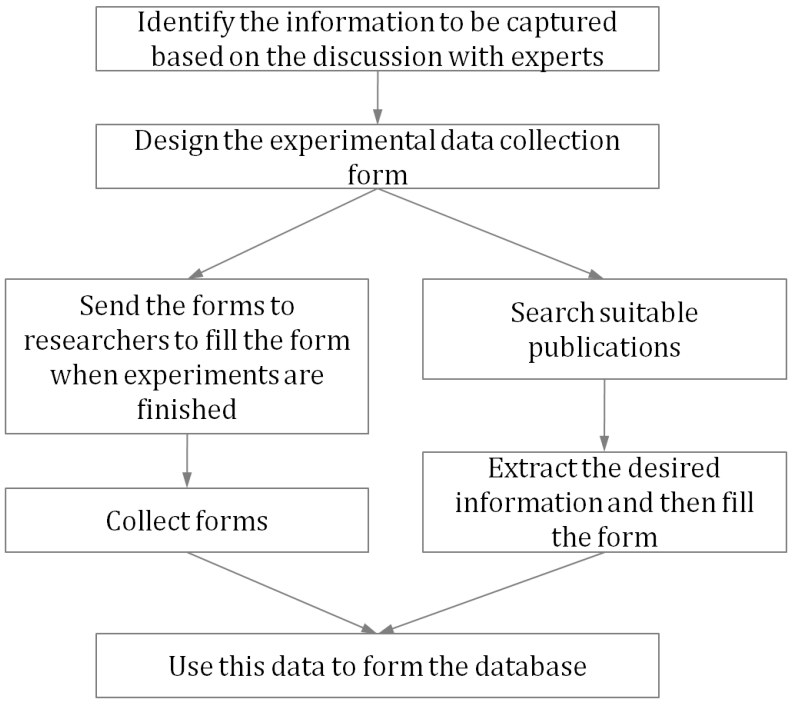


Figure 3.4: Flowchart of capturing experimental data from researchers and journal papers

In order to identify what information should be captured, discussions with the experts were made. For this project, the information captured for centrifugation and chromatography design was discussed with Dr. Andrew Tait, Dr. Jean Aucamp, Dr. Balasundaram Bangaru and Dr. Sunil Chhatre. This information will be explained in Chapter 4, 5 and 6.

In order to collect the raw data, the collection form developed in Excel (Microsoft, US) was used. Each column had a specific parameter, each row had an item of experimental data. The cell would be empty if the information was not available.

For current researchers' data, the forms were sent to the researchers via emails. With the forms, instructions were attached to explain the meaning of each parameter and the associ-

ated unit. For the publications' data, the information was extracted from text, diagram or table, and then used to fill the form. All of the forms were collected for database development.

3.3 ONTOLOGY AND REPRESENTATION

3.3.1 Introduction

Some information included in the experimental data is described by specific terms, e.g. the name of molecule, name of cell culture. Thus, querying these experimental data requires the included terms to be defined. These terms are usually related, if these specific relationships could be defined and used, it may contribute to the effective data searching, i.e. not only the data about the specific terms, but also the related terms data could be found. These data may help users to better access the feasible solutions to the bioprocess design problems. For these aims, the ontology is proposed to be introduced, which is used to define these terms and describe their relationships in order to assist the experimental data searching. In the following, section 3.3.2 introduces the basic definition of ontology; section 3.3.3 describes the general sorts of ontology; section 3.3.4 introduces two specific ontologies used for chemical process development which may help reader to understand the role of ontology in process development; 3.3.5 explains the methods of ontology development and 3.3.6 explains how to develop ontologies in BDKF approach.

3.3.2 Definition of ontology

An ontology is a set of formal, explicit expression of relationship and terminology (Ding and Foo, 2002b,a). 'Explicit' indicates that the types of terminology used and the constraints for using them are explicitly defined. 'Formal' refers to the fact that the terms and relationships have precise notation and meaning. Ontologies describe the semantics of data sources and make the contents explicit. They are used to unify information from database, data warehouses, knowledge bases and keep consistency through inferential reasoning systems.

The main concept in an ontology is *class*. The *class* is a specific term in the domain, and these classes are arranged hierarchically by the *relationships* between the classes. The *relationships* indicate the classes has the properties inherited by their subclasses. Each class can be instantiated with a specific instance that inherit the properties of parent classes and ancestor classes. In order to explain what ontology is and how they would contribute to information searching, a simple example of 'word' ontology is used.

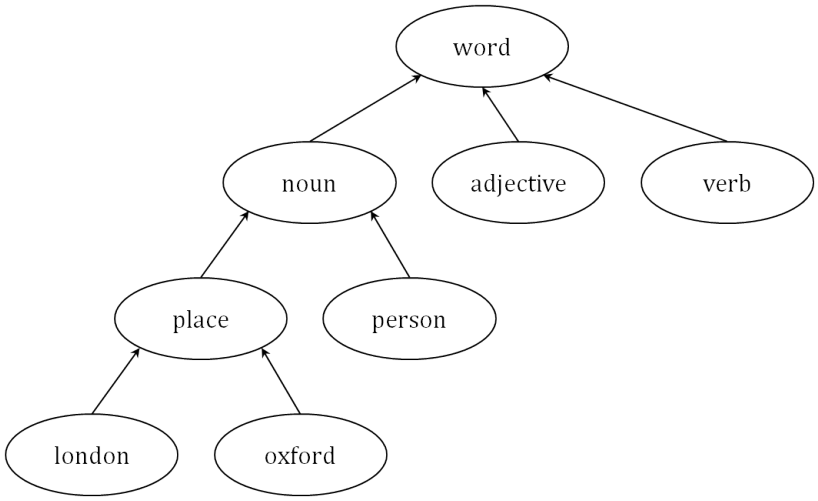


Figure 3.5: Ontology of 'word' domain

The Figure 3.5 gives the hierarchical structure about the 'word' ontologies. Each node in this hierarchy is a class, e.g. noun, verb. Each arrow indicates the relationship between the two classes. The class at the head of the arrow is the parent class (a general term of word domain), while the class at the tail of the arrow is the child class (a specific term of word domain). For instance, the class 'place' is the parent class of 'london' and 'oxford', as well as the child class of 'noun'. The arrow pointing to 'london' from 'place' indicates that the 'london' inherits an affiliation from 'place', i.e. london is a type of place. Similarly, the 'london' also inherits an affiliation from 'noun' which is parent class of 'place', i.e. london is a type of noun.

The 'word' ontology provides a vocabulary about the 'word' domain. It defines what terms could be searched, e.g. none, place, london, and also describes the specific relationships between these terms, e.g. london is a place, place is a none. This vocabulary would

help user to search target term within the 'word'. For instance, if the 'place' is concerned, the term 'place' is available for searching because it has been defined in 'word' ontologies, and the two terms, 'london' and 'oxford', are also suggested to be searched because 'london and oxford are two types of place'.

Before introducing the methods of ontology development, it is necessary to review the current ontology implementations which would help reader to understand why the ontology is a potential and useful technique for bioprocess design.

3.3.3 Types of ontology

The ontology can be divided into two types, namely upper ontology and domain ontology (Mizoguchi, 2003).

3.3.3.1 Upper ontology

The upper ontology (also known as top-level ontology or foundation ontology) describes the general terminologies that are the same across the all domains (Batres et al., 2007). The upper ontology is used to support broad semantic interoperability between a set of ontologies that are under this upper ontology, sometimes it is not a strictly ontology, and it is usually employed as a linguistic for learning domain ontology. Developed by collaborated groups, examples of prominent upper ontology are given in Table 3.1. Interested reader may find more information about the definitions of class as well as downloading and maintenance from the website.

Upper ontology performs as an agent to communicate with different domain ontologies to allow them to work together regarding the specific task.

3.3.3.2 Domain ontology

The domain ontology is the specific domain oriented ontology that characterizes the computational architecture of a knowledge-based system that perform a specific task in that domain, it also characterize the terminologies of the domain where the ontology is performed

Table 3.1: Definitions of upper ontology and website address

Ontology name	**Definition**	**Website**
Dublin Core	A set of terms are used to describe resources for purpose of discovery. It is employed to describe the full range of the web resources and the physical resources	www.dublincore.org
General formal ontology	It is used to integrate the processes and the objects that can be employed to develop custom, domain-specific ontology	www.ontomed.de
Cyc research ontology	It aims at assembling a comprehensive ontology and a knowledge base of the everyday common sense knowledge for AI application to perform human-like reasoning	www.opencyc.org
Suggested upper merged ontology (SUMO)	It is developed for multiple computer information processing system. The SUMO concerns on the general entities that do not belong to any specific domain so that would be a naturally encyclopedia. It is a type of IEEE standard ontology	www.ontologyportal.org
WordNet	It is a free semantic network based on psycholinguistic principles. It groups English words into a set of synsets with the general definitions. The semantic relationship are also recorded to organize the synsets	www.globalwordnet.org

(Kaiya and Saeki, 2006). The domain indicates the specific problem, e.g. monitoring, scheduling, design, and it is usually established by analyzing the structure of real task, e.g. the scheduling task can be described by scheduling recipient, scheduling resource, due date, constraints, goal, priority. Each character is described by specific terms, e.g. constraints can be delineated as strong constraint, constraint satisfaction and constraint predicates. Much ontology was developed by academic researcher and the prominent domain ontologies are given in Table 3.2. The associated publications gave details of the ontology techniques and the specific implementations.

Most domain ontologies are developed for the biology or chemistry disciplines that can contribute to manage or retrieve the complicated biology or chemistry terms efficiently, e.g. manage the heterogeneous biological data. They are not suitable for bioprocess design task because these ontologies were not developed for bioprocess domain. However, two of them, ontoCAPE and POPE, are developed for chemical engineering which may contribute to the development of bioprocess ontology.

3.3.4 Ontology of chemical engineering

In this section, two specific ontologies, ontoCAPE and POPE, are introduced to illustrate how to use ontology for chemical process design. These introductions would help reader to understand the development of bioprocess ontologies.

3.3.4.1 ontoCAPE

The ontoCAPE is a formal ontology for the Computer Aided Process Engineering (CAPE). This model has been developed for chemical process design in ISA88 (Morbach et al., 2007). In this model, the knowledge of chemical process engineering has been organized as four domains, namely *process function*, *process realization*, *process behavior*, and *process performance*.

- *Process function* indicates the desired behavior of a chemical process, which includes processing materials that describes chemicals, physical procedures.

Table 3.2: Definitions of domain ontologies and implementations

Ontology name	**Definition**	**Publication**
BioPAX	An ontology to represent biological pathways at the molecular and the cellular level and to exchange an interoperability of the cellular processes data	(Demir et al., 2010, 2012)
Business model ontology (BMO)	An e-business model ontology based on a review of the enterprise ontologies and the business model	(Zott et al., 2011)
Cell cycle ontology (CCO)	An ontology to provide biologists with a one-stop shop for the cell cycle knowledge	(Antezana et al., 2009)
Disease Ontology	An ontology to facilitate the mapping of diseases and the associated conditions to the specific medical codes	(Schriml et al., 2012).
Gene Ontology	An ontology of major bioinformatics to unify the representation of gene and the gene product across all species	(Chan et al., 2012; Huala, 2012)
Geopolitical Ontology	An ontology to describe, manage and exchange the data of geopolitical information, e.g. countries, territories, regions	(Caracciolo et al., 2012)
Plant Ontology	An ontology to describe the plant anatomy and the morphology, as well as stages of development for all plant species	(Lens et al., 2012)
NIFSTD Ontology	A set of ontologies developed for the neuroscience domain	(Imam et al., 2012)
Ontology for Biomedical Investigations	An open ontology for description for the biological and the clinical investigation, e.g. design, protocols, instrumentation, materials, data generated and analysis methods	(Torniai et al., 2011)
POPE	An ontology developed for the pharmaceutical process development	(Hailemariam and Venkatasubramanian, 2010a,b)
Protein Ontology	An ontology of definitions of the proteins and the relationships between them	(Natale et al., 2011)
Systems Biology Ontology	An ontology to define the terms used in the system biology, especially in the computational modeling	(Courtot et al., 2011)
ontoCAPE	An standard ontology developed for the Computer Aided Process Engineering	(Natarajan et al., 2012)

- *Process realization* represents the physical constitution of chemical process, it includes the equipment operations for this chemical process.
- *Process behavior* describes how the process work under a specific conditions, e.g. materials amount and etc.
- *Process performance* is illustrated which would be used for process performance evaluation and benchmarking.

Each domain is described by specific behavior or requirement. For instance, in the process function domain, three sub-domains are concerned, i.e. process, process control and controller. For the process ontology, each 'process' is a combination of several process steps. Each 'process step' is described by 'reaction' and 'unit operation'. The unit operation is distinguished by four types of classes, i.e. combination, separation, fragmentation and enthalpy change. The hierarchy of process ontology in ontoCAPE is given in Figure 3.6. This ontology covers the functional viewpoint of the chemical process, and this allows the process operation to be represented formally which may give a conceptual view on the desired processing, e.g. which unit should be selected to compose the sequence, what the requirements would be involved in this specific sequence.

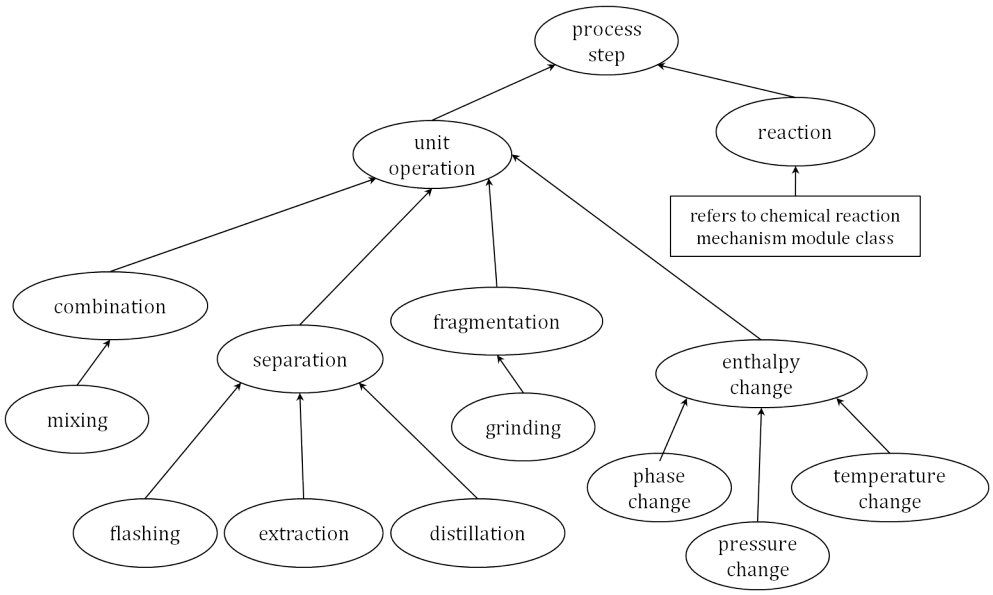


Figure 3.6: Process ontologies of ontoCAPE

The ontoCAPE serves chemical process development, e.g. supervision of the work of instruments and software in an extreme large scale off oil production plant (Natarajan et al., 2012), integration of heterogeneous chemical process information that is generated by various software for process control (Wiesner et al., 2011), representation of complex chemical reaction and multiple scale simulation for functional materials design (Lapkin et al., 2011).

3.3.4.2 POPE

The Purdue Ontology for Pharmaceutical Engineering (POPE) is an ontology to assist the decision making in pharmaceutical process development (Hailemariam and Venkatasubramanian, 2010a). Eight domains of ontology are involved in the POPE, namely *material*, *molecule structure*, *reaction*, *property*, *experiment*, *unit operation*, *equipment* and *value ontology*.

- *Material ontology* describes the materials as two parts, substance and phase system. Substance system indicates the natural composition of material while phase system defines aggregation state. For example, water, H_2O, has the substance of ionic species, H^+; and its phase system is liquid.

- *Molecule structure ontology* uses a set of fragments to represent the molecules used in the chemical reaction.

- *Reaction ontology* indicates the chemical reaction as the interaction between the functional groups and phase systems. Each reaction involved in the reaction ontology should have a physical context, restrictions and the required mathematical calculation.

- *Property ontology* employs the inter-property relations and solid material properties which are the extension of property ontology in ontoCAPE.

- *Experiment ontology* describes the experimentation as time (when), place (where) and identity of people who did this experiment (who).

- *Unit operation ontology* indicates the inlet and outlet stream involved in an operation of equipment and reaction. The streams are described by terminal port, a phase system(Material ontology) and a flowrate.

- *Equipment ontology* organizes the equipment terms referred by unit operation and experiment ontology, e.g. actuating, analytical, flow, processing and structure equipment.
- *Value ontology* is a set of quantitative specifications of material properties or environmental conditions. Five types value are involved in value ontology, i.e. single number with units, a specific range, list, table and picture.

The information of pharmaceutical engineering is represented by the specifications of the eight combinations. For example, the description of value and physical context organized in value ontology are generally used by other seven domains ontologies. POPE provides a general template for integration of data, knowledge and tool in order to benefit the information processing of pharmaceutical development.

POPE has been used in four types of applications, i.e. decision making of product formulation, unit operation model integration, reaction prediction and experiment analysis. Formulation is selecting a manufacturing route and a set of excipients for a drug production (Zhao et al., 2005). The unit operation model integration allows various knowledge of unit operation models to be integrated (Venkatasubramanian et al., 2006). The reaction prediction allows the information of molecule and reaction to be semantically searched and compared (Kayala et al., 2011). The experiment analysis makes the experiments comparable with respect to procedure, equipment settings and the data quality (Chepelev and Dumontier, 2011).

3.3.4.3 Summary

The ontoCAPE and POPE are developed for chemical engineering, hence they are not suitable for bioprocess design tasks due to the fundamental differences between the chemical engineering and the biochemical engineering. For example, the chromatography used in bioprocess is not defined in ontoCAPE or POPE.

These two types of ontology implementations suggest that the ontology would be an open technique that is capable of managing various terminological information. It would be

an useful tool to harness the different biological terms referred by experimental data.

In order to facilitate the bioprocess design by harnessing bioprocess data and knowledge, the appropriate ontologies are needed to be developed. For this, the methods of ontology development are explained in next section.

3.3.5 Methodology of ontology development

Ontology building is an evolutionary process that requires multiple skills, and it is an art rather than technology. METHONTOLOGY, On-To-Knowledge and DILIGENT were most referred development methodologies (Surez-Figueroa et al., 2011). Recently, a new approach called NeOn methodology also has been proposed (Surez-Figueroa, 2010). The information about these four development methodologies is given in the following.

3.3.5.1 METHONTOLOGY

Based on IEEE standards for Developing Software Life Cycle Processes, 1074-1995 (Fernndez-Lpez et al., 1997), the METHONTOLOGY includes: (1) the identification of ontology development, i.e. what tasks should be done for ontology building; (2) life cycle of development activities and techniques of ontology management and supportive activities. Furthermore, it also illustrates the activities to be done for ontology reuse and reengineering process. One of recent implementation is development of herbal medicine ontology (Mustaffa et al., 2012).

3.3.5.2 On-To-Knowledge methodology

The On-To-Knowledge methodology takes into account how these ontologies can be used in knowledge management applications (Sure and Studer, 2002). This methodology consists of five major steps (with 13 sub-steps), namely feasibility study, kickoff, refinement, evaluation and application. The implementation of this methodology for ontology development can be found elsewhere (Palma et al., 2011).

3.3.5.3 DILIGENT methodology

This method focuses on collaborative and distributed ontology development (Casanovas et al., 2007). It includes five aspects, i.e. building, local adaptation, analysis, revision and local update. Practical implementation and evaluation of this methodology is given by Pinto et al. (Pinto et al., 2009).

3.3.5.4 NeOn methodology

NeOn is scenario based methodology that supports the reuse of ontology, and collaborative activities of ontology development and evaluation. It includes nine scenarios of ontology development, identification of processes and activities for developing ontology, general guidelines for ontology reuse and re-engineering. Further details and practical implementations can be found elsewhere (Adamou et al., 2012).

3.3.6 Ontology development in BDKF approach

The common principles of ontology development methodologies are summarized as the four steps, i.e. *determine the domain and scope of the ontology*, *enumerate important terms in the ontology*, *define the classes and class hierarchy*, and *create instance*. These four steps have been validated by research (Venkatasubramanian et al., 2006), and they serve development of ontologies used by BDKF approach.

- *Determine the domain and scope of the ontology*: The domain would help to define the terms and relationships that would be modeled by ontology. It should have been established so that the specific terms can be formally captured and represented.
- *Enumerate important terms in the ontology*: Decisions of which specific classes constitute a particular ontology should be considered in the second step. Usually the classes and concepts in the specific domain ontologies are identified through discussions with domain experts, and the internal relationships could be obtained from the textual material, such as relevant journal papers.

- *Define the classes and class hierarchy*: The class attributes define the properties of concepts in that class. The properties includes two types, objective type and data type (Venkatasubramanian et al., 2006). The data type attribute is primitive value, e.g. integer, float and etc. The object type attribute value is an instance of other classes, e.g. a specific relationship. For the BDKF approach, the ontology is proposed to organize the terms referred by the bioprocess data, thus the objective attribute would be used to describe the properties of concepts in that class, i.e. the relationship between the parent class (a general term) and the child class (a specific term). For demonstration, the classes with the hierarchy are usually expressed as a hierarchical tree that is common form adapted by various ontologies.

- *Create instances*: The actual data associated with a class would be demonstrated by instance of that class, and it illustrates how the terms involved in this domain are organized and constrained, e.g. in the 'word' ontologies, 'london is a type of place'.

The demonstration will be presented in Chapter 3 where the specific terminologies are employed to illustrate how each of four-step is applied.

3.4 KNOWLEDGE AND REPRESENTATION

3.4.1 Introduction

Generally speaking, the knowledge is a type of fact acquired from study or experience which help people to solve or further understand the problem. For bioprocess design, the knowledge of various disciplines are required, e.g. the biological, chemical or medical knowledge. This knowledge often imposes constraints on bioprocess design. For instance, thermostability must be taken in to account when choosing operating temperature for enzyme production, biochemistry knowledge is required to understand the properties of the bioproduct, mechanical engineering knowledge is required to understand the interaction between the bioproduct and the physical forces generated by the specific equipment. In BDKF approach, the knowledge is proposed to be used to assist the experimental data utilization,

e.g. doing calculation required by data analysis, interpreting the experimental data as useful information in order to help users to make decision.

3.4.2 Definitions of knowledge in BDKF approach

For knowledge management, in BDKF approach, two categories are used to tag all of the feasible knowledge. i.e. *theoretical knowledge* and *empirical knowledge*.

- *Theoretical knowledge* is the formal definitions about the bioprocess, e.g. fundamental equations, equipment background information.
- *Empirical knowledge* is the knowledge established from empirical studies, e.g. scale up/down principles.

The theoretical knowledge is the general knowledge introduced in the bioprocess text books, and it would provide the basic assistance to the experimental data utilization, e.g. standardizing different unit setting, managing the background information of equipment. The empirical knowledge represents the newly understanding about the bioprocess which is acquired from the experimental studies or journal papers. This knowledge works under the specific context, and it is proposed to be used as the explicit rules for data utilization. In BDKF approach, the empirical knowledge is the scale down approaches developed for each bioprocess step. The scale down approaches aim at using small scale experiments to predict the performance of the large scale operation, thus they could be used as the rules to harness the experimental data generated from different experimental scales. For each bioprocess step, the scale down approaches are different (Titchener-Hooker et al., 2008), so the empirical knowledge is specific to its bioprocess step.

3.4.3 Definitions of knowledge representation

The theoretical and empirical knowledge is usually in various formats. Representing this knowledge is the precondition to knowledge utilization. The knowledge representation generally includes two elements, *knowledge entity* and *knowledge formalization* (Cadoli et al., 2011; Segev, 2011).

- *Knowledge entity* refers the content that the knowledge describes, and the content is used to solve the specific problem.
- *Knowledge formalization* is a interpretation approach that allows the knowledge entity to be represented by variables and the specific relationships.

The knowledge representation performs as a surrogate that allows the external information to be understood by the BDKF system in order to reason with the experimental data. Currently, there is not an 'optimized and standard' knowledge representation, and it is still under discussing (Rassinoux, 2012; Chua et al., 2012). Therefore, identifying the optimized knowledge representation will not be concerned in this thesis. Different knowledge expression should have different representation. In order to introduce the knowledge representation, the fundamental equation, the background information and the specific scale down approaches are considered as demonstrations.

3.4.3.1 Representation of fundamental equation

For the fundamental equations, they can be represented as the symbols with the specific mathematical relationship. For example, the density of a material is defined as its mass per unit volume, and this definition is the knowledge entity which can be formalized as the symbols, density 'ρ' (output), mass 'm' (input) and volume 'v' (input) and the fact of the mathematical relationship '$\rho=\frac{m}{v}$' (function). This knowledge is proposed to calculate the value of 'ρ' for given values of 'm' and 'v'.

3.4.3.2 Entity relationship model

The background information of equipment is generally provided by the manufacturer. This information is usually in different formats, e.g. diagram, table, and it may tell users the physical properties of equipment that would be used for equipment selection. To represent this information, the Entity-relationship Model (ERM) that can explicitly describe the background information and the equipment is used.

ERM is an abstract and conceptual representation (Chen, 1980). It is widely used to produce conceptual schema about data and knowledge (Samba, 2012; Bollati et al., 2012), e.g. describing the relationship or linkage between different database. The *entity*, *attribute* and *relationship* are the primary concepts about the ERM.

- *Entity* is a particular thing that is capable of an independent existence that can be uniquely identified.
- *Attribute* is an abstract of a domain for the entity, e.g. physical object or a specific concept.
- *Relationship* is a specific connection between the entity and an attribute

Figure 3.7 gives a general form of ERM. If the entity (rectangle) is a specific term of equipment, the attribute1, attribute2 and attribute3 (ellipse) can represent three items of background information about the equipment. The number of attributes with respect to the entity depends on the number of items of background information. The relationship (diamond) indicates the specific 'relation' between the attribute and entity.

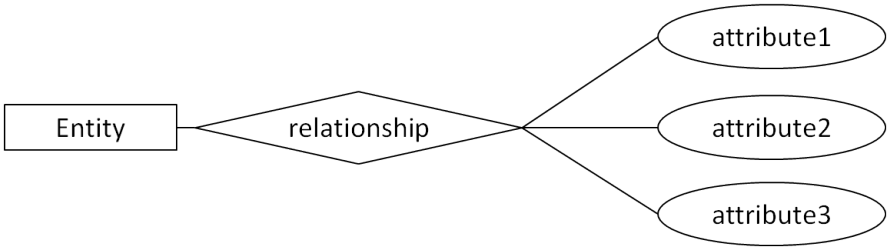


Figure 3.7: Illustration of ERM: entity, relationship and attribute

3.4.3.3 Representation of scale down approach

Different bioprocess steps have the different scale down approaches. Executing the scale down approaches requires to satisfy their specific context. In order to describe the context and the conclusions involved in the scale down approaches, the formation 'If...Then..' is proposed to be used. It is a common formation of rules that has been adapted by various expert systems (Yazdani, 2012).

The 'if' content is the condition, and the 'then' content is the consequent. Only the condition is satisfied, the consequent is available for use. The contents of condition and consequent depend on the knowledge entity, e.g. equations, symbols. These will be explained by practical scale down approaches in Chapter 4.

3.4.4 Summary

The reason why these knowledge representation techniques are considered is that they can reflect the knowledge entity in a nature way. It is difficult to say that only the three types of knowledge representation can be used by BDKF approach, but they are enough to demonstrate how BDKF approach works on the bioprocess design tasks. Different bioprocess steps have different knowledge entities, thus the three types of representation and their roles in bioprocess design task will be explained in the case studies.

For any specific BDKF system, the ontology, theoretical and empirical knowledge comprise the knowledge base which aims at organizing and managing the captured knowledge. With the represented experimental data and knowledge, the next step is to develop the reasoning functionalities to use them to solve design problems.

3.5 REASONING FUNCTIONALITIES

Given a new bioprocess design problem, people usually search related experimental data to understand the design problem, then do data analysis to assess the feasible solutions, and finally people would be interested in studying the detailed information in order to achieve the design target. Based on this logic, harnessing experimental data and knowledge for the bioprocess design problem is proposed to be achieved by the following steps.

Step 1: identify and represent the bioprocess design problem.

Step 2: search the experimental data relevant to the design problem.

Step 3: evaluate the possible performance based on the relevant experimental data.

Step 4: generate suggestions for further experimentation.

In BDKF approach, the four steps are realized by design query formalization, and the three reasoning functionalities, i.e. search, prediction and suggestion. The definitions about the four steps, as well as how they work are introduced in the followings.

3.5.1 Design query

3.5.1.1 Definition of design query

The design query is the bioprocess design problem to be solved. Each bioprocess design can be treated as an experiment that has not been done yet, it may include the information about the processing material, operating conditions and desired performance to be achieved. Same as the experimental data representation, this information can be grouped into three parts, input, step and output.

3.5.1.2 Representation of design query

For each part, each item of information is represented as a parameter associated with specific value, in order to differentiate the specification referred in experimental data representation, it is called a *feature*. For a specific BDKF system, the representation of design query and the experimental data use the same parameter setting. For any requested information to be solved, the value of the feature is indicated by 'X', and this feature is called queried feature, e.g. 'temperature (X)' is a queried feature indicates the question 'what the temperature should be used for this design'. Each design query consists of a set of features about the input, step and output. These features indicate all of the information referred by the design problem. Table 3.3 gives the summary about the representation of design query. Examples will be given in the case studies in the following chapters.

Table 3.3: Structure of design query representation

Input	The features about processing material properties, e.g. cell line.
Step	The features about operating conditions, e.g. scale, equipment.
Output	The features about performance to be achieved, e.g. yield, purity.

3.5.2 Search functionality

3.5.2.1 Definition of search functionality

The search functionality is finding the experimental data which are relevant to the design problem. In the BDKF system, the search functionality aims at returning a set of datapoints from the database which are relevant to the design query. A design query related datapoint is that the specifications of the datapoint satisfy the whole features involved in the design query. There are two types of criteria have been created for relevance judgment, namely numerical criterion and terminological criterion. The numerical criterion is used by a specification or feature whose value is a number, while the terminological criterion is used by a specification or feature whose value is a term.

3.5.2.2 Definition of numerical criterion

For the features of input or step involved in the design query, the numerical criterion can be expressed as follows:$[m \times (1 - \mu), m \times (1 + \mu)]$, where μ is a criterion specified by user, m is the value of a feature. Such criterion can take the experimental error existing in the experimental data into consideration. For instance, if the true pH value of a particular solution is 5, and user defines μ is 10%, the criterion for this pH is $[5 \times (1 - 10\%), 5 \times (1 + 10\%)]$. In addition, this criterion also allows user to study a range of conditions about a specific parameter by giving a big μ. For instance, for the pH 5, if user defines μ equals to 50%, the criterion $[5 \times (1 - 50\%), 5 \times (1 + 50\%)]$ allows the pH conditions ranging from 2.5 to 7.5 to be studied.

How to define the value of μ relies on users' knowledge and experience, or their specific design purpose. The value may impact the results of search functionality, e.g. if μ is small, the numerical range is narrow and the datapoints found by search functionality would be similar to the design query.

The numerical feature of the output indicates the minimal performance requirement. Usually, engineer is interested in the solutions that can achieve better performance, because

the performance of a bioprocess sequence can be significantly changed if the performance of individual bioprocess step is improved. For example, if a bioprocess sequence consists of two bioprocess steps and both of their yield are 90%, then the yield of bioprocess sequence is $90\% \times 90\% = 81.0\%$. If the yield of the two bioprocess steps increase to 93% and 94% respectively, then the whole sequence yield would be $93\% \times 94\% = 87.4\%$, which is better than the initial yield by 6.4%. Therefore, the datapoint which has better performance is considered to be relevant to the design query.

For a bioprocess step, two types of performance would be considered. The first type is production measurement, e.g. separation performance, yield. For this type of performance, given the specific feature n, its numerical criterion is defined as $[n, +\infty)$ which indicate any value not less than n is satisfactory. The second type is product loss measurement. For this type of performance, given the specific value n, its numerical criterion is $(0, n]$ that indicates any value not bigger than n is required.

3.5.2.3 Definition of terminological criterion

For terminological feature, the relationship defined by ontologies is used as the terminological criterion. All of relationships described in ontology are expressed as a list of facts. For example, 13 items of relationship facts are involved in the word ontology (see section 3.3.2), such as noun is a type of word, place is a type of noun. These facts serve as the terminological criteria used by search functionality. For a specific term, any term defined as the child class of this term is satisfied, i.e. their relationship fact appears in the list of relationship facts. For example, in the 'word' ontology (Figure 3.5), if 'place' was used in feature, the 'London' specification satisfies this 'place' feature, because 'London' is a type of 'place'.

3.5.2.4 Pseudo code of search functionality

In order to demonstrate the programming logic of the search functionality, the pseudo code is given in Figure 3.8 which illustrates how the design query constrain the search and how the relevant datapoints are determined.

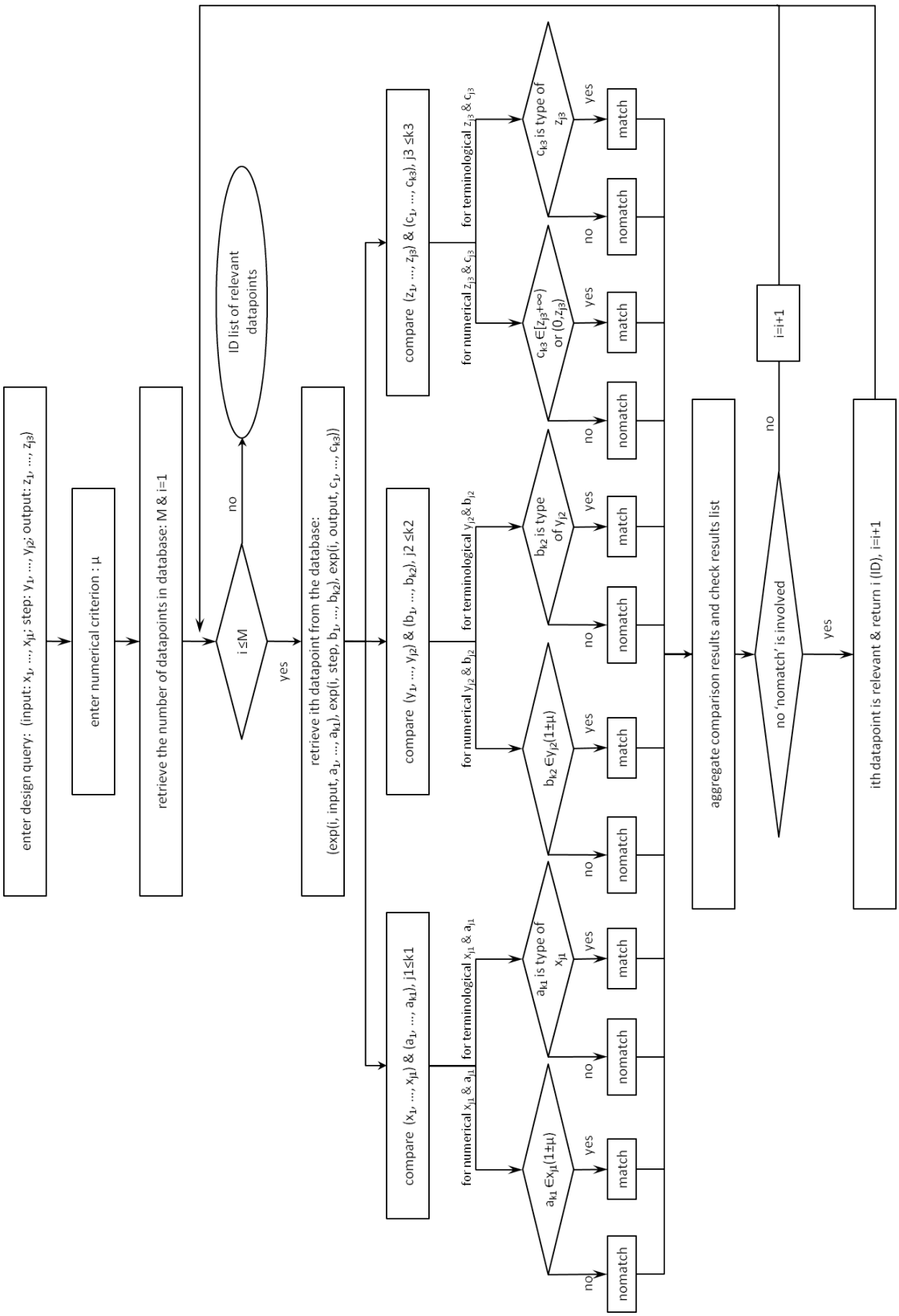


Figure 3.8: Pseudo code of search functionality

This pseudo code is interpreted as the following steps:

1. The design query includes a set of features, x_{j1} or y_{j2} or z_{j3} indicates one feature of input or step or output.

2. The numerical criterion μ given by user is used to constrain the search.

3. The amount of datapoints (M) and the ID (i) are used to constrain the loop of comparison.

4. i$\leqslant$M (start a comparison loop).

   (a) If yes, the search functionality retrieves the specifications of ith datapoint, e.g. a_{k1} or b_{k2} or c_{k3} represents one specification of input or step or output. The comparison is carried out between each of the corresponding feature and specification based on the numerical or the terminological criteria, e.g. the input feature, x_{j1}, is compared with the input specification, a_{k1}.

      i. If the criterion is satisfied, then a 'match' is returned;

      ii. If the criterion is not satisfied, then a 'nomatch' is returned.

      iii. All of the comparison results are aggregated.

         A. If no 'nomatch' presents in the list of comparison results, then this datapoint is relevant to the design query and the ID (i) is recorded.

         B. If 'nomatch' presents in the list of comparison results, then this datapoint is not relevant.

      iv. start a new comparison loop (i=i+1 and go back to step 4).

   (b) If no, the comparison loop is completed, all of the IDs of relevant datapoints are returned as the result.

By examining these relevant datapoints from the database, the user would know what have been done before and how those experiments turned out. Such information may give user a general picture about the design query. These datapoints can also be used by the prediction functionality to evaluate the likely performance that may be achieved for the design query.

3.5.3 Prediction functionality

3.5.3.1 Definition of prediction functionality

The prediction functionality aims at assessing the possible performance for the design problem based on the relevant experimental data. The prediction functionality generates a predicted result to the output feature about the design query. The predicted result is a type of statistical result generated by the relevant datapoints that are returned by the search functionality. In general, the prediction functionality for aggregating the datapoints to generate the predicted result would not be strictly defined. However, simple option could be employed for this purpose, such as arithmetic mean, or weighted mean, or taking a range from those data points that are in some defined way nearest to the features specified in the design query. In the following, the arithmetic mean and weighted mean are used to explain how to generate the predicted results based on the relevant datapoints.

3.5.3.2 Use of arithmetic mean for prediction

For using the arithmetic mean to generate the predicted result, the prediction functionality aggregates the value corresponds to the output feature of the design query from the relevant datapoints, the equation (3.1) is used to generate the predicted result.

$$\bar{X} = \frac{1}{N}(\sum_{i=1}^{N} x_i), \tag{3.1}$$

where $\bar{X}$ is predicted result, x_i indicates the value of the specification included in the ith relevant datapoint, and N represents the number of relevant datapoints.

The arithmetic average would give an approximation on the performance that may be achieved.

3.5.3.3 Weighted arithmetic mean for prediction

The weighted mean is a general type of descriptive statistics to indicates how relevant datapoints can contribute to the final predicted result, and it is calculated by equation (3.2).

$$\begin{aligned} x_i &\in (x_1, x_2, ..., x_n), \\ \omega_i &\in (\omega_1, \omega_2, ..., \omega_n), \\ \bar{X} &= \frac{1}{N}(\sum_{i=1}^{N} \omega_i x_i), \end{aligned} \tag{3.2}$$

where x_i represents the value of the specification included in the ith relevant datapoint, ω_i describes the associate weight to x_i, $\bar{X}$ is the value of weighted mean and N is the number of datapoints.

The associated weight indicates the reliability of the data. For instance, if the data is from the experiments done by post doctoral researchers, then the associated weight would be high; if it is from the graduate students, then the associated weight would be low. Because the post doctoral researchers would have more knowledge and experimental experience than the graduate students, thus the experimental data produced by the post doctoral researchers should be more reliable.

Other statistic algorithms also could be used for prediction. For instance, the algorithm of finding max/min results may be helpful to indicate the possible boundaries of the performance to be achieved, and the boundary also could be used to assess the design solutions. In addition, the Bayesian algorithm may also be applied here in order to assess the probability of the prediction results.

3.5.3.4 Pseudo code of prediction functionality

The pseudo code of prediction functionality is given in Figure 3.9 which would help to understand the logic of prediction functionality.

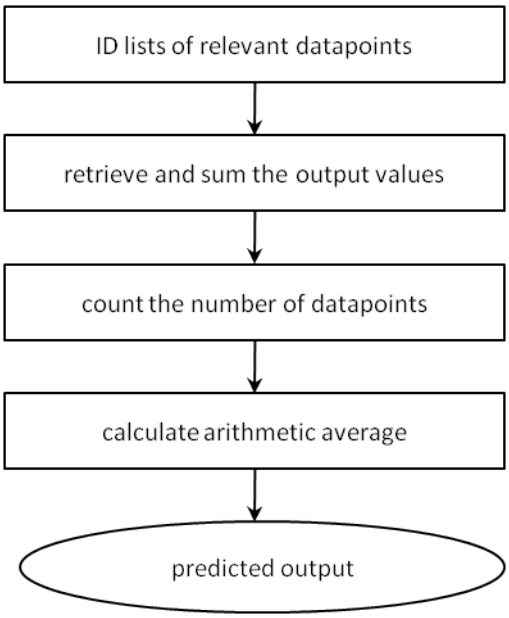


Figure 3.9: Pseudo code of prediction functionality

The search functionality returns the ID list of relevant datapoints, then their specific output are aggregated for predicted performance by arithmetic average. The prediction functionality concerns on analyzing the relevant datapoints , any algorithm that would realize this goal could be used. In this thesis, proving the BDKF approach concepts is primarily concerned, hence the arithmetic mean algorithm is employed due to the simplicity. Although the weighted mean would be also available for demonstration, the 344 experimental data were all provided by postdoctoral researchers which would make the associated weight equal so that the weighted mean would be equivalent to the arithmetic mean.

3.5.4 Suggestion functionality

3.5.4.1 Definition of suggestion functionality

Suggestion functionality is to provide solutions for further experiments. These solutions provided by suggestion functionality are the answers to the questions included in the design query. These questions would be about the input, e.g. the feed stream properties, or the step, e.g. the operation conditions. The techniques for answering these questions would not be strictly defined, In this thesis, retrieving the requested information is considered, e.g. retrieving the temperature information to the queried feature ‘temperature(X)’. Retrieving has

been applied by CBR to solve the new problem by reusing previous similar cases (Aamodt and Plaza, 1994), and its chemical implementation proved that the retrieved information was very useful for the new design task (Stephane and Marc, 2008).

Before introducing the retrieving technique, it is necessary to explain the reason why the retrieved information would useful for the design task. For any bioprocess experiment, the input and step determine the output performance. If the input and step of two experiments are similar, then their performance should be similar. Following this logic, if two experiments' performance are similar, one possible reason is that their input and step should be similar. For example, if the filtration design problem is to identify the flowrate to realize the yield 98%, then the flowrate information from the datapoint whose yield is close to 98% would be a feasible solution.

Given the specific design query, each datapoint returned by search functionality represents a relevant experiment fact, and the predicted result indicates the likely performance to be achieved. The input and step information from a datapoint whose performance is most similar to the predicted result may make the design realize the predicted performance. Therefore, the retrieved information is the reasonable answer to the queried feature specified in the design query.

3.5.4.2 Definition of Euclidian distance

For the retrieving, the 'similar' is determined by the similarity distance, and 'most similar' means the smallest similarity distance. In this thesis, the Euclidian distance is used due to its simplicity. It is a type of general distance measurement. For the L-dimension space, X and Y, the Euclidian distance is given by equation (3.3).

$$d(X, Y) = (\sum_{i=1}^{L} |x_i - y_i|^2)^{1/2}, \tag{3.3}$$

where x_i and y_i represent a ith element of the X and Y that are the L-dimension space.

Considering the X is the a set of output features of a design query, Y is a set of output specifications of a datapoint, then equation 3.3 can be used to calculate the distance between the design query and datapoint.

Considering the simplest situation, i.e. one output feature is concerned in the design query, then the distance calculation can be shown by equation (3.4).

$$d(X, Y) = |x - y|, \tag{3.4}$$

Here is a simple example, X is the yield feature, Y is the yield specification, x is the predicted yield value, e.g. 24 mg, y is the yield value recorded in a datapoint, e.g. 28 mg, the similarity distance, *d(X,Y)*, can be calculated as 4.

3.5.4.3 Pseudo code of suggestion functionality

The pseudo code of suggestion functionality is given in Figure 3.10 which illustrates how to retrieve the requested information.

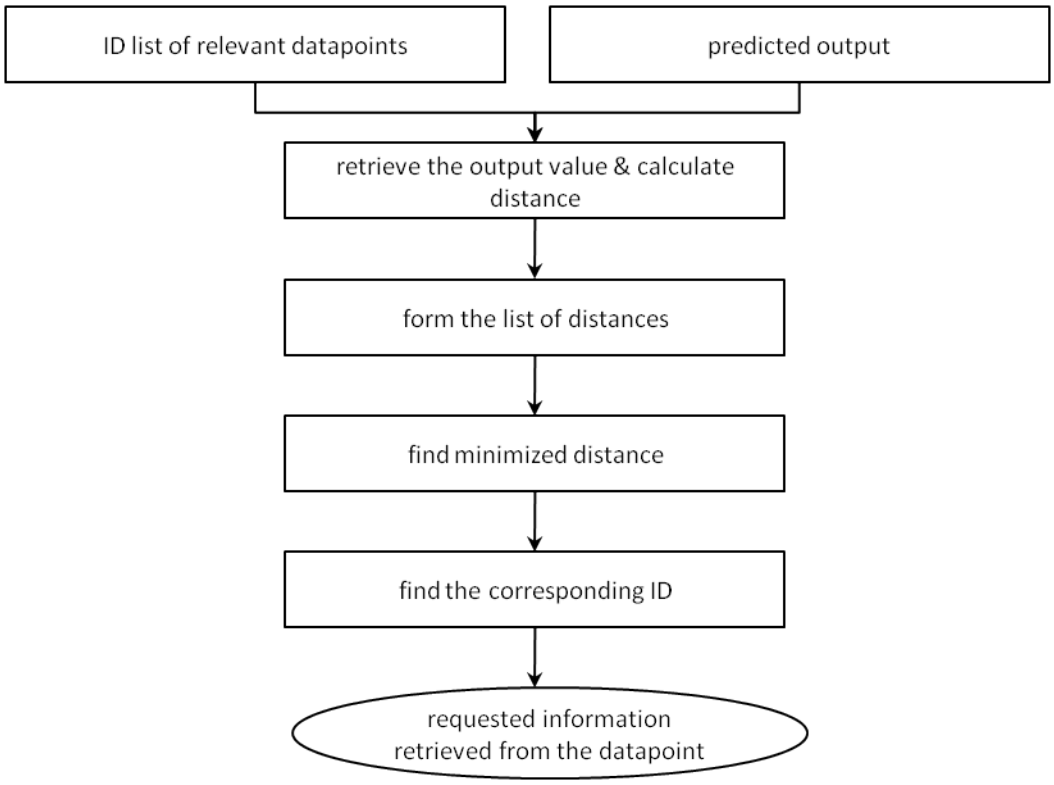


Figure 3.10: Pseudo code of suggestion functionality

For any design query, the similarity distance is calculated for each relevant datapoint returned by search functionality, then the search functionality retrieves the requested information from the datapoint which has the smallest similarity distance. The retrieved information would be used for further experiments for validation.

3.6 FLOWCHART OF BDKF APPROACH FOR BIOPROCESS DESIGN

3.6.1 Flowchart of BDKF approach

In BDKF approach, the datapoints represented by experimental data forms the database, the represented ontologies, theoretical and empirical knowledge constitute the knowledge base. Given the design query that represents the specific bioprocess design problem, the three functionalities harness the data and knowledge to generate the solutions to the design

query. This procedure is shown in Figure 3.11.

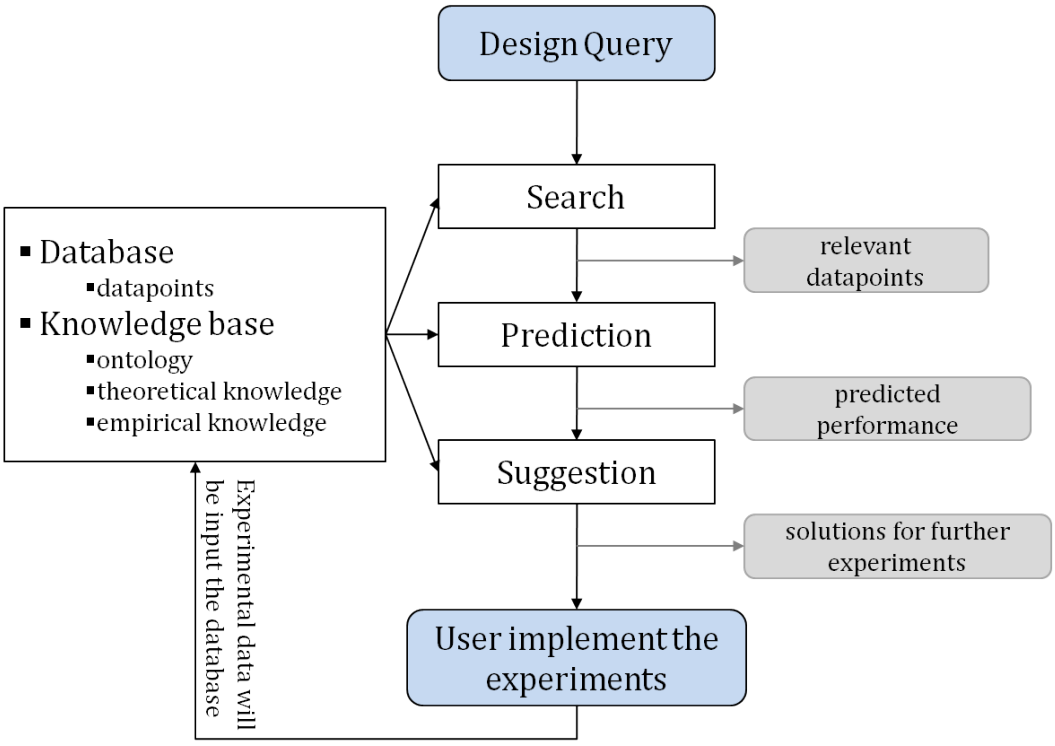


Figure 3.11: Working flowchart of BKDF approach for bioprocess design

The flowchart can be explained in the following steps.

- The design query that represents a specific bioprocess design problem should be given by user, and the design query should consist of one feature of input or step or output at least.
- Give the design query, the search functionality accesses the database to return design query related datapoints, while the necessary knowledge included in the knowledge base may also be utilized, e.g. ontology.
- Based on the relevant datapoints, the prediction functionality aggregates the value about the concerned performance, and then generates the likely performance to the design query.
- With this predicted performance, the suggestion functionality retrieves the requested information from the datapoint whose performance is most close to the predicted one.

- The retrieved information would be used for the new experiments for validation about the predicted performance or the start for further exploration for the design problem. The data generated by the new experiments would be stored in the database for the new bioprocess design problem to use.

3.6.2 Discussion

The three reasoning functionalities would also work independently. For instance, if users want to look at what relevant experiments have been done in the past, then the search functionality can realize this request. If users would like to evaluating the performance of a bioprocess design, then the search and prediction functionality can achieve this task. If users want to retrieve the specific requested information, then the search and suggestion functionality can do this job.

If the implemented experiments show the predicted performance can not be achieved, then user could use more features or narrower criteria in the design query in order to find more relevant datapoints. If there is no datapoint is returned, it indicates no relevant experiment has been done for the design problem, and user may use the less features in the design query to try to re-find relevant datapoints. For this situation, the BDKF approach is proposed to allow the user to screen the feasible solutions about the design problem or the potential design space to be explored.

The Figure 3.11 illustrates one implementation of BDKF approach for the bioprocess design, but other implementations can be developed, e.g. mining the information from the bioprocess experimental data for bioprocess design. For this, the algorithms employed by prediction and suggestion functionality would be changed based on the data mining approach. This would be an interesting research topic about the BDKF approach in the future and it will be introduced in the future work.

3.7 IMPLEMENTATION PLATFORM OF BDKF APPROACH

In order to implement the BDKF approach, two general types of programming languages were considered, i.e. *object-oriented* and *logic*.

The *object-oriented* programming language employs the objects that consists of data filed and methods and the interactions between the objects to realize the design applications. The typical programming languages include Java which is developed by Oracle, and Python which is an open-source language.

The *logic* programming language uses the first order logic to represent the data and methods as facts and rules for the design applications. The prominent programming language is Prolog that is an open-source language.

It is difficult to say which tool would be the best to perform the BDKF approach, since they are all basic programming language to transfer the users' idea as the specific codes. Since the BDKF approach aims at dealing with the inconsistent and incomplete data and knowledge, the syntax is used to determine the platform.

The object-oriented programming languages are good at processing the interactions between the various mathematical models. For example, each specific model can be stored as an object that consists of the variables and the specific mathematical equations. This is the reason why they are adopted to develop the simulation tools which formulate the design task as a set of mathematical models. The logic programming languages represent the information as the declarative sentences consisting of facts and rules (first order logic based), it is suitable to represent and process the terminological information. This is the reason why the Prolog is the first choice for expert system development. Since the data and knowledge considered in BDKF approach are the specific terms and the relations, the Prolog is better for information representation. In addition, the Prolog performs more effectively on reasoning

with terms and relations. For example, comparing the two specific terms, the Prolog can return 'true' directly if they are identical, but the Java or the Python has to transfer the terms as the ASCII(American Standard Code for Information Interchange), and then the codes are compared for answer.

Considering the information to be harnessed, the Prolog is selected to develop the BDKF approach. Although the Prolog is not good at representing the complicated mathematical models, e.g. calculus, it is competent for the demonstration in this thesis because few complex mathematical equations would be used. A specific commercial package, namely WinProlog (LPA, UK), is employed due to its good email support and well-written manual.

Chapter 4

DEVELOPMENT OF BDKF APPROACH ON A REAL BIOPROCESS STEP: CENTRIFUGATION CASE STUDY

4.1 INTRODUCTION

In order to investigate and evaluate the BDKF approach, a centrifugation case study has been established. The centrifugation is a type of solid-liquid separation, and it is a simple bioprocess step comparing with other bioprocess steps used in the downstream processing. Therefore, it would be a proper case to illustrate how to apply the BDKF approach on a real bioprocess step. For this, a BDKF system about centrifugation has been developed in Win-Prolog (LPA, UK), and it is called centrifugation system.

The sections in this Chapter are organized as follows, the theories of centrifugation are presented in section 4.2; the representation of centrifugation experimental data is discussed in section 4.3; the ontology representation is described in section 4.4; the knowledge repre-

sentation is introduced in section 4.5; in section 4.6, the centrifugation design query formalization and reasoning functionalities are explained; the evaluations that were implemented for centrifugation prototype are discussed in section 4.7; finally the discussion and conclusions about the centrifugation case study is presented in section 4.8.

4.2 CENTRIFUGATION INTRODUCTION

Centrifugation is used to separate processing materials with different densities when the force for separation is greater than gravity, and it separates the processing material into two phases, supernate and sediment. Centrifugation is one of the primary downstream steps that removes cells from the fermentation broth, or eliminates cell debris, or collect precipitates.

Generally, the centrifugation design would start at a small scale, e.g. bench top centrifuge. These experimental results would be used to identify a narrowed design space. Then, engineers will use the pilot scale centrifuges to examine this design space. This procedure has accumulated lost of centrifugation experimental data and knowledge, and the centrifugation system is proposed to use these data and knowledge for the new design task. Before discussing how to realize this aim, it is necessary to give a brief introduction about centrifuge equipment, characteristics of centrifugation and ultra scale down approach, because these information would be used for centrifugation experimental data utilization.

4.2.1 Centrifuge equipments

Generally speaking, a centrifuge is classified according to internal structure. The tubular bowl centrifuge and disk stack centrifuge are commonly used for bioprocessing at pilot and manufacture scale (Saite et al., 2006) while the bench top centrifuges are used for the laboratory work.

The tubular bowl centrifuge has the simplest configuration. The feed enters under pressure through a nozzle at the bottom, then is accelerated to rotor speed and move upwards

through the cylindrical bowl.

Disk stack centrifuge is common in bioprocessing, especially for cells harvesting. Various types of disk stack centrifuges are available, and their main difference are the methods used to discharge of solid, i.e. continuous or intermittent discharge. Disk stack centrifuges contain conical sheets of metal discs that are stacked one on top of other with small clearances. The discs rotate with the bowl and the liquid can be split into thin layers. The liquid is discharged from the top of the centrifuge while the sediment is aggregated at the bottom edge of the bowl.

The bench top centrifuges used for laboratory centrifugation are in the similar structure. A specific rotator that is used to hold even tubes, each tube can be filled with small volume of processing materials. The solid and liquid can be separated into two layers as the rotator is spinning in high speed.

These tree types centrifuge are mainly referred in the centrifugation experiments for bioprocessing. After the brief introduction about the equipment, the next step is to introduce how the solids are removed from the liquid by the desired gravity forces generated by centrifuges.

4.2.2 Centrifugation fundamental theories

Generally speaking, centrifugation separates solids from liquid. The solid movement during centrifugation is described by sedimentation velocity. Usually, the sedimentation velocity is determined by the solid liquid physical properties and the internal structure of the centrifuge as well as centrifugal force. Characterizing the internal structure of a centrifuge is realized by sigma factor. Thus, in this section, the theories about the sedimentation velocity and sigma factor are introduced. Although not all of the equations of the theories will be used in the centrifugation system, the parameters and the relationships referred in the theory may help to explain the representation of centrifugation experimental data and knowledge.

4.2.2.1 Sedimentation velocity

The particle velocity achieved in a specific centrifuge and the settling velocity under the gravity force characterize the effectiveness of centrifugation. In the centrifugation, the corresponding particle velocity is determined by equation (4.1).

$$u_c = \frac{\rho_p - \rho_f}{18\mu} D_p^2 \omega^2 r, \tag{4.1}$$

where u_c is the particle velocity in the centrifuge, ρ_p is the density of particle, ρ_f is the density of liquid, μ is the viscosity of the liquid, D_p is the particle diameter, and ω is the angular velocity of the bowl, r is the radius of the centrifuge drum.

Sedimentation happens in a centrifuge as particles moving away from the centre of rotation collide with the walls of the centrifuge bowl. Increasing the velocity of particles will improve the rate of sedimentation to achieve better separation performance.

4.2.2.2 Sigma factor about centrifuges

The properties of specific centrifuge can be characterized by a specific parameter called *sigma factor* (Σ) (Ambler, 1959). It relates the centrifuge geometry and rotational speed to the area of a gravity settling tank capable of performing the same amount of clarification. The Σ is determined by the centrifuge geometries, for the disk stack, tubular and bench top centrifuge, the Σ calculations are given as follows.

For disk stack centrifuges, Σ is determined by equation (4.2) (Tait et al., 2009).

$$\Sigma = \frac{2\pi\omega^2(N-1)}{3g\tan\theta}(r_2^3 - 2_1^3), \tag{4.2}$$

where ω is angular velocity, N is the number of discs in the stack, r_2 is the outer radius of the disc, r_1 is the inner radius of the disc, g is gravitational acceleration and θ is the half-cone angle of the disc.

For the tubular bowl centrifuges, Σ can be determined by the equation (4.3) (Boychyn et al., 2001).

$$\Sigma = \frac{\pi\omega^2 b}{2g}(3r_2^2 + r_1^2), \tag{4.3}$$

where b is the length of the bowl, r_1 is the radius of the liquid surface and r_2 is the radius of the inner wall of the bowl.

For the laboratory batch centrifuges, the Σ is calculated by equation (4.4) (DORAN, 2011).

$$\Sigma = \frac{V\omega^2(3 - 2x - 2y)}{6g \ln \frac{2R_2}{R_2+R_1}}, \tag{4.4}$$

where V is the volume of material in the tube, ω is the angular velocity, R_1 and R_2 are the inner and outer radius, x and y are the fractional times required for acceleration and deceleration respectively.

The $\frac{Q}{C\Sigma}$, named as separator capacity, is commonly used to characterize the performance of centrifugation. In the simplest case of a continuous centrifuge, the separator capacity can be defined as equation (4.5).

$$\Sigma = \frac{Q}{2u_c}, \tag{4.5}$$

where Q is the volumetric feed rate and u_c is the particle velocity in a gravitational field.

The separator capacity allows the separation performance of different centrifuges can be compared. If two centrifuges have equal performance, then the equation (4.6) is generated.

$$\frac{Q_1}{C_1\Sigma_1} = \frac{Q_2}{C_2\Sigma_2}, \tag{4.6}$$

where subscriptions 1 and 2 denote the two different centrifuges, e.g. Q_1 is the flowrate of the centrifuge 1, Q_2 is the flowrate of centrifuge 2. The C_1 and C_2 are the adjustment

constants determined by the centrifuges.

Equation 4.6 gives the relationship of flowrate between the two different centrifuges. For example, if Q_1 and Σ_1 are known for the disk stack centrifuge and Σ_2 is known for a tubular bowl centrifuge, then the Q_2 can be calculated. The separator capacity is a specific parameter concerned by centrifugation experimentation because it can be used for equipment selection by calculating the Σ or the flowrate to the specific centrifuge.

4.2.3 Ultra scale down approach

In order to speed up the centrifugation design, engineers propose to use the small scale experiments to simulate the large centrifugation operation, because the small scale centrifugation allows different operating conditions to be tested in parallel with small volume of processing material. This approach has been validated by practical study (Maybury et al., 2000). However, the physical configuration of the large and the small scale centrifuges are different, this usually leads to great discrepancies about the centrifugation performance between these two scales. In order to accurately predict large scale centrifuge performance based on small scale experiment results, the ultra scale down (USD) approach has been established (Titchener-Hooker et al., 2008).

In large scale centrifuge, the entrance zone where the highest flow stresses are expected to prevail, then the partial particles of the processing material can be split into small particles which may decrease the average particle size, and hence the separation performance would be reduced (Hutchinson et al., 2006). In laboratory centrifugation, the shear force does not exist, thus the average particle size of processing material does not change. Therefore, using these laboratory centrifugation results to predict the performance of large scale centrifugation brings errors. For example, the clarification efficiency of pilot scale centrifugation is worse than the performance of laboratory scale centrifugation (Boychyn et al., 2004). For this issue, the USD approach makes the processing material be treated by a specific shear device (also called rotating disc device) in order to mimic the share damage caused by the

highest flow stresses when the processing material passes through the entrance zone of the large scale centrifuge. The flow stresses caused by the flow pattern in the centrifuge can be measured and simulated by the Computational Fluid Dynamics (CFD) approach, which can characterize the shear environment as energy dissipation respects to the specific geometric configuration of centrifuge (Boychyn et al., 2001). The specific energy dissipation can be simulated by the rotating disc device that is indicated by the shear speed and shear time of the rotating disc device. The simulation is specific to the equipment and the processing material, e.g. cell line, because the shear environment is determined by the centrifuge geometry, and damaging different cell lines require different shear forces. By using the rotating disc device to process the processing material, the small scale centrifugation results can be used to predict the performance of large scale centrifugation.

For the centrifuge in different scales, e.g. pilot and laboratory scale, the equations (4.6) is modified as the equation (4.9).

$$\frac{Q}{C\Sigma} = \frac{V_{lab}}{t_{lab} C_{lab} \Sigma_{lab}}, \tag{4.7}$$

where the C, C_{lab} is the constant, Q is the volumetric feed rate, Σ is sigma factor of pilot scale centrifuge, V_{lab} is volume of material in the tube of laboratory centrifuge, Σ_{lab} is the sigma factor of laboratory scale centrifuge, t_{lab} is the spin time.

By using the equation 4.9, the laboratory centrifugation result can be used to determine the flowrate of the specific large scale centrifugation or the specific centrifuge for the desired flowrate. Several processing materials have been examined by USD approach, e.g. yeast (Maybury et al., 2000), mammalian cell (Zaman et al., 2009; Tait et al., 2009) and *E coli.* (Chan et al., 2006). These studies demonstrate that this approach can make good prediction to the performance of pilot or even larger scale centrifuges.

4.3 REPRESENTATION OF CENTRIFUGATION EXPERIMENTAL DATA

Centrifugation experimental data is the data describing the centrifugation experiments that are primarily concerned by the centrifugation system. To explain how to represent the centrifugation experimental data, the general information included in the centrifugation experiment is introduced, then the parameters used to represent the information of the centrifugation input, step and output are presented respectively. Finally, an example is used to illustrate the representation of the centrifugation experimental data.

The centrifugation experimental data usually includes the information about the processing material properties, e.g. the density, liquid viscosity, the operation conditions and separation performance. The centrifugation experiments are generally implemented to examine the linear interaction between the separator capacity ($Q/c\Sigma$) and clarification efficiency (CE), which serve the maximal throughput identification and the equipment selection (Boychyn et al., 2001). This information is proposed to be used by the centrifugation system, 344 centrifugation experimental data was captured from Dr. Andrew Tait, Dr. Jean Aucamp and Dr. Balasundaram Bangaru in Biochemical Engineering Department, University College London.

Overall the 344 experimental data, 108 experimental data was about separating debris deriving from baker's yeast homogenate, 102 experimental data was about removing debris deriving from *E.coli* homogenate, and 134 experimental data was about CHO cells separation from fermentation broth by using USD approach.

4.3.1 Representation of centrifugation input information

The input information describes the properties of processing material. For 344 experimental data, 10 parameters were identified to represent the information about the properties of processing material which are given in the Table 4.1.

Table 4.1: Parameters for representation of centrifugation input information

Category	Parameter	Definition	Unit	Example
Input	strain	name of cell line or micro-organism that is used to produce the targeted molecule	n/a	CHO
	product	name of target molecule	n/a	IgG
	feed	name of processing material type to the centrifuge	n/a	whole cell
	OD_{feed}	optical density value of processing material at 670nm	n/a	2.16
	pH	pH value of processing material	n/a	7
	solid density	density value of solid part in the processing material	kg/L	1.05 kg/L
	liquid density	density value of liquid part in the processing material	kg/L	1.00 kg/L
	density difference	difference value of solid density and liquid density	kg/L	0.05 kg/L
	particle size	the average particle size value of processing material	μm	50 μm
	solid concentration	the concentration value of solids contained in the processing material	v/v	1.5%
	viscosity	the value of viscosity about liquid	mPa.s	1.3

4.3.2 Representation of centrifugation step information

Information included in step describes the operating conditions about the centrifugation. For these information, 12 parameters were identified to represent the information of centrifugation operation, see Table 4.2.

Table 4.2: Parameters for representation of centrifugation step information

Category	Parameter	Definition	Unit	Example
Step	centrifuge	name of centrifuge	n/a	CSA-1
	function	name of centrifugation function	term	harvest cell
	scale	name of centrifugation experiment scale	n/a	pilot
	temperature	temperature value of processing material	°C	20 °C
	flowrate	flowrate value of processing material through centrifuge	L/h	150 L/h
	sample volume	sample value of small scale centrifugation	ml	1.5 ml
	residence time	residence time value of particles	s	5 s
	sigma Σ	equivalent sediment space value of a centrifuge	m^2	2000 m^2
	separator capacity	value of $Q/c\Sigma$	m^2/s	1.08E-8 m^2/s
	rotation speed $Q/c\Sigma$	value of rotation speed/bowl speed	rpm	5000 rmp
	shear speed	value of shear device rotating speed	rpm	6000 rmp
	shear time	rotation time value of shear device	s	20 s

4.3.3 Representation of centrifugation output information

The information about centrifugation output includes the separation performance. For the 344 experimental data, 4 parameters were identified to represent the separation results that are shown in Table 4.3.

Table 4.3: Parameters for representation of centrifugation output information

Category	Parameter	Definition	Unit	Example
Output	product phase	name of phase that bioproduct exists	n/a	supernate
	OD_{sample}	optical density value of supernate	n/a	0.65
	$OD_{reference}$	optical density value of reference used for eliminating background noise	n/a	0.02
	CE	value of clarification efficiency	%	97.1 %

4.3.4 Illustration of experimental data representation

For the centrifugation experimental data, 26 parameters were identified to represent the information about the processing material properties, operation conditions and separation performance. In order to illustrate the experimental data representation, a centrifugation experiment is used to show how to draw the information and how to represent them.

In this experiment, the laboratory scale centrifugation was used to harvest the CHO cells from the cell culture. The experimental information was extracted from the 'materials and methods section' and 'results section' of a journal paper (Tait et al., 2009), and it was labeled by 'Materials', 'Operations' and 'Results', as shown in Table 4.4.

Table 4.4: Information of laboratory scale centrifugation experiment

Materials: Chinese Hamster Ovary (CHO) cell culture broth as the processing material; the solid density was 1.05 kg/m^3; the liquid density was 1 kg/m^3, mean particle size was 18μm.
Rotating disc device and centrifuge: laboratory scale centrifugation was conducted; the sample was 0.5 ml; the sample was sheared by the specific shear device at 12000 rpm for 20 s; Eppendorf 5810r was used for spinning 620 s; the separator capacity (Q/cΣ) was $1.25 \times 10^8 m^2/s$.
Results: cells were in the sediment and total 91.8% cells were removed.

The information of **Materials**, **Rotating disc device and centrifuge** and **Results** corresponds to the input, step and output separately. Each item of information is represented as a specification. For instance, 'Chinese Hamster Ovary (CHO) cell culture broth as the processing material' describes the strain used in this experiment which is represented as 'strain(CHO)', 'totol 91.8% cells were removed' introduces the result of CE that is represented as 'CE(91.8)'. All of the specifications from a datapoint that is illustrated in Figure 4.1.

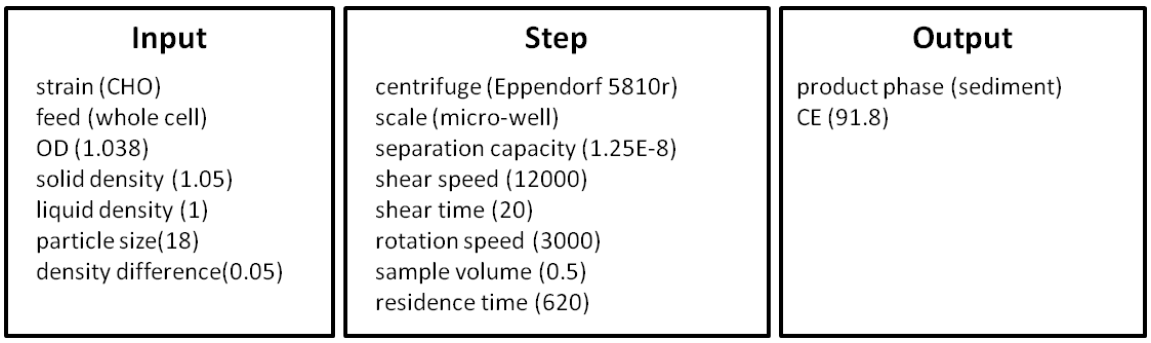


Figure 4.1: Representation of a laboratory scale centrifugation experiment

In this representation, 16 parameters were used which indicates that not all of the 26

parameters would be used for one experimental data representation, e.g. the parameter 'flowrate' is not used.

4.3.5 Summary

For the centrifugation system, each represented experimental data was a datapoint, and the 344 datapoints formed the database of the centrifugation system. Each datapoint was tagged with a unique ID. All of captured experimental data were kept in the form of a table, in which each column represented one parameter of the centrifugation input or step or output, and each row represented one experiment data.

The data representation is expandable. For instance, the discharge time is usually used in the manufacturing data, this information can be represented and captured by adding on a new column, 'discharge time', in the database without modifying the existing datapoints.

4.4 REPRESENTATION OF CENTRIFUGATION ONTOLOGY

The centrifugation ontology systematically describes the relationships between the terms used in the centrifugation experimental data. Ontology development consists of four steps, i.e. determine the domain, enumerate the specific terms, define the classes and class hierarchy and create the instances which have been explained in Chapter 3. Following the four steps, the development of centrifugation ontologies is introduced in this section.

For the centrifugation system, seven domains ontologies were considered, namely *strain*, *product*, *feed*, *equipment*, *scale*, *phase* and *centrifugation function*.

- *Strain ontologies* define the terms about the cell lines or the micro-organisms used for bioproduct production, e.g. baker's yeast.
- *Product ontologies* define the terms about the target bioproduct, e.g. IgG.

- *Feed ontologies* define the terms about processing material characteristics, e.g. whole cell.
- *Equipment ontologies* define the terms about the centrifugation equipments, e.g. disk stack centrifuge.
- *Scale ontologies* define the terms about the experiment scale, e.g. pilot.
- *Phase ontologies* define the physical statement terms of solid and liquid, e.g. sediment.
- *Centrifugation function ontologies* define the terms of centrifugation function in the bioprocessing, e.g. cell harvest.

The seven domains were considered because their terms were referred in 344 centrifugation experimental data. With the defined domains, the following steps are enumerating the terms included in the domains, defining the class and hierarchy and creating the examples. Two domains, i.e. strain and equipment, are used as the example to illustrate the development of ontologies and their roles in the centrifugation system.

4.4.1 Ontologies of strain

Strain is the specific micro-organism used in producing product. Usually, there are many expression systems to produce the biologic substances, and the material properties from different expression systems are also different, e.g. cell shape. These different characteristics would lead to different operating conditions of the centrifugation, even if the same equipment is employed. Furthermore, cells with gene modifications would be given a new name but the physical properties of associated material would be very similar. For instance, CHO-K1 and CHO-RD were genetically modified CHO cell line with different gene expression function, but they have the very similar physical properties, e.g. cell shape, cell size. It is very possible that they would have similar separation behavior with the centrifugation, which is based on the processing material physical properties. Therefore, it is reasonable to use the experimental data about the CHO-RD for the centrifugation design problem of the CHO-K1, and the ontologies of strain can be used for these situations.

For the strain ontologies, the terms were obtained from the 344 datapoints and consolidated through the discussion with the biochemical engineering scientists, and they are commonly used for the biomolecule production. The strain ontology consists of 15 strain terms, the general strain terms are in the parent classes while the specific strain terms are in the child classes. The relationship, 'a type of', is used as the affiliated relationship between the parent and child class which indicates that the specific strain inherits the physical properties from the general strain, e.g. cell shape. All of the terms and the relationships are arranged by a hierarchy that is shown in Figure 4.2.

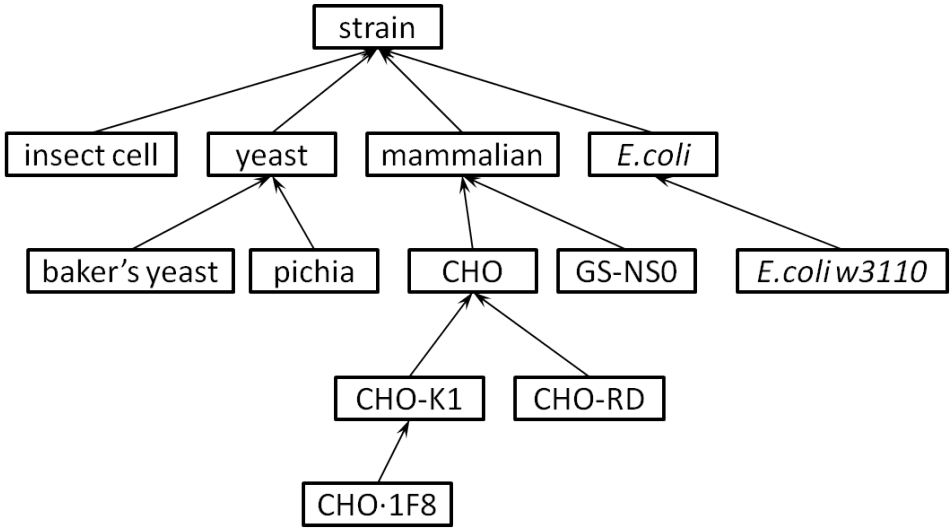


Figure 4.2: Strain ontologies of centrifugation

Each node is a class and represents a strain term, and each arrow indicates the affiliated relationship. The class at the arrow-head is the general strain term and the class at the arrow-tail is a specific strain name. The arrow can be read as 'a type of' between the specific and general strain terms. Two facts (instances) are given for the strain ontologies, e.g. *E. coli* is a type of strain, *E. coli* w3110 is a type of *E. coli*. These facts serve as the terminological criteria for searching functionality, i.e. whether the strain specification of the experimental data satisfies the strain feature of the design query.

The strain ontologies can be used to differentiate specific cell line from others in searching the experimental data. For example, if the CHO cell is interested, the strain ontologies

would not only allow the experimental data about the CHO cell line to be found, but also includes the experimental data about the cell line which is the subset of CHO cell line, e.g. CHO-K1 or CHO-RD or CHO·1F8. The data of three cell lines would contribute to the CHO cell centrifugation design, because they have similar physical properties. This may help to find new design solutions that have not been implemented on the CHO cell line before. It is also an effective way of data utilization, especially for a new strain that few previous experimental data of this particular strain is available. Hence, more related experimental data can be accessed for identifying the feasible design solutions by using this ontologies.

4.4.2 Ontologies of equipment

Usually, there are many centrifuges available ranging from those for laboratory experiments to those for industrial operations. Each equipment has the specific physical configuration, e.g. the Σ, which would impact the operation and separation performance. Thus, it is necessary to differentiate the desired equipment from others when doing the centrifugation design. To realize this purpose, the ontologies of equipment were established.

For the equipment ontologies, the terms were determined based on the laboratory scale and the industrial scale equipments. The four types of equipments are commonly used in centrifugation, i.e. bench top, disk stack, tubular bowl and multi-chamber. The general centrifuge terms are at the parent classes. For each of the general centrifuge, there are a set of the specific centrifuges and the term of each of specific centrifuge is in the child class. 'A type of' is used to indicate the relationship between the parent class and child class which represents the specific centrifuge inherits the similar configuration from the general centrifuge, e.g. all of the disk stack centrifuges contain conical sheets of metal discs, but each specific disk stack centrifuge has the different radius of the metal disc. The equipment ontologies consist of 14 terms that are shown in Figure 4.3.

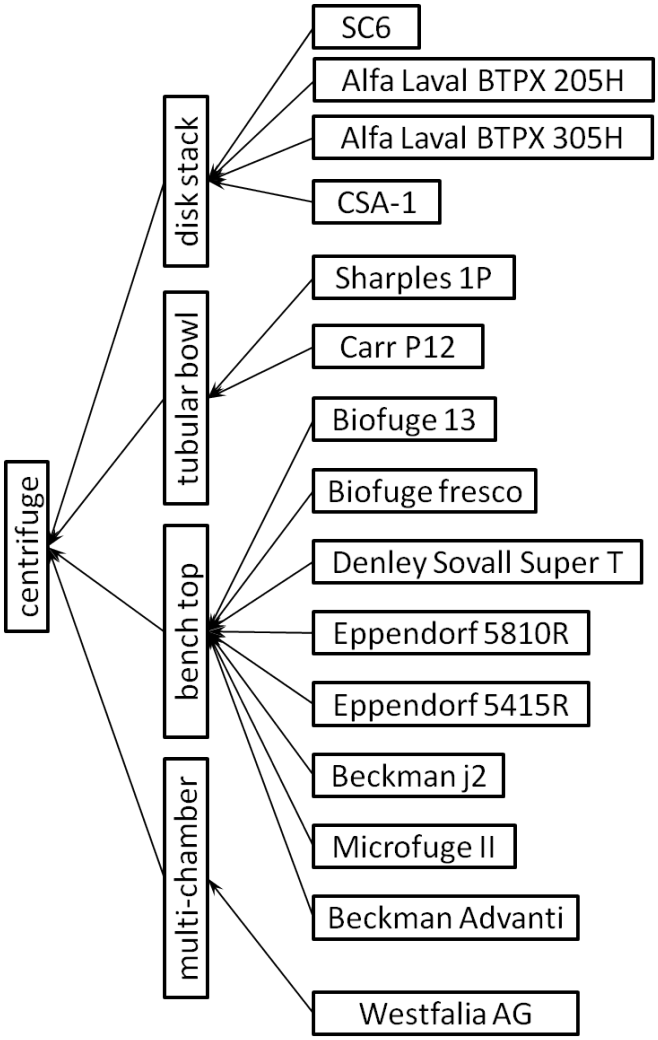


Figure 4.3: Equipment ontologies of centrifugation

In the hierarchy, each node represents a centrifuge term, each arrow is the affiliated relationship, i.e. 'a type of'. The general equipment terms are in the parent level, e.g. disk stack, tubular bowl and etc. Other specific equipment terms are in the child level. Two facts are given, e.g. disk stack is a type of centrifuge, SC6 is a type of disk stack centrifuge. These facts serve as the terminological rules for equipment judgment.

In order to search the experimental data about the disk stack centrifuge, the equipment ontologies can be used to exclude the experimental data related to other types of equipment, i.e. tubular bowl, bench top and multi-chamber, and also to find all experimental data whose equipment is a type of disk stack centrifuge, e.g. SC6, CSA-1. The found data may be

helpful for identification of the feasible operating conditions and selection of the specific disk stack centrifuge.

4.4.3 Other ontologies in centrifugation system

For other five ontologies, i.e. feed, product, scale, phase and function, the development is the same as the ontologies of strain and equipment. The terms included in these domains were captured from the 344 datapoints. These ontologies would allow the centrifugation system to recognize and differentiate the terms used in the experimental data and knowledge. The established ontologies of the five domains are shown in the following.

4.4.3.1 Feed ontologies

The feed describes the forms of solids in the processing material. The whole cell, homogenate and precipitate were the three common forms used in the bioprocessing. For the three terms, the ontologies of the feed are shown in Figure 4.4.

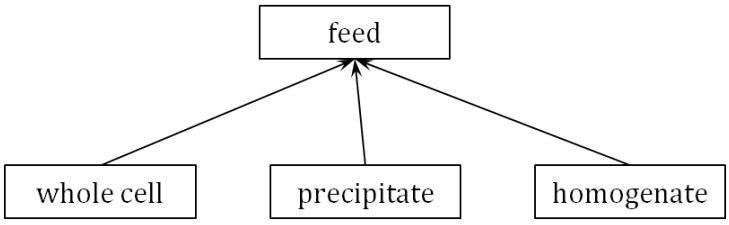


Figure 4.4: Feed ontologies of centrifugation

4.4.3.2 Product ontologies

Currently, the product terms, e.g. monoclonal antibody (mAb), IgG, IgA, IgD, polyclonal antibody (pAb), enzyme and lipase, were used. The ontologies of product are given by Figure 4.15. It defines the product terms that can be searched, and it also describes the relationship between these products, e.g. IgG is a type of mAb.

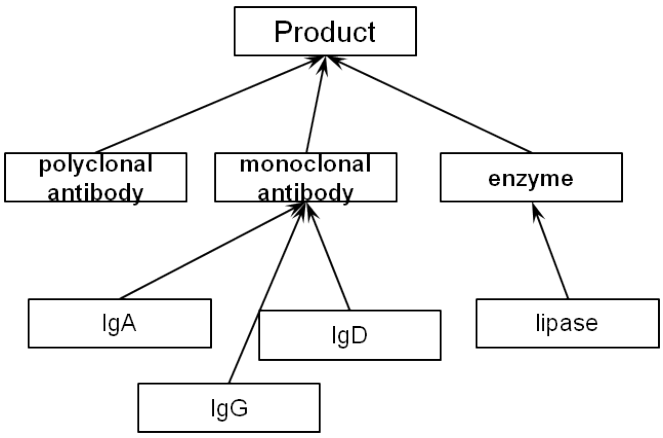


Figure 4.5: Product ontologies of centrifugation

4.4.3.3 Scale ontologies

The 344 datapoints included three types of experimental scale, e.g. laboratory, pilot and manufacturing. The ontologies about the four experimental scale terms are shown in Figure 4.6.

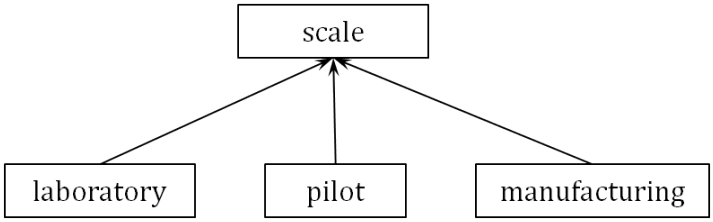


Figure 4.6: Scale ontologies for centrifugation

4.4.3.4 Phase ontologies

Sediment and supernate were used to describe the solid or liquid formation after the centrifugation. The phase ontologies are given in Figure 4.7

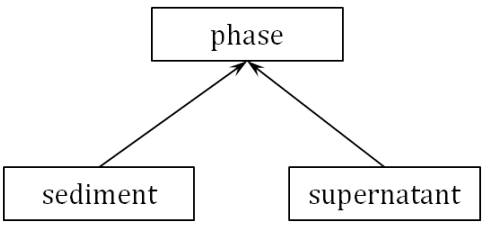


Figure 4.7: Phase ontologies for centrifugation

4.4.3.5 Centrifugation function ontologies

Cell harvest, debris elimination and precipitate collection are the three roles of centrifugation in downstream processing. Therefore, the three terms were captured by the centrifugation function ontologies that are shown in Figure 4.8.

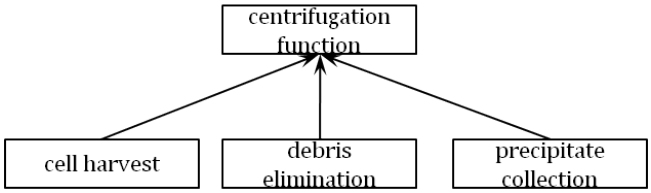


Figure 4.8: Function ontologies for centrifugation

4.4.4 Summary

For the centrifugation system, 40 ontologies have been captured in total that include 8 strain ontologies, 16 centrifuge ontologies, 3 feed ontologies, 4 scale ontologies, 2 product ontologies and 2 phase ontologies and 3 centrifugation function ontologies. Each ontology was a fact that consisted of the child class term and the parent class term as well as the 'a type of' relationship. These 40 facts formed the ontology base in the centrifugation system.

The ontology is an open technique that it can be expanded. For instance, if the new term was used in the experimental data, it can be represented as a new class that was linked to other classes in the corresponding domain with the defined relationship. Therefore, the ontologies could be further developed for the newly captured experimental data.

4.5 REPRESENTATION OF CENTRIFUGATION KNOWLEDGE

Two types of knowledge are proposed to be harnessed by centrifugation system for experimental data utilization, namely theoretical and empirical knowledge. The theoretical and empirical knowledge usually has various expressions, e.g. equations, plain text and etc.

Using a way to represent this knowledge is required for harnessing it. In this section, the representation of these two types of knowledge as well as their roles in experimental data utilization are discussed.

4.5.1 Representation of centrifugation theoretical knowledge

The theoretical knowledge of centrifugation includes the fundamental equations and the background information of the centrifuge. In the following, two examples are given to illustrate the theoretical knowledge representation.

4.5.1.1 Fundamental equation

The CE calculation serves as an example to illustrate the representation of fundamental equation. The CE is calculated by the optical density (OD) measurements of the processing material and the supernate which is shown in equation (4.8).

$$\%CE = \frac{OD_{feed} - OD_{sample}}{OD_{feed} - OD_{reference}} \times 100, \tag{4.8}$$

Equation (4.8) gives the knowledge entity of CE calculation, the four variables and the mathematical relationship referred by this entity were used to formalize this knowledge. The variable 'CE' was the output of this knowledge, while the three variables, i.e. 'OD_{feed}', 'OD_{sample}' and '$OD_{reference}$', were the input of this knowledge. To execute this knowledge, the three input variables are required to be defined, then the value of output variable would be generated by the mathematical relationship.

The role of this knowledge aims at generating the CE value when it is not available. For instance, if the CE value is missing in a datapoint, this knowledge allows the CE value to be completed for further use.

All of the required fundamental equations can be formalized in this way, e.g. separator capacity calculation (see equation (4.9)). To represent the equation, the 7 variables and the mathematical relationship were used. If the 'Q' is the output variable, it can be calculated

by defining the values of other six input variables. Harnessing the fundamental equations allows the centrifugation system to have the potential to use the mathematical modes for further data analysis.

$$\frac{Q}{C\Sigma} = \frac{V_{lab}}{t_{lab}C_{lab}\Sigma_{lab}}, \tag{4.9}$$

4.5.1.2 ERM of centrifuge background information

Equipment selection is an important goal of centrifugation design, and it requires to harness the background information of centrifuge to decide whether this equipment is suitable choice. The ERM is proposed to be used to represent the background information of the specific centrifuge. Each specific centrifuge is an entity, each item of background information of the centrifuge is an attribute, the relationship between the entity and attribute is defined as 'has' which indicates the entity owns the attribute.

Currently, 5 items of background information about the centrifuge were considered, i.e. maximum rotation speed, maximum flowrate, settling area, bowl volume and adjustment constant. The maximum rotation speed and flowrate would influence the solutions on throughput of centrifugation. The settling area, bowl volume and the adjustment constant are required by separator capacity calculation or ultra scale down approach for the centrifugation design. The ERM of 'Alfa Laval BTPX 305H' is used for demonstration, and it is shown in Figure 4.9.

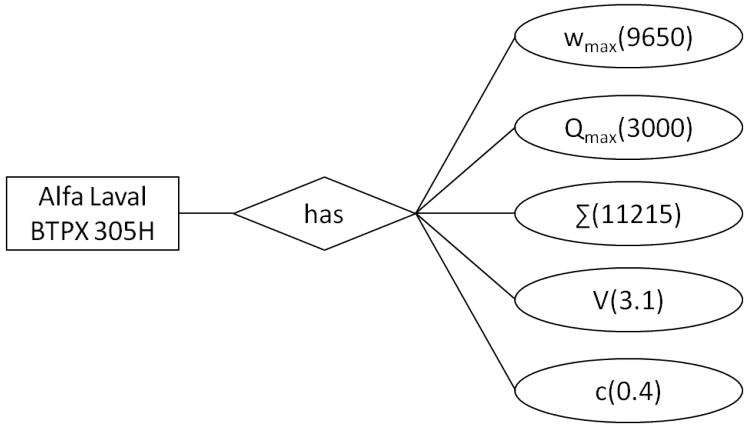


Figure 4.9: ERM of Alfa Laval BTPX 305H background information

In this ERM, 5 items of background information were captured from the Alfa Laval BTPX 305H manual. The rectangle represents the entity, Alfa Laval BTPX 305H; the diamond denotes the relationship between the entity and the attribute; and each ellipse is an item of background information represented as a parameter with value. For these attributes, ω_{max}(9650) represents 'the max rotation speed is 9650 rpm'; Q_{max}(3000) defines 'the max flowrate is 3000 L/h'; Σ(11215) introduces 'the settling space of Alfa Laval BTPX 305H is 11215 m^2'; 'bowl volume is 3.1 L' and 'adjust constant is 0.4' are described by V(3.1) and c(0.4) respectively.

If the entity is concerned, all of the attributes are available for the centrifugation system to use. These attributes could be used to decide whether the centrifuge is reasonable solution or not to the design problem. For example, the Q_{max}(3000) defined the maximum flowrate is 3000 L/h, if the flowrate is greater than 3000 L/h, then the Alfa Laval BTPX 305H would not be suggested as the centrifuge to choose. These attributes can also provide the information required by centrifugation system. For example, the Σ value required by separator capacity calculation can be obtained from the corresponding attribute.

The attributes of the specific centrifuge can be expanded. For instance, the general utilities consumption is also provided by manufacturer, e.g. operating water, it can be represented as an attribute that can be used to estimate the utilities cost about the centrifugation.

4.5.2 Representation of centrifugation empirical knowledge

The empirical knowledge focuses on the principles obtained from the empirical studies that have been published on journals. For the centrifugation system, the USD approach is considered as the empirical knowledge. The USD approach establishes a link of the separation performance between the small scale and large scale centrifugation that has been introduced in section 4.2.3. Thus, it can be used as a rule by centrifugation system to harness the experimental data generated from different experimental scales.

4.5.2.1 Representation of USD rules

One practical USD approach was captured from the Hutchinson's study (Hutchinson et al., 2006). If the CHO cell broth is processed by the rotating device at 6000 rpm for 20 s, the CE results done at the laboratory scale can be used to predict the CE performance of Alfa Laval BTPX 305H centrifuge which is a type of disk stack centrifuge. For this USD approach (knowledge entity), the variables about cell line, scale, centrifuges, shear time and shear speed are specified. The fact that the laboratory scale centrifugation results can be used to predict the pilot scale performance indicates an equivalent relationship. With the 5 variables and the relationship, the USD rule is formalized as follows (italic font).

For two experiment data A and B, namely Data A and Data B;

IF strain(CHO) & scale(pilot) & centrifuge(Alfa Laval BTPX 305H) & CE(x) ⊂ Data A;

strain(CHO) & scale(laboratory) & shear time(6000) & shear time(20) & CE(y) ⊂ Data B;

THEN x=y.

This USD rule indicates that the CE performance in the datapoint which consists of the specifications of strain(CHO), scale(pilot) and centrifuge(Alfa Laval BTPX 305H) is equal to the CE performance in the datapoint which includes the specifications of strain(CHO), scale(laboratory), shear speed(6000) and shear time(20). It can be used for experimental data searching. For instance, if users want to search the pilot scale experimental data about using Alfa Laval BTPX 305H centrifuge to separate CHO cells, then the USD rule allows the experimental data generated by USD approach to be searched.

The USD rules of the centrifugation system is specific to the rotating disk device which was designed and fabricated in UCL. The shear speed and shear time specified in the rules are the operating conditions of the rotating disk device. For centrifugation system, all of the USD experimental data was generated by using the same rotating disk device. Therefore, other new USD approaches to be captured should be developed by using this specific rotating disk device.

Other empirical knowledge about USD approach can be represented as the USD rules in the same way. However, they are not going to be demonstrated, because capturing all of USD rules is not primary concern in this thesis.

4.5.3 Summary

For the centrifugation system, two fundamental equations were captured, five ERMs about centrifuge background information were established, i.e. Alfa Laval BTPX 305H, Carr P12, Eppendorf 5810R, Alfa Laval BTPX 205H, and SC6, and two USD rules were used to harness the experimental data generated from the laboratory and pilot scale.

The ontologies, theoretical and empirical knowledge introduced in this chapter served as examples to illustrate the knowledge representation and how they are harnessed to solve design problem. Due to the limited time, not all of the knowledge has been included. Therefore, studying the representation of other available knowledge would be desirable in the future.

4.6 REASONING OF CENTRIFUGATION DATA AND KNOWLEDGE

With the four types of data and knowledge, the next question is how to use them for centrifugation design. The design query and three reasoning functionalities are developed for these problems. In order to illustrate how they work, in this section, the representation of centrifugation design query and how the three reasoning functionalities work regarding the design query are presented. For this, formalizing the centrifugation design problem is introduced first, then the reasoning of search, prediction and suggestion is explained by using the example of centrifugation design query. Finally, a realistic case is used to illustrate how the theoretical and empirical knowledge can be harnessed by centrifugation system.

4.6.1 Representation of centrifugation design query

The centrifugation design query represents the centrifugation design problems that users may have, e.g. to predict a separation performance for giving the information of processing material and operating conditions, to identify the suitable operating conditions in order to achieve the desired separation performance for the specific processing material.

The centrifugation design query representation is the same as the centrifugation experimental data representation. The centrifugation design query has three parts, i.e. input, step and output. In each part, each item of information is represented as a feature including a parameter with value. The parameters employed by features are from the 26 parameters that are used to represent the centrifugation experimental data.

In order to illustrate the design query representation, a practical design problem about centrifugation is used. In this design problem, users want to know what the flowrate about a pilot scale disk stack centrifuge should be in order to achieve at least 90% CE in yeast cell harvest. The information about this design problem and the design query is shown in Table 4.5.

Table 4.5: Representation of design query about case study on yeast cell harvest by centrifugation

Design information	**Design query**
Processing material: Yeast cell, optical density is 2.42 **Centrifuge:** Pilot scale disk stack would be used, and the flowrate is queried **Requirement:** At least 90% clarification efficiency	**Input:** strain(yeast), OD_{feed}(2.42) **Step:** scale(pilot) centrifuge(disk stack) flow rate(X) **Output:** CE(90)

Six items of information about the processing material, centrifuge and requirement are used to formalize the six features about the input, step and output. For instance, the feature

'scale(pilot)' represents 'the scale is pilot'. The queried feature 'flowrate(X)' denotes 'what the flowrate should be', and it requests the solution of flowrate from the centrifugation system.

The number of features included in the design query depends on the information involved in the design problem. The more information the design problem has, the more features the design query includes. However, for any design query, the feature of strain should be included at least, and other parameters which are not included in the 26 parameters should not be used in the centrifugation design query. The design query provides a way to allow the user to access the datapoints by using the limited information. For example, if use only specifies the *E. coli* in the design query, then all of the datapoints contained *E. coli* would be examined.

The design query is used to constrain the three reasoning functionalities on harnessing the data and knowledge. In order to demonstrate how the reasoning functionalities work, the design query shown in Table 4.5 is used to explain the reasoning of search, prediction and suggestion functionality in the following.

4.6.2 Reasoning of search

For the centrifugation system, the search functionality is designed to find a set of experimental data that are relevant to the centrifugation design query. These relevant data may answer user's question that what similar centrifugation experiments have been done before. In Chapter 3, the finding relevant datapoints for the design query is realized by using the two types of criterion, namely numerical and terminological criterion, to judge the relevance between the specification and the feature. This section explains the reasoning process by using the practical design design query.

4.6.2.1 Definition of search criteria

There are six features in the design query (see Table 4.5). For the input feature of OD_{feed}(2.42), the numerical criterion is defined as [2.42(1-10%), 2.42(1+10%)], i.e. [2.18, 2.66], 10% here is used as demonstration only. For the output feature of CE(90), the numerical criterion is [90, $+\infty$), because 90% is the minimal requirement of CE. For the three features, i.e. strain(yeast), centrifuge(disk stack) and scale(pilot), the terminological criteria have been defined by the ontologies of strain, equipment and scale. For the queried feature, flowrate(X), no criterion is required.

4.6.2.2 Pseudo code of search functionality

To illustrate how search functionality access the database and knowledge base, the pseudo code is used and given in Figure 4.10.

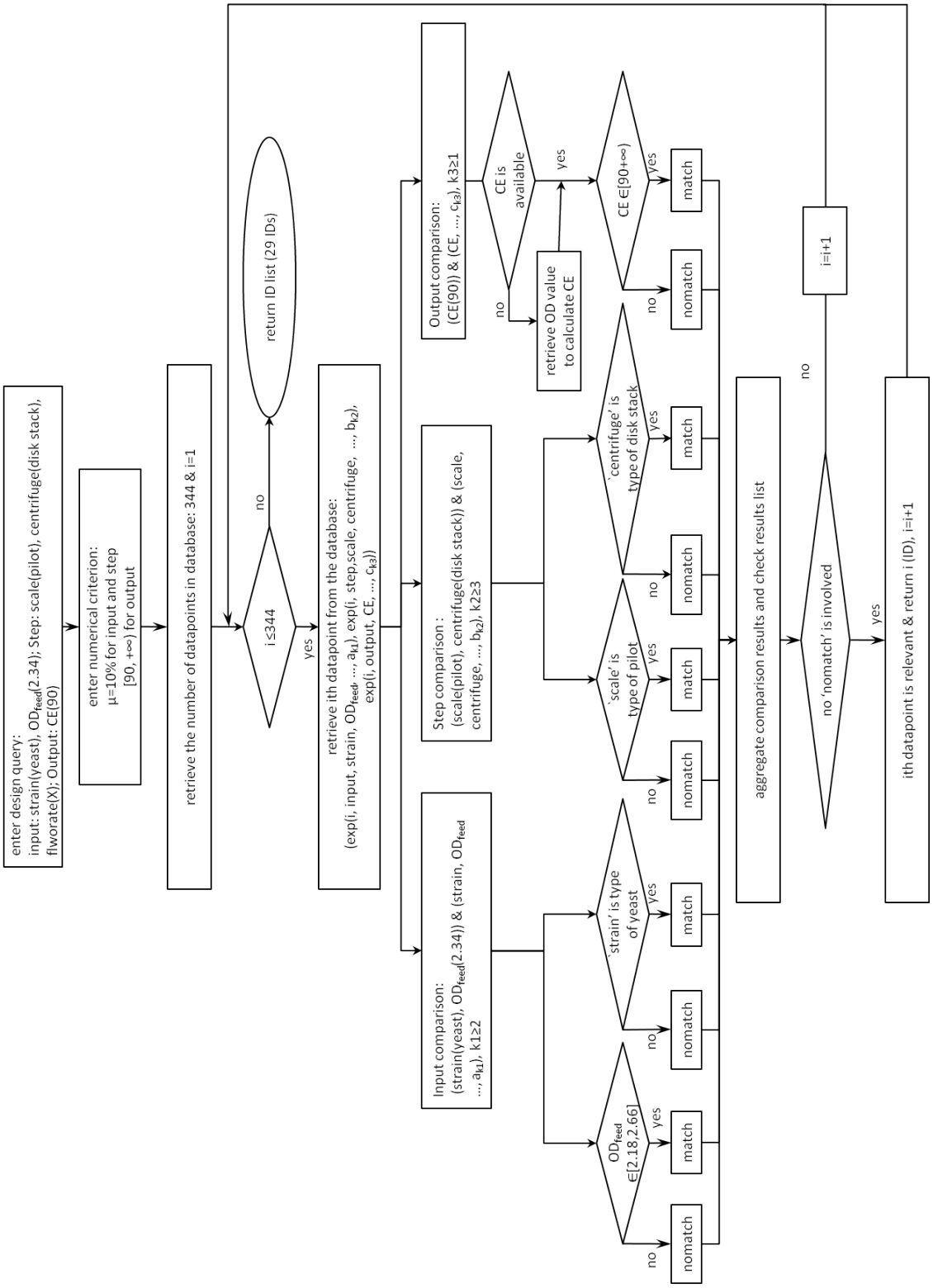


Figure 4.10: Pseudo code of search datapoints for the design query on yeast cell harvest, where the 'strain', 'scale' and 'centrifuge' indicate the specification of strain, scale and centrifuge involved in any of the 344 datapoints.

The logic of programming implementation is explained as follows:

1. The design query consists of six features (the requested feature is not used for comparison).

2. Define the numerical criteria to constrain the data search.

3. 344 comparison loops (M=344 and i=1) would be implemented for this design query.

4. i⩽M (start a comparison loop).

   (a) If yes, the corresponding specifications of input, step and output, i.e. k_1, k_2 and k_3, from the ith datapoint are used to compare with the five features of design query.

      i. For features of OD_{feed} and CE:

         A. The 'match' is returned, if the specification is satisfactory.

            - If the CE value of a datapoint is not available, the theoretical knowledge is accessed for CE value generation, e.g. the OD information (i.e.$OD_{reference}$, OD_{sample} and OD_{feed}) is retrieved from this datapoint and harnessed by CE calculation for CE value.

         B. The 'nomatch' is returned, if the numerical feature is not satisfied.

      ii. For features of strain, scale and centrifuge:

         A. The 'match' is returned, if the relation fact of the two terms exists in the ontology base (the ontology base is accessed), e.g. 'baker's yeast is a type of yeast' is one item of strain ontologies.

         B. The 'nomatch' is returned, if the terminological criterion is not satisfied.

      iii. Check the aggregated comparison results,

         A. If there is no 'nomatch', the datapoint is relevant to the design query and the ID is recorded.

         B. Otherwise, the datapoint is not relevant.

      iv. start a new comparison loop (i=i+1 and go back to step 4).

   (b) if no, the comparison loop stops and return the ID list of relevant datapoints.

4.6.2.3 Reasoning results in search

To illustrate the reasoning of the judgment about the feature and the specification under the defined criteria, one datapoint is selected as an example. The reason why the specification satisfies the feature is given in Table 4.6 by italic font.

Table 4.6: Reasoning on a datapoint and design query about case study on yeast cell harvest by centrifugation

Design query	**Datapoint**	**Reasoning**
Input: strain(yeast) OD_{feed}*(2.42)*	Input: strain(baker's yeast) OD_{feed}*(2.34)*	*- strain(yeast) is satisfied by strain(bakers yeast), because bakers yeast is a type of yeast defined in strain ontology.* *- OD_{feed}(2.42) is satisfied by OD_{feed}(2.34), because 2.34 is with in [2.18, 2.66]*
Step: centrifuge(disk stack) scale(pilot) flowrate(X)	Step: centrifuge(CSA-1) scale(pilot) flowrate(75) separator capacity(2.05E-8) rotation speed(9800)	*- disk stack is satisfied by CSA-1, because CSA-1 is a type of disk stack defined in centrifuge ontology.* *- pilot is satisfied, because pilot is a type of pilot defined in scale ontology.*
Output: CE(90)	Output: CE(91.8) product phase(sediment) OD_{sample}(0.202) $OD_{reference}$(0.11)	*- 90 is satisfied, because 91.8 is within [90, +).*

For this design query, 29 datapoints were returned from the 344 datapoints included in the centrifugation system database which form the basis for prediction and suggestion functionality.

4.6.3 Reasoning of prediction

For centrifugation design, user wants to know how the centrifugation performs under a given a set of operating conditions. This would form a base for use to select process operating conditions that lead to safe, efficient and robust separation solution. The prediction functionality is designed for this purpose. In this section, an example is used to illustrate how the prediction functionality generate the likely achievement for CE.

For prediction, the arithmetic mean algorithm is selected to produce the likelihood performance. For the design query shown in Table 4.5, based on 29 datapoints returned by search functionality, the prediction functionality aggregates all of CE value from the returned datapoints and uses the arithmetic mean algorithm to generate mean CE value. This calculation is shown by equation (4.10).

$$\overline{CE} = \frac{1}{29} \times (95.9\% + 93.9\% + ... + 91\%), \tag{4.10}$$

The result of $\overline{CE}$ is 91.8%. It represents the likely CE performance that may be achievable. It means that if using pilot scale disk stack centrifuge to harvest yeast cells, the CE may be 91.8% according to the relevant experimental data.

4.6.4 Reasoning of suggestion

With the predicted performance, the next step is to draw conclusion from the relevant experimental data and suggest how to realize the predicted performance in reality. The suggestion functionality is designed to achieve that goal. The solutions generated by suggestion functionality will be used for the further centrifugation experimentation, and these experiments are expected to achieve the predicted performance. In this section, an example is used to illustrate how the suggestion functionality retrieve the specific information requested by the design query.

For the centrifugation design query used in section 4.6.1, the flowrate is the operating condition that is required to realize at least 90% CE. According to the 29 relevant datapoints,

the centrifugation process is to likely perform 91.8% CE based on the prediction functionality. As it is better than 90% and has high chance to be realized, the flowrate that achieves 91.8% is the one wanted. Therefore, the datapoint that has the closest CE to 91.8% needs to be identified and its flowrate can be retrieved.

4.6.4.1 Pseudo code of retrieving flowrate

For better explanation, the pseudo code of retrieving the requested flowrate performed by suggestion functionality is given in Figure 4.11.

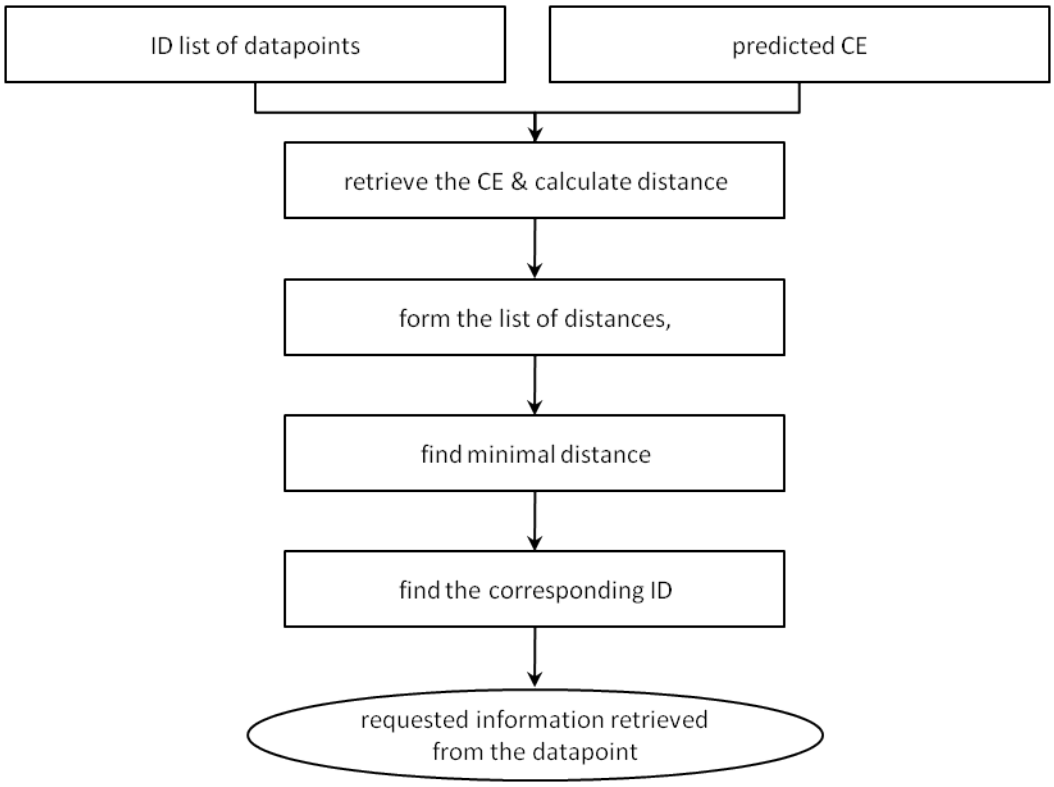


Figure 4.11: Pseudo code of retrieving flowrate from one of 29 datapoint

The logic of retrieving flowrate is explained as follows:

1. Given the 29 datapoints ID list returned by search functionality.
2. Given the predicted CE (91.8%) returned by prediction functionality.
3. Retrieve the CE from each of 29 datapoint and calculate the distance by equation (4.11).

4. 29 distances form a list, [-4.0, -2.0, 0.1,, 1.4].

5. Minimal distance is located, 0.1, and corresponding ID is returned from ID list, 22.

6. Retrieve the flowrate from the 22nd datapoint.

$$CE\ distance = |practical\ CE - predicted\ CE|, \tag{4.11}$$

where the *practical CE* denotes the CE value recorded in the datapoints, the *predicted CE* is the CE value generated by the prediction functionality.

Based on the distance calculation, the flowrate of 90 L/h, is retrieved from the datapoint whose distance is 0.1. It is the smallest comparing with other 28 distance that indicates the CE of this datapoint is most similar to the predicted CE, and the flowrate in this datapoint may achieve the 91.8% CE performance.

4.6.4.2 Discussion of retrieved flowrate

In the 29 datapoints, all of the CE values are better than the 90% requirement and various flowrate conditions are included ranging from 20 L/h to 270 L/h (see Figure 4.12). These different experiments provide a general design space to the design query. The predicted CE generated by arithmetical average is the statistical expectation. For example, each of the 29 experimental data includes other conditions that were specified in the design query, e.g. solid concentration, rotation speed, these conditions could impact the CE performance, therefore the average 91.8% is the possible achievement. Furthermore, the arithmetical average would help to eliminate the experimental errors included in the experimental data, e.g. the CE 95.9% appearing around flowrate 50 L/h that is much higher than other CE values which may be caused by experimental errors. The datapoint indicated by the rectangle is a centrifugation experimental fact that has CE 91.7% which is most close to predicted CE 91.8%, therefore the information of input and step included in this datapoint would be more potential than other datapoints to make the centrifugation achieve the 91.8%.

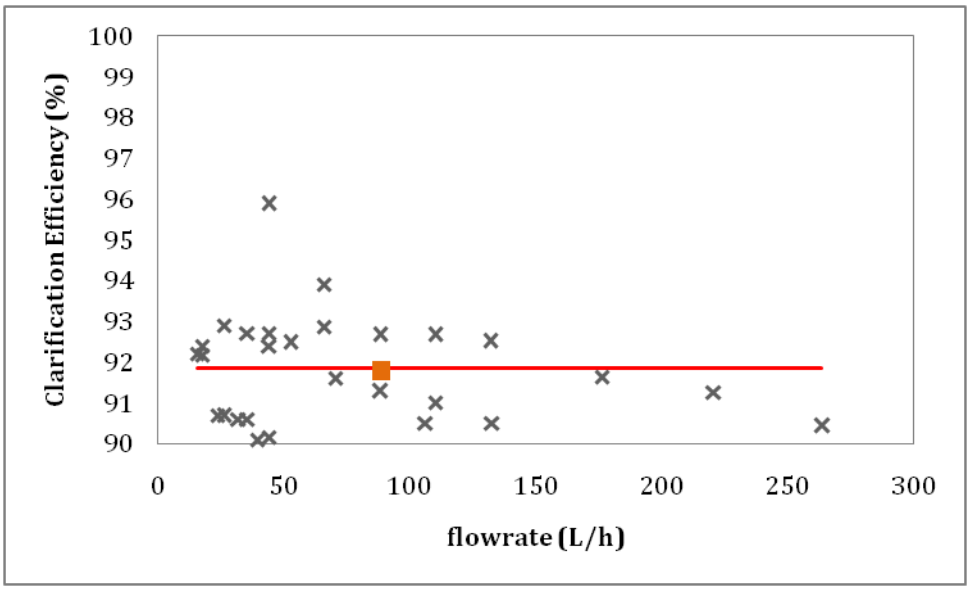


Figure 4.12: 29 datapoints related to the design query of yeast harvest by centrifugation. The solid line represents the predicted CE value, the rectangle represents the datapoint which the flowrate was retrieved from and the cross represents the other 28 datapoints.

This case study illustrates that the centrifugation system can provide an effective way for users to access the relevant experimental data based on the limited information. These relevant experimental data provides a set of solution candidates that can help users to gather useful information efficiently.

4.6.5 Knowledge utilization in reasoning functionalities

The previous case study does not show how the theoretical and empirical knowledge can be systematically harnessed in the three reasoning functionalities. Therefore, in this section, a realistic case about harvesting CHO cells by centrifugation is used to demonstrate how these knowledge can be harnessed for design problem solving. For this, the background information about the specific case is introduced first, then the results generated by each reasoning functionality are given to explain the knowledge utilization.

4.6.5.1 Design query and criteria of case study

This case study was about using the centrifugation to remove the CHO cells from the broth while the product IgG remained in the supernate. The pilot scale disk stack centrifuge,

namely Alfa Laval BTPX 305H, was used. The processing material physical properties have been identified, where pH was 6.95, temperature was 20 $°C$, solid density was 1.051 kg/L and liquid density was 1.0 kg/L. At least 95% cell removal was required. The flowrate of centrifuge was required to be determined for the desired CE performance.

The information about the processing material properties, operating conditions and separation requirement are formalized as the features about input, step and output of the design query that is shown in Table 4.7.

Table 4.7: Design query of the case study on harvest CHO cells by centrifugation

Input	**Step**	**Output**
strain(CHO) feed(whole cell) temperature(20) product(IgG) solid density(1.051) liquid density(1) pH(6.95)	scale(pilot) centrifuge(alfa Laval BTPX 305H) flowrate(X)	product phase(supernate) CE (95)

This design query consists of 12 features. The queried feature 'flowrate(X)' represents the value of flowrate is unknown. While for searching relevant datapoints, each of other 11 features requires a criterion to constrain the search, but the 'flowrate' feature does not need any criterion. For the 6 terminological features, i.e. 'strain', 'feed', 'product', 'scale', 'centrifuge', 'product phase', their criteria have been defined in the corresponding ontology. For the other 5 numerical features, the numerical criteria are given in the Table 4.8 that are used for demonstration purpose.

Table 4.8: Numerical criteria of the case study on harvest CHO cells by centrifugation

feature	μ	**numerical range**
temperature(20)	10%	[18, 22]
solid density(1.051)	15%	[0.89335, 1.20865]
liquid density(1)	15%	[0.85, 1.15]
pH(6.95)	10%	[6.255, 7.645]
CE(95)		[95, +∞)

4.6.5.2 Results generated by reasoning functionalities

Given the design query, the relevant datapoints, predicted CE and the value of flowrate can be found by search, prediction and suggestion functionality as explained in section 4.6.2 to 4.6.4. The results generated by three reasoning functionalities are given directly in Table 4.9.

Table 4.9: Results of three reasoning functionalities about the case study on harvest CHO cells by centrifugation

Functionality	**Results**	**Introductions**
Search	8 datapoints	The 8 datapoints were found by using empirical knowledge of centrifugation
Prediction	96.0%	The 96.0% indicates the possible CE achievement of using centrifugation to harvest CHO cell
Suggestion	576 L/h	The flowrate was calculated by harnessing theoretical knowledge of centrifugation

It is noticed that the 8 datapoints found by search functionality were from the ultra scale down experiments and the suggested flowrate was calculated by fundamental equation. How these theoretical and empirical knowledge are harnessed by reasoning functionalities are explained in the following.

4.6.5.3 Use of USD rule for experimental data search

The USD rules established in the centrifugation system is that :

For Data A and B;
IF strain(CHO) & scale(pilot) & centrifuge(Alfa Laval BTPX 305H) & CE(x) ⊂ Data A;
strain(CHO) & scale(laboratory) & shear time(6000) & shear time(20) & CE(y) ⊂ Data B;
THEN x=y.

The design query includes the features of 'strain(CHO), scale(pilot) and centrifuge(Alfa Laval BTPX 305H)'. Therefore, the USD rule indicates the CE performance of the design query is equal to the CE in the datapoints which have the specifications of 'strain(CHO), scale(laboratory), shear time(20) and shear speed(6000)'. Hence, these datapoints should be searched and used for solving the design query. For explanation, the datapoint found by the USD rule will be called USD datapoint.

In order to find the USD datapoints, the four specifications, i.e. 'strain(CHO), scale(laboratory), shear time(20) and shear speed(6000)', were used to formalize a new design query to constrain the search functionality. This new design query kept the input and output features from the original design query in Table 4.7, but it replaced the feature of 'scale(pilot)' with 'scale(laboratory)' and deleted the feature of 'centrifuge(Alfa Laval BTPX 305H)'. The new design query is shown in Table 4.10, called USD design query.

Table 4.10: USD design query of case study on CHO cell harvest by centrifugation

Input	**Step**	**Output**
strain(CHO) feed(whole cell) temperature(20) product(IgG) solid density(1.051) liquid density(1) pH(6.95)	scale(laboratory) flowrate(X)	product phase(supernate) CE (95)

It is deliberately not to include the features of shear speed and shear time in the USD design query. This is because of the centrifugation system will search all of the relevant datapoints based on the USD design query, and these relevant datapoints may include different specifications of shear time and shear speed, then the system will use the USD rules to pick up the desired USD datapoints automatically. In this case, the centrifugation system retrieved the datapoints including 'shear time(20)' and 'shear speed(6000)'.

Meanwhile, the initial design query is also processed by search functionality to find the relevant datapoints. Thus the datapoints found by search functionality includes USD datapoints and the datapoints related to the initial design query. For this specific design query, 8 datapoints were returned in all, and all of them were USD datapoints. The pseudo code of using USD rule to search datapoints for this case study is given in Figure 4.13

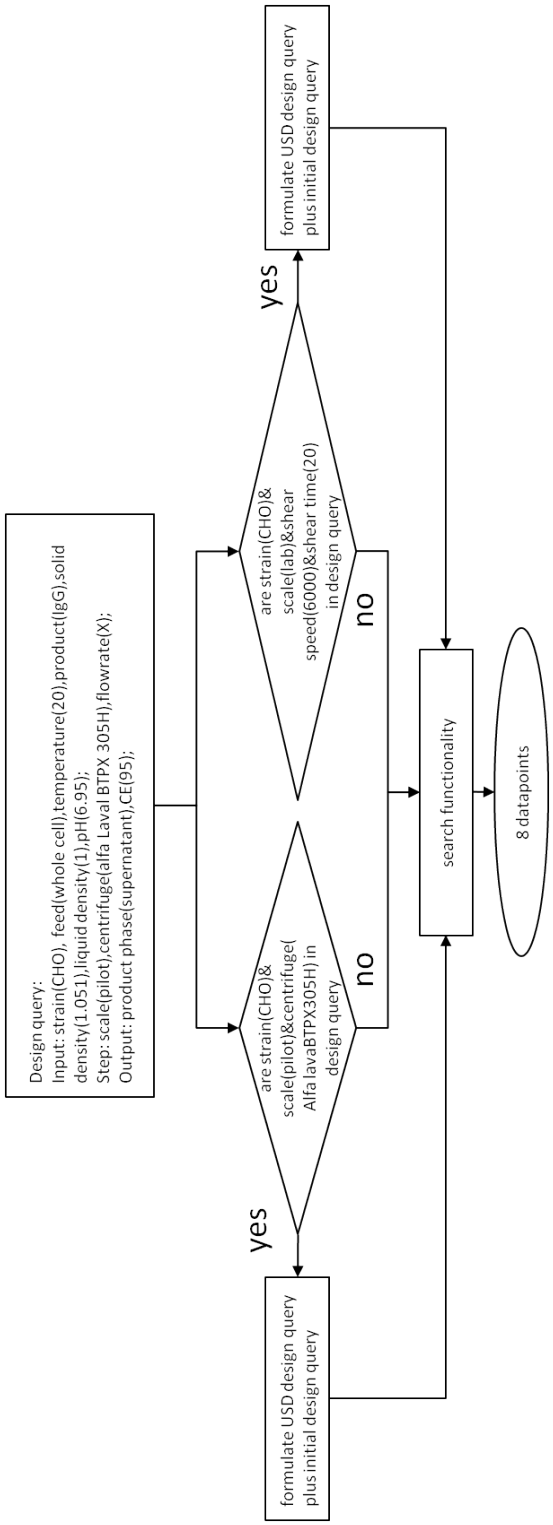


Figure 4.13: Pseudo code of using USD rule to find 8 relevant USD datapoints

This captured USD rule can also be used for the other situation. For example, if the features, 'strain(CHO), scale(laboratory), shear time(20) and shear speed(6000)', were included

in the design query, then the centrifugation system would use the 'strain(CHO), scale(pilot) and centrifuge(Alfa Laval BTPX 305H)' to formalize the USD design query to find the corresponding USD datapoints.

4.6.5.4 Use of fundamental equation and ERM for solution

The flowrate was requested by the design query, thus the suggestion functionality would retrieve the flowrate from the datapoint whose CE value is most similar with the predicted CE. However, the 8 datapoints were generated by USD approach which were done by bench top centrifuge at batch mode, therefore, the flowrate information was not available to retrieve.

Each of the 8 datapoints had the separator capacity that was usually measured in the USD approach, and the separator capacity can be used to calculate the flowrate with the settling area (Σ), adjust constant (C) (see equation (4.9)). In this case, the centrifuge Alfa Laval BTPX 305H was specified, thus the information, Σ(11215) and C(0.4), represented by its ERM would be available for flowrate calculation. Based on the predicted CE 96%, the separator capacity, $3.64 \times 10^{-8} m/s$, was retrieved from datapoint whose CE was 95.8 %. Then, the flowrate to the centrifugation design was calculated as 576 L/h. The pseudo code of flowrate calculation by harnessing the ERM is given in Figure 4.14.

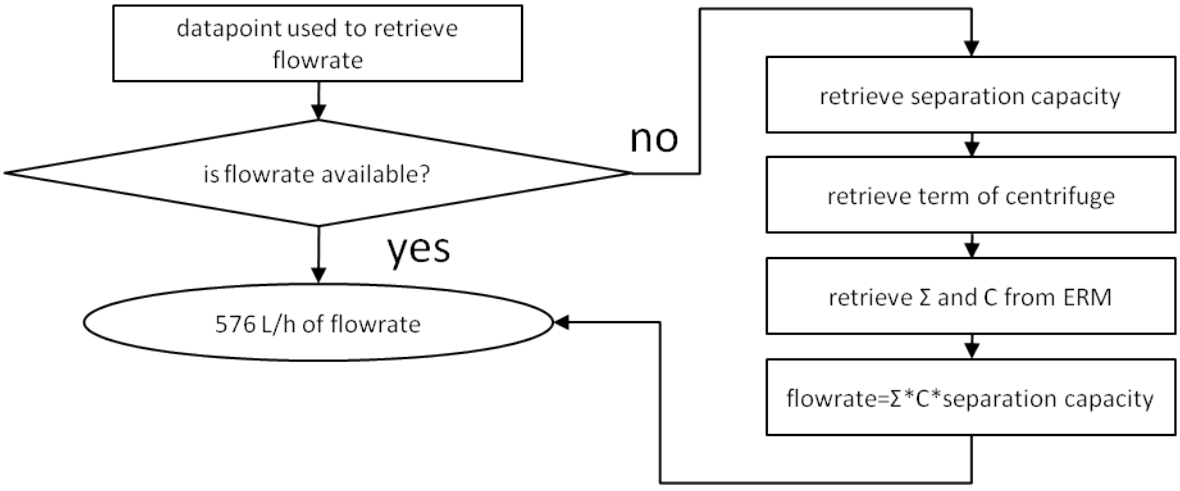


Figure 4.14: Pseudo code of generation of flowrate by harnessing the knowledge

The maximal flowrate represented by ERM can also be used to validate the flowrate

solution. For instance, in this case, the maximal flowrate of Alfa Laval BTPX 305H was 3000L/h, so 576 L/h was a feasible solution.

4.7 A FLOWCHART OF FLOWRATE GENERATION

Identifying the flowrate to the desired CE performance was demonstrated in these two cases discussed above. Base on these, a flowchart to generate flowrate solution regarding the desired CE performance in centrifugation system is summarized in Figure 4.15.

In this flowchart, the results generated by the centrifugation system is indicated by the ellipse, and the logic of this flowchart can be described as the following five steps.

1. Formalizing the design information as the design query which should include the feature about strain and centrifuge, the queried flowrate as well as the minimal CE requirement. The numerical criteria are also required to be defined.

2. Given the design query, the centrifugation system examines whether the USD rules are available to the design query or not.

   (a) If no, the design query proceeds to search functionality to find relevant datapoints.

   (b) If yes, the USD rule is used to formalize the USD design query in order to find corresponding USD datapoints, then both the initial design query and USD design query proceed to search functionality.

3. The search functionality uses the numerical and terminological criteria to find the relevant datapoints, which would include the USD datapoints if the USD rule was available to the design query.

4. The prediction functionality aggregates the CE value from relevant datapoints and produce the predicted CE.

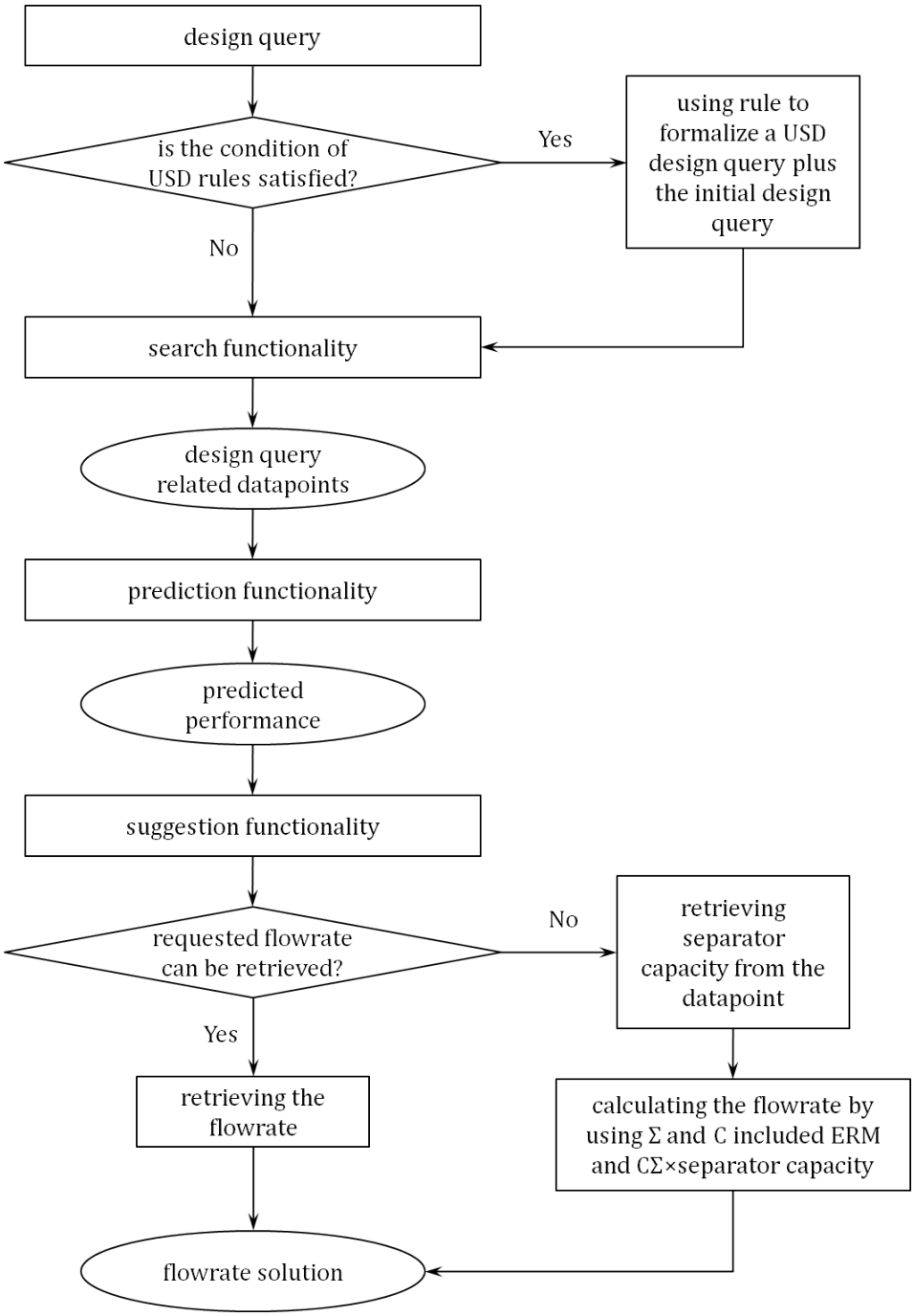


Figure 4.15: Flowchart of generating flowrate solution regarding the performance requirement in the centrifugation system

5. The suggestion functionality locates the datapoint whose CE is most similar to the predicted CE, and then examine whether the flowrate information can be retrieved from this datapoint or not.

   (a) If yes, the suggestion functionality retrieves the flowrate information from the datapoint.

   (b) If no, the suggestion functionality retrieves the separator capacity from the datapoint and

      i. Obtain the adjust constant (C) and settling area (Σ) information from the ERM model about the centrifuge specified in the design query.

      ii. Calculate the flowrate by equation (4.9)

This flowchart illustrates how the centrifugation system harness the four types of data and knowledge to generate solutions to the queried variable. Following this way, any condition of centrifugation input or step can be identified by centrifugation system.

4.8 EVALUATIONS OF CENTRIFUGATION SYSTEM

After the case study, it is important to know how accurate the results generated by centrifugation system are, and what benefits can be achieved by the solutions from the centrifugation system. For this, the evaluations of the centrifugation system will be carried out in this section. In order to assess the performance of the centrifugation system, evaluations were undertaken for prediction and suggestion functionality. The two functionalities are based on the results of search functionality, therefore if the evaluation results of prediction and suggestion were satisfactory, they would indirectly validate the search functionality.

To ensure a fair evaluation, a set of new centrifugation experimental data were captured from journal papers (Hutchinson et al., 2006; Zaman et al., 2009). All of these centrifugation experimental data was about mammalian cell harvest, because the mammalian cell is effective working horse used for the therapeutic protein production (Walsh, 2010). Since the

mammalian cells are sensitive to shear environment, the interaction between the shear force and the mammalian cells harvest is highly concerned in the centrifugation studies. Thus, this experimental data is easily captured from the recent journal papers. 35 batches of centrifugation experiments, each of which included a set of experiments, so total 152 experiments were captured. These experiments studied how the different shear forces at low, medium and high levels impact the separation performance. The rotation disk device was used to simulate the shear environment in these experiments, so the USD rules can be applied for these experimental data.

Each of the 152 individual experimental data is called an *evaluation datapoint*, and each batch of centrifugation experiment is called an *evaluation dataset*. Figure 4.16 gives an example of evaluation dataset that consisting of five evaluation datapoints that were used to study the interaction between the separator capacity and CE performance with the same processing material. Each evaluation datapoint has a specific CE value that will be used to assess the accuracy of predicted CE. The evaluation datapoints included in the evaluation dataset provides a range of separator capacity which will be used to evaluate the results generated by suggestion functionality. In this example, the range of separator capacity is 5.6×10^{-9} m/s to 1.5×10^{-7}m/s .

Figure 4.16: An evaluation dataset including five evaluation datapoints that were used to investigate the interaction between the separator capacity and CE performance with the processing material presheared at 10500 RPM. Source: Zaman et al.,2009. The figure restricts access and has been removed.

The evaluation datapoint, unlike the experimental data captured from the post doctoral researchers, has less information about input and step due to the different research objectives. Hence, the information included in each evaluation dataset was incomplete, e.g. some datasets included the information about the viscosity of processing material and others did not. For the centrifugation system, it is flexible to decide how much information should be employed to formalize the design query. The more information the user has, the more features the design query have. The more features included in the design query mean that more constraints are used to search relevant datapoints, therefore the predicted and suggested re-

sults are closer to the real situations. In order to compare the evaluation results, the same information of these evaluation datapoints was used, i.e. the information of strain, feed, separator capacity, scale, shear speed and CE. In the following sections, the evaluations for prediction and suggestion functionality are discussed separately.

4.8.1 Prediction evaluation

The evaluation of prediction functionality is to examine how accurate the predicted performance is as well as how the design query and criterion impact the predicted performance.

4.8.1.1 Methods of prediction evaluation

The CE was considered in the prediction evaluation. Each evaluation datapoint consists of a real CE value while the other information in this evaluation datapoint will be used to form the design query for the centrifugation system to produce the predicted CE value. Hence, equation (4.12) is used to assess the accuracy of the prediction functionality, i.e. how close between the predicted result and the real result.

$$Prediction\ Error = |Predicted\ CE - Real\ CE|, \tag{4.12}$$

where the Predicted CE indicated the CE value generated by prediction functionality, Real CE represented the CE value recorded in the evaluation datapoint.

To illustrate the impact of the design query on prediction functionality, the three scenarios that examine different features and criteria were designed, namely Evaluation A, B and C. As shown in Table 4.11, four features were used to formalize the design query in Evaluation A with a specific low criterion setting of 10% on the separator capacity which was used as the base case. One extra feature, i.e. shear speed, was added to the design query in Evaluation B. By comparing the prediction errors of Evaluation A and B, how much impact of this change in the design query on the prediction results could be examined. To illustrate the influence of criterion setting on the accuracy of prediction result, the separator capacity's search range has been increased from 10% to 50% in Evaluation C while the other features

were kept the same as those in Evaluation B.

Table 4.11: Features used in the design queries and the criterion setting of the three scenarios for the prediction evaluation

Evaluation A:
Input: strain, feed
Step: separator capacity, scale
criterion on separator capacity: 10%
Evaluation B:
Input: strain, feed
Step: separator capacity, scale, shear speed
criterion on separator capacity and shear speed: 10%
Evaluation C:
Input: strain, feed
Step: separator capacity, scale, shear speed
criterion on separator capacity and shear speed: 50%

For each scenario, 152 evaluation datapoints were used to formalize 152 design queries, but not all of these design queries can find relevant datapoints due to limited relevant datapoints included in the centrifugation system database. Therefore, some design queries may not have the predicted CE, and hence no prediction error would be generated. This made the number of prediction error different in each scenario.

4.8.1.2 Results and analysis

The results of the three scenarios are shown in Figure 4.17, where each point indicates a prediction error and the solid line indicates the average of prediction errors.

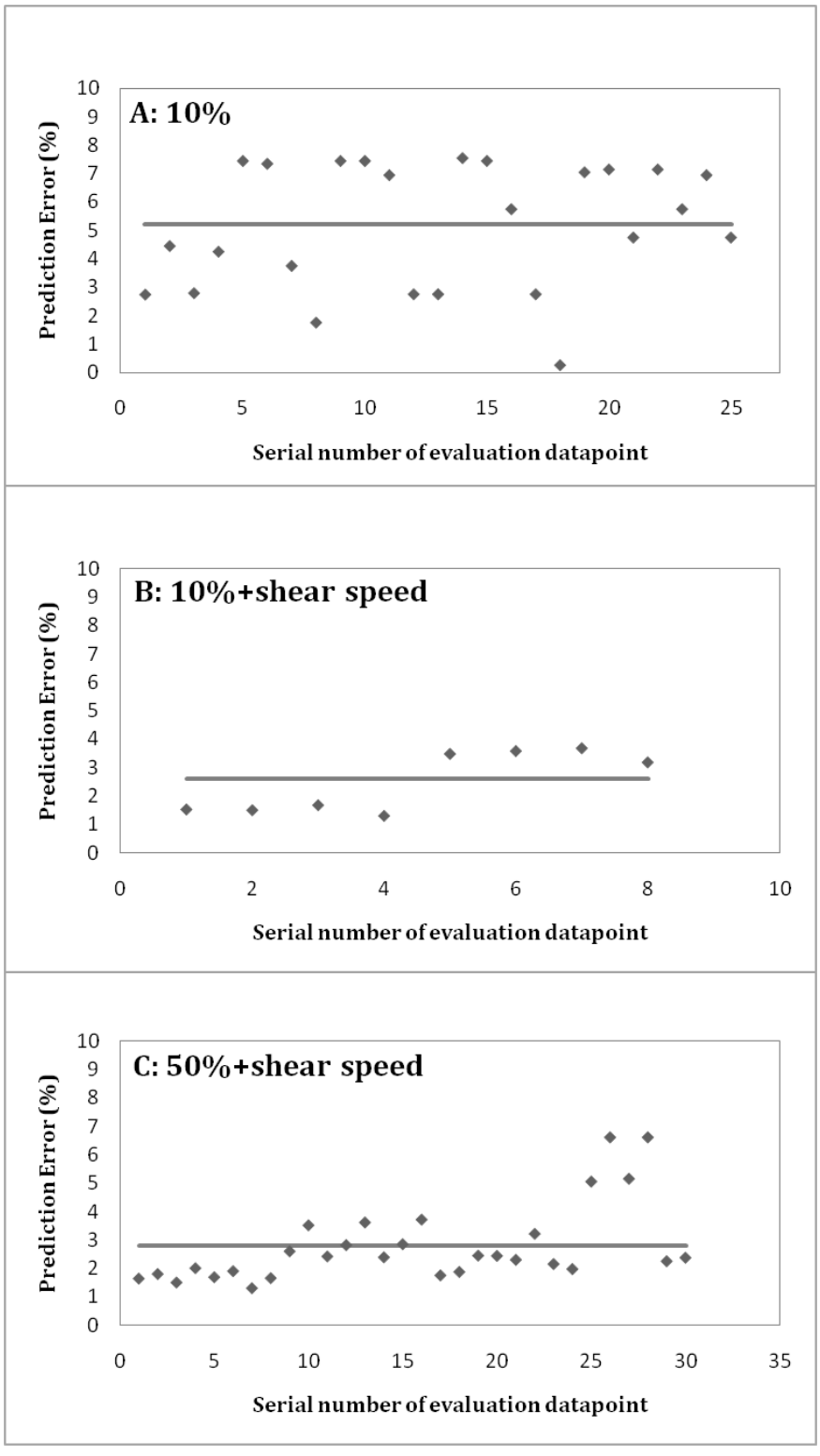


Figure 4.17: Prediction errors of Evaluation A, B and C

For Evaluation A, 25 prediction errors were generated while other 127 evaluation data-

points cannot find any relevant datapoints so that the prediction errors can not be produced. The prediction errors are ranging from 0.25% to 7.5% and average is 5.2%. The 5.2% indicates that if the predicted CE was 90%, then the real CE would be within the range of 90±5.2%.

For Evaluation B, 8 prediction errors were obtained when a further feature of shear speed was added. All of prediction errors are within 5%, and the average prediction error is 2.5%. The feature of shear speed makes the design queries of Evaluation B be more specific so that the searched datapoints were more relevant and fewer relevant datapoints from the database were found. This is why the number of prediction error was less than the Evaluation A. However, this extra feature made the average prediction error have reduced from 5.2% to 2.5%. It is also shown in the Figure 4.17 that the big prediction errors ranging from 5% to 8% in Evaluation A disappeared in Evaluation B, so the accuracy of the predicted CE values were much better than that in Evaluation A. It is expected that the more features in the design query, the more relevant datapoints will be searched. The average prediction error in Evaluation B is less than half of that in Evaluation A.

When the search criterion of separator capacity was relaxed from 10% to 50% in Evaluation C, 30 evaluations were successful. The average prediction error is 2.8%. These relaxation made search functionality find more relevant datapoints, hence the number of prediction error rises. The average prediction error has been risen up from 2.5% to 2.8%. As seen in Figure 4.17, four big errors ranging from 5% to 8% appeared in Evaluation C. This indicates that as criterion is more relax, the difference between the predicted CE and real CE is bigger.

Since the design queries and criteria used in the three scenarios were different, the number of successful evaluations in each scenario was different. In order to examine the three scenarios in details, the eight datapoints which were successful in Evaluation B were considered again as they are successful in Evaluation A and C. The evaluation results of the eight prediction errors generated from three scenarios are shown in Figure 4.18, where the

prediction errors from different scenario were indicated by different colour.

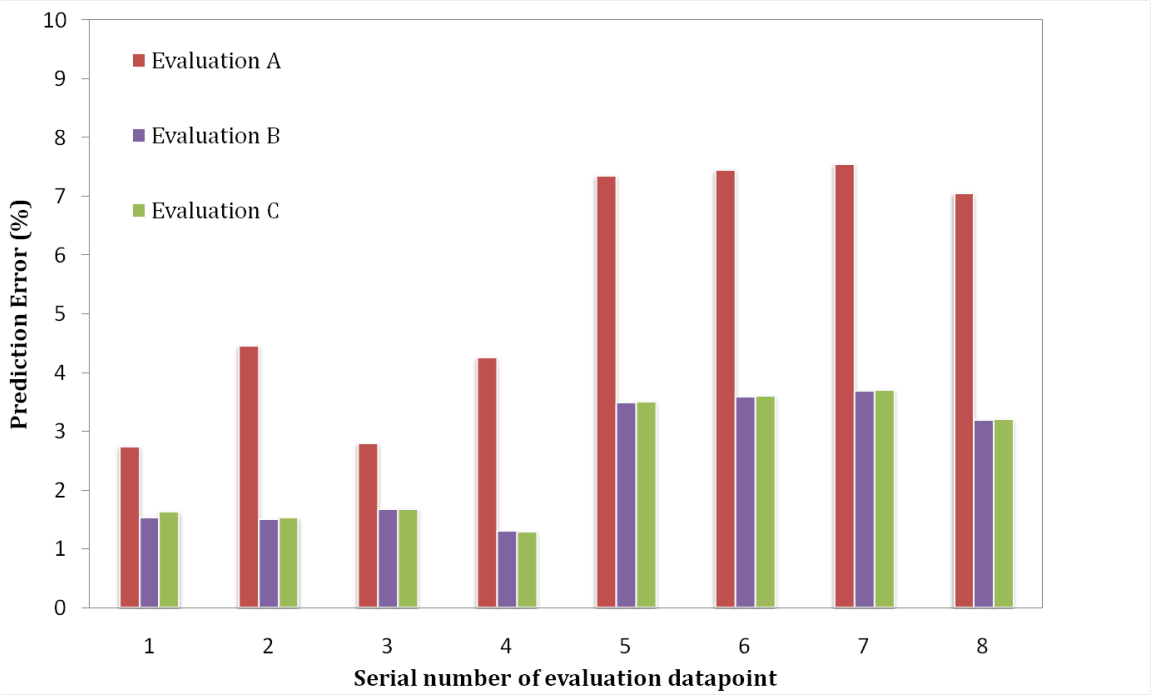


Figure 4.18: Prediction errors generated from the same 8 evaluation datapoints of Evaluation A, B and C

The comparison of eight prediction errors corresponds with the analysis of Figure 4.17. All of the prediction errors generated from the three scenarios are within the 8%, this indicates that the centrifugation system can effectively use the limited information to predict good results with respect to the design problems.

The accurate prediction results could be obtained by using more specific design query and narrow criterion setting. However, for some circumstances, the specific design query and the narrow criterion setting may not get relevant datapoints from the database, therefore user should select the features or criteria flexibly in order to obtain the predicted results. Without running experiments, the fact that the centrifugation system can provide such a good prediction demonstrates the effectiveness and benefit of the centrifugation system.

4.8.2 Suggestion evaluation

Separator capacity is a key design parameter in centrifugation to determine the centrifuge equipment and flowrate (King et al., 2007), so it is employed in the evaluation of suggestion functionality. The suggestion functionality in the centrifugation system retrieves the separator capacity from a relevant datapoint whose CE is the closest to the predicted CE. The retrieved separator capacity is a potential solution to achieve the predicted CE. Such a suggestion can be used to narrow down the design space for the further experimentation in order to realize the predicted CE.

4.8.2.1 Methods of suggestion evaluation

The suggestion result depends on the relevant datapoints found by search functionality. Using more features to formalize the design query allows more relevant datapoints to be found, hence the suggestion result would be closer to the real situation. However, it is difficult to know how the criterion setting will impact the results of suggestion functionality. Therefore, the impact of the criterion setting on the suggestion functionality measured by the reduction of the design space will be evaluated.

For this, two scenarios were developed, namely Evaluation D and E. Both of the scenarios used the design query that contained the features of strain, feed, shear speed and CE. In Evaluation D, μ=50% was used as the criterion setting for searching relevant datapoints, while μ=10% was used in Evaluation E. The information of design queries and criteria are shown in Table 4.12.

Table 4.12: Features used in the design query and the criteria of the two scenarios for the suggestion evaluation

Evaluation D:
Input: strain, feed
Step: scale, shear speed
Output: CE
criterion on shear speed: 50%
Evaluation E:
Input: strain, feed
Step: scale, shear speed
Output: CE
criterion on shear speed: 10%

Given the design query formalized by each evaluation datapoint, the suggestion functionality will retrieve a separator capacity. The error of retrieved rate (ERR) referred in the equation (4.13) is used to assess the retrieved separator capacity.

$$ERR = \frac{|suggested\frac{Q}{c\Sigma} - real\frac{Q}{c\Sigma}|}{feasible\ range}, \tag{4.13}$$

where the $suggested\frac{Q}{c\Sigma}$ is the separator capacity retrieved from the datapoint, $real\frac{Q}{c\Sigma}$ is the separator capacity included in the evaluation datapoint, the feasible range is the width of the separator capacity range included in each batch of datapoints .

The ERR represents the proportion of the distance between the real and suggested separator capacity to the width of the feasible range. For instance, if the ERR was 20%, it means the distance between the real and suggested separator capacity takes 20% of the width of the feasible range of separator capacity. The ERR indicates the quality of the suggestion, the less the ERR is, the more accurate the suggestion would be.

Each batch of centrifugation experiments had a specific range of separator capacity, thus 35 batches of centrifugation experiments had 35 feasible ranges. For this evaluation, 30

feasible ranges were used, because each of other 5 batches only contained one evaluation datapoint that can not be regarded as a range.

4.8.2.2 Results and analysis

The ERR results of Evaluation D and E are shown in Figure 4.19, where each point represents a ERR and solid line is the average ERR.

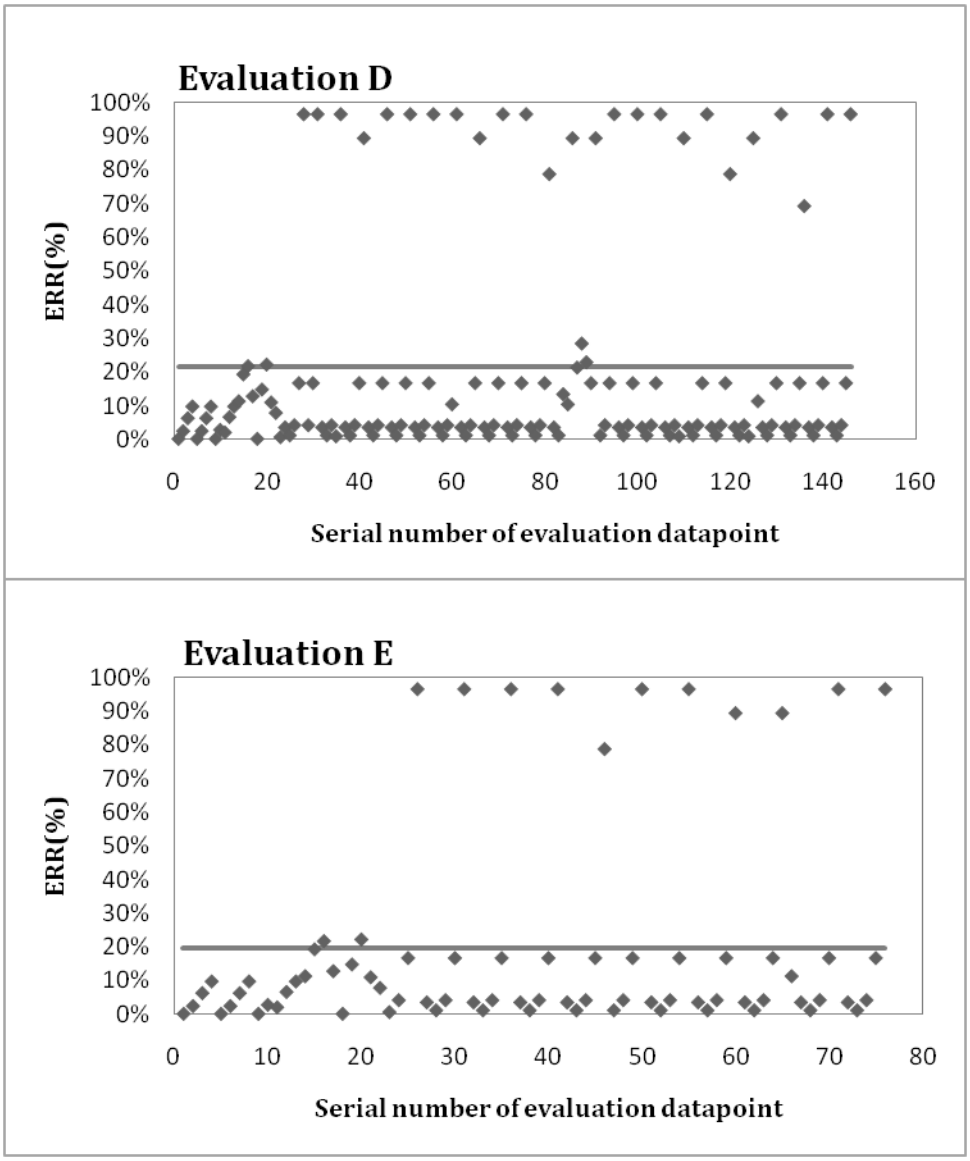


Figure 4.19: ERR Results of Evaluation D and E

For the Evaluation D, 147 ERRs were generated ranging from 0.23% to 96.8%. The average ERR is 21.57%. For Evaluation E, 76 ERRs were generated, the average ERR is

19.71%. The average ERRs of two scenarios are around 20% means the distance between the suggested and real separator capacity takes 20% of the whole separator capacity range. All of the results generated by using five features indicates the suggestion functionality can produce the good suggestions based on limited information.

In both of scenarios, several obvious ERRs were presented. For these ERRs, all of the separator capacity extracted from the evaluation datapoints were around $1.5 \times 10^{-7} m/s$, and their CE performance were from 86% to 90%. But in the database of centrifugation system, the experimental data generated by corresponding USD approach whose CE performance ranged from 86% to 90% had the separator capacity around $2.0 \times 10^{-8} m/s$, and these experimental data was consistent with results generated in the other study (Hutchinson et al., 2006). Thus, the conflicts exist between the evaluation datapoints and the experimental data involved in the database which would be the reason to these obvious ERRs. However, these ERRs only take small proportion (25 out of 147 for Evaluation D and 11 out of 76 for Evaluation E), which would not influence the conclusion that the suggestion functionality has capacity to produce good suggestions.

The reduction of the number of ERRs in the two scenarios illustrates that using the narrow criterion setting may not obtain the suggestion due to no relevant datapoint can be found. The same ERRs presented in both scenarios indicates the narrow criterion setting may not improve the suggestion results. According to these observations, it is suggested that user should select the criterion setting flexibly in order to obtain the useful information for the design problem. For example, if no suggestion results were generated, then the criterion setting could be widened in order to get the information from the previous experimental data.

If the database of centrifugation system was populated with more experimental data, the sensitivity analysis about the impact of the number of features in the design query and the criterion setting on the results generated by centrifugation system can be further examined. This would help to generate the conclusion of impact quantitatively, however, due to the limited time, this work will not be discussed in this thesis.

4.9 CONCLUSIONS

In this chapter, the centrifugation system was developed to illustrate how to apply the BDKF approach for a real bioprocess step. The experimental data representation illustrates a way to reproduce the centrifugation experimental facts as a set of specifications about input, step and output. The ontologies representation can organize the terms within a specific domain as a hierarchy that can be used by search functionality. The representation of theoretical and empirical knowledge allows the fundamental equations, background information of centrifuge equipments and USD rules to be systematically formalized and utilized by searching experimental data and producing solutions. The three reasoning functionalities demonstrate a logic way to reuse the centrifugation experimental data and knowledge to solve the new design problem. The search functionality allows the design query related datapoints to be found, the prediction functionality generates predicted performance and the suggestion functionality retrieves requested information. A general flowchart of using centrifugation system to generate flowrate solution has been developed based on two case studies which shows how to use the BDKF approach to solve a type of simple bioprocess design problem, i.e. identify one variable with respect to the desired performance.

The evaluations about the prediction and suggestion functionality demonstrate that the centrifugation system can make good prediction about CE performance and generate useful suggestions about the separator capacity. This indicates the BDKF approach is an effective approach that allows users to gain useful information from the previous experimental data with limited design information.The evaluation results also prove that using more specific design query with narrower criteria, more accurate predicted performance with more useful suggestion can be obtained. However, it is unable to give the defined suggestion about how much information should be involved in the design query as well as how narrow the criterion setting should be used, because the quality of prediction and suggestion functionality results depends on the experimental data involved in the database of centrifugation system. As a general guideline, it is suggested that user could obtain the accurate solutions by using more specific design query and more narrow criterion setting. But if there were no results

generated, then user may reduce the number of the features included in the design query or wider criterion setting in order to obtain the results. This would be applicable to other BDKF systems, because each of BDKF system uses the same reasoning functionalities to harness the experimental data and knowledge to generate solutions.

Chapter 5

DEVELOPMENT OF BDKF APPROACH TO SOLVE GENERAL DESIGN PROBLEM: CHROMATOGRAPHY CASE STUDY

5.1 INTRODUCTION

Chromatography is an essential purification step in downstream processing to achieve the high purity required for manufacturing the target molecules. The operation is complicated and involving its equipment and consumables that represents large proportion of the cost of goods (Doran, 1995). The issue of chromatographic process design is the intrinsic complexity due to the various interacting variables which may need copious standard experimentation to explore the whole feasible space. These experiments have accumulated substantial experimental data and knowledge. In order to reduce the number of experimentation required by the chromatographic process design, the chromatography system has been developed in WinProlog (LPA,UK).

Unlike the centrifugation system that focuses on generating solution to one variable, the chromatography system aims at providing solutions to the multiple interrelated variables. For this, the content in this chapter is expanded along the following steps. First, the typical procedure of chromatographic process design is introduced in section 5.1 which would help user to understand the challenges of the chromatographic process design. Second, the fundamental theories about the chromatography are described in section 5.2, such as the general resin types, column operation and adsorption theories. Third, the representations of chromatography experimental data and knowledge are explained in section 5.3, 5.4 and 5.5. Fourth, an approach called hierarchical heuristic approach (HHA) is introduced in section 5.7 to allow the chromatography system to generate the solutions to the internal related variables. Case studies discussed in section 5.8 are used to illustrate how the HHA works with search, prediction and suggestion functionality and how it can be used for bioprocess design problems.

5.2 INTRODUCTION ON CHROMATOGRAPHY

Before discussing how to build up the chromatography system, it is necessary to give a basic introduction about the chromatography. Two common concepts will be used to explain the isolation that occurs in the chromatography, namely *mobile phase* and *stationary phase*.

- *Mobile phase* is the fluid carrying molecules (solute) through the column or for elution.
- *Stationary phase* is the resin that stays inside the column and impacts the separation.

5.2.1 Chromatography roles in purification

The processing material contains substantial substances, therefore it is difficult to extract the bioproduct from the processing material with sufficient purity and quantity by using one-step chromatography. Generally, the bioprocess sequence may include two or more units of chromatography, and these units may perform three roles of bioproduct purification, i.e. *capture*, *intermediate purification* and *polishing*.

- The objectives of *capture* are to isolate and concentrate the bioproducts from the processing material and remove the harmful contaminants.
- The *intermediate purification* is to remove the most of impurities, e.g. nucleic acids, viruses. If the capture was efficient, then the intermediate purification could be replaced by more polishing steps.
- The *polishing* is to remove the remaining impurities with trace amounts, since most impurities have been removed by capture and intermediate purification. After the polishing step, the bioproduct will be formulated to suitable conditions for storage or packaging.

The three roles do not mean only three chromatography units will be used in a bioprocess sequence, the number of units used in the specific bioprocess sequence depends on the purity requirements.

5.2.2 Resin types used in chromatography

Chromatography has been employed successfully for purification of therapeutics and pharmaceuticals, such as protein, peptides, amino acids, nucleic acids, alkaloids, vitamins, steroids and other biological materials. Different molecule consists of specific physical properties, e.g. negative/positive charge, functional group. These properties determine the different protocols of purification and specific resins to be developed in order to isolate these molecules effectively. Four types of resin are mainly concerned for the chromatographic process design, namely ion exchange, affinity, hydrophobic and multimode (Fischer, 2011). Table 5.1 gives a summary of the molecule properties used in the different resins. These four types of rein are considered as the candidates of resin selection in the chromatography system [1].

[1]The size exclusion resin is not considered here because it is difficult to find suitable experimental data from the recent papers.

Table 5.1: Molecule properties and the purification techniques

Molecule properties	**Chromatography techniques**
Electrostatic charge	Ion exchange resin
Biospecific or nonbiospecific ligand recognition	Affinity resin
Hydrophobicity	Hydrophobic interaction resin
Combination of two properties above	Multimode resin

5.2.2.1 Ion exchange chromatography

The ion exchange chromatography (IEX) separates the molecules based on electrostatic attraction between the molecule and dense clusters of charged groups on the resin. The IEX is one of the most frequently used techniques for purification of proteins, peptides, nucleic acides and other charged molecules (Anand et al., 2001). The molecules can be separated by performing different degrees of interaction with the charged IEX resins based on the differences in their overall charge, charge density and surface charge distribution (Zhou et al., 2007). The molecule net surface charge is pH dependent and their ionizable groups can be titrated (Staby et al., 2000). For example, each protein has the specific relationship between the net charge and pH that is called titration curve. This curve indicates how the overall net charge of this protein changes according to the pH change and the isoelectric point (pI) that is the pH at which the protein carries no net surface charge (Marcus and Sengupta, 2001). At a pH equivalent to pI, the protein will not interact with a charged resin, however, at a pH above the pI, the protein will bind to anion resin or at a pH below its pI, a protein will bind to cation resin. The cationic and anionic resins are the general types of ion exchange resins. In the ion exchange separation, the reversible interactions between charged molecule and oppositely charged resin are controlled in order to bind or elute specific molecules by changing pH or the ionic strength in the mobile phase (Snyder et al., 2010).

5.2.2.2 Affinity chromatography

The affinity chromatography (AC) separates molecules on the basis of reversible interaction between the specific molecule, e.g. proteins or a group of proteins, and a specific

ligand attached to a insoluble support (Janson, 2011). This technique offers high selectivity with high resolution for separating molecules, like proteins, enzymes, hormones, antibodies, antigens, whole cells. Thus, the purification that would otherwise be time consuming or difficult using other purification techniques can be achieved by affinity easily. In AC, the target molecule is reversibly bound by a ligand. Unbound molecules will be washed out of the column. The bound molecule is recovered by changing the conditions. The biological interactions between ligand and target molecule are a result of electrostatic or hydrophobic interactions, van der Waaals's froces and hydrogen bonding (Hober et al., 2007). To elute the target molecule from the resin, the biological interaction can be reversed by changing the pH, ionic strength or polarity. Currently, two types of AC are frequent used for antibody purification, namely protein A and protein G. These two ligands can bind to the fragment crystallized region (Fc region) which is the constant part of antibody. Therefore, these two types of resins are employed to isolate the low concentration antibodies from the other contaminates.

5.2.2.3 Hydrophobic interaction chromatography

The hydrophobic interaction chromatography (HIC) separates molecules by using the reversible interaction between the molecules and the hydrophobic surface of the resin, and it is widely used in protein purification as a complement to other techniques that separate based on charge or size (Janson, 2011). In hydrophobic interaction chromatography, the target molecules are retained by a weak hydrophobic force at high salt concentration, and the bound molecules are eluted with a descending salt gradient. This technique has grown increasingly popular in the last decade for analysis and purification of proteins, and it can be used for capture, intermediate purification and polishing steps in a purification protocol (Kramarczyk et al., 2008). There is no universally accepted theory on the mechanisms for hydrophobic interaction, although several theories have been developed (Queiroz et al., 2001). The interaction between the molecule and the hydrophobic resin is impacted significantly by salt concentration in running buffer. A high salt concentration would enhance the interaction and the low salt concentration would weaken this interaction. When the salt concentration changes, the interaction will be reversed and the molecule with the lowest degree

of hydrophobicity is eluted first and the most hydrophobic molecule comes out last. The hydrophobic interaction can be determined by various operation variables, such as the salt type, salt concentration, gradient steepness, pH, temperature. (Szepesy and Rippel, 1992).

5.2.2.4 Multimode chromatography

The multimode chromatography (multimode) consists of more than one form of interactions between the resin and molecules (Zhao et al., 2009). There are existing situations that one separation mode can not completely resolve the mixture. In this case, the multimode may provide desired performance of such purification. For instance, the weak mixed-mode cation resin, Capto MMC, contains a cross-linked agarose based resin that tolerates much higher flowrate used with conventional agarose resin. The ligand includes an aliphatic backbone with number of carbon atoms being substituted by oxygen, sulphur and nitrogen atoms and the groups of hydroxyl, carbolic acid and aromatic groups. Therefore, depend on the suitable conditions, this resin would preform hydrogen bounding, ion exchange and hydrophobic interaction properties (Oehme and Peters, 2010). All of these interactions can be manipulated by using specific buffer compositions, pH and salt concentrations.

5.2.3 Column operation

The selected resin is packed into a column as a packed bed with a fixed height (stationary phase), then the mobile phase is forced by the pressure to flow through the stationary phase where the bioproduct is isolated from the other contaminates. The column operation is a dynamic process with batch mode. In each batch, the column operation consists of following steps in a sequential order, i.e. *equilibration*, *loading*, *washing*, *elution* and *regeneration*. The column operation data is considered as the chromatography experimental data.

- *Equilibration* is to set up the desired conditions so that the resin filled in the column is ready for adsorption,
- *Loading* is to load liquid solution containing target molecule and contaminates onto the column,

- *Washing* is to remove residual unabsorbed material after loading phase,
- *Elution* desorbates the molecule from the resin with a specific buffer, generally, the bound molecule are eluted by changing the ionic strength of buffer (Jungbauer, 2005), or occasionally by changing the pH (Dai et al., 2005),
- *Regeneration* makes the adsorption resin back to the original conditions for next adsorption.

For some circumstances, the washing step is required to remove the tightly bound molecule after the elution step. In manufacturing, the sterilization step is required to eliminate the virus or other micro-organisms remained in the column. In each of five steps, the specific buffer is used to manipulate the adsorption or desorption between the molecules and resins in order to isolate the target molecule from other contaminates. Adsorption usually has high selectivity but small capacity (Vijayalakshmi, 2002) which is realized by physical or ionic forces. The desorption is opposite process of adsorption, thus the buffer conditions of desorption should overcome these forces that make the molecule unbind from the resin. This is usually realized by feeding a new buffer or changing the current buffer conditions.

Both of the adsorption and desorption are the primary phenomena of chromatography. In the following, the fundamental theories about the adsorption is introduced, these information would serve as the knowledge for chromatography experimental data utilization, e.g. which variables should be considered and captured to represent the experimental data.

5.2.4 Equilibrium of adsorption

Adsorption analysis may start with the equilibrium relationships that determine the extent to which materials can be adsorbed onto a particular adsorbent surface. As the adsorbate and adsorbent are at equilibrium, there is a defined distribution of solute between solid, i.e. resin and fluid phases.

The adsorption equilibrium can be represented as adsorption isotherms. Two types of

isotherms are employed to explain the chromatography adsorption, namely *Freundlich isotherm* and *Langmuir isotherm* (Zimmer, 2003).

- *Freundlich isotherm* has been used to explain the adsorption of antibiotics, steroids and hormones. The mathematical relationship is given as $q^* = K_d C^{*n}$, where q^* is the amount of solute adsorbed per unit of adsorbent, C^* is the solute concentration in solute, K_d is the equilibrium dissociation constant, n is a constant.

- *Langmuir isotherm* is used to describe the adsorption of proteins. The isotherm is represented as $q^* = (Q_m C^*)/(K_d + C^*)$, where Q_m is the maximum amount of solute that can be adsorbed per unit of adsorbent.

The adsorption isotherms for specific resins have been also studied, e.g. IEX (Cornelissen et al., 2008), AC (Rusch et al., 1997), HIC (Queiroz et al., 2001). Over all of studies of adsorption isotherms, the molecule concentration and the properties of resin, e.g. pore size, are considered as the variables that can determine the mass of molecule bound by resin when the equilibrium has been achieved.

5.2.5 Performance of adsorption and elution

The fixed bed is the common chromatography equipment form. When the feed material is loaded to the column at certain flowrate, target product and some impurities will be bound to the resin while other impurities will flow through. When the reins reach certain level of saturation, the product breaks through the column. The breakthrough curve is used to indicate when the column is fully saturated. Figure 5.1 gives a theoretical breakthrough curve, '$\frac{C}{C_0}$' ranges from zero to one and represents the fraction of the molecule concentration in the processing material that flow through the column. The areas above and below the breakthrough curve represent the mass of the molecules that are bound (molecule adsorbed) and flow through the column (molecule lost). Excessive loading would give rise to loss of bioproduct, whereas insufficient loading would make the column not fully utilized. For manufacturing, less than 10% breakthrough point is often used depending the value of the product. It means the loading would be suspended when the effluent concentration achieves

10% of the initial concentration of the product.

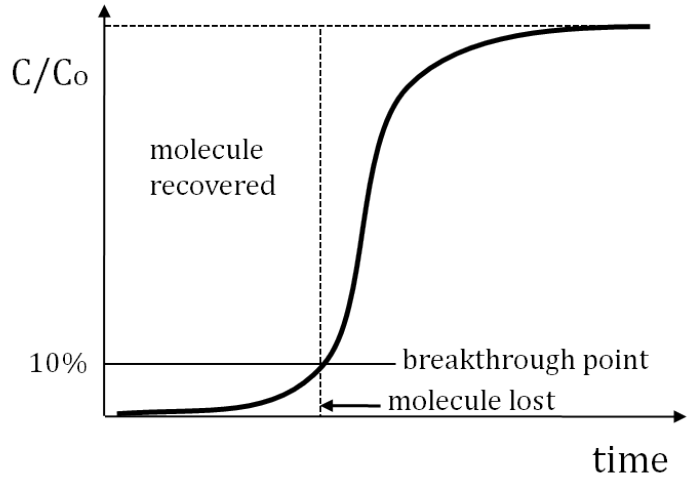


Figure 5.1: Theoretical breakthrough curve and breakthrough point

Given the defined breakthrough point, the performance of adsorption and elution in column is indicated by three variables, i.e. dynamic binding capacity (DBC), yield and purity.

The DBC indicates the quantity of target molecules have been adsorbed by selected resin while the liquid solution continues flowing through the column, usually it is expressed by the mass of bounded molecules per unit of column volume (mg/ml). DBC is an important factor for resin selection (Janson, 2011) and generally calculated by equation (5.1).

$$DBC = \frac{M_{molecule\ bounded}}{V_{bed}}, \tag{5.1}$$

where the $M_{molecule\ bounded}$ represents the mass of target molecule has been captured by column after the loading phase, V_{bed} indicats the volume of resin packed into the column.

Unlike the binding capacity that is determined when the whole bed resins are saturated, the DBC is a function of operating conditions, including the residence time, molecule concentration, wash and elution protocols (Bergander et al., 2008). For instance, the DBC can be increased by either decreasing the flowrate and hence increasing the residence time or by increasing the bed height that still increases the residence time.

Given the specific breakthrough point, the volume of processing material to be loaded onto the column and the mass of target molecule can be calculated. Some target molecules will be lost during the loading phase because the specific breakthrough point indicates partial target molecules that do not have enough time to bind to the resins will flow though the column with the loading flowrate. Some target molecules that have not been tightly bounded may be washed out of the column during the washing phase. The bounded molecules and impurities can be eluted from the column in elution phase. Usually, it is difficult to expect the target molecule and impurities can be eluted at a different instant in time, the overlapping would happen between the molecules next to each other, especially when the molecules structures are similar (Doran, 1995). The yield and purity are used to quantify the target molecules collected from the elution phase.

Yield is used to indicate how much target molecule has been collected from the column operation, usually it is a portion number that is calculated by the equation (5.2).

$$Yield = \frac{M_{molecule\ eluted}}{M_{molecule\ loaded}} \times 100, \tag{5.2}$$

where the $M_{molecule\ eluted}$ is the mass of target molecule that presents in elute, the $M_{molecule\ loaded}$ is the total mass of target molecule loaded to the column.

Purity describe the absence of impurities in collected target molecules, i.e. how much target molecule would present in the collection of the elution. The purity is calculated by equation (5.3).

$$Purity = \frac{M_{molecule\ eluted}}{M_{molecule\ eluted} + M_{impurity\ eluted}} \times 100 \tag{5.3}$$

where the $M_{contaminant\ eluted}$ represents the mass of impurity presents in elute.

The yield and purity depend on the point of starting and stopping the elute collection, namely cutting point. If the setting of the cutting points allows more overlapping can be

collected that indicates more target molecules and impurities would be contained, this leads to a high yield but low purity. Contrarily, if the stetting of cutting points make the less overlapping collected which would produce a low yield and high purity. The trade off between purity and yield has been proposed to be optimized by a graphical method called fractionation diagram (Ngiam et al., 2001).

The DBC is highly concerned at the early stage of chromatography development, because it is used to determine the column operating conditions, e.g. resin type, buffer components, while the yield and purity would be considered as an optimization problem of column operation (Ngiam et al., 2003).

5.2.6 Scale up principles

Same as the centrifugation development, the column operating conditions are screened at laboratory scale, then the selected conditions are scaled up to the desired large scale column for validation, e.g. whether the desired DBC can be reproduced by using the selected buffer, pH, flowrate. The principles of scale up are shown in the following, which have been validated by the practical studies (Al-Jibbouri, 2006; Vijayalakshmi, 2002).

The following variables should be maintained to ensure the conformity in performance between the laboratory scale and large scale:

- residence time of processing material on column,
- maintenance of gradient slope, e.g. gradient volume/resin volume,
- processing material concentration and composition,
- ratio of processing material volume to resin volume.

The scale up is achieved by increasing the following variables:

- column diameter,

- volumetric flowrate, i.e. volumetric flowrate in proportion to column cross-sectional area,
- processing material volume proportionally, i.e. increase the processing material volume in proportion to the resin volume,
- gradient volume proportionally, i.e. gradient volume in proportion to the column cross-sectional area.

Increasing the resin volume is achieved by increasing the column diameter, volumetric flowrate, processing material load volume and gradient volume, and it would ensure the same performance performed at laboratory scale.

5.2.7 Chromatographic process development

The chromatographic process design design usually starts at using laboratory scale column to conduct experiments in order to determine the anticipate the feasible conditions. The conditions that may play the key roles in the chromatographic process design are given in the following.

- the nature of mobile and stationary phases, reins and buffer types
- bed volume and height, column capacity (molecular loads for adsorbing products and contaminates)
- column diameter and mobile phase velocity
- methodology of column elution, washing, regeneration and equilibration associated with liquid volume

Table 5.2 illustrates the typical development procedure that is used for identifying the conditions referred above (Chhatre and TitchenerHooker, 2009).

Table 5.2: Procedure of chromatographic process development

Design steps	**Explanations**
1. Resin selection	For a given model of interaction, this is often chosen from known likely resin candidates that have been shown in the past to achieve the adequate binding capacity for similar bioproduct concentrations and impurity profiles
2. Loading and washing buffer selection	possibilities of buffer components, concentration, ionic strengths and pH values are more often determined by previous experience rather than by a systematic investigation of these variables, these information are usually validated by the laboratory scale experiments
3. Elution buffer selection	the buffer composition, gradient type and length are selected based upon prior knowledge, and using laboratory scale experiments for testing
4. Select column dimensions, loading volume and flowarte	The height of scale-down bed, the liner operating velocity and the feed load are fixed at values expected for larger columns. Relatively sample volumes are required for validation.

5.2.8 Conclusions

In chromatographic process development, the decision making highly relies on the previous experience rather than the systematic experimentation, because the multiple variables included in the chromatographic process design would require thousands of experiments to examine the whole design space which would be difficult due to the limited time and processing material. In addition, the decisions made by experience may not be optimal or have errors that would make wrong solutions for the design task, e.g. if wrong resin is selected, then the conditions identified in the following steps are hardly to achieve the desired perfor-

mance.

The chromatography system is proposed to harness the chromatography experimental data and knowledge to generate solutions with respect to the design problems. Same as the centrifugation system, the chromatography experimental data, ontology, theoretical and empirical knowledge were considered in the chromatography system. A new approach, namely hierarchical heuristic approach, has been established to allow the three reasoning functionalities to harness the four types of data and knowledge to provide solutions to the internal related variables with respect to the specific performance requirement. The potential solutions can narrow down the design space to be explored. Therefore, in the following, the representation of the four types of data and knowledge will be introduced first, then following the explanation of hierarchical heuristic approach and case studies.

5.3 REPRESENTATION OF CHROMATOGRAPHY EXPERIMENTAL DATA

The experimental data to be harnessed by chromatography system was generated from the column operation, called column data. These column data may be kept in different places and in different formats, therefore, in order to reuse it, the data representation is required to formalize these column data structurally.

For the chromatography system, 1021 column data was captured. 941 column data was captured from the journal papers (Do et al., 2008; Hahn et al., 2003, 2005; McCue et al., 2003; Staby et al., 2000; Staby and Jensen, 2001; Staby et al., 2004, 2006, 2007; Swinnen et al., 2007; Verdoliva et al., 2002) and 80 experimental data was kindly provided by Dr. Sunil Chhatre in Biochemical Engineering Department, UCL. This column data covered four types resin, including 456 IEX experimental data, 130 AC experimental data, 355 HIC experimental data and 80 multimode experimental data. The 941 column data captured from the journal paper was implemented on the column volume ranging from 1 to 2 mL. The other

80 experimental data was generated from the microwell chromatography study.

Same as the representation of centrifugation experimental data, the information included in the column data was grouped into three parts, input, step and output. For each part, each item of information was represented as a specification which consisted of parameter, value and unit. The original unit in the data were kept because different researchers may use various units to record the information. In the following sections, the parameters used for representing the information of input, step and output are explained.

5.3.1 Chromatography input

The input includes the information about processing material to be loaded onto the column. The properties of target molecule and impurities in the processing material are concerned, because these would affect the column performance. The 'molecule weight' and 'p*I*' are the two common variables to describe the target molecule properties (Protein purification handbook, GE, Document ID: 28-9833-31) . For impurity properties, the 'total protein concentration' is the quantitative description while the 'impurities' is the qualitative indicator.

The impurities may be categorized as two types, process related impurities and product related impurities. The process related impurities may derive from the production process, e.g. fermentation ingredients, host cell components. The product related impurities may be the variants of the bioproduct that do not have the desired biological activities, e.g. degradation, misfolded isomers of the protein. This type of impurities may be indistinguishable from the target bioproducts, such as the Benzyl Ezetimibe which is a type of degradation Ezetimibe that has an extra phenyl (Gajjar and Shah, 2011). A number of analytical techniques have been established for impurities analysis, e.g. using high performance liquid chromatography (HPLC) to quantify the host cell proteins (process related impurities). These analytical techniques used to identify the impurities will not be discussed in this thesis, but the reference is available for the interested reader (Ahuja, 2000). Therefore, the impurities can be qualitatively identified by using the corresponding analytical techniques. The identified impurity

names would be represented by the parameter 'impurities', and its value would be a string of specific molecule terms. This item of information would tell users what impurities were included in the processing material for this column operation. It may help users to assess whether this specific column data is useful to the current purification design. For instance, if the impurities included in the column data are similar to that contained in the processing material, then the column operating conditions, e.g. elution strategy, buffer compositions, may be the feasible solution to the current design problem.

For the captured 1021 column data, 9 parameters were identified to represent the information included in input, and they are given in Table 5.3.

Table 5.3: Parameters for representation of chromatography input information

Parameter	**Definition**	**Example**
product	name of target bioproduct to be purified	pAb (n/a)
strain	name of cell lines or micro-organism	bovine serum (n/a)
molecule weight	value of the target molecule weight	69 (kDa)
p*I*	the pH value at which the molecule carries no net surface charge	6.5 (n/a)
feed volume	value of feed volume to be fed to the column per batch	5 (cv)
feed viscosity	value of feed viscosity	1.0 (Pa·s)
product concentration	value of target bioproduct concentration of processing material	1 (mg/mL)
total protein concentration	value of total protein concentration of processing material	2 (mg/mL)
impurities	names of impurities included in the processing material	albumin (n/a)

For example, 'The polyclonal antibody is the target molecule' is represented as 'product(pAb, n/a)', where the'n/a' indicates the unit for this parameter is not applicable; 'the protein concentration is 2 mg/ml' is represented as 'product concentration(2, mg/ml)'.

The parameter setting of chromatography input is expandable, if other information is available, then it could be captured and represented by the defined parameter. For example, if the critical impurity can be quantified, the results can be represented by the parameter 'impurity concentration'.

5.3.2 Chromatography step

The step includes the information about column operation. The column operation includes five steps that has been introduced in section 5.2.2. The parameters identified to represent the information in each of five steps are introduced as followings.

5.3.2.1 Information of column and resin

For chromatography experimentation, it is natural to know the column and resin setting that will be used for purification, e.g. the column diameter, column length, resin type, bed height. 12 parameters were identified to represent the information of column and resin contained in the column data, and they are given in Table 5.4.

Table 5.4: Parameters for representation of column and resin setting information

Parameter	**Definition**	**Example**
chromatography function	term of chromatography role in the bioprocessing	capture (n/a)
scale	scale term of chromatography experiment	lab (n/a)
temperature	value of column temperature	20 (°C)
chromatography type	term of resin type used for purification	IEX (n/a)
column volume	value of column volume specified by manufacturer	10 (mL)
column height	value of column height specified by manufacturer	10 (cm)
column diameter	value of inner diameter of column	1 (cm)
column manufacturer	name of column manufacturer	GE (n/a)
column model	term of column defined by manufacturer	BPG (n/a)
resin volume	packed resin volume	1.78 (mL)
resin manufacturer	name of resin manufacturer	GE (n/a)
resin model	term of resin defined by manufacturer	Q Sepharose Fast Flow (n/a)

5.3.2.2 Equilibration step

Equilibration aims at using specific buffer to set up the desired conditions so that the resins filled in the column are ready for binding the target molecules. 6 parameters were used to represent the information of equilibration included in the column data (see Table 5.5).

Table 5.5: Parameters for representation of equilibration step information

Parameter	Definition	Example
equilibration buffer chemical	chemical names of buffer	Tris, Bis-Tris-Propane (n/a)
equilibration buffer pH	pH value of equilibration buffer	8 (n/a)
equilibration buffer chemical concentration	concentration value of buffer chemicals	25,25 (mM)
equilibration volume	value of buffer volume used in equilibration	15 (cv)
equilibration flowrate	value of flowrate applied for equilibration	5 (mL/min)
equilibration back pressure	value of column back pressure in equilibration	1.5 (kPa)

5.3.2.3 Loading step

Loading step is pumping the processing material onto the column to make the molecule be bound by resin. The buffer used in this phase is generally the same as the equilibration buffer, therefore the buffer information need not be included. Table 5.6 gives three parameters that were identified to represent the loading step information.

Table 5.6: Parameters for representation of loading step information

Parameter	Definition	Example
loading flowrate	value of feed loading flowrate	0.6 (mL/min)
loading volume	value of feed volume to be loaded onto column	2 (cv)
loading back pressure	value of column back pressure in loading step	1.5 (kPa)

5.3.2.4 Washing step

The washing step is using specific buffer to wash all unbound molecule away from the column. Table 5.7 lists 6 parameters that were used to represent the washing step information.

Table 5.7: Parameters for representation of washing step information

Parameter	**Definition**	**Example**
wash buffer chemical	name of wash buffer chemicals	sodium chloride (n/a)
wash buffer pH	pH value of wash buffer	7 (n/a)
wash buffer chemical concentration	value of wash buffer chemicals	500 (mM)
wash volume	value of buffer volume used for wash	20 (cv)
wash buffer flowrate	value of flowrate wash flowrate	0.5 (cm/s)
wash back pressure	value of column back pressure in washing step	1.5 (kPa)

5.3.2.5 Elution step

The elution step is using the specific buffer to elute the bound molecules from the resins for collection, and the mechanisms of eluting bound molecule from the different resins have been introduced in section 5.2.2. The bound molecules can be eluted in two ways, namely gradient elution and step elution (Fischer, 2011). In gradient elution, the ionic strength of elution buffer changes linearly with time, and in the step elution, the ionic strength of elution buffer changes stepwise, e.g. 100 mM, 150 mM, 200 mM. Figure 5.2 illustrates the gradient and step elution by NaCl buffer.

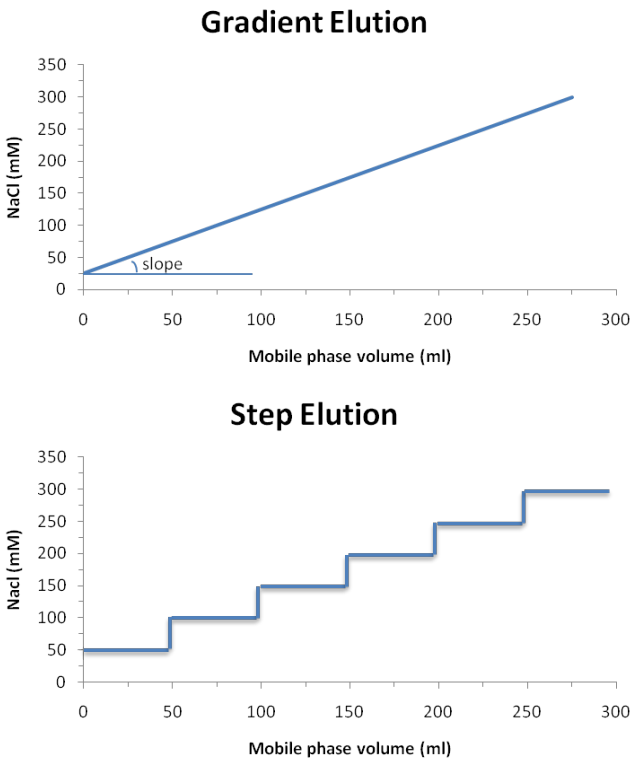


Figure 5.2: Demonstrations of ionic strength changes in gradient elution and step elution

For the gradient elution, the ionic strength is indicated by the salt (NaCl) concentration. The initial concentration is 25 mM, and then it increases linearly as the mobile phase volume increases. The change of ionic strength by gradient elution can be profiled as a line with a specific slope that can be represented by the parameters 'gradient slope'.

For the step elution, the initial salt concentration is 50 mM, and it is kept constantly until the mobile phase volume increases to 50 ml. At that point, the salt concentration is switched to 100 mM. This stepwise switch repeats when the mobile phase volume is increased by 50 mL. The variation of ionic strength in step elution can be profiled as the 'switch number' and 'concentration interval'.

Table 5.8 lists 11 parameters that were used to represent the information about elution step.

Table 5.8: Parameters for representation of elution step information

Parameter	Definition	Example
elution strategy	term of elution approach	gradient/step elution
elution buffer chemical	name of buffer chemicals	sodium chloride (n/a)
elution buffer pH	pH value of elution buffer	7.5 (1)
start buffer concentration	value of start concentration of buffer in gradient elution	0 (mM)
final buffer concentration	value of final concentration of buffer in gradient elution	1000 (mM)
gradient slope	slope value of salt concentration variation line	0.5 (n/a)
switch number	value of elution buffer concentration switch number	2 (n/a)
chemical concentration interval	value of concentration variation between switches	50 (mM)
elution volume	value of buffer volume used for elution	15 (cv)
elution buffer flowrate	value of elution flowrate	1 (ml/min)
elution back pressure	value of column back pressure	1.0 (kPa)

If the 'elution strategy' is gradient, the 'gradient slope' should have the specific value while the 'switch number' and 'chemical concentration interval' should not have values.

5.3.2.6 Regeneration step

The regeneration phase plans to re-equilibrium the column for next run. The buffer employed in this step is usually the same as the buffer used in the equilibration step, but other buffers may be used occasionally. For this, the information about the buffer in regeneration step is represented by 6 parameters shown in Table 5.9.

Table 5.9: Parameters for representation of regeneration step information

Parameter	Definition	Example
regeneration buffer chemical	name of regeneration buffer chemicals	sodium chloride (n/a)
regeneration buffer concentration	value of chemical concentration	1000 (mM)
regeneration buffer pH	value of buffer pH	7 (n/a)
regeneration volume	value of buffer volume used in regeneration	200 (ml)
regeneration buffer flowarte	value of flowrate in regeneration	1.5 (cm/s)
regeneration back pressure	kPa value of column back pressure in regeneration	2.0 (kPa)

In total, 44 parameters were identified to represent the step information of chromatography experimental data.

5.3.3 Chromatography output

The chromatography output includes the column performance, e.g. how much target molecule has been bound to the column. Based on the captured experimental data, four parameters were used to represent the related information, i.e. breakthrough point, yield, dynamic binding capacity (DBC) and purity (see Table 5.10).

Table 5.10: Parameters for representation of chromatography output information

Parameter	**Definition**	**Example**
breakthrough point	value of breakthrough point	10 (%)
yield	ratio of target molecule collected from elution to the mass of target molecule in feed	85 %
DBC	value of bound molecule mass to resin volume	25 (mg/ml)
purity	ratio of target molecule collected from elution to all molecule mass in feed	75 (%)

For the captured experimental data, the DBC, yield and purity were given directly. If they were not available, the equations and definitions about the three parameters introduced in section 5.2.4 could be used for calculation.

58 parameters were identified to represent the input, step and output of the captured column data. Each experimental data was represented as a datapoint that consisted of a subset of 58 specifications. Each datapoint had a unique ID number ranging form 1 to 1021. All of the datapoints were kept in a table, in which each column represented one parameter and each line represented one experimental data. These 1021 datapoints formed the chromatography system database.

5.3.4 Mapping experimental data representation to chromatogram

The chromatogram is a visual output of the column operation, which is a profile of the molecules generally used to determine the cutting points and assess the purification efficiency. To illustrate how to extract the information of the chromatogram and represent it structurally, a demonstration is used and shown in Figure 5.3.

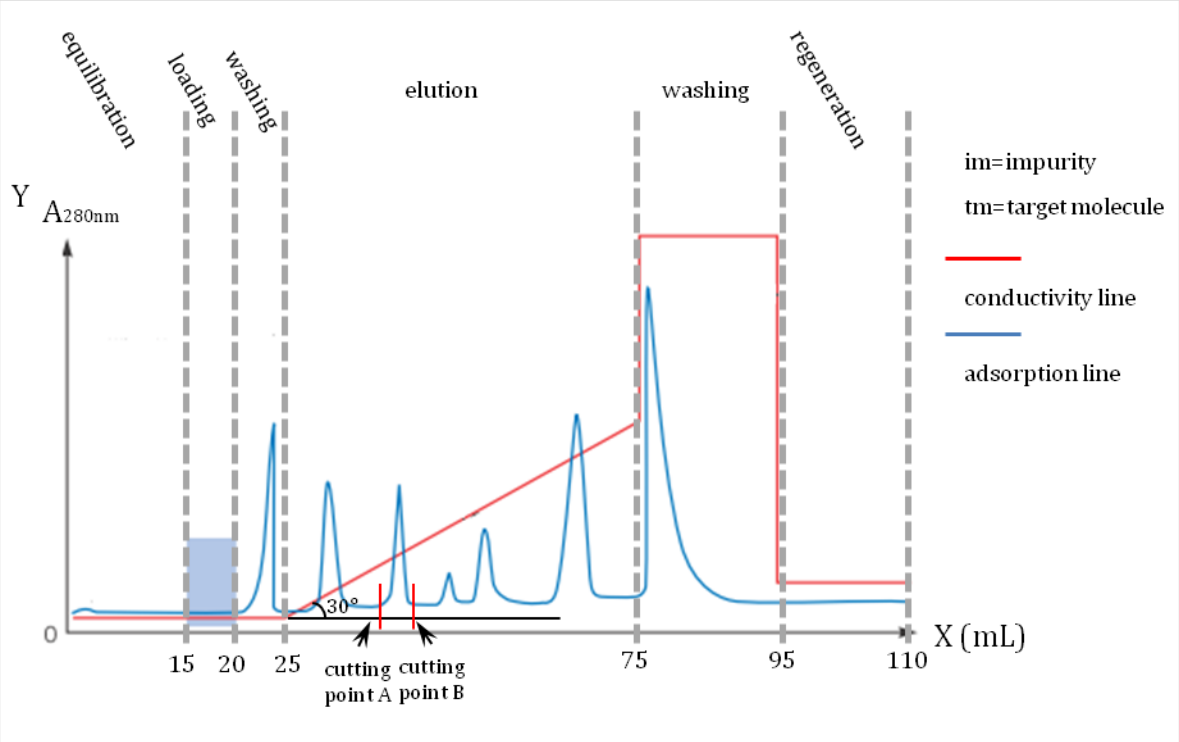


Figure 5.3: Chromatogram of isolating the target molecule from impurities

The chromatogram illustrates the isolation of seven molecules. Y axis is the molecule concentration measured by the spectrophotometer using a wavelength of 280 nm, X axis is cumulative volume of mobile phase that flows out of the column. The adsorption line indicates the molecule concentration eluted out of the column. The six impurities and one target molecule are indicated by seven peaks respectively. The conductivity line indicates the ionic strength of the mobile phase.

This column operation includes six steps, i.e. equilibration, loading, washing, elution, washing and regeneration, which form six areas of the chromatogram. Each area can be represented by the buffer volume used in this step. The ionic strength change in each step can be profiled by the conductivity line, e.g. elution. The cutting points indicate when should collect the mobile phase. In this chromatogram, the information of cutting points are represented by cumulate volume of the mobile phase, i.e. 38 and 42 mL. The distance between the two points is the volume collected from the elution, e.g. 4 mL. The cutting point determines

the yield and purity that has been explained in section 5.2.5. The collected impurities and target molecule can be quantified as total protein concentration and product concentration. Table 5.11 illustrates the representation of information extracted from the chromatogram.

Table 5.11: Representation of extracted information of six step involved in the chromatogram

Operation step	**Information**
Equilibration	equilibration volume(15, mL)
Loading	loading volume(5, mL)
Washing	washing volume(5, mL)
Elution	elution volume(50, mL)
	gradient slope(30, n/a)
	cutting point(38,42, mL)
Washing	washing volume(20, mL)
Regeneration	regeneration volume(15, mL)

The information of operation steps and buffer volume of each step may tell users the operation procedure of this column operation. The information of collecting target molecule from elution may tell users the trade off between the yield and purity. However, for the captured 1022 datapoints, the information of cutting point is not applicable, therefore, the parameter 'cutting point' was not shown in the parameter list (Table 5.8).

In addition to the column data, batch adsorption experimental data can be also used by chromatographic process design. The batch adsorption experiment is mixing the adsorbent and molecule solution and leaving it for an extended time, e.g. 24 hours. During this time, samples will be taken at specific intervals to measure the molecule concentration and the results will illustrate the uptake kinetics of the adsorbent (Mller, 2005). The data produced by these experiments are described the Freundlich or Langmuir isotherm and they are used to study the binding capacity or the equilibrium dissociation constant of the adsorption. It is called 'static' experiment because there is no 'mobile' phase in this procedure, therefore the binding capacity identified by this approach is called static binding capacity. The static

binding capacity generally poorly match with the dynamic binding capacity. For example, the SuperQ 650C which is a type of IEX resin, its statistic binding capacity was 100 mg/mL while the dynamic binding capacity at 10% breakthrough point was only 70 mg/mL (Yao and Lenhoff, 2006). Usually, the batch adsorption experiment focuses on one molecule that makes it not suitable to solve the design problem of isolating the target molecule from impurities. In addition, the results of batch adsorption experiment are usually used by general rate model to calculate the dynamic binding capacity, hence it is not evidence based data. Therefore, the batch adsorption experimental data is not considered in chromatography system.

To harness the chromatography experimental data, the ontology, theoretical and empirical knowledge were captured and formalized for the chromatography system. Thus, in the following, the representations of the three types of knowledge are introduced.

5.4 REPRESENTATION OF CHROMATOGRAPHY ONTOLOGIES

For the chromatography system, 9 domains ontologies were considered, namely product, strain, chromatography function, scale, manufacturer, column, resin, buffer chemical and elution strategy. All of the terms referred by captured column data belonged to these 9 domains, and these ontologies served as terminological criteria for searching relevant datapoints. The ontologies of product and strain were used in the centrifugation system, thus expanding the existing ontologies were concerned for these two domains. The ontologies about the other 7 domains can be developed by following the approach of the ontology development that has been introduced in Chapter 3.

5.4.1 Ontologies of product

Different bioproducts have different physical properties, e.g. molecule weight, Isoelectric point (pI), and these properties may perform different binding strength to the resin (Walsh, 2003). In order to differentiate the bioproducts, the product ontologies that describe the

product names and their affiliated relationship ares used. Shown in Figure 5.4, the product ontologies are extension of the one that has been built up for the centrifugation system in section 4.4.3. The extra terms employed in the product ontologies were from the captured column data, these new captured bioproduct terms would be helpful to illustrate how to expand the established ontologies.

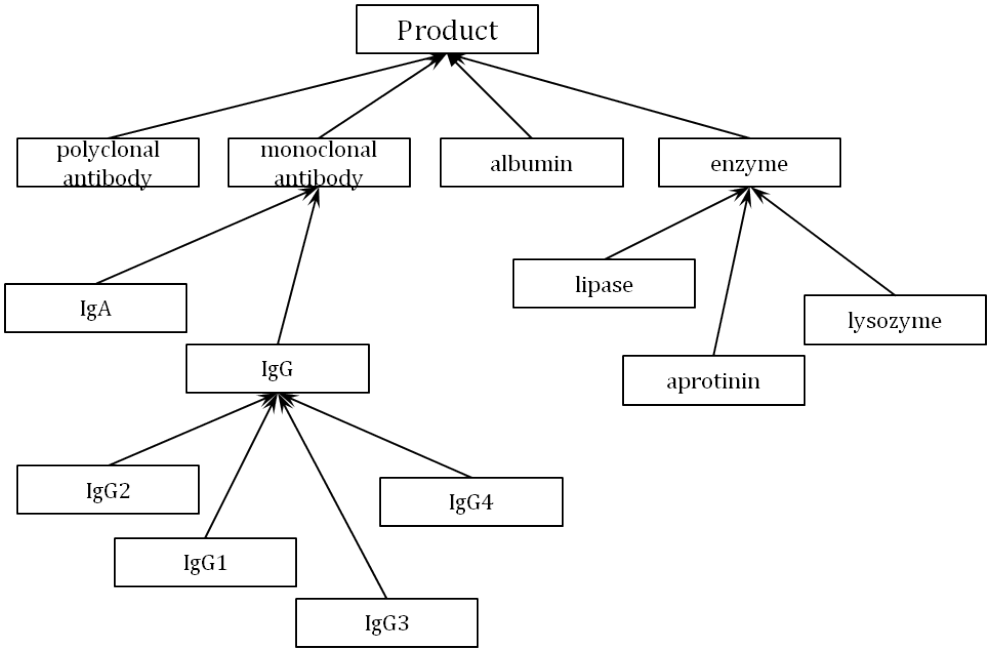


Figure 5.4: Product ontologies of chromatography

Each node represents a term of bioproduct, and the arrow indicates the relationship, 'a type of', between the two terms. In order to expand the product ontologies, the general bioproduct term is put into the parent class, while the specific bioproduct term is in the child class. For example, the enzyme is a general term that indicates the molecules which can catalyse the biological reactions. The lysozyme is a specific enzyme can be found in secretion which is the child class of the enzyme. The lysozyme inherits properties from the enzyme, e.g. catalysing the biological reactions. Using these new captured terms to expand the product ontologies can be a general landscape for users who want to arrange the new bioproduct terms used by the column data generated from the experiments.

5.4.2 Ontologies of strain

The molecules generated from different cell lines would have different chemical-physical properties, e.g. surface charge, which would require different conditions for adsorption/desorption, e.g. pH (Klinkenberg, 1955). To differentiate the cell lines, the strain ontologies are used. Figure 5.5 shows the strain ontologies that has been further developed for the chromatography system. The extra strain terms were from the column data.

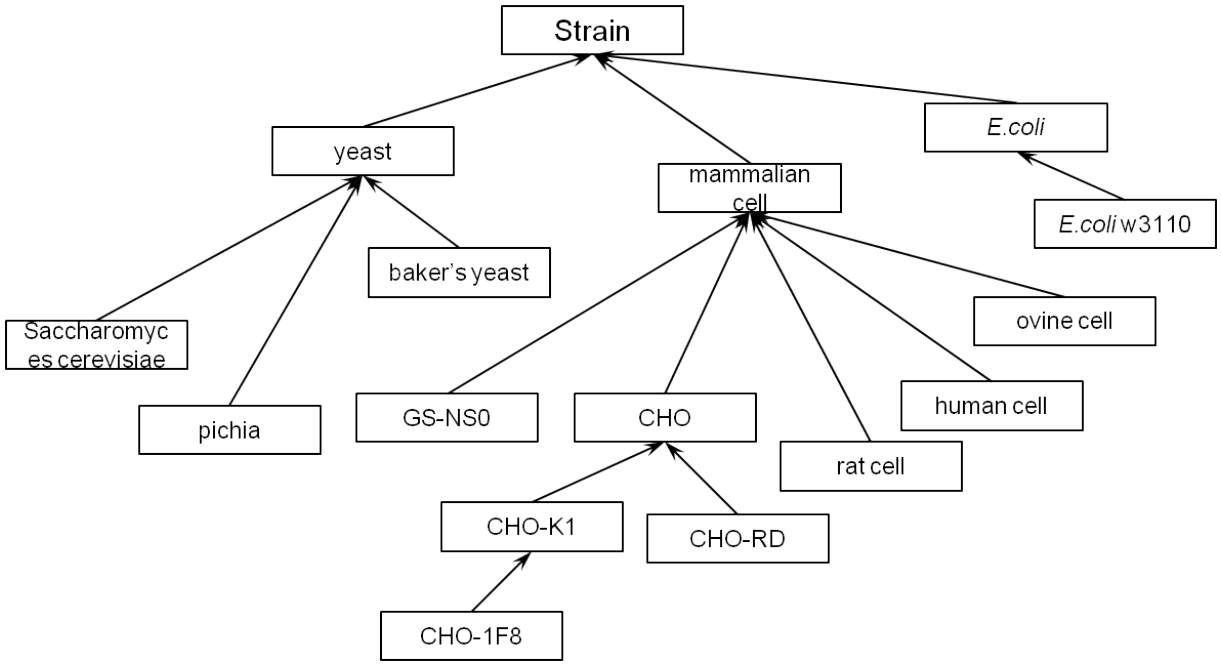


Figure 5.5: Strain ontologies of chromatography

Expanding the strain ontologies is the same as the further development of product ontologies, i.e. linking the specific term and the general term by the relationship 'a type of'. The product and strain ontologies illustrate a way to further develop the current ontologies which does not need to modify the structure of the existing ontologies.

The product and strain ontologies can describe the bioproduct precisely which will be used by search functionality to find relevant datapoints. For instance, the IgG1 generated from the human cells and the IgG1 came from the rat cells have different physical properties that perform different binging strength to the protein A resin which requires different operat-

ing conditions, e.g. pH, buffer chemicals (Affinity chromatography, principles and methods, Healthcare, GE, 18-1037-46). The product and strain ontologies can be used to differentiate the column data of IgG1 of human cells from the column data of IgG1 of rat cells.

5.4.3 Other ontologies in chromatography system

For other seven ontologies, i.e. chromatography function, scale, manufacturer, column model, resin model, buffer chemical and elution strategy, terms in each domain were arranged in a simple hierarchical structure. Since the development of ontologies have been discussed in Chapter 3, these domains ontologies will be given directly. All of ontologies introduced in the following are used to define and manage the terms referred by captured column data which allows them to be recognized and searched by chromatography system.

5.4.3.1 Chromatography function ontologies

The chromatography are used for three purposes of purification, i.e. capture, intermediate purification and polishing, which have been introduced in section 5.2.1. The function ontologies that consist of three terms are given in Figure 5.6.

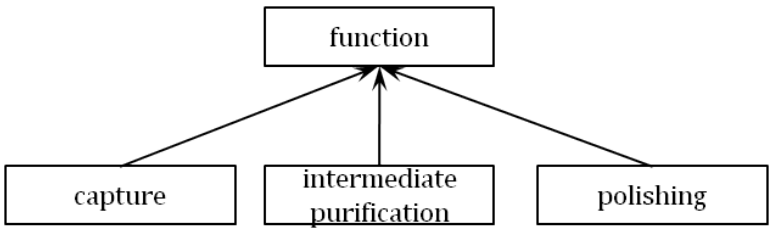


Figure 5.6: Function ontologies of chromatography

5.4.3.2 Scale ontologies

The scale ontologies used in the chromatography system are the same as the scale ontologies used in the centrifugation system (see Figure 4.6), thus it will not be repeated here.

5.4.3.3 Manufacturer ontologies

The equipments and consumables used in the chromatography, e.g. resins and columns, are provided by different manufacturers. The ontologies of manufacturer are used to manage the names of the different manufacturers, and they are given by Figure 5.7.

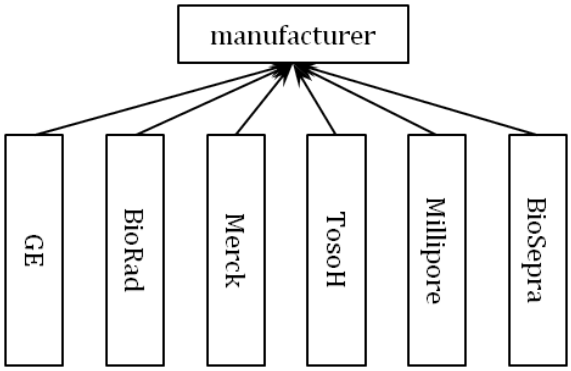


Figure 5.7: Manufacturer ontologies of chromatography

5.4.3.4 Column ontologies

The column ontologies describe the column model used in the chromatography experiments. Currently, there are two column models recorded in the 1021 datapoints, which are shown in Figure 5.8.

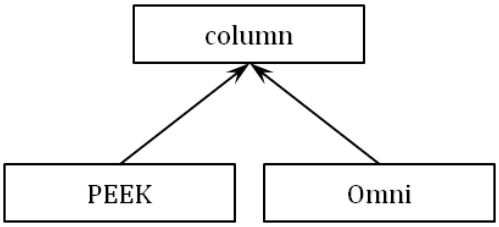


Figure 5.8: Column ontologies of chromatography

5.4.3.5 Resin ontologies

Various resins have been developed for purification. To manage these terms, the resin ontologies are used. In resin ontologies, IEX, AC, HIC and Multimode are the general terms in the parent classes. Anion and cation are the specific resin types of IEX, while protein A and protein G are the specific resin types of AC. Over all of the 1021 datapoints, there

were 37 IEX resin terms that included 23 anion resin terms and 14 cation resin terms, 16 HIC resin terms, 6 AC resin terms that included five protein A rein terms and one protein G resin term, and 1 Multimode resin term. In the resin ontologies, the specific resin inherits the same working mechanism from the general one, for example, Poros 50D is a type of anion resin that inherits using positively charged group to bind the molecules. The size exclusion chromatography (SEC) is not involved in the resin ontologies, because the SEC experimental data has not been captured. The SEC can be located at parent class while the terms of specific SEC resin are the child classes. The resin ontologies are shown in Figure 5.9.

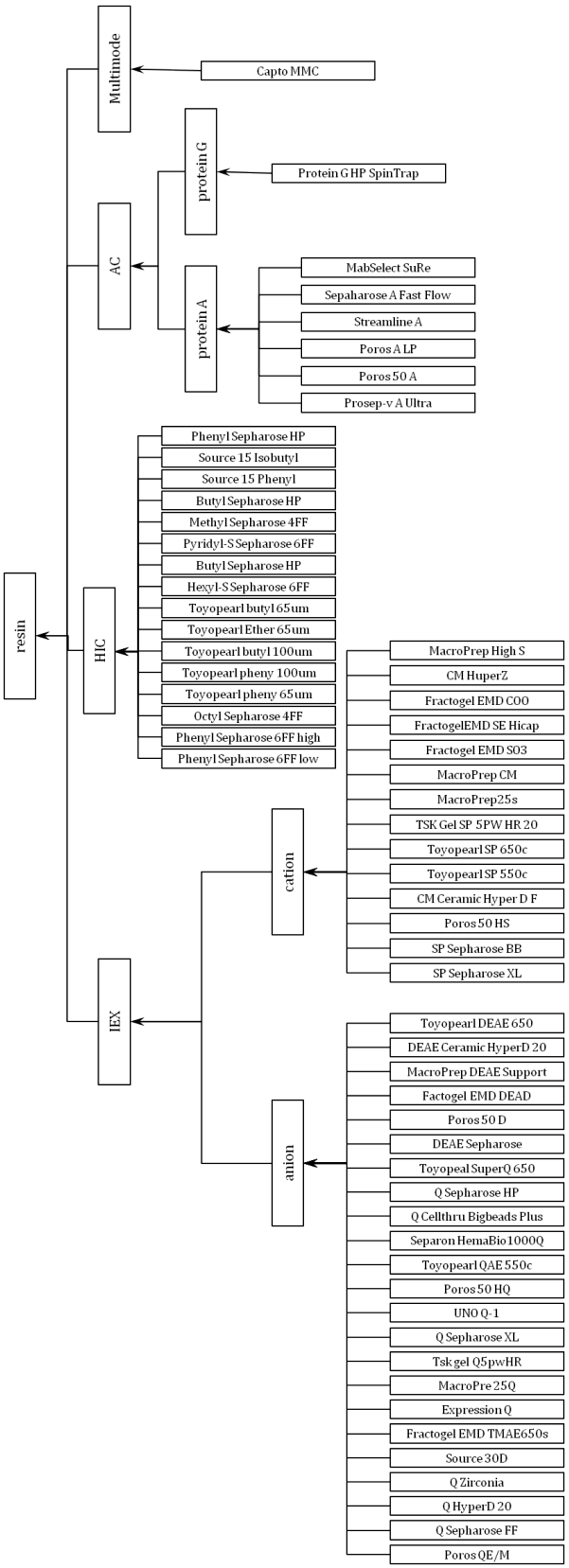


Figure 5.9: Resin ontologies of chromatography

5.4.3.6 Buffer chemical ontologies

The buffer chemical ontologies manage the chemical terms appeared in the column data. For a specific resin, the buffer chemicals are usually recommended by manufacturer. For example, sodium chloride and sodium phosphate are recommended by manufacturer for using Capto MMC resin (GE, UK). For the buffer chemical ontologies, each combination of buffer chemical terms serves as a class, e.g. sodium chloride and sodium phosphate is a class. Based on the 1021 column data, 10 classes (buffer chemical combinations) were arranged as a hierarchy shown in Figure 5.10.

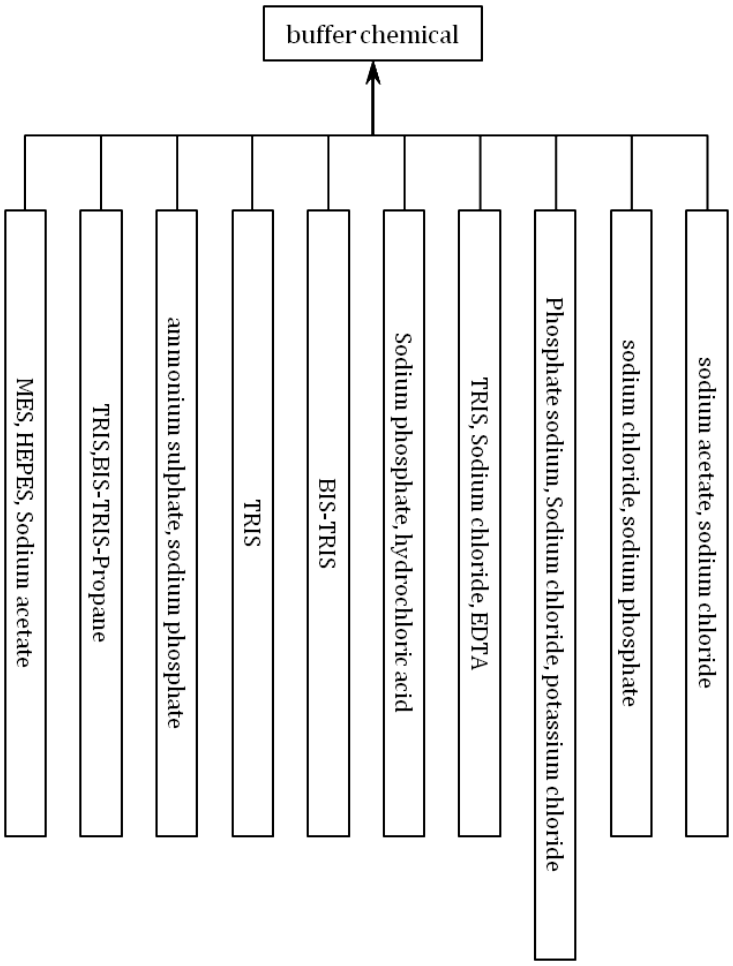


Figure 5.10: Buffer chemical ontologies of chromatography

5.4.3.7 Elution strategy ontologies

Two elution strategies were introduced in section 5.3.2, i.e. gradient elution and step elution. The elution strategy ontologies are given by Figure 5.11.

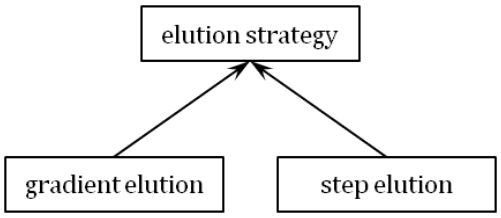


Figure 5.11: Elution strategy ontologies of chromatography

For the chromatography system, 132 ontologies within the 9 domains were established. Over all of these ontologies, 25 ontologies were about product, 17 ontologies were about strain, 64 ontologies were about resin, and 23 ontologies about chromatography function, manufacturer, column, buffer chemical and elution strategy.

5.5 REPRESENTATION OF CHROMATOGRAPHY KNOWLEDGE

5.5.1 Representation of chromatography theoretical knowledge

For the chromatography system, the theoretical knowledge includes the unit conversion, fundamental equations and background information of resin which would help to utilize the column data regarding the design problem. In the following, the representation of theoretical knowledge and its role in chromatography system are introduced.

5.5.1.1 Unit conversion

Usually, the column operation information may have different unit settings, e.g. the unit of loading flowrate can be linear velocity (cm/s) or volume velocity (ml/s). Using this information properly and correctly requires the various units to be standardized. For this,

the unit conversion aims at converting the different units into the standard units. Table 5.12 gives the standard units of the parameters which were used for data representation.

Table 5.12: Standard unit setting for the parameters included in data representation

Category	**Parameter**	**Standard unit**
Input	molecule weight	kDa
	feed volume	mL
	feed viscosity	Pa·s
	product concentration	mg/mL
	total protein concentration	mg/mL
Step	molecule weight	kDa
	temperature	^{o}C
	column volume	mL
	column heigh	cm
	column diameter	cm
	resin volume	ml
	equilibration buffer chemical concentration	mM
	equilibration buffer volume	mL
	equilibration flowrate	mL/s
	equilibration back pressure	kPa
	loading flowrate	mL/s
	loading volume	mL
	start buffer concentration	mM
	final buffer concentration	mM
	chemical concentration interval	mM
	elution buffer volume	mL
	elution flowrate	mL/s
	elution back pressure	kPa
	wash buffer chemical concentration	mM

continued on next page

continued from previous page

Category	**Parameter**	**Standard unit**
	wash buffer volume	mL
	wash flowrate	mL/s
	wash back pressure	kPa
	regeneration buffer chemical concentration	mM
	regeneration volume	mL
	regeneration flowrate	mL/s
	regeneration back pressure	kPa
Output	breakthrough point	%
	yield	%
	dynamic binding capacity	mg/ml
	purity	%

For any specification of a datapoint, if its unit was not a standard unit, then it would be converted as the standard unit automatically. The conversion of mL/h to mL/s serves as an example.

$$Y(mL/s) = \frac{X(mL/h)}{3600} \tag{5.4}$$

Equation (5.4) illustrates the knowledge entity of converting the mL/h to mL/s. The three variables and the mathematical relationship were used to represent this knowledge. The Y is knowledge output that has the unit mL/s, the X is the knowledge input that has the unit mL/h. The Y will be calculated by giving the X value, for example, if a datapoint includes the specification 'flowrate(4, mL/h)', then the specification 'flowrate(0.0011, mL/s)' will be generated by using this knowledge. But, the initial specification 'flowrate(4, mL/s)' will not be replaced by the new specification 'flowrate(0.0011, mL/s)'.

5.5.1.2 Fundamental equation

Fundamental equations are captured for doing the calculations that may be required by solving the chromatographic process design problem. The residence time is used as an example to demonstrate the how the fundamental equations work in the chromatography system.

The residence time is the time of the mobile phase passing through the stationary phase (from the column inlet to outlet detector). It is an primary consideration of chromatography scaling up (see section 5.2.5). Usually, the residence time is calculated by resin volume, column diameter and related flowrate, which is shown by equation (5.5)

$$residence\ time = \frac{resin\ volume}{loading\ flowrate} \tag{5.5}$$

This equation was represented by the four variables and the mathematical relationship. The *residence time* was the output while the *resin volume* and *loading flowrate* were the input. To use this knowledge, the values of the resin volume and loading flowrate are standardized by unit conversion first, and then the residence time is calculated. The residence time is used to generate the scale up solutions with respect to the desired resin volume.

5.5.1.3 Background information of resin

Different resins have different physical properties, e.g. mean particle size, surface chemistry, max flowrate, binding capacity, which would impact the adsorption performance. These information would contribute to resin selection. To manage this information, the ERM is used.

In an ERM, the specific resin is represented as an entity, each item of physical properties is represented as an attribute, the 'has' is used to indicate the relationship between the entity and the attribute. Figure 5.12 shows the ERM of Q Sepharose XL which is a type of IEX resin.

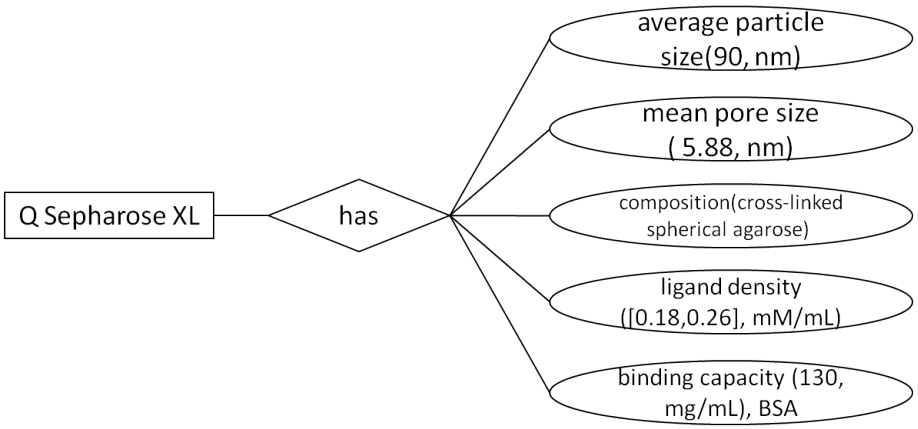


Figure 5.12: ERM of physical properties of Q Sepharose XL

In this ERM, the entity 'Q Sepharose XL' has five attributes which are the 'the average particle size is 90nm', 'the man pore size is 5.88 nm', 'the composition is cross-linked spherical agarose', 'the ligand density ranges from 0.18 mM/mL to 0.26 mM/mL', and 'the binding capacity for BSA is 130 mg/mL'. The five attributes are considered because they are key variables to the adsorption performance.

The 'average particle' and 'mean pore size' would influence the performance of binding capacity that have been explained by pore diffusion and intraparticle diffusion theory (Tsai et al., 1990; Yao and Lenhoff, 2006). The smaller pore size would lead to better DBC performance. This relation has been validated by two resins, Q Sepharose XL and Poros 50HQ (Staby and Jensen, 2001).

The particle size would impact the mass transfer. The smaller particle size indicates that the internal beads can be accessed more easily which would make better mass transfer. However, the small particle size would make the packed bed incompact so that the bed could be compressed easily during the column operation, this would cause the high pressure drop which may reduce the throughput of the column (Fahrner et al., 1999).

The composition indicates the support material which determines the compressibility. If the support material is soft, then it would be easy to be compressed which would increase

the pressure drop and reduce the flowrate during the operation. For instance, the agarose is more compressible than porous glass so that the flowrate of agarose column is slower than the porous glass column, the slower flowrate would reduce the column performance(Stickel and Fotopoulos, 2001).

The ligand density and binding capacity of BSA can be used as the reference for users to estimate the possible DBC performance. The greater the ligand density is, the more molecule can be bound. The BSA is a type of standard molecule which is generally used to quantify the binding capacity of the resin, this information may help to determine which resin may have higher binding capacity regarding the target molecule.

If the specific resin was concerned, the information organized by ERM would be available. For the chromatography system, 9 items of theoretical knowledge have been developed which include 4 items of unit conversion, i.e. L to mL, cm/h to cm/s, cv to mL and mL/h to mL/s; three fundamental equations, i.e. residence time calculation, conventing volumetric flowrate to linear velocity (cv/s to cm/s and cv/min to cm/s); and one ERM, i.e. Q Sepharose XL.

5.5.2 Representation of chromatography empirical knowledge

The empirical knowledge of centrifugation system is the USD rules that allow the centrifugation system to harness the centrifugation experimental data generated from different experimental scales. For the chromatography, similar approach has been developed by scientists. For example, the conditions of adsorption and desorption between a given protein and resin can be screened by using a pre-packed pipette tips. However, the results have relatively large errors, e.g. DBC (Chhatre et al., 2009). In addition, other information, such as elution strategy, elution flowrate, can not be mimicked by such experimentation. Thus, this approach is not used as the empirical knowledge for the chromatography system for now, since it is not well developed yet.

5.5.2.1 Representation of scale up principles

For chromatographic process development, the small scale column experiments, e.g. laboratory scale, are used to identify the column operating conditions, and then the results will be scaled up to the desired large scale column for manufacturing purpose by principles, i.e. scale up principles. These principles have been verified by empirical studies, therefore, they are considered as the empirical knowledge. The eight scale up principles explained in section 5.2.6 were represented as eight rules. The eight rules involved nine variables to be considered. Table 5.13 gives the five variables which would keep constant for scale up, where the **Generation** indicates how the variable is generated. Table 5.14 shows the other four variables that would be increased, where the **Generation** indicates how it is calculated to address the desired resin volume. Implementing these rules requires the desired resin volume, e.g. v1, given by user.

Table 5.13: 5 variables value are kept constant in scale up

Variable name	**Generation**
residence time	$\frac{resin\ volume}{loading\ flowrate}$
gradient slope	gradient slope specification
product concentration	product concentration specification
total protein concentration	total protein concentration specification
ratio of processing material volume to resin volume	$\frac{loading\ volume}{resin\ volume}$

Table 5.14: 4 variables value are increased to achieve the desired volume requirement

Variable name	**Generation**
column diameter (d_1)	$\sqrt{\frac{V_1}{V_2}} \times d_2$
volumetric flowrate	$loading\ flowrate \times \frac{d_1^2}{d_2^2}$
processing material volume	$V_1 \times \frac{loading\ volume}{resin\ volume}$
gradient elution buffer volume	$V_1 \times \frac{elution\ buffer\ volume}{resin\ volume}$

Given the desired volume, the value of corresponding column diameter, volumetric flowrate,

processing material volume and gradient volume are generated respectively by using the scale up principles. The example will be explained in the case study.

5.6 USING HIERARCHICAL HEURISTIC APPROACH FOR CHROMATOGRAPHIC PROCESS DESIGN

In the centrifugation system, a case study was used to demonstrate how to use the search, prediction and suggestion functionality to identify one operating condition with respect to the desired performance. However, the challenge of chromatographic process design is that multiple internal related variables require to be identified. To address this challenge, the hierarchical heuristic approach (HHA) is proposed. For this, the introduction of HHA will be given first, then the case study and the further analysis about using HHA for chromatographic process design problem are followed.

5.6.1 Introduction about HHA

Douglas (Douglas, 1985) introduced the hierarchical approach and refined it for solving the chemical process synthesis problem which has been reviewed in Chapter 2. In that situation, the chemical process synthesis problem is decomposed into 5 levels (5 sub-problems), the decision of each level is terminated by economic analysis, and the upper level decision would be used to make the decision for the lower level. Enlightened by this application, the HHA is adopted for the chromatography system to determine the solutions to a set of internal related variables.

The HHA can decompose the chromatographic process design problem into a set of *hierarchical levels* in a specific *hierarchical order*. In each hierarchical level, only one variable needs to be determined. The solution to this particular variable serves as a condition that will be used to determine the solution to the variable at next level. This means all feasible solution candidates to be examined at next level should satisfy the solution generated from the previous level. For each level, the solution is determined by the chromatography per-

formance, e.g. yield. The solutions generated from all of the hierarchical levels form the answer to this specific chromatographic process design problem. The specific hierarchical order which reflects the user's understanding as well as priority about the design problem should be defined by users, the higher level is, the more important the variable is. It also guides the solution generation based on user's understanding about the specific design problem. For instance, if the user think the resin type is the most important consideration, then the resin type should be put at the first level to be determined.

The HHA is a structured technique to organize and analyse complicated problems. Unlike other design solution approaches that focus on prescribing a correct solution, the HHA helps the users to generate the solutions in the way that best suits their goal and problem understanding.

5.6.1.1 Pseudo code of HHA implementation

In order to illustrate how the HHA works with the three reasoning functionalities, i.e. search, prediction and suggestion functionality, the pseudo code that describes the programming logic is given in Figure 5.13.

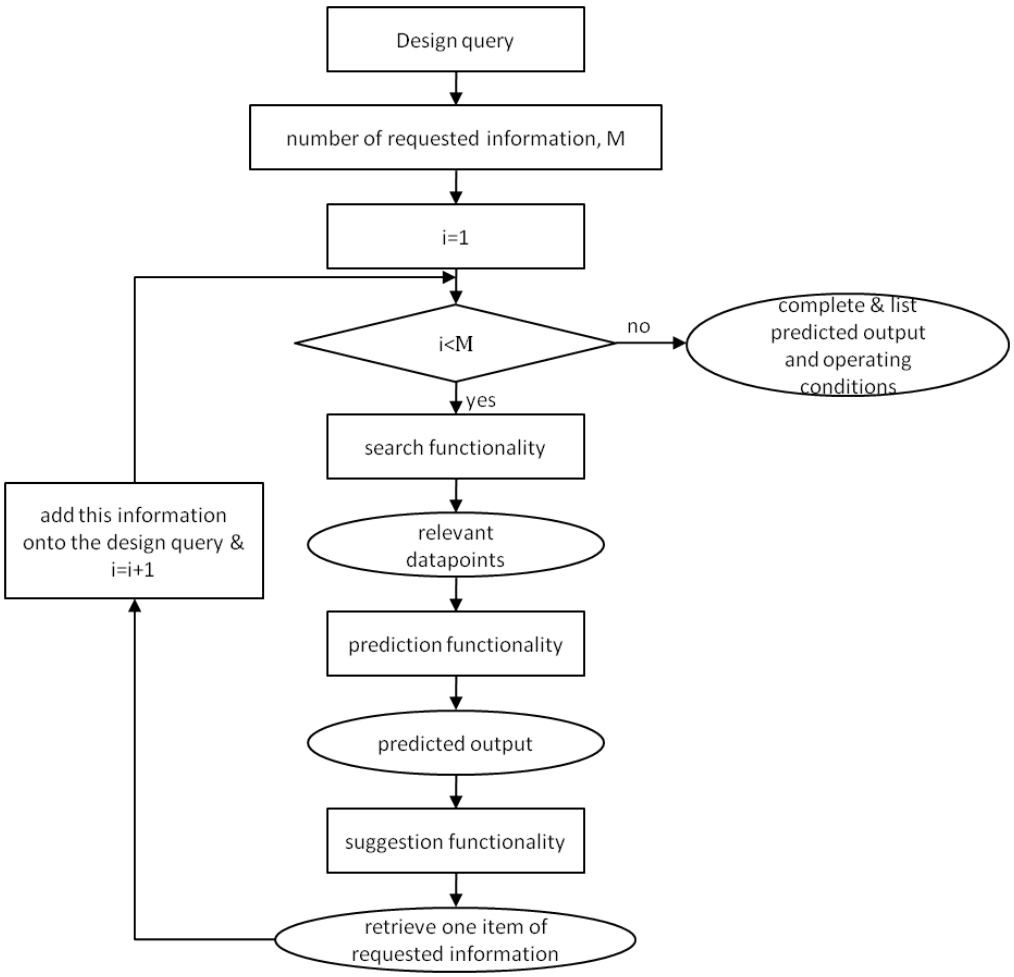


Figure 5.13: The pseudo code of use of HHA and three reasoning functionality for design problem

The procedure is explained as follows:

1. The design query is used to constrain the reasoning functionalities.

2. The number of requested variables defines the number of hierarchical levels (M) and i indicates the ith hierarchical level, they are used to control the loop.

3. i<M (start the loop, i.e. generate solution for ith hierarchical level),

   (a) if yes

      i. Return the design query related datapoints by search functionality (database and knowledge base are accessed).

ii. Generate the predicted performance by prediction functionality, e.g. predicted DBC (database is accessed).

iii. Retrieve the requested information by suggestion functionality (database and knowledge is accessed).

A. For numerical information, e.g. buffer concentration, pH and flowrate, the value is retrieved from the datapoint whose performance is most close to the predicted performance.

B. For terminological information, e.g. resin type and buffer chemical, the value is retrieved from the group of datapoints which has the highest performance.

iv. The solution is added onto the design query as a feature and generate solution for next hierarchical level (i=i+1 and go back to step 3).

(b) if no, the loop stops and all of solutions are shown.

5.6.2 Implementation of HHA on DBC design problem

In order to illustrate how to use the HHA for the chromatographic process designproblem, the DBC identification is used as an example. The DBC is an essential factor concerned in the chromatography development, because it determines the purity and yield of the chromatography (Fischer, 2011), and it is also a critical economic factor for cost of goods estimation (Larson et al., 2003).

5.6.2.1 Establishment of hierarchical levels

For a specific processing material, the DBC is mainly determined by the operating conditions (Snyder et al., 2010). For the case study, the variables including resin type, equilibration buffer chemical, equilibration buffer chemical concentration and loading flowrate are considered. By using the HHA, these four variables can form the four hierarchical levels. The solution to each variable generated in each level will answer the four following questions, which column should be selected, what buffer should be used, what the buffer composition

should be and what the loading flowrate is. Figure 5.14 illustrates the hierarchical levels generated by HHA that involves four variables.

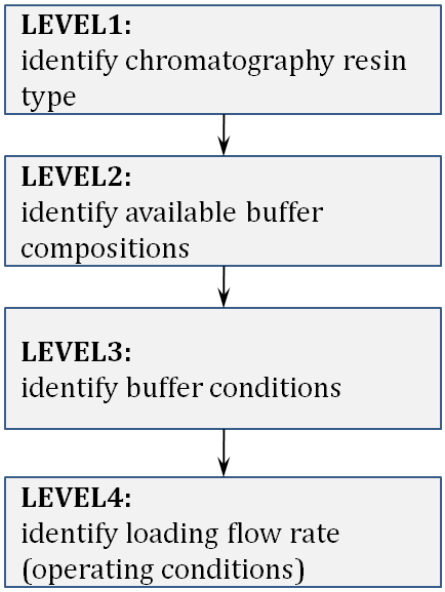


Figure 5.14: Four hierarchical levels about resin type, buffer compositions, buffer conditions and loading flowrate used for the DBC design problem

5.6.2.2 Formalization of design tree

In order to explain how the HHA generates solutions to internal related variables, the design tree is used. For the four hierarchical levels, the formalized design tree is given in Figure 5.15 which consists of the feasible solution candidates.

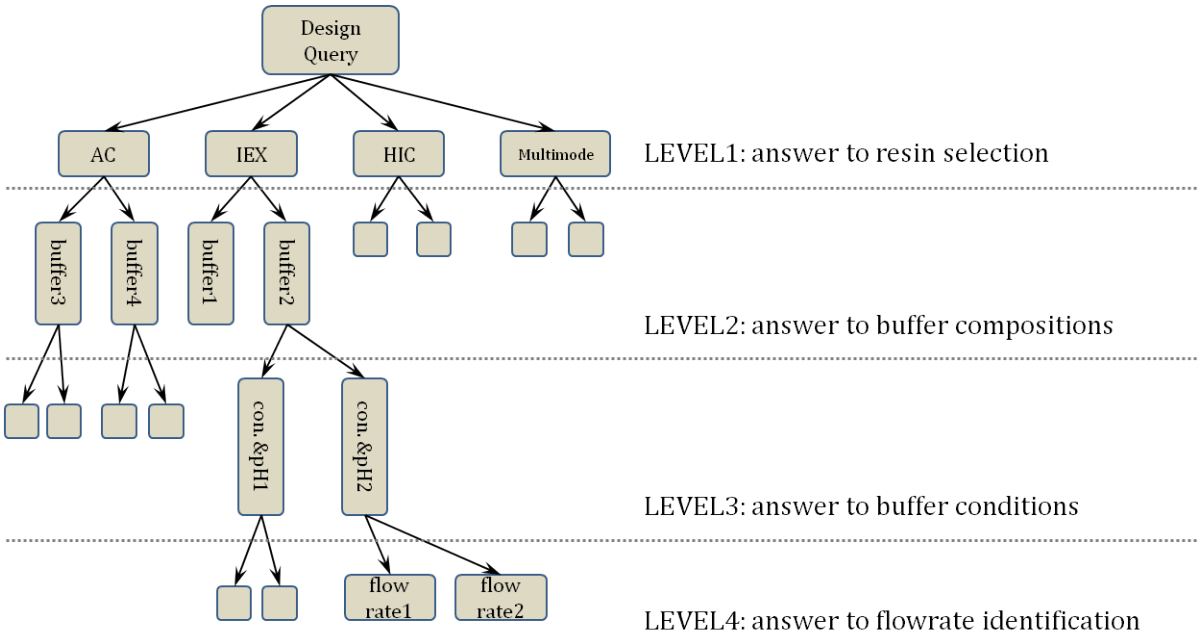


Figure 5.15: Design tree formalized by the solution candidates of the four hierarchical levels. The root represents the design problem that consists of the four queried variables, each node represents a feasible solution candidate to the current level.

The solution generated from the upper level is used as an constraint for the solution to the following level. For instance, if the IEX is generated to the first level, then it is used as a constraint to search relevant datapoints for the second level that is about the buffer chemicals selection, i.e. at the second level, the resin type included in the relevant datapoints are specific IEX. This make only the buffer 1 and 2 be examined while the buffer 3 and 4 are not searched since they are solution candidates to AC. Hence, the lower hierarchical level is, the more constraints will be used to search the relevant datapoints. This indicates that the design space to be explored is reduced along the hierarchical level, and the solution is a branch of the tree that consists of four nodes determined from the four hierarchical levels.

5.6.2.3 Generation of solutions by reasoning functionalities

In each hierarchical level, the requested variable is determined by the possible DBC achievement produced by three reasoning functionalities to harness the chromatography experimental data and knowledge.

For the buffer compositions (concentration, pH) and the loading flowrate, the search functionality finds the relevant datapoints and the prediction functionality generates the possible DBC performance, then the suggestion functionality retrieves the requested information from the datapoint whose DBC is most close to the predicted result.

For the resin types and the buffer chemicals, the search functionality finds the relevant datapoints and sorts them out as a set of groups, then the prediction functionality generates the possible DBC for each group and the suggestion functionality retrieves the solution from the group which has the highest predicted DBC. For instance, for the resin type at the first level, the search functionality finds the related datapoints and then sorts them out as four groups, i.e. IEX, AC, HIC and Multimode. The prediction functionality generates four possible DBC performance for the four resin types. Then the suggestion functionality retrieves the resin type whose predicted DBC is the highest. The predicted DBC indicates the DBC that would be possibly achieved by this resin, and the highest one represents the optimized performance achieved by this resin.

5.6.2.4 Establishment of hierarchical levels with different hierarchical orders

The order of hierarchical levels naturally reflects the user's understanding and emphasis about the design problem. In this case, the order of the four hierarchical levels indicates the common protocol that is used to identify the DBC regarding the specific processing material (Chhatre et al., 2009).

Different people may have different emphasis about the same problem, therefore the order of the hierarchical levels can be changed according to users' understanding. For instance, for the same design problem, the order of the four hierarchical levels can also be defined as identifying the buffer type first, then following by the column type, buffer conditions and loading flowrate (Figure 5.16). It allows the user to address the specific interest on buffer selection, e.g. which resin can work with this particular buffer?

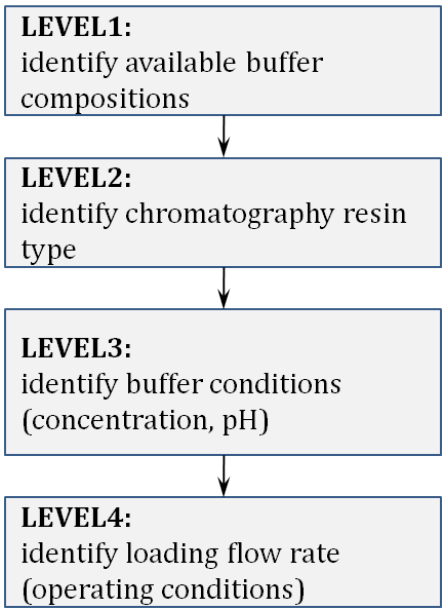


Figure 5.16: Four hierarchical levels of DBC design problem in another hierarchical order

By using the HHA, a complicated DBC design problem can be decomposed into a set of simple design problems. The solution to each sub-design problem can be generated by using the three reasoning functionalities. In the following, a case study is used to illustrate the implementation of HHA.

5.6.3 Case study and results analysis

5.6.3.1 Background information of the design problem

A clarified solution containing polyclonal antibody (pAb) after harvesting is used as the feed material purification. The product concentration was 1 mg/mL and total protein concentration was 2 mg/ml. A column needed to be identified to capture the pAb prodcut. The desired DBC requirement was at least 10 mg/ml at 10% breakthrough point. A serials of questions require to be answered: what was likely DBC that can be achieved? which resin should be selected in order to achieve the best DBC? what was the suitable loading buffer? what was the loading flowrate? In addition, if the resin packing volume was suggested as 1000 mL, what were the column size and loading flowrate?

5.6.3.2 Design query formalization

The formalization of the design query has been explained in Chapter 3, and the same approach is used in the chromatography system. Based on the information above, the represented design query is shown in Table 5.15, and the unit included in each feature is the standard unit that has been defined in Table 5.12

Table 5.15: Design query representation of case study on capturing pAb by chromatography

Input:
product(pAb, n/a),
product concentration(1, mg/mL),
total protein concentration(2, mg/mL)
Step:
function(capture, n/a),
chromatography type(X1, n/a),
equilibration buffer chemical(X2, n/a),
equilibration buffer chemical concentration(X3, mM),
equilibration buffer pH(X4, n/a),
loading flowrate(X5, cm/s)
Output:
breakthrough point(10, %),
DBC(10, mg/mL)

In the design query, the queried variable is indicated by 'X'. For instance, 'chromatography type(X1, n/a)' represents 'which resin should be selected', 'what is the loading buffer' is described by 'X2' (the equilibration buffer is same as the loading buffer in the column operation). The questions about equilibration buffer chemical concentrations, pH and loading flowrate are introduced by 'X3', 'X4', and 'X5' respectively.

When the design query has been formalized, the numerical criteria should also be defined to constrain the search functionality to find relevant datapoints. For this design query, the

numerical criteria are given in Table 5.16.

Table 5.16: Numerical criteria for design query of case study on capturing pAb by chromatography

Specification	**Numerical criterion**
product concentration (1, mg/mL)	[1(1-10%), 1(1+10%)]
total protein concentration (2, mg/ml)	[2(1-10%, 2(1+10%)]
breakthrough point (10, %)	[10(1-0%), 10(1+0%)]
DBC(10, mg/mL)	[10, $+\infty$)
equilibration buffer chemical concentration(X3, mM)	[X3(1-20%), X3(1+20%)]
pH(X4, n/a)	[X4(1-10%), X4(1+10%)]
loading flowrate(X5, cm/s)	[X5(1-10%), X5(1+10%)]

Please note that the equilibration buffer chemical concentration, pH and loading flowrate also require criteria definition, where the 'X3', 'X4' and 'X5' are the solution generated by the chromatography system. The criteria used here serve as demonstration only.

Given the design query, the next step requires user to define the hierarchical levels in order to generate the solutions. For this case study, the two hierarchical levels introduced in sections 5.7.2 (Figure 5.16 and 5.14) were employed, namely scenario A and B. Since the scale up solutions were also required by this design problem, and it can be determined when all of the operating conditions have been identified, therefore the scale up solution was added onto both of the scenarios as the fifth hierarchical level.

5.6.3.3 Pseudo code of scenario A and B

In order to illustrate the programme logic of the two scenarios, the pseudo codes are given in Figure 5.17 and 5.18

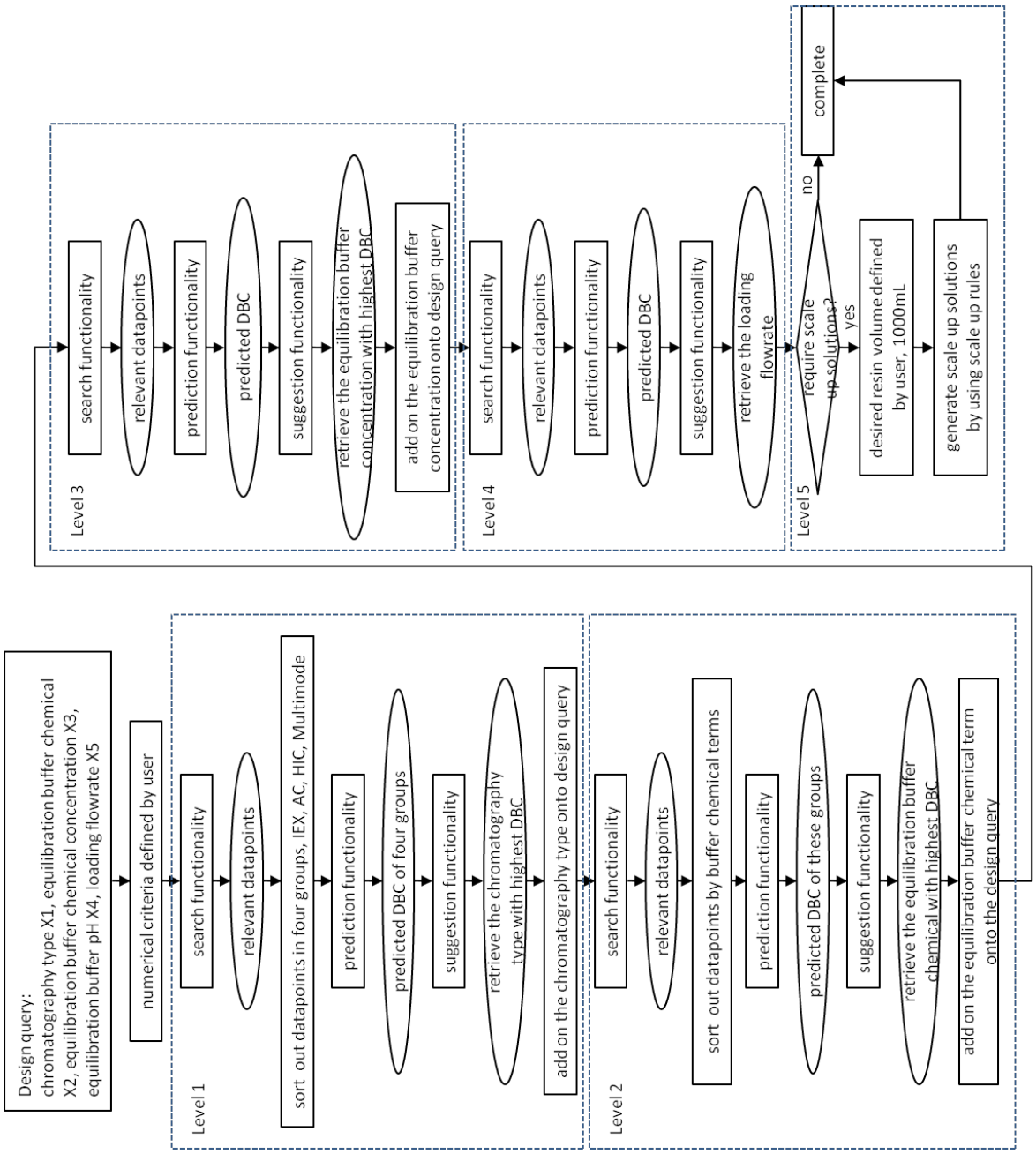


Figure 5.17: Pseudo code of scenario A for solutions about column type, buffer composition, conditions, flowrate and scale up solutions

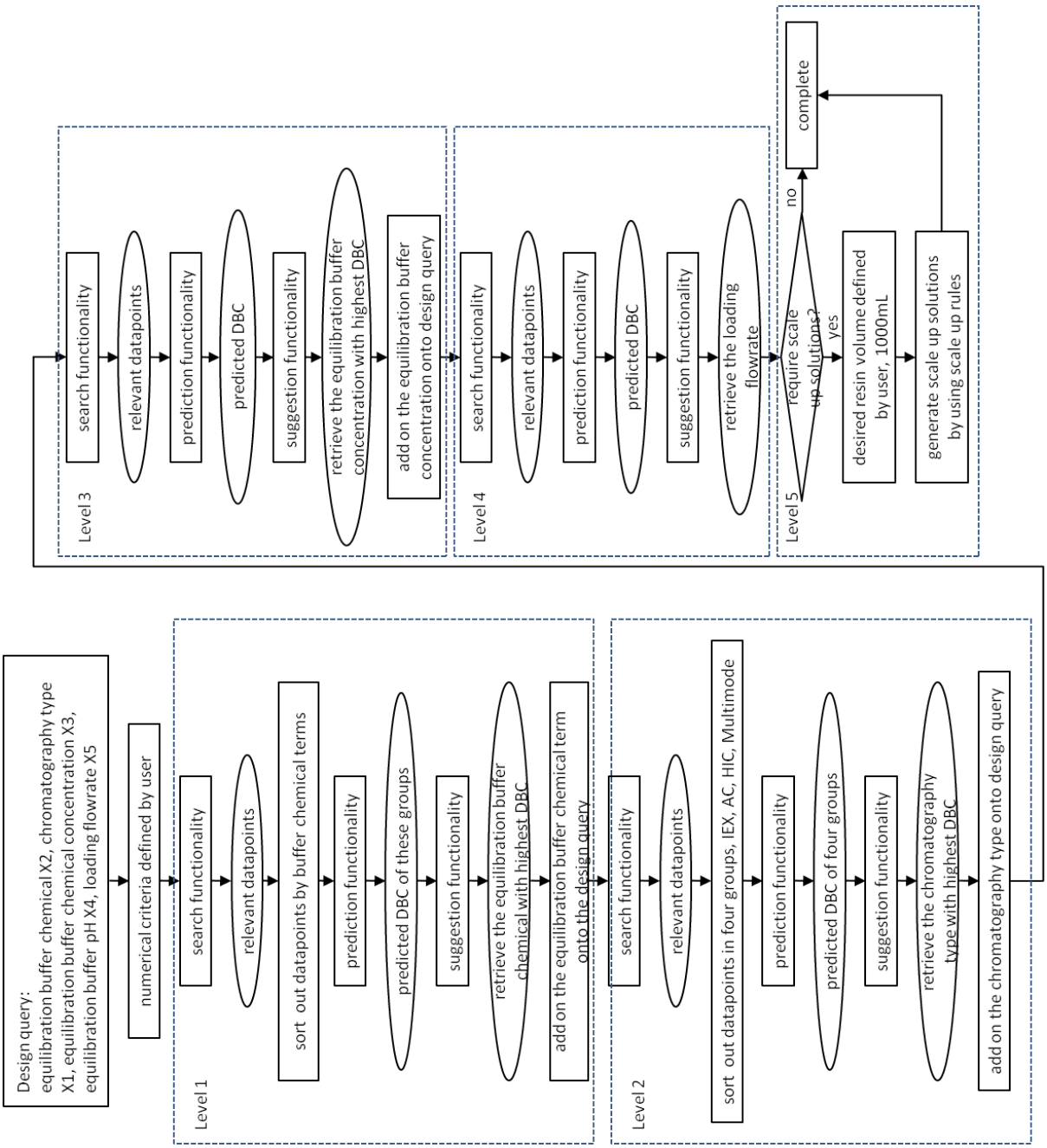


Figure 5.18: Pseudo code of scenario B for solutions about buffer composition, column type, buffer conditions, flowrate and scale up solutions

The two pseudo codes illustrate the programming logic of two types of hierarchical levels. For each level, the solution was used as the constraint for searching relevant datapoints in next level; in level 5, whether the scale up solution was required or not is asked first, if yes, the desired resin volume needed to be defined by user, then the scale up principles (empirical knowledge) were harnessed for solutions; if no, the programme completed and the results were shown to users.

5.6.3.4 Results for scenario A

For scenario A, the LEVEL 1 to 4 produced the answers to the resin type [X1], buffer chemicals[X2], buffer compositions [X3, X4] and loading flowrate [X5], and the LEVEL 5 was to generate the scale up solutions for the desired resin volume (i.e. 1000 mL) by using the scale up rules. The datapoints found in each of five hierarchical levels are shown in Figure 5.19 and the results are given in Table 5.17 with explanations.

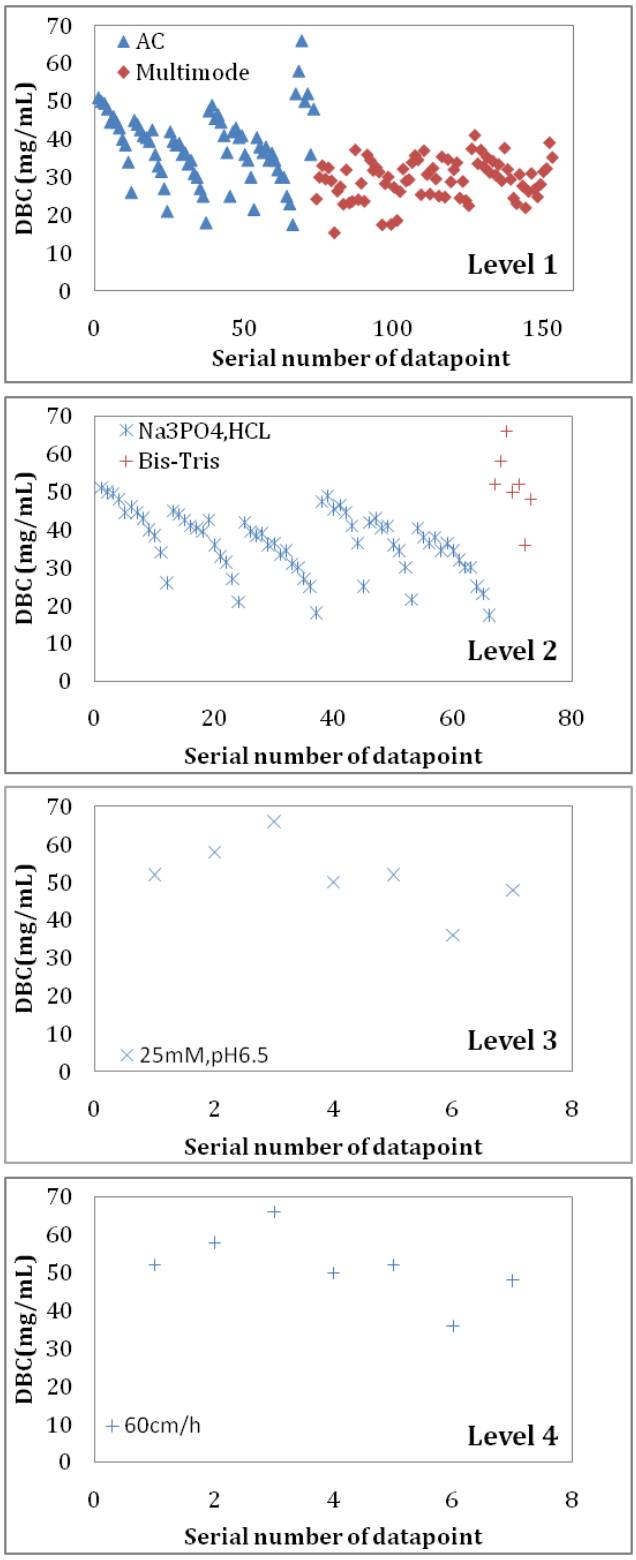


Figure 5.19: Relevant datapoints found for each hierarchical level of scenario A of case study about capturing pAb by chromatography. The number of relevant datapoints was reduced along the hierarchical levels. The lower the hierarchical level was, the less points were found.

Table 5.17: Results generated from each of five hierarchical levels of scenario A. The bold font is the solution generated from the hierarchical level.

LEVEL	Solution candidates	Number of relevant datapoints	Predicted DBC (mg/mL)
1	**AC**	73	38.2
	Multimode	80	29.5
2	**Bis-Tris**	7	51.7
	Na_3PO_4, HCL	66	36.8
3	**25 mM, 6.5**	7	51.7
4	**0.017**	7	51.7
5	The residence time was 172 s. The gradient slope was not available The pAb concentration was 1 mg/mL and the total protein concentration was not available. The ratio of processing material volume to resin volume was 6.4 mL. The column diameter should be increased to 20.8 cm. The volumetric flowrate should be increased to 5.7 mL/s. The loading volume should be increased to 6400 mL. The increased elution volume was not available.		

LEVEL1:

153 datapoints were found in level 1. 73 datapoints were about AC resins and 80 data points were about Multimode resins. No datapoint was relevant to HIC and IEX. The AC was suggested as the resin type, because it may realize the highest DBC performance, i.e. 38.2 mg/mL.

LEVEL2:

With the new constraint of AC resin, 73 datapoints were found for this level. 7 datapoints used the buffer Bis-Tris and 66 datapoints used the buffer Na_3PO_4, HCL. The Bis-Tris should be used as the equilibration buffer chemical, because it may produce the highest DBC performance.

LEVEL3:

Given the resin and buffer chemical that have been identified in the first two levels, 7 datapoints were found by search functionality. All of the 7 datapoints used the same buffer composition setting, i.e. Bis-Tris, 25 mM and pH 6.5. Thus, this composition was selected as the decision for the buffer composition.

LEVEL4:

With the solutions generated by previous three decision levels, 7 datapoints were found. The flowrate information was then retrieved from the datapoint whose the DBC performance was most similar to the predicted DBC 51.7 mg/mL.

LEVEL5:

For generating scale up solutions for the 1000 mL packed resin, the scale up rules were used. To use the rules, the specifications referred in the rules were retrieved from the datapoint whose flowrate was taken as the answer, because this datapoint satisfied the design query and the solutions generated from LEVEL 1 to 4, In this case, the specifications, product concentration(1, mg/mL), resin volume(1, mL), loading flowrate(60, cm/h), column diameter(0.66, cm) and loading volume(6.4, mL) were retrieved, while other specifications, gradient slope, total protein concentration and elution volume were not applicable because they were not presented in the data source. Therefore, some rules were not able to generate results. The results generated by using the scale up rules are shown in Table 5.17.

5.6.3.5 Results for scenario B

For scenario B, the buffer chemicals [X2] was studied in the LEVEL 1, then the resin selection [X1] was considered in LEVEL 2, followed by the buffer compositions [X3, X4] and the loading flowrate [X5] that were examined LEVEL 3 and 4 respectively, the scale up solutions for 1000 mL were generated at LEVEL 5. The datapoint found in each hierarchical level are shown in Figure 5.20.

Based on this hierarchical order, the results generated from the five levels are summarized

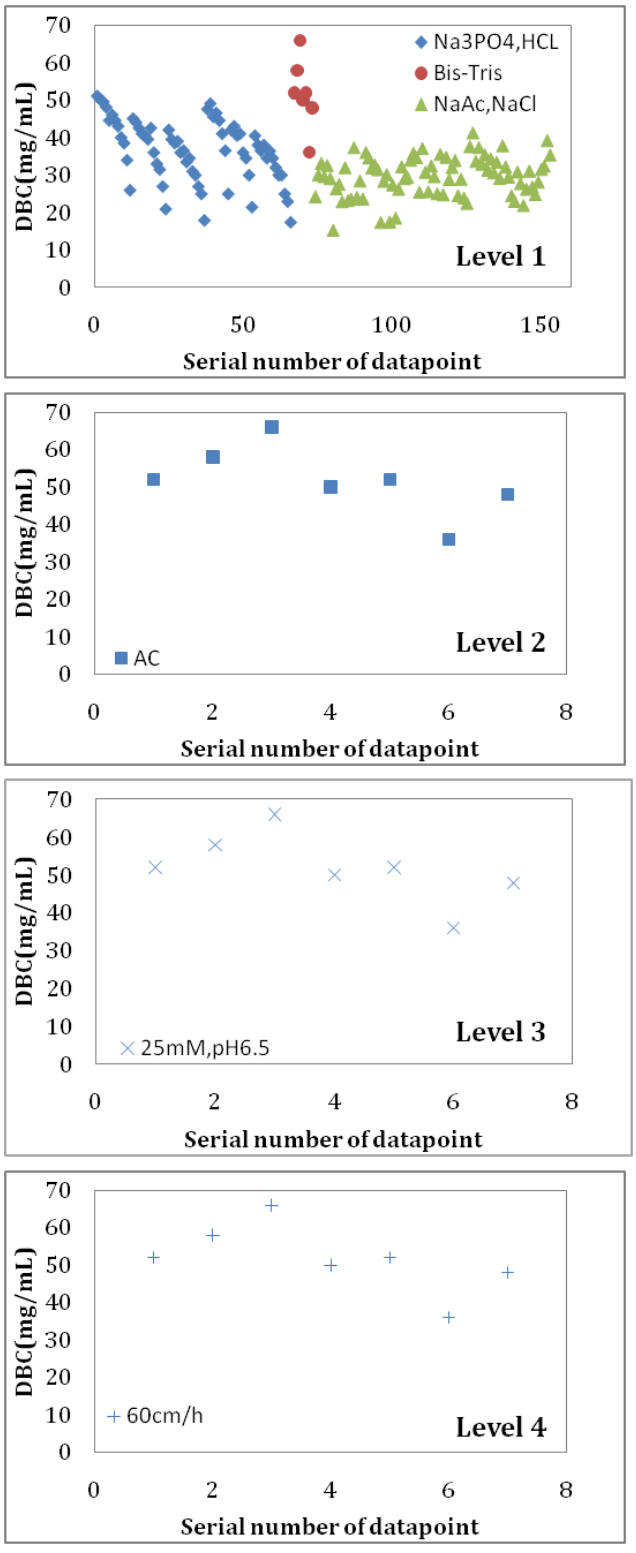


Figure 5.20: Relevant datapoints found for each hierarchical level of scenario B of case study about capturing pAb by chromatography

in Table5.18.

Table 5.18: Results generated from each of five hierarchical levels of scenario B. The bold font is the solution generated from the hierarchical level.

LEVEL	Solution candidates	Number of relevant datapoints	Predicted DBC (mg/mL)
1	NaAc, NaCl	80	29.5
	Bis-Tris	7	51.7
	Na_3PO_4, HCL	66	36.8
2	**AC**	7	51.7
3	**25 mM, 6.5**	7	51.7
4	**0.0167 cm/s**	7	51.7
5	The residence time was 172 s.		
	The gradient slope was not available.		
	The pAb concentration was 1 mg/mL and the total protein concentration was not available.		
	The ratio of processing material volume to resin volume was 6.4 mL.		
	The column diameter should be increased to 20.8 cm.		
	The volumetric flowrate should be increased to 5.7 mL/s.		
	The loading volume should be increased to 6400 mL.		
	The increased elution volume was not available.		

5.6.3.6 Analysis of scenario A and B

In order to analysis the generated results, the reasoning of the design trees regarding the two scenarios are shown in Figure 5.21.

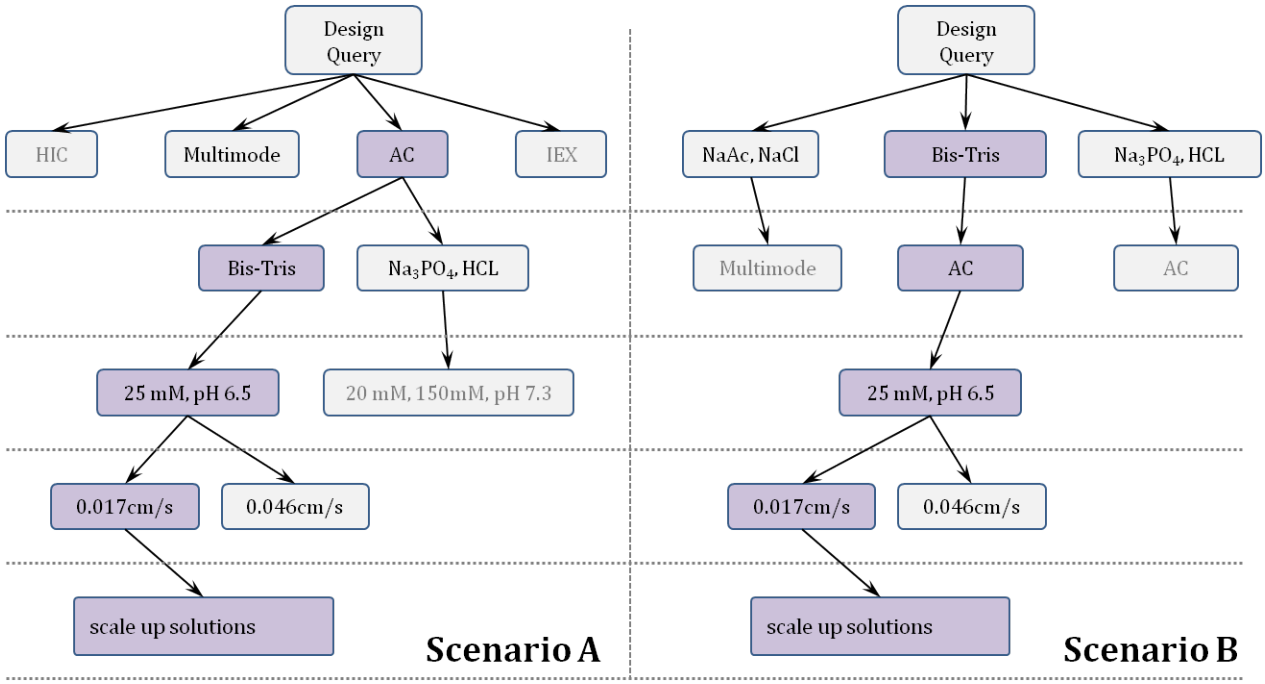


Figure 5.21: Reasoning of design trees for scenario A and B. Each node represents a solution candidate for current hierarchical level, the selected solution is coloured darkly.

Different hierarchical order allows the different branch of the design tree to be explored which would generate different solutions. For instance, in the scenario A, the four types of resin were examined for pAb purification first, then based on the selected AC resin, the two types of buffer chemicals were studied; for scenario B, the three types of buffer chemicals used for pAb purification were explored at first and then only one type of resin, i.e. AC, was analysed under Bis-Tris buffer (other three types of resins were not considered because Bis-Tris buffer was not used for the experimentation of other three types of resin yet).

Usually, the different hierarchical levels would generate the different solutions regarding the same chromatographic process design problem. However, in this case study, the solutions of the two scenarios are the same. The reason is that the datapoints included in the chromatography database are not enough to cover all types of situations, for example, in scenario B, only the AC resin is available for the buffer Bis-Tris and this made the solution of resin type same as the scenario A. If more datapoints are involved, the solutions to the scenario A and B would be different. Further analysis about this situation will be discussed

in section 5.8.

5.6.3.7 Discussion

In this case study, the equilibration buffer conditions, i.e. chemical concentration and pH, were generated in the same hierarchical level, because the value of these two variables included in the captured datapoints were fixed with respect to the target molecule and resin. However, they certainly can be determined separately by two hierarchical levels.

The knowledge can be harnessed by HHA, e.g. if the IEX resin is suggested, then the resin ontologies can list all of available IEX resins for users to determine which specific resin should be selected; if one specific resin is concerned, its background information represented by ERM model can be harnessed for further consideration.

The solutions to other variables referred in Table 5.2 can be generated by conducting more hierarchical levels, e.g. elution buffer chemical, pH, elution strategy.

The two scenarios demonstrate that the HHA allows user to assess all of the feasible solutions regarding the specific DBC requirement. For instance, in this case study, the four operating conditions represent P_4^4=24 types of scenarios to be studied (scenario A and B are two of them). The chromatography system allows the users to examine the 24 scenarios quickly, and these results may help user to further understand the DBC design problem in order to find the optimal solution.

In addition to DBC, the yield and purity are also important column performance. It will be more relevant if the three variables were considered as the criteria to generate solutions to each hierarchical level of a general chromatographic process design problem. For this, a possible way is to establish a weighted objective function based on the three variables. Due to the limited time, this situation will not be discussed in this thesis.

5.7 FURTHER ANALYSIS ABOUT HHA

The HHA decomposes the design problem as a set of hierarchical levels with specific order. Different orders may generate different solutions. Different criteria may impact the number of relevant datapoints found for each hierarchical level that may influence the solution. In this section, a case study is used to illustrate how the criterion setting and the order of hierarchical levels impact the results generated by HHA.

5.7.1 Method of HHA further analysis

Different hierarchical order allows the different branches of design tree to be explored. If the design space only contained partial solution candidates that would limit the number of design tree branches to be explored, then the solutions generated from different hierarchical orders may be the same. For example, in the design tree of scenario A (Figure 5.21), the node 'AC' has two buffer candidates 'Bis-Tris' and 'Na_3PO_4, HCl' while the node 'Multimode' has one buffer candidate 'NaAc, NaCl'. If 'AC' is selected, then the buffer 'NaAc, NaCl' could not be examined. Contrarily, in scenario B, if 'Bis-Tris' is selected, then only the 'AC' could be analysed. This indicates that each node does not have equivalent solution candidates to examine, and it is the reason why the same results were produced in the two scenarios. Therefore, in order to do a fair analysis, a design space that allows each node to have the equivalent solution candidates at the same hierarchical level is required, and it is called a *well-design space*.

Generally, the design space obtained by design of experiments (DoE) is considered as a well-design space. The DoE is a statistical controlled approach of experiment design. In the DoE, the variables to be studied are varied in a carefully structured pattern so that the individual and interactive effects of these variables can be examined simultaneously and equally.

For further analysis of HHA, the well-design space formalized by the experimental data about using Multimode resin to purify the pAb antibody was used (Chhatre et al., 2009). These experimental data generated by DoE was used to study the interactions between DBC

performance and three buffer conditions, i.e. concentrations of sodium acetate (NaAc), sodium chloride (NaCl) and pH. The concentrations of two chemicals were defined as four-level, i.e. ranging from 5 mM to 20 mM with 5 mM interval and the pH was 5-level, i.e. ranging from 4 to 5 with 0.25 interval. Therefore, 4×4×5=80 experimental data was included in this well-design space. The 80 chromatography experimental data with a design tree of pH as well as the NaAc and NaCl concentrations are shown in Figure 5.22. In this design tree, each node has the equivalent solution candidates, e.g. each specific pH value has 4 NaAc concentration candidates, i.e. 5, 10,15 and 20 mM. This indicates that no matter what specific pH value is, the same setting of NaAc concentration setting will be examined in the next hierarchical level, even with the different hierarchical order. Hence, this design space is appropriate to study the impact of different hierarchical order on solutions.

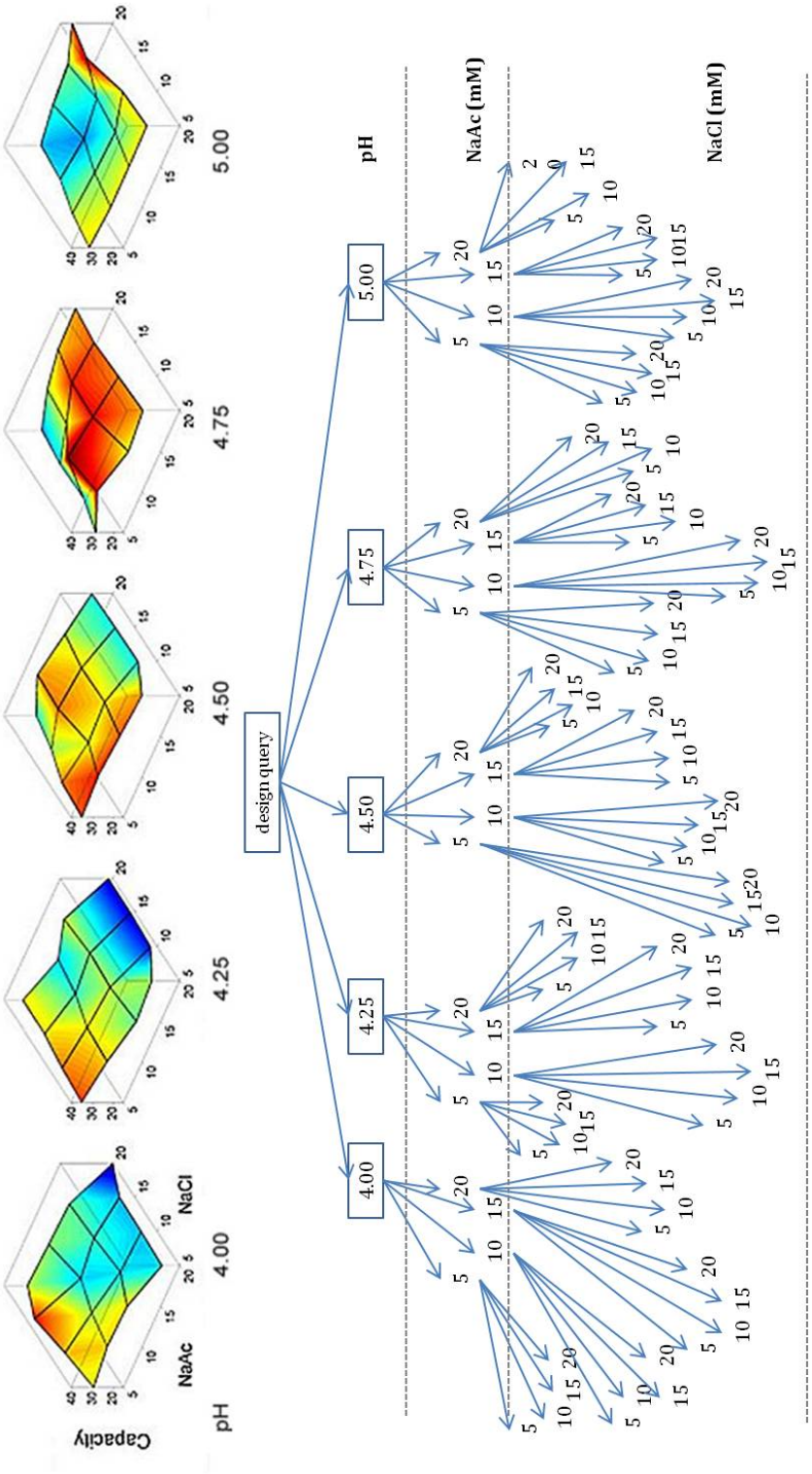
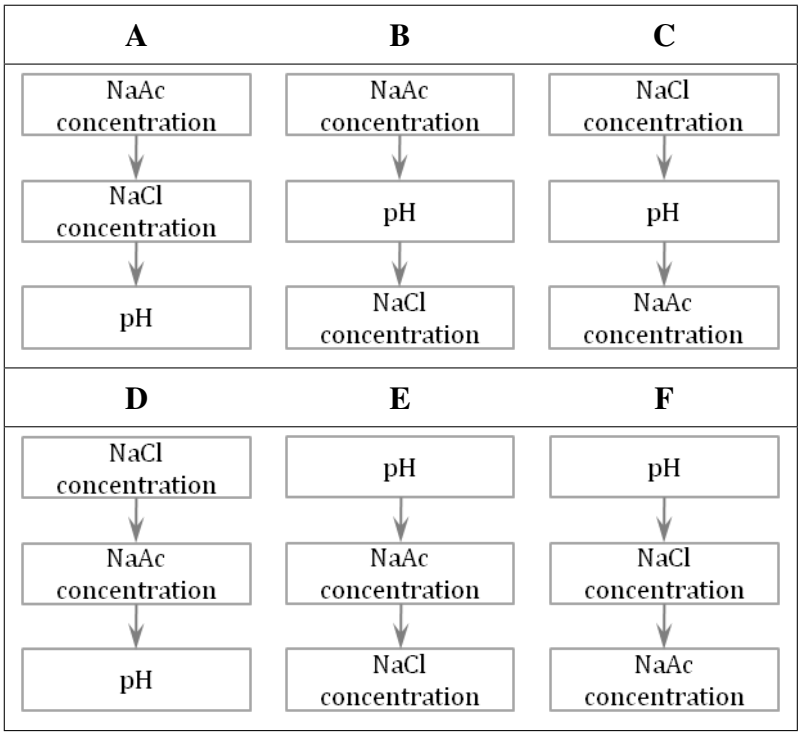


Figure 5.22: A well-design space and one type of design tree. The space consists of 80 experimental data generated by DoE, each plot illustrates the interactions about the concentrations of NaAc, NaCl, pH and DBC performance, each combination of the four variables forms a branch of the design tree, source: Chhatre, S., et al. 2009. Permission to reproduce the partial content of this figure has been granted by Journal of Chromatography A.

5.7.2 Scenarios of different hierarchical orders

The well-design space gives 80 solution candidates to the design problem which aims at identifying the buffer conditions (three variables) to realize the desired DBC performance. The three variables represent three hierarchical levels that can be arranged by a specific hierarchical order, each hierarchical order forms a scenario. There are P_3^3=6 scenarios in total and they are shown in Table 5.19.

Table 5.19: 6 scenarios for the three variables of NaAc concentration, NaCl concentration and pH



Each scenario has a specific design tree. Each hierarchical order represents a specific strategy to explore the 80 branches of a design tree. For instance, in scenario A, the NaAc concentration is generated first, followed by NaCl concentration and pH value. The results generated from the six scenarios may be different, and investing these results would help to draw conclusions about the impact of different hierarchical order on solutions.

For each of six scenarios, the solutions are determined by the relevant datapoints found by search functionality with defined criteria. Therefore, the criteria used to constrain the data

searching should be examined. In addition, the performance requirement also impacts the number of the relevant datapoints. The higher performance is required, the less datapoints would be found from the database. Therefore, the requirement of DBC performance also should be considered.

In order to study the results of the six scenarios with respect to the different criteria or DBC requirement, a new version of chromatography system has been established in WinProlog called analysis system. In this system, the database consists of 80 datapoints captured from the well-design space, and the knowledge base contains the ontologies, theoretical and empirical knowledge that have been established in the chromatography system. Given the specific criteria of the three variables and the DBC requirement, the analysis system will produce the results in regard to the six scenarios. These results will be discussed in the following.

5.7.3 Impact of criterion setting on HHA

In order to illustrate how these criteria impact the results generated by HHA, two sets of criterion were considered, namely wide criterion and narrow criterion, as shown in Table 5.20 and 5.21

Table 5.20: Wide criteria of NaAc concentration, NaCl concentration and pH

Feature	**Numerical criterion**
NaAc concentration (X)	[X-15, X+15]
NaCl concentration (Y)	[Y-15, Y+15]
pH (Z)	[Z-1, Z+1]

Table 5.21: Narrow criteria of NaAc concentration, NaCl concentration and pH

Feature	**Numerical criterion**
NaAc concentration (X)	[X-4.9, X+4.9]
NaCl concentration (Y)	[Y-4.9, Y+4.9]
pH (Z)	[Z-0.24, Z+0.24]

The chemical concentrations range from 5 mM to 20 mM and the pH ranges from 4 to 5. For each hierarchical level, this wide criterion setting allows all of 80 datapoints can be searched, i.e. all types of solution candidates can be examined for a specific node. The chemical concentrations interval is 5 and the pH interval is 0.25. For each hierarchical level, this narrow criterion setting only allow the datapoints which have the equal value of concentration of NaAc and NaCl as well as the pH generated from previous hierarchical level to be returned for solution generation. For example, if the NaAc was 10 mM generated from the first hierarchical level, then the narrow criterion would only allow the datapoint whose NaAc is 10 mM to be returned. For any specific node, the narrow criterion setting only allows the solution candidates deriving from this node to be explored.

Given the two types of criterion setting and the DBC requirement, i.e. 0mg/mL, the analysis system generated 12 results. All of the results generated from six scenarios are correct solutions to the design problem to accommodate the user's specific requirement about the design problems and they are shown in Table 5.22.

Table 5.22: Results of 6 scenarios under two types of criterion setting

Results for the wide criterion setting						
Parameter	A	B	C	D	E	F
NaAc concentration (mM)	5	5	5	5	5	5
NaCl concentration (mM)	20	20	20	20	20	20
pH	4	4	4	4	4	4
predicted DBC (mg/mL)	29.5	29.5	29.5	29.5	29.5	29.5
Results for the narrow criterion setting						
Parameter	A	B	C	D	E	F
NaAc concentration (mM)	5	5	15	15	15	15
NaCl concentration (mM)	5	10	20	20	20	20
pH	5	4	4	4.5	4	4
predicted DBC (mg/mL)	32.6	32.4	25.2	29.6	27.5	25.2

The six results generated under the wide criterion setting are the same. They indicate the

very like DBC performance 29.5 mg/mL can be obtained by using NaAc 5 mM, NaCl 20 mM and pH 4. This indicates that the wide criterion will not allow the solution generated from current hierarchical level to constrain the relevant datapoints searching at next level. Hence, the relevant datapoints found for each hierarchical level of the six scenarios were same, and the solutions of the three variables and predicted DBC were same.

However, when the narrowed criterion setting was used, the six scenarios produced the five different results (the scenario C and F have the same result). The narrow criterion setting makes the solution generated from current hierarchical level as the constraint for finding relevant datapoints at next hierarchical level. Taking the scenario E as an example, the pH 4 was generated from the first level, then only the datapoints whose pH value was 4 would be returned to determine the NaAc Concentration. Therefore, the relevant datapoints found for the second and third hierarchical level included in the six scenarios were different so that the different solutions were produced. The reason that the scenario C and F have the same solution is that the predicted DBC generated in their first and second hierarchical level were quite close, i.e. 29.5 mg/mL and 27.5 mg/mL for scenario C, 29.5 mg/mL and 27.5 mg/mL for scenario F. These predicted DBCs were all close to the datapoint whose DBC was 27.5 mg/mL, hence the pH and the NaCl concentration were same in the two scenarios since they were retrieved from the same datapoint. However, the predicted DBC generated in their second hierarchical level were different which indicated that the different datapoints were found.

Based on the results, the narrow criteria allow the different branches of design tree to be explored with different hierarchical orders while the wide criteria can not. Therefore, using narrow criteria allows the solutions to be generated based on users' understanding about the design problem. The general suggestions on criterion setting for HHA can be summarized as: if the criteria were wider, then the results generated from different scenarios would be more similar; if the criteria were narrower, then the results from different scenarios would be more specific. The six different results generated by using the narrower criterion setting indicate that the solution generated by using the protocol (see Table 5.2) is only one of feasible solutions regarding the chromatographic process design problem, other different

solutions would be generated by changing the identification order of variables.

5.7.4 Impact of performance requirement on HHA

The DBC performance generated from the different scenarios represents the possible DBC to be realized. The possible DBC would be impacted by the DBC requirement defined in the design query. This section aims at examining how the specific DBC requirement impact the results generated by HHA. For this, three types of DBC requirements were used, i.e. 0 mg/mL, 30 mg/mL, 40 mg/mL.

The 0 mg/mL may indicate that the user have no specific expectation about the DBC to be realized, e.g. users may have no experience about this design problem so that the '0 mg/mL' allows all related experimental data can be screened. The 30 mg/mL is obtained by averaging the DBC values included in the 80 datapoints, it represented the central tendency of the 80 DBC values. From the profile of the well-design space (Figure 5.22), the highest DBC value was around 40 mg/mL and using 40 mg/mL as the DBC requirement would help to demonstrate what results can be generated by HHA if the maximum performance is concerned.

For each of the specific DBC requirement, the narrow criteria was used by the analysis system to generate results, because this type of criterion would allow the different branch to be explored with different hierarchical orders. All of the 18 results with respect to the three DBC requirements are shown in Table 5.23, and they were all correct solutions.

Table 5.23: Results of 6 scenarios regarding three types of DBC requirement and narrow criteria

Results for the DBC 0 mg/mL						
Parameter	A	B	C	D	E	F
NaAc concentration (mM)	5	5	15	15	15	15
NaCl concentration (mM)	5	10	20	20	20	20
pH	5	4	4	4.5	4	4
predicted DBC (mg/mL)	32.6	32.4	25.2	29.6	27.5	25.2
Results for the DBC 30 mg/mL						
Parameter	A	B	C	D	E	F
NaAc concentration (mM)	10	10	20	15	20	15
NaCl concentration (mM)	20	20	20	20	5	5
pH	4.5	4.5	4.75	4.75	4.5	4.5
predicted DBC (mg/mL)	33.9	34.1	33.2	31.5	33.9	35.6
Results for the DBC 40 mg/mL						
Parameter	A	B	C	D	E	F
NaAc concentration (mM)	10	10	10	10	10	10
NaCl concentration (mM)	10	10	10	10	10	10
pH	4.75	4.75	4.75	4.75	4.75	4.75
predicted DBC (mg/mL)	41.1	41.1	41.1	41.1	41.1	41.1

For DBC requirement at 0 mg/mL, different scenarios generated different possible DBC performance. These predicted DBC results varied in value, e.g. minimal predicted DBC is 25.2 mg/mL while the maximal DBC is 32.6 mg/mL. As the DBC requirement was 30 mg/mL, the predicted DBC results generated from the different scenarios were stable. As the DBC requirement increases to 40 mg/mL, all of the six scenarios generated the same solutions and the predicted DBC was the highest DBC involved in the well-design space.

Based on the results above, the observations about the impact of different performance requirements can be described in the following. If no specific performance requirement is

given, e.g. DBC 0 mg/mL, the predicted performance generated by HHA for each scenario would vary. Although they are all satisfied solutions, the possible DBC achievement would be greatly impacted by the hierarchical order. If the specific performance requirement is given, then the predicted DBC performance of different scenarios would be similar. This indicates that if users has clear expectation about the DBC performance to be achieved, the possible DBC achievement would not be greatly influenced by hierarchical order. As the performance requirement is rising up, the predicted DBC performance for different scenarios would be more similar. The extreme situation is that only the DBC requirement only allows one datapoint in the design space can be found for solutions.

Based on the analysis results, there are no absolute regulations to be defined for the criteria and performance requirement to be used by HHA. Actually, the criteria and performance should be defined by users based on their studying objectives. However, the analysis results may tell a clue that using narrow criteria and specific performance requirement may be a better way to explore the design space.

Obviously, the more variables are included in the design query, the more solutions would be generated by HHA. Therefore, it is interested to investigate how to select the desired solution. The pairwise comparison may achieve this target. This approach was introduced by Fechner in 1860 (Fechner et al., 1966), and then further developed by Saaty to solve the multiple preferences decision making (Saaty, 2008). By using this approach, all of the solutions generated by HHA can be assessed as different priorities with respect to the users preferences about the design target, e.g. highest DBC requirement, long resin life time, low resin cost. Ranking the solutions based on the priorities allows users to know which solution is the best to their needs, and it would also help users to know the design problem better. This study may conduct a bridge that allows the scientific decision making approach can be used to solve the bioprocess design problem.

5.8 CONCLUSIONS

This chapter presents the development of chromatography system for the chromatographic process design problem. In this system, 57 parameters were used to represent the column operation experimental data that is mainly considered in the chromatographic process development. Following the approach introduced in Chapter 3, ontologies, theoretical and empirical knowledge were captured and developed for the chromatography system. These represented column data and chromatographic knowledge illustrate that the data and knowledge representation of BDKF approach can be applied for any bioprocess step.

The hierarchical heuristic approach (HHA) was introduced to solve the chromatographic process design problem. It decomposes the complicated design problem into multiple sub-design problems and then generates solution in a specific hierarchical order. Two scenarios about the pAb purification were used to illustrate that the HHA is an effective way to generate solutions to multiple variables of chromatography. It also indicates that the HHA is a promising way to solve the general bioprocess design problem.

Further analyse about how the criteria and the output performance requirement impact the solutions generated by HHA suggests that if users have specific expectations of the output performance or use narrow criterion setting, then the different solutions will be generated from the different hierarchical orders. These solutions may help users to further understand the design problem in order to find the optimized solution.

Chapter 6

DEVELOPMENT OF BDKF APPROACH FOR BIOPROCESS SEQUENCE DESIGN

6.1 INTRODUCTION

generally speaking, the bioprocess sequence design consists of sequence construction and sequence synthesis. Sequence construction is selecting a set of bioprocess steps to produce and purify the target biomolecule. Sequence synthesis is to determine the operating conditions of each bioprocess step and then combine them for the optimized performance, e.g. cost of good.

Usually, the sequence construction is achieved by engineers' experience. It is proposed that the sequence construction can be finished by BDKF approach automatically. However, due to the time limitation, this work will not be discussed in this thesis, but suggestions of how to do it are given in Chapter 7 as future work.

For sequence synthesis, the operating conditions of individual bioprocess steps need to

be identified first. Centrifugation case study explained that how to use the BDKF approach to solve the specific bioprocess design problem, e.g. what is the equipment? what is flowrate? Chromatography case study illustrated that how to use the BDKF approach to solve the general bioprocess design problem, e.g. what are solutions to the purification of IgG? These two case studies proved that the BDKF approach is a promising approach to generate solutions to bioprocess step by harnessing the bioprocess data and knowledge. Once the operating conditions of bioprocess steps have been identified, the next step is to combine them. The combination has two steps, first is to allow the selected bioprocess steps work systematically, and second is to synthesis the sequence operation based on the operating conditions of individual bioprocess steps.

The sequence operation synthesis is constrained by specific requirements, e.g. high yield, high purity, short time, low cost. These requirements are internal restrict, e.g. high yield brings high cost that attacks the low requirement. So the synthesis aims to achieve the trade-off between these requirements. It is a complicated work that has been addressed by several techniques, e.g. mathematical modelling, simulator (see section). It will not be discussed in this thesis since it may require another 3 years work. But the feasible solutions are given in Chapter 7.

This chapter focuses on the question: how to make the selected bioprocess steps work together? For this, a specific three step sequence consisting of centrifugation, filtration and chromatography is used for illustration. Then, the general conclusions of using BDKF approach for bioprocess sequence design are given.

For this, in this chapter, the information of the case study will be introduced in section 6.2, then the methods of coordinating the different BDKF systems work and conjoining the solutions generated by each of BDKF system will be explained in section 6.3, following the results of case study as well as the discussion and conclusion in section 6.4 and 6.5 respectively.

6.2 INFORMATION OF CASE STUDY

6.2.1 Background information

A three-step sequence that includes centrifugation, filtration and chromatography to recover the IgG product from the mammalian cell culture broth was considered in this case study. The sequence is shown in Figure 6.1

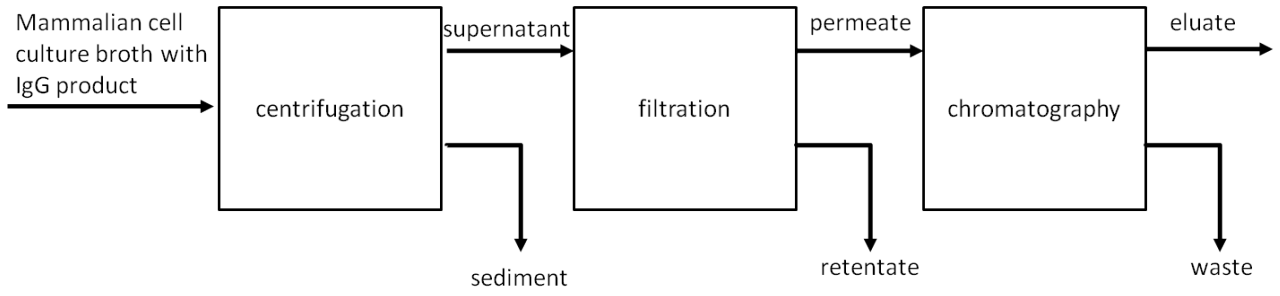


Figure 6.1: A three-step sequence consisting of centrifugation, filtration and chromatography

Given the mammalian cell culture broth, the centrifugation removed the whole cells in the broth, then the supernatant stream with the IgG product flowed through the filtration where the remaining solids was removed by filtration, finally the permeate stream containing the IgG product was loaded on to the column and the elute was collected and used for the next bioprocess step.

6.2.2 Information of feed material

The information of the feed material is given in Table 6.1.

Table 6.1: Information of mammalian cell culture broth components

Components	**Value**
mass of IgG product	100 g
mass of total protein	200 g
mass of mammalian cell	10.5 kg
mass of liquid	90 kg
volume of mammalian cell	10 L
volume of liquid	90 L
feed volume	100 L
product concentration	1.0 mg/mL
total protein concentration	2.0 mg/mL

6.2.3 Operating information of centrifugation, filtration and chromatography

The operating conditions and the performance requirement of centrifugation, filtration and chromatography are given in Table 6.2.

Table 6.2: Information about centrifugation, filtration and chromatography

Bioprocess step	**Operation information**	**Performance requirement**
Centrifugation	CSA-1 centrifuge pilot scale flowrate(X1)	at least 90% CE 70% dewatering level
Filtration	dead-end filtration foundabac nutsche filter constant rate operation mode 0.012 m^2 membrane area	no more than 2% liquid loss
Chromatography	chromatography type(X2) equilibration buffer chemical(X3) equilibration buffer chemical concentration(X4) equilibration buffer pH(X5) loading flowrate(X6)	at least 10 mg/mL DBC 10% breakthrough point

The dewatering level indicates how much liquid remains in the sediment, the 70% is a general dewatering level of disk stack centrifuge for intermittent discharge (Wang, 2007). Liquid loss indicates the mass of liquid loss when the processing material flows through the filtration step and 2% is specified in the output. Generally, the higher filtrate mass loss is, the more products are lost, therefore the criterion used to constrain this feature would be [0, 2%]. The filtrate mass loss included in relevant datapoints should be less than 2%, hence the predicted filtrate mass loss generated by averaging these values should be less than 2%. However, in this case study, the predicted filtrate mass loss is assumed as exact 2% because no filtration experimental data has been captured yet. This assumed value will be used for mass balance analysis of the filtration. The DBC 10 mg/mL is assumed as the minimal performance of the column. The 10% breakthrough means that the loading step is suspended when the effluent concentration achieves the 10% of initial IgG concentration.

The 'X' indicates the design variable to be determined. In this sequence design problem, there are six design questions. For the centrifugation, the question is what the flowrate should be. For the chromatography, the column type, equilibration buffer components and loading flowrate are concerned.

The solutions to the design of centrifugation and chromatography can be obtained by the centrifugation and chromatography system that have been developed. The information about filtration is proposed to be harnessed by filtration system which has not been developed. The development of filtration system is same as the centrifugation and chromatography system, e.g. identifying the parameters for filtration experimental data representation, capturing and representing the ontologies, theoretical and empirical knowledge. 18 parameters identified to represent the filtration experimental data are given in Table 6.3.

Table 6.3: Parameters of input, step and output for filtration experimental data representation

	Parameter	**Definition**
Input	strain	name of cell line
	product	name of target biomolecule
	liquid viscosity	viscosity value of processing material
	solid density	value of solid density
	liquid density	value of liquid density
	product concentration	value of product concentration
	solid concentration	value of solid concentration
Step	filtration type	name of filtration type
	filter	name of filter used in filtration
	pressure drop	value of pressure difference before and after the filter
	filtration flowrate	value of the flowrate in filtration
	scale	name of filtration scale
	membrane area	value of membrane area used in filtration
	retention time	value of retention time in filtration
	operation mode	name of filtration operation mode
Output	flux	value of the average flux
	membrane capacity	value of membrane capacity
	filtrate volume loss	value of filtrate volume loss

6.3 SEQUENCE SYSTEM FOR THE THREE-STEP PROCESS DEVELOPMENT

A BDKF system, named as *sequence system*, is proposed to be developed to coordinate centrifugation, filtration and chromatography system to solve the three-step sequence design problem. Same as the other BDKF systems, the sequence system requires a design query, called *sequence design query*, to start. In the following, the logic of how the sequence system works is given first, followed by the representation of sequence design query and the

workflow of sequence system.

6.3.1 Pseudo code of sequence system

To illustrate the programming logic of the sequence system, the pseudo code is employed and shown in Figure 6.2

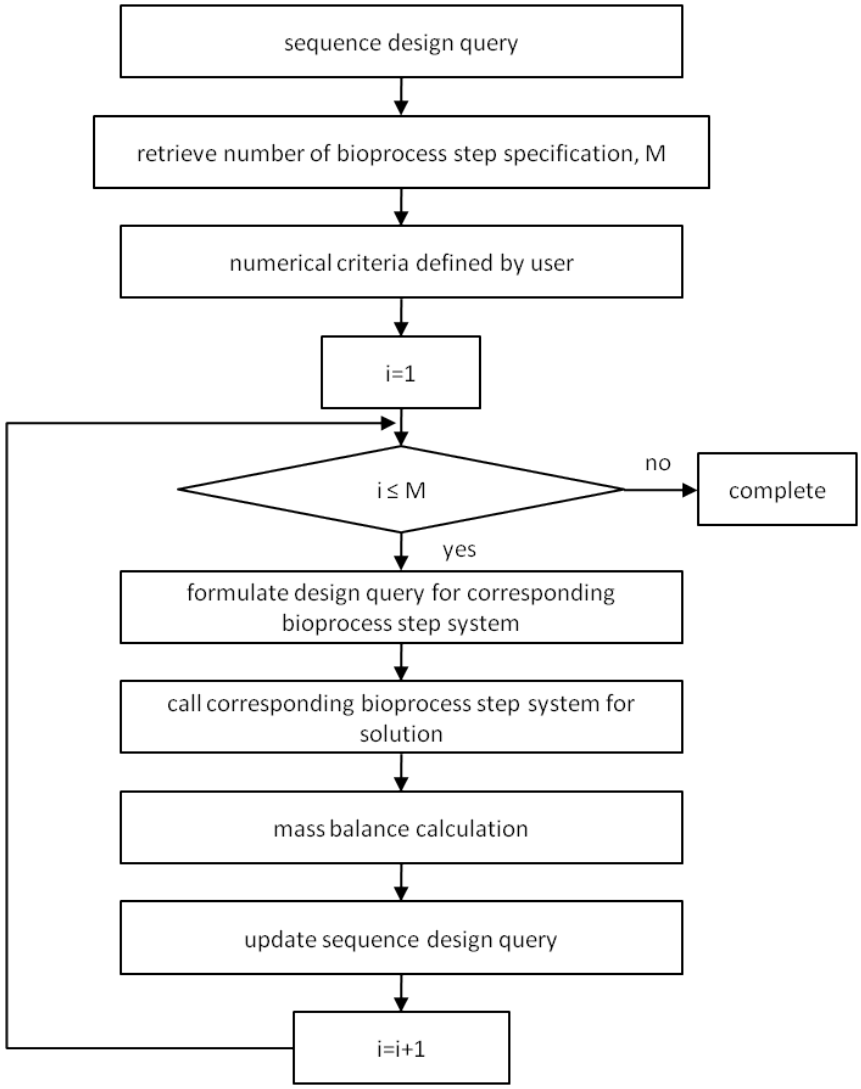


Figure 6.2: Pseudo code of sequence system to coordinate the required BDKF systems

The explanations are given in the following:

1. The sequence design query defines which BDKF system should be called.

2. The number of specifications of bioprocess steps determines how many BDKF systems should be called (M=3), i indicates the ith bioprocess step in the sequence.

3. The numerical criteria are used to constrain the search relevant datapoints in each of BDKF system.

4. i⩽M (start a loop of using BDKF system for solution to ith bioprocess step).

   (a) If yes

      i. Formalize the design query to the ith bioprocess step.

      ii. Call the corresponding BDKF system for solutions.

      iii. Mass balance analysis based on the predicted performance.

      iv. Update the features of sequence input.

      v. Start a new loop to generate solution to the next bioprocess step (i=i+1 and go to step 4).

   (b) If no, the loop stops and all results are shown.

6.3.2 Representation of sequence design query

The sequence design query represents the sequence design problem to be solved and it consists of three parts, i.e. *sequence input*, *sequence step* and *sequence output*.

The *sequence input* indicates the properties of processing material, e.g. strain, product. The volume information of processing material also should be specified in the sequence input. For instance, the sequence design query of the three-step sequence includes the features: 'total volume(100, L)', 'solid volume(10, L)' and 'liquid volume(90, L)'.

The *sequence step* indicates the information of bioprocess steps included in the sequence, i.e. the name of bioprocess steps and the operating conditions of each bioprocess step. The names of bioprocess steps consisted in the sequence should be in the order that they present in the sequence, such as 'step(centrifugation, n/a), step(filtration, n/a), step(chromatography,

n/a)' for the three-step sequence. The operating conditions of each bioprocess step should be represented by the parameters defined in the corresponding BDKF system. For instance, the information of centrifugation operating conditions should be represented by the parameters used by the centrifugation system.

The *sequence output* indicates the performance requirement of each bioprocess step. For instance, the CE, liquid loss, breakthrough point and DBC.

The requested information should be represented by the corresponding parameter associated with 'X'. The sequence design query of the three-step sequence is shown in Table 6.4, where each item of information included in the sequence input, sequence step and sequence output is represented as a feature consisting of parameter, value and unit.

Table 6.4: Representation of design query about the three-step sequence case study

Sequence input:	**Sequence step**:
strain(mammalian, n/a)	step(centrifugation, n/a)
product(IgG, n/a)	scale(pilot, n/a)
solid density(1.05, kg/m^3)	centrifuge(CSA-1, n/a)
liquid density(1.00, kg/m^3)	flowrate(X1, L/h)
product concentration(1, mg/mL)	step(filtration, n/a)
total protein concentration (2, mg/mL)	filtration type(dead end, n/a)
total volume (100, L)	filter(foundabac nutsche, n/a)
solid volume (10, L)	operation mode(constant rate, n/a)
liquid volume (90, L)	membrane area(0.012, m^2)
Sequence output:	step(chromatography, n/a)
CE(90, %)	function(capture, n/a)
filtrate mass loss (2, %)	chromatography type(X2, n/a)
breakthrough point (10, %)	equilibration buffer chemical(X3, n/a)
DBC(10, mg/ml)	equilibration buffer chemical concentration(X4, mM)
	equilibration buffer pH(X5,1)
	loading flowrate(X6, cm/s)

6.3.3 Coordination of three BDKF systems

Given the sequence design query, the sequence system coordinates the required BDKF systems to solve the design problem. The features contained in the sequence step will tell the sequence system which BDKF system should be harnessed. The order of the bioprocess step features presented in the sequence design query indicates the order of the BDKF systems to be coordinated. In the sequence design query of the three-step sequence, the features of 'step(centrifugation, n/a)', 'step(filtration, n/a)' and 'step(chromatography, n/a)' indicate the centrifugation system was used first, followed by the filtration and chromatography system respectively.

When the corresponding BDKF system is called, the following three steps are executed by the sequence system.

- Formalizing a design query to the BDKF system by selecting the suitable features from the sequence design query;
- Analysing the mass balance based on the predicted performance generated by the BDKF system;
- Updating the corresponding features of the sequence input.

When the three steps have been completed, the sequence system coordinates the next BDKF system, until all of required BDKF systems have been called. For the three-step sequence, the three BDKF systems are coordinated by sequence system and the procedure is shown in Figure 6.3.

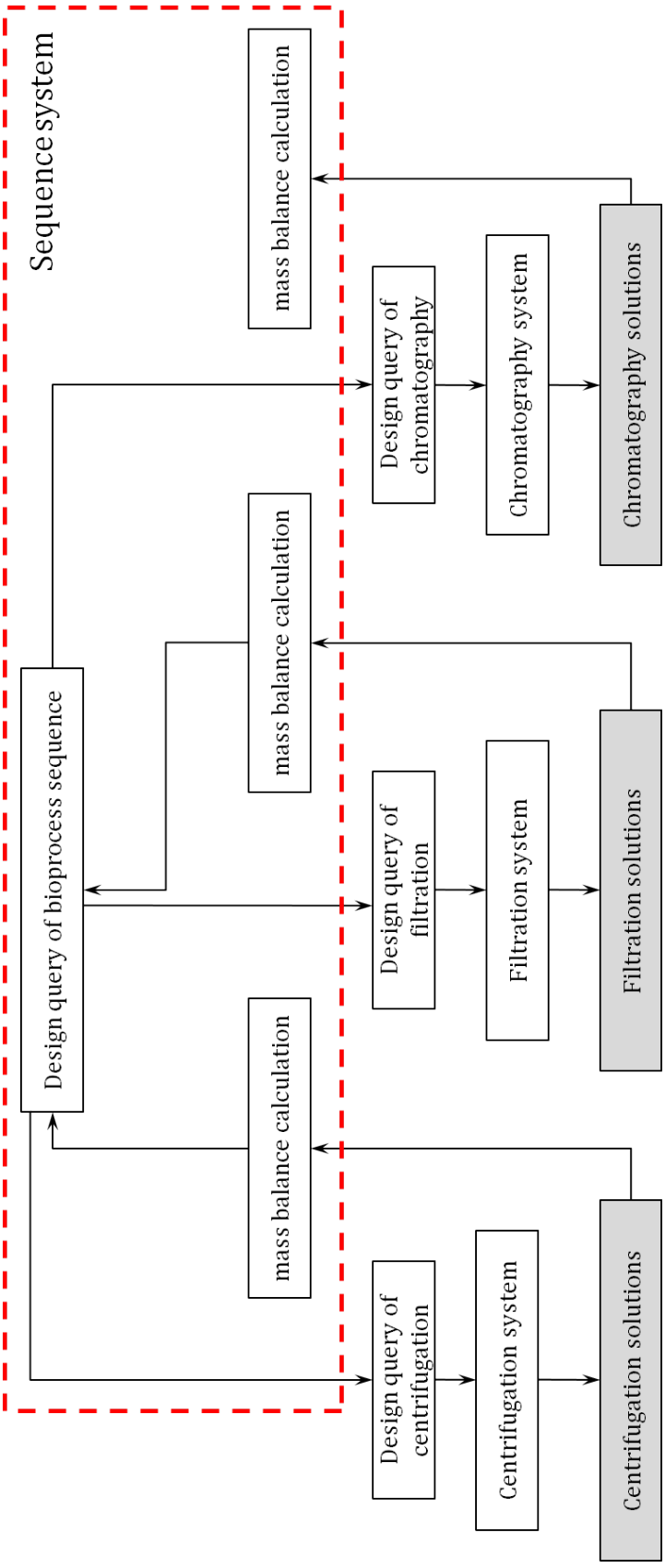


Figure 6.3: Flowchart of using sequence system to coordinate the three BDKF systems

For the centrifugation system, the sequence design query was used to formalize the design query, then the predicted CE was used for mass balance calculation and the results of mass balance updated the sequence input. Same procedure was applied to the filtration system and the chromatography system. The sequence system completed the work when solutions were generated from the chromatography system. The details of each of three steps are introduced in the following.

6.3.3.1 Design query formalization

The design query is formalized by selecting the suitable features from the sequence design query, and it is given to the BDKF system to generate solutions. Selecting the suitable features is proposed to be realized by the parameters employed by the corresponding BDKF system for experimental data representation. For example, the 34 parameters employed by centrifugation system are used to select the features from the sequence design query to formalize the design query to the centrifugation system.

In the sequence system, the parameters of each BDKF system are proposed to be arranged as a string list, called a parameter list. Each BDKF system has a parameter list. For this three-step sequence, the three parameter lists were contained in the sequence system, namely centrifugation parameter list, filtration parameter list and chromatography parameter list. The centrifugation parameter list consisted of 34 parameters (Table 4.1, 4.2 and 4.3;), while the filtration and chromatography parameter list included 18 parameters (Table 6.3) and 57 parameters (Table 5.3, 5.4, 5.5, 5.6, 5.7, 5.8, 5.9 and 5.10) respectively.

The corresponding parameter list is concerned when a specific BDKF system is called by the sequence system. For example, when the centrifugation system is called, only the centrifugation parameter list is considered to select the suitable features. For any specific feature, if its parameter is included in the parameter list, then this feature should be selected to formalize the design query. For instance, the feature centrifuge(disk stack, n/a) included in the sequence step was selected to formalize the design query to the centrifugation system, because the ‘centrifuge’ appeared the centrifugation parameter list. The features from the

sequence input, step and output are used as the input, step and output features of the design query.

For the three-step sequence, the design query of centrifugation system serves as an example shown in Table 6.5, where 5 features about the input were selected from the 9 features included in the sequence input, 3 features came from the 13 features of the sequence step, and 1 feature in output was from the 4 features consisted in the sequence output. The parameters of the 9 features were all in the centrifugation parameter list.

Table 6.5: 9 features selected from the sequence design query to formalize the design query to centrifugation system

Input	**Step**	**Output**
strain (mammalian, n/a)	scale (pilot, n/a)	CE (90, %)
product (igg, n/a)	centrifuge (CSA-1, n/a)	
product concentration (1, mg/ml)	flowrate (X1, L/h)	
solid density (1.05, kg/m^3)		
liquid density(1.00, kg/m^3)		

Based on the formalized design query, the BDKF system provides the possible performance and requested information, e.g. the predicted CE and requested flowrate provided by centrifugation system for this design query.

6.3.3.2 Mass balance analysis

The mass of processing material compositions is changed after each bioprocess step, e.g. solid mass. These changes would impact the operation setting of the following bioprocess steps. The mass changes of processing material between different bioprocess steps can be processed by mass balance. Mass balance is a simple way to analysis the interactions between different bioprocess steps, because many complex situations are simplified by investigating the movement of mass and equating what comes out (Walsh, 2003) (see equation

(6.1)).

$$mass\ in = mass\ out \tag{6.1}$$

The *mass in* represents the input mass of the bioprocess step, e.g. processing material, while the *mass out* describes output mass of the bioprocess step, e.g. supernatant after the centrifugation.

For any specific bioprocess step, the mass in equals to the mass out, e.g. in centrifugation, the mass of product of the feed stream equals to the mass of product contained in the sediment and supernatant. For any specific component of the processing material, the mass in equals to the mass output, e.g. the protein mass in equals to the protein mass out, the liquid mass in equals to the liquid mass out.

The input mass of bioprocess step can be calculated by harnessing the corresponding features of sequence input, e.g. the mass of product can be calculated by the features of total volume and product concentration. The output mass of the bioprocess step can be calculated by the predicted performance, e.g. the solid mass in the supernatant after centrifugation can be calculated by the predicted CE.

The equations used by the mass balance analysis can be captured as the theoretical knowledge. The representations of these calculations requires the mathematical relationship and the parameters referred in the mathematical relationship. For this three-step sequence case study, all of mass balance equations are be given in Appendix, where the variables on left side of equal sign are the output of equation and the variables on the right side of equal sign are the input of equation.

6.3.3.3 Update of sequence input features

The results generated from the mass balance is used to replace the initial value of the feature included in the sequence input. For example, the centrifugation removes the solid

particles from the liquid, hence the volume of feed material would reduce. The new volume of feed material can be calculated by mass balance, and it is used to replace the initial feed volume in the sequence input. The new feature is available for the following BDKF system to use.

6.4 RESULTS OF THE CASE STUDY

This section shows the results generated by each of three BDKF systems and the sequence system, which includes a design query, a predicted output and a mass balance table.

6.4.1 Centrifugation results

The design query of centrifugation system formalized by sequence design query has been shown in Table 6.5. The numerical criteria required for finding relevant datapoints were set as 10% (Table 6.6).

Table 6.6: Numerical criteria of the formalized design query for the centrifugation system about three-step sequence case study

specification	**numerical range**
product concentration (1, mg/mL)	[1×(1-10%), 1×(1+10%)]
solid density (1.05, kg/m^3)	[1.05×(1-10%), 1.05×(1+10%)]
liquid density (1.00, kg/m^3)	[1.00×(1-10%), 1×(1+10%)]
CE (90, %)	[90, +∞]

The results from the centrifugation system include the predicted and the flowrate which are:

- predicted CE is 97.2%
- flowrate for CSA-1 centrifuge is 180.7 L/h

After centrifugation, the feed material was separated into two phases, supernatant and sediment. The mass of components with respect to each phase were calculated by using the predicted CE and dewatering level. The results are shown in Table 6.7.

Table 6.7: Mass balance sheet of centrifugation

Component	**Feed stream** (g)	**Supernatant stream** (g)	**Sediment stream** (g)
Product	100	95.4	4.6
Solid	10500	294	10206
Liquid	90000	85834.3	4165.7
Total protein	200	190.7	9.3

The compositions of feed material after the centrifugation are shown in Table 6.8.

Table 6.8: Volume information of feed material compositions after centrifugation

	Supernatant stream	**Sediment stream**
Solid volume (L)	0.3	9.7
Liquid volume (L)	85.8	4.2
Total volume (L)	86.1	13.9

The supernatant stream was the product stream that would be processed by filtration. Based on the mass of product and total protein as well as the volume of supernatant stream, the product concentration and total protein concentration were calculated as 1.1 mg/mL and 2.2 mg/mL respectively. Five features: 'product concentration(1.1, mg/mL)', 'total protein concentration(2.2, mg/mL)', 'liquid volume(85.8, L)', 'solid volume(0.3, L)' and 'total volume(86.1, L)', replaced the initial features: 'product concentration(1, mg/mL)', 'total protein concentration(2.0, mg/mL)', 'liquid volume(90, L)', 'solid volume(10, L)' and 'total volume(100, L)', in the sequence design query which were used by the filtration system.

6.4.2 Filtration results

The design query for the filtration system is shown in Table 6.9, where some of the features were inherited from the Table 6.4.

Table 6.9: Design query consisting of 11 features selected from the sequence design query for the filtration system

Input	**Step**	**Output**
strain (mammalian, n/a) product (igg, n/a) product concentration (1.1, mg/ml) total protein concentration (2.2, mg/ml) solid density (1.05, kg/m^3) liquid density(1.00, kg/m^3)	filtration type(dead end, n/a) filter (foundabac nutsche, n/a) operation mode(constant rate, n/a) membrane area(0.012, m^2)	filtrate mass loss (2, %)

Because the filtration system has not been established yet, the queried parameter was not be involved. But the 8 step parameters and 3 output parameters in Table 6.3 can be used to represent the queried operating conditions and specific performance requirement for any specific filtration design query. After the filtration, the feed stream was separated into two streams, i.e. retentate and permeate. The feed stream components in these streams had different mass which were determined by the filtration specific output, i.e. filtrate mass loss that was assumed as 2%.

Table 6.10: Mass balance sheet of filtration

Components	**Feed stream** (g)	**Retentate stream** (g)	**Permeate stream** (g)
Product	95.4	1.9	93.5
Solid	294	294	0
Liquid	85834.3	1716.7	84117.6
Total protein	190.7	3.8	187.0

The compositions of feed material after the filtration are shown in Table 6.11.

Table 6.11: Volume information of feed material compositions after filtration

	Retentate stream	**Permeate stream**
Solid volume (L)	0.28	0
Liquid volume (L)	1.7	84.1
Total volume (L)	1.98	84.1

After the filtration, the retentate stream was discarded and the permeate stream was carried forward to chromatography step as feed stream. Based on the mass balance results and the volume information, the product concentration and the total protein concentration in permeate stream remained the same. The three features: 'liquid volume(84.1, L)', 'solid volume(0, L)' and 'total volume(84.1, L)', were used to replace the corresponding features in sequence input to be used for chromatographic process development.

6.4.3 Chromatography results

The design query to chromatography system that was formalized by the sequence system is shown in Table 6.12.

Table 6.12: Design query consisting of 13 features selected from the sequence design query for the chromatography system

Input	**Step**	**Output**
strain (mammalian, n/a) product (IgG, n/a) product concentration (1.1, mg/ml) total protein concentration (2.2, mg/ml)	function (capture, n/a) chromatography type (X2, n/a) equilibration buffer chemical (X3, n/a) equilibration buffer chemical concentration (X4, mM) equilibration buffer pH(X5, n/a) loading flowrate(X6, cm/s)	DBC(10, mg/mL) breakthrough point(10, %)

The criteria of numerical specifications are given in Table 6.13.

Table 6.13: Numerical criteria for design query of chromatography system about three-step sequence case study

parameter	**numerical range**
product concentration (1, mg/mL)	[1.12×(1-10%), 1.12×(1+10%)]
total protein concentration(2.23, mg/mL)	[2.23×(1-10%), 2.23×(1+10%)]
breakthrough point (10, %)	[10×(1-0%), 1×(1+0%)]
DBC (10, mg/mL)	[10, +∞]

For this design query, the HHA was employed for generating solutions that are summarized in the following:

- the predicted DBC is 33.6 mg/ml
- the IEX column is selected as chromatography type
- the MES, HEPES, NaAc is selected as equilibration buffer
- 16.7mM, 16.7mM, 16.7mM are selected as equilibration buffer concentrations
- 5.5 is selected as pH of equilibration buffer
- 0.0424 cm/s is retrieved as loading flowrate

For chromatography mass balance analysis, the yield and purity are usually considered to calculate the mass of product and total protein product in the feed stream. However, the experimental data involved in the chromatography system concerned on studying the interactions between the chromatography operation conditions and DBC, hence the yield and purity data was not available. The calculations used to instead of the specific value are shown in the mass balance sheet of chromatography(Table (6.14)). The *yield*, *purity* and *cuttingvolume* can be retrieved from the specific relevant datapoint.

Table 6.14: Mass balance sheet of chromatography

Components	**Feed stream** (g)	**Elute** (g)	**Waste** (g)
Product	84.1	$84.1 \times yield$	$84.1 \times (1 - yield)$
Liquid	85.8	$\frac{cuttingvolume}{liquiddensity}$	$85.8 - \frac{cuttingvolume}{liquiddensity}$
Total protein	168.2	$\frac{84.1 \times yield \times (1-purity)}{purity}$	$168.2 - \frac{84.1 \times yield \times (1-purity)}{purity}$

The total volume used in the sequence input would be replaced by the cutting volume, because the contaminates included in other liquid would be discarded while the liquid collected from the cutting point would flow through the following bioprocess step.

Since the chromatography was the last step specification in sequence step, the sequence system stopped working after the solutions have been generated by the chromatography system. All of the results, i.e. solutions regarding centrifugation, filtration and chromatography with the mass balance results were shown to the user to assess the bioprocess sequence performance and efficiency.

6.4.4 Summary

The sequence system only coordinates required BDKF systems and do the mass balance analysis, the solutions of each bioprocess step are generated by corresponding BDKF systems. The Chapter 4 proves that the BDKF system can make good prediction and useful suggestion, therefore the results provided by sequence system expect to be accurate. However, it is suggested that users should use the individual BDKF systems to identify the required information and then use experiments to evaluate these solutions. These information may help sequence system to generate better results. In addition, using the narrow numerical criteria would also help to produce the accurate results.

For the three-step sequence, 29.06 milliseconds (ms)[1] were cost by centrifugation system for results generation while 574.76 ms were cost by chromatography system for exploring a

[1]The running time was generated based on the laptop computer (Intel Core2 Duo 2.53 GHz, 4GB)

larger database. The time required by filtration system was not available because it has not been established. The time of sequence system used for the three-step sequence expects to be less than 1 second which shows the BDKF is an efficient way to solve the sequence design problem.

6.5 DISCUSSIONS

Based on this case study, the elements of the sequence system as well as the guidelines for using sequence system to solve the specific bioprocess sequence design problem are introduced in the following.

6.5.1 Elements of sequence system development

In this section, the primary elements of the sequence system development are introduced which include three types of data and knowledge.

6.5.1.1 Parameters for representation of volume information

The mass balance processed by sequence system requires the volume information of feed material. Therefore, it is necessary to specify the volume features of processing material in the sequence design query. To represent the volume information, any parameter that is not included in the BDKF system but referred by the mass balance equations should be captured. For the three-step sequence, three parameters were captured, i.e. total volume, solid volume and liquid volume. These three parameters were appeared in the mass balance equations, but not included in centrifugation system or filtration system or chromatography system. The parameter setting of volume information representation is expandable when more volume information is specified in the sequence design problem.

6.5.1.2 Ontologies of bioprocess steps

The bioprocess steps included in the bioprocess sequence are required to be specified in the sequence design query. These bioprocess step features allow the sequence system to

know which BDKF system should be coordinated. In order to allow the sequence system to recognize these features, the bioprocess step ontologies that define the bioprocess step terms should be developed. In the ontologies, each class is a specific bioprocess step term, e.g. centrifugation. For the protein purification, the bioprocess step ontologies are given in Figure 6.4, where the arrow indicates the relationship 'a type of'.

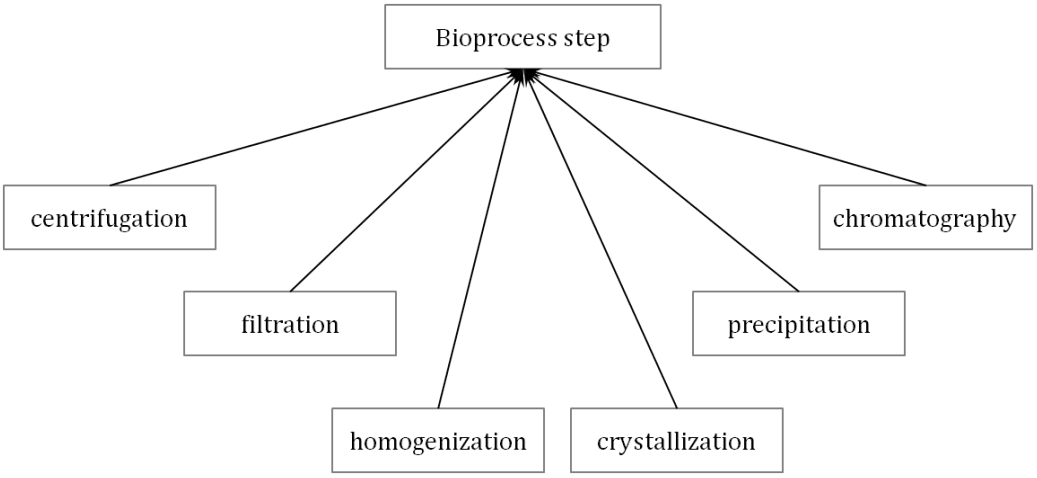


Figure 6.4: Bioprocess step ontologies for bioprocess sequence

6.5.1.3 Fundamental equations for mass balance

The key task of sequence system is to calculate the mass balance of each bioprocess step included in the bioprocess sequence. Any equation used by mass balance analysis should be captured and stored in the knowledge based of sequence system. For the three-step sequence case study, 25 mass balance equations have been captured (see Appendix). All of these captured equations aim at calculating the mass of processing material components based on the predicted performance.

6.5.2 Guidelines of sequence system implementation

Based on the three-step sequence, the guidelines of sequence system development and utilization can be summarized in the following.

1. Guidelines for sequence design query formalization:

- The volume information of processing material should be specified in the sequence input;
- The name of bioprocess steps contained in the sequence should be specified in the sequence step while their order should be exactly same as the order of the bioprocess steps presented in the sequence;
- Each bioprocess step should give at least one defined performance requirement that should be presented in the sequence output;

2. Guidelines for sequence system implementation:

   - The sequence system aims at coordinating the BDKF systems to provide solutions to the sequence design problem, hence no database is required to be established;
   - For each coordinated BDKF system, the three steps introduced in section 6.3.2 are repeated. This procedure stops when the mass balance calculation of last bioprocess step is finished;
   - The solutions generated by each BDKF system as well as the results of mass balance analysis in regard to each bioprocess step are presented together when the coordination finishes.

For some specific bioprocess sequence, some materials, e.g. buffer, are required to be fed in the bioprocess step in purpose, e.g. dilute the processing material. The quantitative information of these extra material changes the feed stream properties, e.g. mass of liquid. For this, volume of extra buffer would be specified in the sequence step and should follow the corresponding step specification. For instance, if the 50 L extra buffer is used to dilute the product in chromatography, then the specification, e.g. buffer (50, L), should be presented after the step specification, step(chromatography). This indicates the 50 L buffer is used in chromatography step and should be used by the mass balance analysis of chromatography step. This situation should be further studied by using different scenarios in order to conduct a well defined rule.

Besides the mass balance analsyis, other analysis about sequence design may be included in the sequence system, such as the time scheduling and cost analysis. The time consumed by specific bioprocess step be calculated based on theoretical knowledge.

6.6 CONCLUSIONS

For bioprocess sequence design problem, the sequence system which is the implementation of BDKF approach for bioprocess sequence design is discussed. The sequence system is considered as an agent that can systematically coordinate the different BDKF systems to work together. If all of the BDKF systems are established, then the sequence system can solve any sequence design problem. The three-step sequence is used as an example to demonstrate how the sequence system can solve the bioprocess sequence design problem. It demonstrates that the BDKF approach can be also used to solve the sequence design problem efficiently.

Chapter 7

CONCLUSIONS AND FUTURE WORK

7.1 INTRODUCTION

The bioprocess development requires substantial experiments to identify the operating conditions of each individual bioprocess step involved in the bioprocess sequence regarding performance requirement. These experiments have accumulated considerable experimental data and knowledge. A computational tool for exploring this data and knowledge could help to narrow down the design space to be investigated. This chapter summarizes the efforts achieved in this research project to harness the bioprocess data and knowledge systematically for bioprocess design problem so as to enhance efficiency of decision-making in bioprocess development. It also introduces the future work that will further enhance bioprocess data and knowledge utilization.

7.2 CONCLUSIONS

The whole thesis has developed a general computational framework called bioprocess data and knowledge framework (BDKF). It can systematically represent and reason with the inconsistent and incomplete bioprocess data and knowledge for solutions to the bioprocess step or sequence design problems.

Four types of data and knowledge were considered in BDKF approach, namely exper-

imental data, ontology, theoretical and empirical knowledge introduced in Chapter 3. The experimental data came from the previous experiments, and its representation can organize the inconsistent and incomplete bioprocess data in a defined structure to benefit data utilization. The ontology described the terms appeared in bioprocess data and knowledge that was represented as the parent-child class in the hierarchical tree, and the terms involved in the hierarchical tree can be searched by associated relationships. The theoretical knowledge was the general bioprocessing principles introduced in text books. Mathematical relationship and entity relation model (ERM) were used to represent the fundamental equations and background information of equipment respectively, and this type of knowledge could assist the data analysis. The empirical knowledge was the new findings published in journal papers. It served as the specific rules for searching heterogeneous experimental data. The experimental data formed the database while the other three types of knowledge constructed the knowledge base.

The design query which indicated the design problems was the input of BDKF approach. Its representation was same as that of experimental data. The information required in the design query depended on the questions. The design query allowed the users to express any design problem based on the available information. Three reasoning functionalities were developed to access the database and knowledge base, namely search, prediction and suggestion functionality. The search functionality returned a set of experimental data that was relevant to the design query. Based on the results, the prediction functionality provided a predicted performance and the suggestion functionality generated solutions for further experimentation to realize the desired performance. The benefits of the design query and reasoning functionalities were highlighted: the design query conferred maximal flexibility since it allowed users to use limited information to access the database and knowledge base, especially at the early stage of bioprocess development; the search, prediction and suggestion functionality performed a natural logic reasoning to reuse previous data and knowledge for new design problem. The prediction and suggestion functionality also can be carried out by other explicit and complicated algorithms in order to extract useful information from accumulated data and knowledge.

With such system, the accumulated experimental data can be utilized directly for bioprocess design or answer process related quarries by any researchers, which provide initial evidence for decision-making. No such tools are available for the industry to harness these accumulated experimental data and knowledge in such efficient way.

The challenge of implementing the conceptual work of BDKF approach into a practical system was explained in Chapter 4. The representation of centrifugation experimental data promoted understanding about how the inconsistent experimental data be formalized. The ontologies built for centrifugation illustrated how the centrifugation terms were formulated as a hierarchical tree with the affiliated relationship and how they were harnessed for experimental data searching, e.g. if yeast is considered, all of the experimental data whose cell line terms were underneath the yeast will be searched. The clarification efficiency (CE) calculation and centrifuge background information showed what the centrifugation theoretical knowledge was and how it generated required information. The representation of ultra scale down (USD) approach demonstrated what the empirical knowledge was and how it was represented as well as how it was harnessed for searching experimental data generated from the different experimental scales. A case study about retrieving flowrate to the desired CE performance was used to explain how to turn the limited design information and question into a specific design query to access the centrifugation data and knowledge. With the same design query, the reasoning process of search, prediction and suggestion functionality showed how to select the relevant experimental data and how to analysis the data. The other case study about pilot scale centrifugation design was used to illustrate how the theoretical and empirical knowledge can be automatically harnessed for solutions. The centrifugation system created a more flexible environment for screening the feasible solutions to facilitate the bioprocess design than that found in conventional bioprocess design tools. The flowchart of using centrifugation system to generate the flowrate solution proved that any specific operating condition can be queried and retrieved from the relevant experimental data. This would facilitate the simple design task so as to promote the efficiency of bioprocess development, e.g. which equipment is best? what is the possible performance of the design?

Evaluations were conducted to assess the performance of centrifugation system. 152 experimental data collected from different sources was used in evaluation. For prediction functionality, the results showed that the good predictions can be made and a clue of using prediction functionality was generated in the mean time, i.e. the more specific design query was given and the narrower criteria were used, the better prediction could be produced. For suggestion functionality, the results demonstrated that the useful suggestions can be produced to narrow down the design space to be explored. They also advised that the narrower criteria may not improve the quality of suggestions but reduce the chance of finding relevant experimental data. The evaluation results provided solid evidence that the BDKF approach can mining useful information to facilitate the design decision making, even if the design has limited information. The prediction would tell people what is the possible achievement, and it may help to formalize the objective of the design. The suggestion would tell people how to do the design, e.g. which design space should not be examined, and it may help to enhance the efficiency of bioprocess design by excluding the unnecessary experimentation.

A technique used to solve the general bioprocess design problem which consisted of multiple variables to be determined was discussed in Chapter 5. This technique, namely hierarchical heuristic approach (HHA), was illustrated by the chromatography system. The chromatography system had the same development procedure as the centrifugation system, but it included more complicated experimental data, ontologies, theoretical and empirical knowledge. Capturing and representing this data and knowledge showed how to further develop the established database and knowledge base. It showed that the BDKF approach provided an easy-development environment for further development. The problem about column selection, buffer conditions and flowrate with respect to the specific dynamic binding capacity (DBC) requirement was used to demonstrate how HHA approach works. This was an attempt to extend the BDKF approach to the general situation so that encompass all types of design problems faced during the bioprocess development. Comparing the results generated from all types of hierarchical order was used as the evaluation of HHA approach. It illustrated that the HHA approach can break down the complicated design problem into

a set of simple problems and the results may help users better understand the bioprocess design which may be helpful for finding optimized and robust solution.

Applying the BDKF approach on sequence design problem was considered and discussed in Chapter 6. The three-step downstream sequence including centrifugation, filtration and chromatography was employed as a case study. The sequence system was developed to represent the sequence design problem and to coordinate the required bioprocess step systems. The coordination was achieved by doing mass balance based on the results generated by individual systems. The three-step sequence illustrated how the information generated by mass balance was used to communicate with these four systems. The results and further development guidelines were given that would help to further apply the sequence system on bioprocess synthesis and optimization problem.

The work in this thesis highlights the benefits of developing a general computational framework to harness the bioprocess data and knowledge for the bioprocess design task. This has been illustrated by BDKF approach involving the data and knowledge representation as well as the reasoning functionalities. The BDKF approach serves as an agent for communication between engineers and all available bioprocess information. Furthermore, development of BDKF approach also aims at understanding about how a bioprocess step or sequence works and how to formulate the design task as first order logic problem. Effective use of results can lead to efficient bioprocess development efforts, more effective use of valuable processing materials, faster time-to-market, and improvement of competition strength and economic performance.

7.3 FUTURE WORK

The BDKF approach discussed in this thesis is an important contribution to the emerging field of computer-aided design tool that integrate the representation and utilization of inconsistent and incomplete bioprocess data and knowledge. It provides a solid base for further research topics, and several opportunities are highlighted and discussed in the following.

In addition to the centrifugation and chromatography discussed in this thesis, other bioprocess steps, e.g. fermentation, should be considered. The representation of data and knowledge will be challenge when applying the BDFK approach on the other bioprocess steps. For example, the oxygen concentration during the fermentation is a type of continuous data. How to represent different types of information is the key to success for BDKF to develop a whole bioprocess under rigorous constraints. While the sequence system can provide solution to a known sequence, the further development should concentrate on bioprocess synthesis.

This thesis focused on demonstrating the BDKF approach plainly so that the algorithms employed were simple, e.g. the performance prediction was generated by arithmetic average algorithm. Comprehensive but complicated algorithms may be needed for mining more useful information from bioprocess data and knowledge. For example, the clustering algorithm can arrange the relevant data into a set of groups, and each group can provide a specific design space with a reasonable performance prediction. Using complicated algorithms will not change the role of reasoning functionalities, but gain more information to help engineers to understand the bioprocess design better.

The mass balance was used as the simple example to illustrate how the sequence system coordinates the results generated by individual bioprocess step systems. More other models, e.g. economic analysis model, energy analysis model, should be involved and developed. Development of these models would be very helpful since they can extend the BDKF approach in optimization and synthesis study, e.g. which sequence is best and robust. Having linked the process and business models could formulate the optimization and synthesis problems to take account not only of the bioprocess performance but also the materials utilization and cost of goods.

The development language used in this thesis is WinProlog. Since it is not good at dealing with complicated mathematical calculations, hence other language, e.g. Java, could be used to process the complicated mathematical model. In addition, the design query perform

as an agent to formalize the users specifications of bioprocess design and communicate with data and knowledge base; hence a friendly user interface (UI) is required to facilitate the usage of BDKF approach. This UI should not only allow users to explicitly present the design information, but also offer a friendly environment to manipulate the requested reasoning functionalities and list the results in a comprehensive way, e.g. using vivid diagram to present the predicted and suggested results.

The case studies referred in this thesis served for demonstration of BDKF approach, and most of them had limited specifications. These case studies may give ideas about using BDKF approach for early stage bioprocess development when limited information is available. What needed is to apply the method to a whole process development case and evaluate the approach accordingly.

In conclusion, the future work has been outlined based on the results and limitations of the work discussed in this thesis. Using complicated data mining algorithm may improve the quality of prediction and suggestion while capturing more experimental data or industrial data would help to do sophisticated case study that can prove the efficiency and effectiveness of BDKF approach. In the following years, it is expected that using bioprocess data and knowledge will become a general and necessary tool for bioprocess design, especially for the automation of biopharmaceutical manufacturing.

Reference

Aamodt, A., Plaza, E., 1994. Case-based reasoning - foundational issues, methodological variations, and system approaches. Ai Commun 7 (1), 39–59.

Adamou, A., Palma, R., Haase, P., Montiel-Ponsoda, E., Cea, G., Gmez-Prez, A., Peters, W., Gangemi, A., 2012. The neon ontology models. Ontology Engineering in a Networked World, 65–90.

Ahuja, S., 2000. Handbook of bioseparations. Vol. 2. Academic Pr.

Al-Jibbouri, S., 2006. Scale-up of chromatographic ion-exchange processes in biotechnology. J. Chromatogr. A 1116 (1-2), 135–142.

Alavi, M., Leidner, D., 2001. Review: Knowledge management and knowledge management systems: Conceptual foundations and research issues. Mis Quart, 107–136.

Alford, J. S., 2006. Bioprocess control: Advances and challenges. Computers & Chemical Engineering 30 (1012), 1464–1475.

Ambler, C. M., 1959. The theory of scaling up laboratory data for the sedimentation type centrifuge. J Biochem Microbiol 1 (2), 185–205.

Anand, V., Kandarapu, R., Garg, S., 2001. Ion-exchange resins: carrying drug delivery forward. Drug Discov Today 6 (17), 905–914.

Anderson, J., 1996. The architecture of cognition. Lawrence Erlbaum Assoc Inc.

Antezana, E., Egaa, M., Blond, W., Illarramendi, A., Bilbao, I., De Baets, B., Stevens, R.,

Mironov, V., Kuiper, M., 2009. The cell cycle ontology: an application ontology for the representation and integrated analysis of the cell cycle process. Genome Biol 10 (5), R58.

Asenjo, J. A., Montagna, J. M., Vecchietti, A. R., Iribarren, O. A., Pinto, J. M., 2000. Strategies for the simultaneous optimization of the structure and the process variables of a protein production plant. Computers & Chemical Engineering 24 (910), 2277–2290.

Aslam Bhutta, M., Hayat, N., Bashir, M., Khan, A., Ahmad, K., Khan, S., 2012. Cfd applications in various heat exchangers design: A review. Appl Therm Eng 32, 1–12.

Astrom, K., 2011. A perspective on modeling and simulation of complex dynamical systems. In: Integrated Modeling of Complex Optomechanical Systems. International Society for Optics and Photonics, pp. 833602–833602–10.

Awan, M., Awais, M., 2011. Predicting weather events using fuzzy rule based system. Applied Soft Computing 11 (1), 56–63.

Bard, J., Rhee, S., Ashburner, M., 2005. An ontology for cell types. Genome Biology 6 (2), R21.

Batres, R., West, M., Leal, D., Price, D., Masaki, K., Shimada, Y., Fuchino, T., Naka, Y., 2007. An upper ontology based on iso 15926. Comput Chem Eng 31 (5-6), 519–534.

Bergander, T., Nilsson-Vaelimaa, K., Oberg, K., Lacki, K. M., 2008. High-throughput process development: Determination of dynamic binding capacity using microtiter filter plates filled with chromatography resin. Biotechnol Progr 24 (3), 632–639.

Bollati, V., Atzeni, P., Marcos, E., Vara, J., 2012. Model management systems vs. model driven engineering: A case study. In: Proceedings of the 27th Annual ACM Symposium on Applied Computing. ACM, pp. 865–872.

Boychyn, M., Yim, S. S. S., Bulmer, M., More, J., Bracewell, D. G., Hoare, M., 2004. Performance prediction of industrial centrifuges using scale-down models. Bioproc Biosyst Eng 26 (6), 385–391.

Boychyn, M., Yim, S. S. S., Shamlou, P. A., Bulmer, M., More, J., Hoare, A., 2001. Characterization of flow intensity in continuous centrifuges for the development of laboratory mimics. Chem Eng Sci 56 (16), 4759–4770.

Bulmer, M., Clarkson, A. I., Titchener-Hooker, N. J., Dunnill, P., 1996. Computer-based simulation of the recovery of intracellular enzymes and its pilot-scale verification. Bioprocess Eng 15 (6), 331–337.

Bulsari, A., Saxen, H., 1991. System identification of a biochemical process using feedforward neural networks. Neurocomputing 3 (3), 125–133.

Bussieck, M., Pruessner, A., 2003. Mixed-integer nonlinear programming. SIAG/OPT Newsletter: Views & News 14 (1), 19–22.

Cadoli, M., Donini, F., Liberatore, P., Schaerf, M., 2011. Space efficiency of propositional knowledge representation formalisms. arXiv preprint arXiv:1106.0233.

Canovas, M., Maiquez, J., Obon, J., Iborra, J., 2002. Modeling of the biotransformation of crotonobetaine into l()carnitine by escherichia coli strains. Biotechnol Bioeng 77 (7), 764–775.

Caracciolo, C., Heguiabehere, J., Gangemi, A., Baldassarre, C., Keizer, J., Taconet, M., 2012. Knowledge management at fao: A case study on network of ontologies in fisheries. Ontology Engineering in a Networked World, 383–405.

Casanovas, P., Casellas, N., Tempich, C., Vrandei, D., Benjamins, R., 2007. Opjk and diligent: ontology modeling in a distributed environment. Artificial Intelligence and Law 15 (2), 171–186.

Chabrier-Rivier, N., Fages, F., Soliman, S., 2005. The biochemical abstract machine biocham computational methods in systems biology. Vol. 3082 of Lecture Notes in Computer Science. Springer Berlin / Heidelberg, pp. 172–191.

Chan, G., Booth, A. J., Mannweiler, K., Hoare, M., 2006. Ultra scale-down studies of the effect of flow and impact conditions during e-coli cell processing. Biotechnol Bioeng 95 (4), 671–683.

Chan, J., Kishore, R., Sternberg, P., Van Auken, K., 2012. The gene ontology: enhancements for 2011. Nucleic Acids Res. 40 (D1), D559–D564.

Charaniya, S., Hu, W. S., Karypis, G., 2008. Mining bioprocess data: opportunities and challenges. Trends Biotechnol 26 (12), 690–699.

Charaniya, S., Le, H., Rangwala, H., Mills, K., Johnson, K., Karypis, G., Hu, W., 2010. Mining manufacturing data for discovery of high productivity process characteristics. J Biotechnol 147 (3), 186–197.

Chen, H., Stadtherr, M., 1985. A simultaneous modular approach to process flowsheeting and optimization. part i: Theory and implementation. Aiche J 31 (11), 1843–1856.

Chen, P., 1980. Entity-relationship approach to systems analysis and design. North-Holland.

Chepelev, L., Dumontier, M., 2011. Chemical entity semantic specification: Knowledge representation for efficient semantic cheminformatics and facile data integration. Journal of cheminformatics 3 (1), 1–19.

Chhatre, S., Bracewell, D. G., Titchener-Hooker, N. J., 2009. A microscale approach for predicting the performance of chromatography columns used to recover therapeutic polyclonal antibodies. J. Chromatogr. A 1216 (45), 7806–7815.

Chhatre, S., TitchenerHooker, N., 2009. Review: Microscale methods for highthroughput chromatography development in the pharmaceutical industry. J Chem Technol Biot 84 (7), 927–940.

Chua, C., Storey, V., Chiang, R., 2012. Knowledge representation: A conceptual modeling approach. Journal of Database Management (JDM) 23 (1), 1–30.

Cornelissen, E. R., Moreau, N., Siegers, W. G., Abrahamse, A. J., Rietveld, L. C., Grefte, A., Dignum, M., Amy, G., Wessels, L. P., 2008. Selection of anionic exchange resins for removal of natural organic matter (nom) fractions. Water Res. 42 (1-2), 413–423.

Cosenza, B., Galluzzo, M., 2012. Nonlinear fuzzy control of a fed-batch reactor for penicillin production. Computers & Chemical Engineering 36 (0), 273–281.

Courtot, M., Juty, N., Knpfer, C., Waltemath, D., Zhukova, A., Drger, A., Dumontier, M., Finney, A., Golebiewski, M., Hastings, J., 2011. Controlled vocabularies and semantics in systems biology. Molecular systems biology 7 (1).

Dai, J., Shieh, C., Sheng, Q., Zhou, H., Zeng, R., 2005. Proteomic analysis with integrated multiple dimensional liquid chromatography/mass spectrometry based on elution of ion exchange column using ph steps. Anal Chem 77 (18), 5793–5799.

Daichendt, M., Grossmann, I., 1998. Integration of hierarchical decomposition and mathematical programming for the synthesis of process flowsheets. Comput Chem Eng 22 (1), 147–175.

Demir, E., Cary, M., Paley, S., Fukuda, K., Lemer, C., Vastrik, I., Wu, G., D'Eustachio, P., Schaefer, C., Luciano, J., 2010. The biopax community standard for pathway data sharing. Nature biotechnology 28 (9), 935–942.

Demir, E., Cary, M., Paley, S., Fukuda, K., Lemer, C., Vastrik, I., Wu, G., D'Eustachio, P., Schaefer, C., Luciano, J., 2012. Corrigendum: The biopax community standard for pathway data sharing. Nature biotechnology 30 (4), 365–365.

Dhaliwal, J., Benbasat, I., 1996. The use and effects of knowledge-based system explanations: theoretical foundations and a framework for empirical evaluation. Information Systems Research 7 (3), 342.

DiMasi, J. A., Hansen, R. W., Grabowski, H. G., 2003. The price of innovation: new estimates of drug development costs. J Health Econ 22 (2), 151–185.

Ding, Y., Foo, S., 2002a. Ontology research and development. part 2 - a review of ontology mapping and evolving. J Inform Sci 28 (5), 375–388.

Ding, Y., Foo, S., 2002b. Ontology research and development. part i - a review of ontology generation. J Inform Sci 28 (2), 123–136.

Do, T., Ho, F., Heidecker, B., Witte, K., Chang, L., Lerner, L., 2008. A rapid method for determining dynamic binding capacity of resins for the purification of proteins. Protein Expres Purif 60 (2), 147–150.

Doran, P., 1995. Bioprocess engineering principles. Vol. 439. Academic Press London.

DORAN, P., 2011. Bioprocess engineering principles,(paper).

Douglas, J. M., 1985. A hierarchical decision procedure for process synthesis. Aiche J 31 (3), 353–362.

Duran, M., Grossmann, I., 1986. A mixedinteger nonlinear programming algorithm for process systems synthesis. Aiche J 32 (4), 592–606.

El-Halwagi, M., Srinivas, B., Dunn, R., 1995. Synthesis of optimal heat-induced separation networks. Chem Eng Sci 50 (1), 81–97.

Faeder, J., Blinov, M., Goldstein, B., Hlavacek, W., 2005. Rulebased modeling of biochemical networks. Complexity 10 (4), 22–41.

Fahrner, R. L., Whitney, D. H., Vanderlaan, M., Blank, G. S., 1999. Performance comparison of protein a affinity-chromatography sorbents for purifying recombinant monoclonal antibodies. Biotechnol Appl Bioc 30, 121–128.

Fan, Z., Lynd, L., 2007. Conversion of paper sludge to ethanol, ii: process design and economic analysis. Bioproc Biosyst Eng 30 (1), 35–45.

Farid, S. S., Washbrook, J., Titchener-Hooker, N. J., 2007. Modelling biopharmaceutical manufacture: Design and implementation of simbiopharma. Computers & Chemical Engineering 31 (9), 1141–1158.

Fechner, G., Howes, D., Boring, E., 1966. Elements of psychophysics. Holt, Rinehart & Winston.

Fernndez-Lpez, M., Gmez-Prez, A., Juristo, N., 1997. Methontology: from ontological art towards ontological engineering.

Fischer, R., 2011. Purification methods in biomanufacturing. Genet Eng Biotechn N 31 (17), 56–58.

Gajjar, A., Shah, V., 2011. Impurity profiling: A case study of ezetimibe. The open conference proceeding J (2), 108–112.

Gassner, M., Marchal, F., 2010. Combined mass and energy integration in process design at the example of membrane-based gas separation systems. Comput Chem Eng 34 (12), 2033–2042.

Gentner, D., 1983. Structure-mapping: A theoretical framework for analogy*. Cognitive science 7 (2), 155–170.

Ghose, A., Herbertz, T., Hudkins, R., Dorsey, B., Mallamo, J., 2011. Knowledge-based, central nervous system (cns) lead selection and lead optimization for cns drug discovery. ACS chemical neuroscience 3 (1), 50–68.

Grossmann, I., 1985. Mixed-integer programming approach for the synthesis of integrated process flowsheets. Comput Chem Eng 9 (5), 463–482.

Grossmann, I., Caballero, J., Yeomans, H., 1999. Mathematical programming approaches to the synthesis of chemical process systems. Korean J. Chem. Eng. 16 (4), 407–426.

Gruninger, M., Fox, M., 1994. The design and evaluation of ontologies for enterprise engineering. Citeseer.

Hahn, R., Bauerhansl, P., Shimahara, K., Wizniewski, C., Tscheliessnig, A., Jungbauer, A., 2005. Comparison of protein a affinity sorbents ii. mass transfer properties. J. Chromatogr. A 1093 (1-2), 98–110.

Hahn, R., Deinhofer, K., Machold, C., Jungbauer, A., 2003. Hydrophobic interaction chromatography of proteins ii. binding capacity, recovery and mass transfer properties. J Chromatogr B 790 (1-2), 99–114.

Hailemariam, L., Venkatasubramanian, V., 2010a. Purdue ontology for pharmaceutical engineering: Part i. conceptual framework. Journal of Pharmaceutical Innovation 5 (3), 88–99.

Hailemariam, L., Venkatasubramanian, V., 2010b. Purdue ontology for pharmaceutical engineering: Part ii. applications. Journal of Pharmaceutical Innovation 5 (4), 139–146.

Hirose, H., Takayama, T., Hozawa, S., Hibi, T., Saito, I., 2011. Prediction of metabolic syndrome using artificial neural network system based on clinical data including insulin resistance index and serum adiponectin. Computers in Biology and Medicine.

Hober, S., Nord, K., Linhult, M., 2007. Protein a chromatography for antibody purification. J Chromatogr B 848 (1), 40–47.

Huala, E., 2012. Go and po facilitate integration and mining of arabidopsis data. In: Plant and Animal Genome XX Conference (January 14-18, 2012). Plant and Animal Genome.

Hutchinson, N., Bingham, N., Murrell, N., Farid, S., Hoare, M., 2006. Shear stress analysis of mammalian cell suspensions for prediction of industrial centrifugation and its verification. Biotechnol Bioeng 95 (3), 483–491.

Imam, F., Larson, S., Bandrowski, A., Grethe, J., Gupta, A., Martone, M., 2012. Development and use of ontologies inside the neuroscience information framework: a practical approach. Frontiers in Genetics 3.

Ishii, Y., Otto, F., 2008. Novel and fundamental strategies for equation-oriented process flowsheeting:: Part i: A basic algorithm for inter-linked, multicolumn separation processes. Comput Chem Eng 32 (8), 1842–1860.

Jain, A. K., Murty, M. N., Flynn, P. J., 1999. Data clustering: A review. Acm Comput Surv 31 (3), 264–323.

Janson, J., 2011. Protein purification: Principles, high resolution methods, and applications. Wiley.

Jayaswal, P., Verma, S., Wadhwani, A., 2011. Development of ebp-artificial neural network expert system for rolling element bearing fault diagnosis. Journal of Vibration and Control 17 (8), 1131–1148.

Jewaratnam, J., Zhang, J., Hussain, A., Morris, J., 2012. Batch-to-batch iterative learning control using updated models based on a moving window of historical data. Procedia Engineering 42, 232–240.

John, M., Lhoussaine, C., Niehren, J., Versari, C., 2011. Biochemical reaction rules with constraints programming languages and systems. Vol. 6602 of Lecture Notes in Computer Science. Springer Berlin / Heidelberg, pp. 338–357.

Jungbauer, A., 2005. Chromatographic media for bioseparation. J. Chromatogr. A 1065 (1), 3–12.

Kaiya, H., Saeki, M., 2006. Using domain ontology as domain knowledge for requirements elicitation. In: Requirements Engineering, 14th IEEE International Conference. IEEE, pp. 189–198.

Kamimura, R. T., Bicciato, S., Shimizu, H., Alford, J., Stephanopoulos, G., 2000. Mining of biological data i: identifying discriminating features via mean hypothesis testing. Metab Eng 2 (3), 218–27.

Karuppiah, R., Grossmann, I., 2006. Global optimization for the synthesis of integrated water systems in chemical processes. Comput Chem Eng 30 (4), 650–673.

Kayala, M., Azencott, C., Chen, J., Baldi, P., 2011. Learning to predict chemical reactions. Journal of chemical information and modeling 51 (9), 2209–2222.

King, J. M. P., Titchener-Hooker, N. J., Zhou, Y., 2007. Ranking bioprocess variables using global sensitivity analysis: a case study in centrifugation. Bioproc Biosyst Eng 30 (2), 123–134.

Kling, J., 2011. Fresh from the biologic pipeline-2010. Nature biotechnology 29 (3), 197–200.

Klinkenberg, A., 1955. A molecular dynamic theory of chromatography. J Phys Chem-Us 59 (11), 1184–1184.

Knapp, R., Rabe, J., Storhas, W., Wolf, M., 2012. A laboratory experiment for teaching bioprocess control–part 2: Bioprocess design, modelling, simulation, and fermentation execution. In: Advances in Control Education. Vol. 9. pp. 384–389.

Kolodner, J., 1997. Educational implications of analogy: A view from case-based reasoning. American Psychologist 52 (1), 57.

Kong, G., Xu, D., Body, R., Yang, J., Mackway-Jones, K., Carley, S., 2011. A belief rule-based decision support system for clinical risk assessment of cardiac chest pain. European Journal of Operational Research.

Kramarczyk, J. F., Kelley, B. D., Coffman, J. L., 2008. High-throughput screening of chromatographic separations: Ii. hydrophobic interaction. Biotechnol Bioeng 100 (4), 707–720.

Kumar, S., Wittmann, C., Heinzle, E., 2004. Review: minibioreactors. Biotechnol. Lett 26 (1), 1–10.

Lapkin, A. A., Voutchkova, A., Anastas, P., 2011. A conceptual framework for description of complexity in intensive chemical processes. Chemical Engineering and Processing: Process Intensification 50 (10), 1027–1034.

Larson, T. M., Davis, J., Lam, H., Cacia, J., 2003. Use of process data to assess chromatographic performance in production-scale protein purification columns. Biotechnol Progr 19 (2), 485–492.

Lee, J., Wankat, P., 2009. Optimized design of recycle chromatography to isolate intermediate retained solutes in ternary mixtures: Langmuir isotherm systems. J. Chromatogr. A 1216 (41), 6946–6956.

Lens, F., Cooper, L., Gandolfo, M., Groover, A., Jaiswal, P., Lachenbruch, B., Spicer, R., Staton, M., Stevenson, D., Walls, R., 2012. An extension of the plant ontology project supporting wood anatomy and development research. Notes.

Li, L., Zhou, R., Dong, H., Grossmann, I., 2011. Separation network design with mass and energy separating agents. Comput Chem Eng 35 (10), 2005–2016.

Li, Q., Aucamp, J., Tang, A., Chatel, A., Hoare, M., 2012. Use of focused acoustics for cell disruption to provide ultra scaledown insights of microbial homogenization and its bioprocess impactrecovery of antibody fragments from rec e. coli. Biotechnol Bioeng.

Liang, P., Tao, H., Li, Z., Tang, K., Peng, Z., 2009. Full process simulation software for natural gas treatment system based on sequential modular method [j]. Natural Gas Industry 29 (1), 100–102.

Liao, S., 2005. Expert system methodologies and applications–a decade review from 1995 to 2004. Expert Syst Appl 28 (1), 93–103.

Lim, J., Sinclair, A., Shevitz, J., Bonham-Carter, J., 2011. An economic comparison of three cell culture techniques. BioPharm International 24 (2), 54–60.

Liu, G., Xiao, X., Mei, C., Ding, Y., 2012. A review of learning algorithm for radius basis function neural network. In: Control and Decision Conference (CCDC), 2012 24th Chinese. IEEE, pp. 1112–1117.

Liu, S., 2012. Bioprocess Engineering: Kinetics, Biosystems, Sustainability, and Reactor Design. Elsevier.

Lopez, M., 1999. Overview of methodologies for building ontologies. pp. 26–34.

Marcus, Y., Sengupta, A., 2001. Ion Exchange and Solvent Extraction: A Series of Advances. Vol. 15. CRC.

Marin, G., 2011. Multiscale Simulation and Design. Vol. 40. Academic Press.

Marquardt, W., 1996. Trends in computer-aided process modeling. Comput Chem Eng 20 (6), 591–609.

Maybury, J. P., Hoare, M., Dunnill, P., 2000. The use of laboratory centrifugation studies to predict performance of industrial machines: Studies of shear-insensitive and shear-sensitive materials. Biotechnol Bioeng 67 (3), 265–273.

McCue, J. T., Kemp, G., Low, D., Quinones-Garcia, I., 2003. Evaluation of protein-a chromatography media. J. Chromatogr. A 989 (1), 139–153.

Mizoguchi, R., 2003. Part 1: Introduction to ontological engineering. New Generat Comput 21 (4), 365–384.

Mizoguchi, R., 2004. Tutorial on ontological engineering - part 2: Ontology development, tools and languages. New Generat Comput 22 (1), 61–96.

Mizoguchi, R., Ikeda, M., Seta, K., Vanwelkenhuysen, J., 1995. Ontology for modeling the world from problem solving perspectives. Proc. of IJCAI-95 WS on Basic Ontological Issues in Knowledge Sharing, 1–12.

Mller, E., 2005. Properties and characterization of high capacity resins for biochromatography. Chem Eng Technol 28 (11), 1295–1305.

Morbach, J., Yang, A., Marquardt, W., 2007. Ontocape–a large-scale ontology for chemical process engineering. Eng Appl Artif Intel 20 (2), 147–161.

Morris, K., Venugopal, S., Eckstut, M., 2005. Making the most of drug development data. PharmaManufacturing 4 (10), 16–23.

Morris, N., Rouse, W., 1985. The effects of type of knowledge upon human problem solving in a process control task. IEEE Transactions on Systems, Man, & Cybernetics.

Mustaffa, S., Ishak, R., Lukose, D., 2012. Ontology model for herbal medicine knowledge repository. Knowledge Technology, 293–302.

Natale, D., Arighi, C., Barker, W., Blake, J., Bult, C., Caudy, M., Drabkin, H., DEustachio, P., Evsikov, A., Huang, H., 2011. The protein ontology: a structured representation of protein forms and complexes. Nucleic Acids Res. 39 (suppl 1), D539–D545.

Natarajan, S., Ghosh, K., Srinivasan, R., 2012. An ontology for distributed process supervision of largescale chemical plants. Comput Chem Eng.

Ngiam, S., Zhou, Y., Turner, M., Titchener-Hooker, N., 2001. Graphical method for the calculation of chromatographic performance in representing the trade-off between purity and recovery. J. Chromatogr. A 937 (1), 1–11.

Ngiam, S. H., Bracewell, D. G., Zhou, Y. H., Titchener-Hooker, N. J., 2003. Quantifying process tradeoffs in the operation of chromatographic sequences. Biotechnol Progr 19 (4), 1315–1322.

Nomikos, P., MacGregor, J. F., 1994. Monitoring batch processes using multiway principal component analysis. Aiche J 40 (8), 1361–1375.

Oehme, F., Peters, J., 2010. Mixed-mode chromatography in downstream process development.

Olugu, E., Wong, K., 2012. An expert fuzzy rule-based system for closed-loop supply chain performance assessment in the automotive industry. Expert Syst Appl 39 (1), 375–384.

Pajula, E., Seuranen, T., Koiranen, T., Hurme, M., 2001. Synthesis of separation processes by using case-based reasoning. Comput Chem Eng 25 (4), 775–782.

Palma, R., Corcho, O., Gmez-Prez, A., Haase, P., 2011. A holistic approach to collaborative ontology development based on change management. Web Semantics: Science, Services and Agents on the World Wide Web 9 (3), 299–314.

Papavasileiou, V., Siletti, C., Petrides, D., 2008. Systematic evaluation of single-use systems using process simulation tools.

Papoulias, S., Grossmann, I., 1983. A structural optimization approach in process synthesisi: Utility systems. Comput Chem Eng 7 (6), 695–706.

Perkins, J., Sargent, R., Thomas, S., 1982. Speed-up: A computer program for steady-state and dynamic simulation of chemical processes. In: I Chem E. Jubilee Symposium.

Petrides, D., Cooney, C. L., Evans, L. B., Field, R. P., Snoswell, M., 1989. Bioprocess simulation - an integrated approach to process-development. Comput Chem Eng 13 (4-5), 553–561.

Petrides, D., Koulouris, A., Lagonikos, P., 2002a. The role of process simulation in pharmaceutical process development and product commercialization. Pharmaceutical Engineering 22 (1), 56–65.

Petrides, D., Koulouris, A., Siletti, C., 2002b. Throughput analysis and debottlenecking of biomanufacturing facilities. Chimica Oggi 20 (5), 10–17.

Pingaud, H., Le Lann, J., Koehret, B., Bardin, M., 1989. Steady-state and dynamic simulation of plate fin heat exchangers. Comput Chem Eng 13 (4-5), 577–585.

Pinto, H., Tempich, C., Staab, S., 2009. Ontology engineering and evolution in a distributed world using diligent. Handbook on Ontologies, 153–176.

Pinto, J. M., Montagna, J. M., Vecchietti, A. R., Iribarren, O. A., Asenjo, J. A., 2001. Process performance models in the optimization of multiproduct protein production plants. Biotechnol Bioeng 74 (6), 451–465.

Potvin, G., Ahmad, A., Zhang, Z., 2012. Bioprocess engineering aspects of heterologous protein production in pichia pastoris: A review. Biochem Eng J 64, 91–105.

Queiroz, J., Tomaz, C., Cabral, J., 2001. Hydrophobic interaction chromatography of proteins. J Biotechnol 87 (2), 143–159.

Rahimpour, M., Shayanmehr, M., Nazari, M., 2011. Modeling and simulation of an industrial ethylene oxide (eo) reactor using artificial neural networks (ann). Ind Eng Chem Res 50 (10), 6044–6052.

Rassinoux, A., 2012. Knowledge representation and management: Benefits and challenges of the semantic web for the fields of krm and nlp. Methods of Information in Medicine, 51.

Rippin, D., 1993. Batch process systems engineering: a retrospective and prospective review. Comput Chem Eng 17, S1–S13.

Roger, S., 2010. Biosimilars: current status and future directions. Expert opinion on biological therapy 10 (7), 1011–1018.

Rosen, E., Pauls, A., 1977. Computer aided chemical process design: the flowtran system* 1. Comput Chem Eng 1 (1), 11–21.

Ross, B., 1989. Some psychological results on case-based reasoning. Proceedings of the DARPA, 144–147.

Rouf, S. A., Douglas, P. L., Moo-Young, M., Scharer, J. M., 2001. Computer simulation for large scale bioprocess design. Biochem Eng J 8 (3), 229–234.

Rusch, U., Borkovec, M., Daicic, J., van Riemsdijk, W. H., 1997. Interpretation of competitive adsorption isotherms in terms of affinity distributions. J. Colloid Interface Sci. 191 (1), 247–255.

Saaty, T., 1980. Analytic hierarchy process. Wiley Online Library.

Saaty, T., 2008. Decision making with the analytic hierarchy process. International Journal of Services Sciences 1 (1), 83–98.

Saite, H., King, J. M. P., Baganz, F., Hoare, M., Titchener-Hooker, N. J., 2006. A methodology for centrifuge selection for the separation of high solids density cell broths by visualisation of performance using windows of operation. Biotechnol Bioeng 95 (6), 1218–1227.

Samba, A., 2012. Logical data models for cloud computing architectures. IT Professional 14 (1), 19–26.

Savola, T., Tveit, T., Fogelholm, C., 2007. A minlp model including the pressure levels and multiperiods for chp process optimisation. Appl Therm Eng 27 (11), 1857–1867.

Schaub, J., Clemens, C., Kaufmann, H., Schulz, T., 2012. Advancing biopharmaceutical process development by system-level data analysis and integration of omics data. [Without Title], 1–31.

Schriml, L., Arze, C., Nadendla, S., Chang, Y., Mazaitis, M., Felix, V., Feng, G., Kibbe, W., 2012. Disease ontology: a backbone for disease semantic integration. Nucleic Acids Res. 40 (D1), D940–D946.

Segev, A., 2011. Multilingual crisis knowledge representation. Crisis Response and Management and Emerging Information Systems: Critical Applications, 85.

Sinclair, A., 2010. How geography affects the cost of biomanufacturing. BioProcess Int 8 (6).

Sinclair, A., Monge, M., 2010. Measuring manufacturing cost and its impact on organizations. BioProcess Int 8 (6).

Snyder, L., Kirkland, J., Dolan, J., 2010. Introduction to modern liquid chromatography. John Wiley & Sons Inc.

Staab, S., Studer, R., Schnurr, H., Sure, Y., 2001. Knowledge processes and ontologies. Intelligent Systems, IEEE 16 (1), 26–34.

Staby, A., Jacobsen, J. H., Hansen, R. G., Bruus, U. K., Jensen, I. H., 2006. Comparison of chromatographic ion-exchange resins - v strong and weak cation-exchange resins. J. Chromatogr. A 1118 (2), 168–179.

Staby, A., Jensen, I. H., 2001. Comparison of chromatographic ion-exchange resins ii. more strong anion-exchange resins. J. Chromatogr. A 908 (1-2), 149–161.

Staby, A., Jensen, I. H., Mollerup, I., 2000. Comparison of chromatographic ion-exchange resins i. strong anion-exchange resins. J. Chromatogr. A 897 (1-2), 99–111.

Staby, A., Jensen, R. H., Bensch, M., Hubbuch, J., Dunweber, D. L., Krarup, J., Nielsen, J., Lund, M., Kidal, S., Hansen, T. B., Jensen, I. H., 2007. Comparison of chromatographic ion-exchange resins vi. weak anion-exchange resins. J. Chromatogr. A 1164 (1-2), 82–94.

Staby, A., Sand, M. B., Hansen, R. G., Jacobsen, J. H., Andersen, L. A., Gerstenberg, M., Bruus, U. K., Jensen, I. H., 2004. Comparison of chromatographic ion-exchange resins iii. strong cation-exchange resins. J. Chromatogr. A 1034 (1-2), 85–97.

Stephane, N., Marc, L. L. J., 2008. Case-based reasoning for chemical engineering design. Chem Eng Res Des 86 (6A), 648–658.

Stephanopoulos, G., Han, C., 1996. Intelligent systems in process engineering: A review. Comput Chem Eng 20 (6-7), 743–791.

Stickel, J., Fotopoulos, A., 2001. Pressureflow relationships for packed beds of compressible chromatography media at laboratory and production scale. Biotechnol Progr 17 (4), 744–751.

Stonier, A., Smith, M., Hutchinson, N., Farid, S. S., 2009. Dynamic simulation framework for design of lean biopharmaceutical manufacturing operations. In: Jacek, J., Jan,

T. (Eds.), Computer Aided Chemical Engineering. Vol. Volume 26. Elsevier, pp. 1069–1073.

Sure, Y., Studer, R., 2002. On-to-knowledge methodology-final version.

Surez-Figueroa, M., 2010. Neon methodology for building ontology networks: specification, scheduling and reuse. Ph.D. thesis.

Surez-Figueroa, M., Garca-Castro, R., Villazn-Terrazas, B., Gmez-Prez, A., 2011. Essentials in ontology engineering: Methodologies, languages, and tools.

Swinnen, K., Krul, A., Van Goidsenhoven, I., Van Tichelt, N., Roosen, A., Van Houdt, K., 2007. Performance comparison of protein a affinity resins for the purification of monoclonal antibodies. J Chromatogr B 848 (1), 97–107.

Szepesy, L., Rippel, G., 1992. Comparison and evaluation of hic columns of different hydrophobicity. Chromatographia 34 (5-8), 391–397.

Tait, A. S., Aucamp, J. P., Bugeon, A., Hoare, M., 2009. Ultra scale-down prediction using microwell technology of the industrial scale clarification characteristics by centrifugation of mammalian cell broths. Biotechnol Bioeng 104 (2), 321–331.

Titchener-Hooker, N. J., Dunnill, P., Hoare, M., 2008. Micro biochemical engineering to accelerate the design of industrial-scale downstream processes for biopharmaceutical proteins. Biotechnol Bioeng 100 (3), 473–487.

Torniai, C., Brush, M., Vasilevsky, N., Segerdell, E., Wilson, M., Johnson, T., Corday, K., Shaffer, C., Haendel, M., 2011. Developing an application ontology for biomedical resource annotation and retrieval: challenges and lessons learned. In: Proceedings in International Conference on Biomedical Ontology, Buffalo, NY. pp. 101–108.

Toumi, A., Jrgens, C., Jungo, C., Maier, B., Papavasileiou, V., Petrides, D., 2010. Design and optimization of a large scale biopharmaceutical facility using process simulation and scheduling tools. Pharmaceutical Engineering.

Tsai, A. M., Englert, D., Graham, E. E., 1990. Study of the dynamic binding-capacity of 2 anion-exchangers using bovine serum-albumin as a model protein. J Chromatogr 504 (1), 89–95.

Venkatasubramanian, V., 2009. Drowning in data: Informatics and modeling challenges in a data-rich networked world. Aiche J 55 (1), 2–8.

Venkatasubramanian, V., Zhao, C. H., Joglekar, G., Jain, A., Hailemariam, L., Suresh, P., Akkisetty, P., Morris, K., Reklaitis, G. V., 2006. Ontological informatics infrastructure for pharmaceutical product development and manufacturing. Comput Chem Eng 30 (10-12), 1482–1496.

Verdoliva, A., Pannone, F., Rossi, M., Catello, S., Manfredi, V., 2002. Affinity purification of polyclonal antibodies using a new all-d synthetic peptide ligand: comparison with protein a and protein g. J Immunol Methods 271 (1-2), 77–88.

Vijayalakshmi, M. A., 2002. Biochromatography : theory and practice. Taylor & Francis, London ; New York.

Wai, P., Bogle, I., Bagherpour, K., Gani, R., 1996. Process synthesis and simulation strategies for integrated biochemical process design. Comput Chem Eng 20, S357–S362.

Walsh, G., 2003. Biopharmaceuticals: biochemistry and biotechnology. Wiley-Blackwell.

Walsh, G., 2010. Biopharmaceutical benchmarks 2010. Nature biotechnology 28 (9), 917–924.

Wang, L., 2007. Biosolids treatment processes. Vol. 6. Humana Pr Inc.

Watson, I., Marir, F., 1994. Case-based reasoning - a review. Knowl Eng Rev 9 (4), 327–354.

Weininger, D., 1988. Smiles, a chemical language and information system. 1. introduction to methodology and encoding rules. J Chem Inf Comp Sci 28 (1), 31–36.

Wiesner, A., Morbach, J., Marquardt, W., 2011. Information integration in chemical process engineering based on semantic technologies. Comput Chem Eng 35 (4), 692–708.

Winston, W., Venkataramanan, M., Goldberg, J., 2003. Introduction to mathematical programming. Vol. 1. Thomson/Brooks/Cole.

Yao, Y., Lenhoff, A. M., 2006. Pore size distributions of ion exchangers and relation to protein binding capacity. J. Chromatogr. A 1126 (1-2), 107–119.

Yazdani, M., 2012. Intelligent tutoring systems: an overview. Expert systems 3 (3), 154–163.

You, F., Grossmann, I., 2010. Integrated multiechelon supply chain design with inventories under uncertainty: Minlp models, computational strategies. Aiche J 56 (2), 419–440.

Zaman, F., Allan, C. M., Ho, S. V., 2009. Ultra scale-down approaches for clarification of mammalian cell culture broths in disc-stack centrifuges. Biotechnol Progr 25 (6), 1709–1716.

Zhang, D., Cheng, B., Wu, A., 2012. Prediction of biomass concentration with hybrid neural network. Advances in Neural NetworksISNN 2012, 638–644.

Zhang, H., Kitchenham, B., Pfahl, D., 2008. Reflections on 10 years of software process simulation modeling: a systematic review. Making Globally Distributed Software Development a Success Story, 345–356.

Zhang, H., Kitchenham, B., Pfahl, D., 2010. Software process simulation modeling: an extended systematic review. New Modeling Concepts for Todays Software Processes, 309–320.

Zhao, C., Bhushan, M., Venkatasubramanian, V., 2003. Roles of ontology in automated process safety analysis. Computer Aided Chemical Engineering 14, 341–346.

Zhao, C. H., Joglekar, G., Jain, A., Venkatasubramanian, V., Reklaitis, G. V., 2005. Pharmaceutical informatics: A novel paradigm for pharmaceutical product development and manufacture. Comp Aid Ch 20a-20b, 1561–1566 1680.

Zhao, G., Dong, X., Sun, Y., 2009. Ligands for mixed-mode protein chromatography: Principles, characteristics and design. J Biotechnol 144 (1), 3–11.

Zhao, Y., Jiang, C., Yang, A., 2012. Towards computer-aided multiscale modelling: A generic supporting environment for model realization and execution. Comput Chem Eng 40, 45–57.

Zhou, J. X., Dermawan, S., Solamo, F., Flynn, G., Stenson, R., Tressel, T., Guhan, S., 2007. ph-conductivity hybrid gradient cation-exchange chromatography for process-scale monoclonal antibody purification. J. Chromatogr. A 1175 (1), 69–80.

Zimmer, D., 2003. Introduction to quantitative liquid chromatography-tandem mass spectrometry (lc-ms-ms). Chromatographia 57, S325–S332.

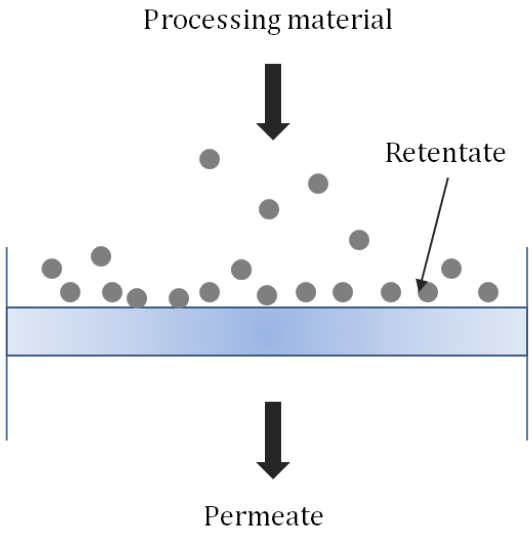
Zott, C., Amit, R., Massa, L., 2011. The business model: Recent developments and future research. Journal of management 37 (4), 1019–1042.

Appendix A

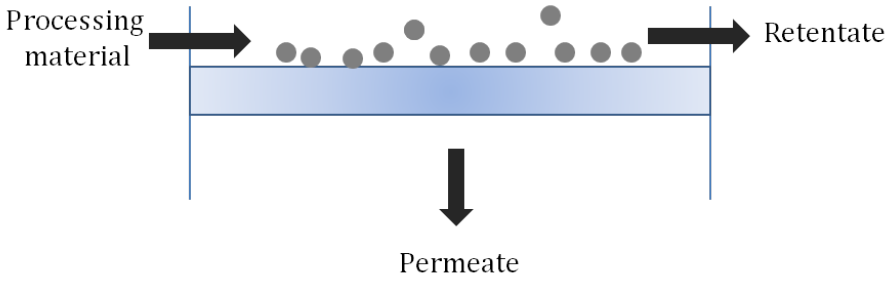
DEVELOPMENT OF FILTRATION SYSTEM

A.1 FILTRATION INTRODUCTION

The filtration separates the solid particles from the liquid-solid mixture by using pressure to force the fluid through a filter medium which is used to keep the solid particles. Solids are accumulated on the filter to form filter cake. The liquid after the filter medium is called filtrate. The theories of filtration have been reviewed well elsewhere that can provide further information of filtration (**?**). Figure A.1 illustrates two types of filtration used in bioprocessing, namely dead end filtration (a) and cross flow filtration (b).



a. dead end filtration



b. cross flow filtration

Figure A.1: Illustration of filtration procedure

These two types of filtration are differentiated by the fluid flow. In dead end filtration, the fluid flows directly towards the filter under the influence of pressure; while the cross flow filtration is based on the pressurized flow of the fluid flowing tangentially over the surface of the filter membrane, with a portion of the feed pushed through the filter and the remaining being swept sweeping tangentially along the membrane to exit the system without being filtered.

As the mixture continues being pumped in, the particles will be cumulated as a cake and bring the resistance for mixture to pass through the filter further. The rate of cake formation and concentration of slurry are determined by the processing material properties, e.g. density. The resistance is indicated by the pressure drop which is the pressure difference between the pressures before and after the filter. The pressure drop of the liquid passing though porous bed of solid spheres is proportional to the flowrate, filtration area, bed permeability. This relationship can be described by Darcy law (**?**). The bed permeability is a property of filters that can be calculated (**?**). Substituting the permeability calculation into the Darcy law will obtain a new equation that the filtration performance relates to the porosity, particle size and shape, distribution and packing, rate of cake formation and concentration of slurry. The porosity, particle size and shape, distribution and packing are the properties of filter, so they will be determined when the filter is determined.

The filtration is usually operated by two modes called constant-pressure filtration and constant-rate filtration. The constant-pressure model defines the pressure drop across the filter keeps constant and the data generated from one batch of this model can determine the resistance of filter and cake. The constant-rate makes the flowrate be constant so that the pressure drop required for any desired flowrate can be found in this model. For representing the filtration operation mode, the parameter called operation mode would used.

For the output of filtration, the flux and membrane capacity are commonly used for characterize the filtration performance. The flux is defined that the volume of mixture flowing through a given membrane area during a given time. Flux is determined by the hydraulic resistance, e.g. membrane resistance, boundary layer resistance and etc. The membrane capacity is determined by the cumulated filtrate volume that is measured through a small area test filter until the flowrate drops to 10% of its initial value (**?**). In addition, the cake formed by the solid particles will not be completely dry, therefore some liquid would loss during the filtration.

A.2 FILTRATION SYSTEM

The filtration system is proposed to harness the filtration experimental data and knowledge in order to solve the filtration design problem. The filtration system includes four kinds of data and knowledge i.e. filtration experimental data, ontology, theoretical and empirical knowledge, and the three reasoning functionalities, i.e. search, prediction and suggestion. Due to the time limitation, it is difficult to capturing the enough filtration experimental data and knowledge to develop the filtration system in this project. However, the feasible parameters that could be used for representation of filtration experimental data and the feasible knowledge will be introduced in the following.

A.2.1 Filtration experimental data representation

18 parameters are proposed to be used to represent the filtration experimental data, and they are given in Table A.1. This parameter setting is expendable if they were not enough to represent all feasible information included in the filtration experimental data. It may be also polished by discussing with filtration experts.

Table A.1: Parameters for filtration experimental data representation

	parameters	**definitions**
Input	strain	name of cell line
	product	name of target biomolecule
	liquid viscosity	viscosity value of processing material
	solid density	value of solid density
	liquid density	value of liquid density
	product concentration	value of product concentration
	solid concentration	value of solid concentration
Step	filtration type	name of filtration type
	filter	name of filter used by filtration
	pressure drop	value of pressure difference before and after the filter
	filtration flowrate	value of the flowrate in filtration
	scale	name of filtration scale
	membrane area	value of membrane area used in filtration
	retention time	value of retention time in filtration
	operation mode	name of filtration operation mode
Output	flux	value of the average flux
	membrane capacity	value of membrane capacity
	filtrate volume loss	value of filtrate volume loss

A.2.2 Ontologies of filtration

The terms appeared in the filtration experimental data are proposed to be defined and organized by the ontologies. Taking the filter domain as an example, the names of general filter used for protein purification will be located at the parent class, e.g. Nutsche, candle, plate, frame tubular, leaf filters and microfiltration system (Doran, 1995), and the term of the specific filter type will be placed at the child class, such as the AcroPrep 24 filter(Pall, UK) which is a type of filter plate.

The ontologies of the domains which have been developed in centrifugation and chromatography system can also be used by filtration system, e.g. strain, product. Based on the parameter setting of filtration experimental data representation, five domains ontologies are proposed to be used by filtration system, i.e. strain, product, filtration type, filter and operation mode.

A.2.3 Theoretical and empirical knowledge of filtration

For the theoretical knowledge of filtration, the fundamental equations and background information of filter should be considered. The fundamental equations would be used for essential calculations, such as membrane capacity calculation. The representation is the same as the fundamental equation representation referred by centrifugation and chromatography system. The background information of filter is proposed to be used to support the design, e.g. using porosity of filter to analyse the resistance during the filtration. The filter background information can be represented by the ERM model, where the filter name is used as the entity and the filter properties are used as the attributes.

For the empirical knowledge, the scale up/down principles are proposed to be considered. The scaling of filtration is based on the same achievement of pressure profile, pore size and ratio of filtrate volume to surface area. Usually, the laboratory scale experimentation is used to characterize the filter geometry, then the average flux and cake resistance will be identified as functions of the pressure which can be used to select the filter for the desired large scale, e.g. size. The experimental generated from the laboratory scale experiments can be used to calculate the pressure profile required by the industrial scale filtration operation. These knowledge can be captured as the rules which can be harnessed by filtration system to generate the solutions to the desired scale filtration, e.g. filter surface. In large scale filtration, the shear effects caused by the ancillary equipment will impact the physical properties of the processing material that would influence the filtration performance. To this issue, a rotating vertical leaf filter has been developed based on a standard Nutsche filter in order to

mimic the performance of large scale filters (**?**). This mimic can be captured as the rules for searching relevant datapoints. For instance, when the large scale filtration experimental data is concerned, the corresponding small scale filtration experimental data can also be searched.

A.2.4 Reasoning with the filtration experimental data and knowledge

Each filtration design problem is represented as a design query which has the same representation of filtration experimental data. Given this design query, the three reasoning functionalities will find relevant datapoints, generate the predicted performance and retrieve the requested information. The HHA introduced by chromatography system is also available for filtration system to deal with the complicated design problem that consists a set of internal related variables.

With the three BDKF systems, i.e. centrifugation, filtration and chromatography system, the next step is to introduce the method of harnessing the three BDKF systems in regarding the three-step sequence design problem.

Appendix B

MASS BALANCE CALCULATION

B.1 CENTRIFUGATION MASS BALANCE CALCULATION

The 97.2%CE indicates the 97.2% of cells would be removed from the feed stream as the sediment, 2.8% of cells would remain in the supernatant. The results can be calculated by equation (B.1 and B.2).

$$mass\ of\ cells\ in\ sediment = mass\ of\ cells \times CE; \tag{B.1}$$

$$mass\ of\ cells\ in\ supernatant = mass\ of\ cells \times (1 - CE); \tag{B.2}$$

The cells in the sediment contain liquid, and the mass of liquid depends on the dewatering level of centrifuge. 70% dewatering level has assumed for CSA-1 centrifuge, thus the mass of liquid included in the sediment and supernatant should be calculated by equation (B.3 and B.4).

$$mass\ of\ liquid\ in\ sediment = \frac{1 - dewatering}{dewatering} \times \frac{mass\ of\ cells\ in\ sediment}{cell\ density} \times liquid\ density; \tag{B.3}$$

$$mass\ of\ liquid\ in\ supernatant = mass\ of\ liquid - mass\ of\ liquid\ in\ sediment; \quad \text{(B.4)}$$

The product and total protein concentration have been specified in the sequence design query, by harnessing these specifications, the mass of product and protein in sediment and supernatant should be calculated by equation (B.5, B.6, B.7 and B.8).

$$mass\ of\ product\ in\ sediment = mass\ of\ liquid\ in\ sediment \times product\ concentration; \quad \text{(B.5)}$$

$$mass\ of\ product\ in\ supernatant = mass\ of\ liquid\ in\ supernatant \times product\ concentration; \quad \text{(B.6)}$$

$$mass\ of\ total\ protein\ in\ sediment = mass\ of\ liquid\ in\ sediment \times total\ protein\ concentration; \quad \text{(B.7)}$$

$$mass\ of\ total\ protein\ in\ sediment = mass\ of\ liquid\ in\ supernatant \times total\ protein\ concentration; \quad \text{(B.8)}$$

After the centrifugation, the sediment stream should be discarded, and the supernatant stream is carried forward as the feed stream to the filtration. Based on the results of mass balance, the volume of liquid and cells in the supernatant stream can be calculated by equation (B.9 and B.10), and the total volume of feed stream for filtration can be obtained by equation B.11

$$liquid\ volume = \frac{mass\ of\ liquid\ in\ supernatant}{liquid\ density}; \quad \text{(B.9)}$$

$$solid\ volume = \frac{mass\ of\ cell\ in\ supernatant}{cell\ density}; \tag{B.10}$$

$$total\ volume = liquid\ volume + solid\ volume; \tag{B.11}$$

Based on these calculated volume, the new protein concentration and total protein concentration can be obtained by equation (B.11 and B.12).

$$protein\ concentration = \frac{mass\ of\ protein}{volume}; \tag{B.12}$$

$$total\ protein\ concentration = \frac{mass\ of\ total\ protein}{volume}; \tag{B.13}$$

B.2 FILTRATION MASS BALANCE CALCULATION

After the filtration, the feed stream is separated into two streams, i.e. retentate and permeate. The feed stream components in these streams have different mass which are determined by the filtration specific output, e.g. filtrate mass loss. In order to demonstrate the calculations of components mass in the two streams, the related equations are given in the following.

The mass of liquid in the stream of retentate and permeate after filtration can be calculated by equation(B.14 and B.15).

$$mass\ of\ liquid\ in\ retentate = mass\ of\ liquid \times filtrate\ mass\ loss; \tag{B.14}$$

$$mass\ of\ liquid\ in\ permeate = mass\ of\ liquid \times (1 - filtrate\ mass\ loss); \tag{B.15}$$

The volume of liquid in stream of retentate and permeate can be calculated by equa-

tion(B.16 and B.17).

$$volume\ of\ liquid\ in\ retentate = \frac{mass\ of\ liquid\ in\ retentate}{liquid\ density}; \tag{B.16}$$

$$volume\ of\ liquid\ in\ permeate = \frac{mass\ of\ liquid\ in\ permeate}{liquid\ density}; \tag{B.17}$$

The mass of product in stream of retentate and permeate can ben calculated by equation(B.18 and B.19).

$$mass\ of\ product\ in\ retentate = volume\ of\ liquid\ in\ retentate \times product\ concentration; \tag{B.18}$$

$$mass\ of\ product\ in\ permeate = volume\ of\ liquid\ in\ permeate \times product\ concentration; \tag{B.19}$$

Similarly, the mass of total protein in stream of retentate and permeate can be calculated by equation(B.20 and B.21).

$$mass\ of\ total\ protein\ in\ retentate = volume\ of\ liquid\ in\ retentate \times total\ protein\ concentration; \tag{B.20}$$

$$mass\ of\ total\ protein\ in\ permeate = volume\ of\ liquid\ in\ permeate \times total\ protein\ concentration; \tag{B.21}$$

After the filtration, the retentate stream was discarded and the permeate stream is carried forward to chromatography step as feed stream. The properties of this feed stream has changed, and the equations used to calculate these properties are given in the following. The volume of liquid in permeate stream could be calculated in equation(B.22) that would be

used as the feed volume for the chromatography.

$$mass\ of\ total\ protein\ in\ permeate = volume\ of\ liquid\ in\ permeate \times total\ protein\ concentration; \tag{B.22}$$

B.3 CHROMATOGRAPHY MASS BALANCE CALCULATION

After the chromatography, the feed stream is split into two phase, i.e. elute and waste. The product mass included in the elute can be calculated by equation B.23.

$$product\ mass = mass of loaded product \times yield; \tag{B.23}$$

The product concentration of feed stream after the chromatography can be calculated by equation B.24.

$$product\ concentration = \frac{product\ mass}{cutting\ volume}; \tag{B.24}$$

where the cutting volume is the volume of liquid collected from the elution phase.

The mass of total protein included in the elute and concentration can be calculated by equation B.25 and B.26 respectively..

$$mass\ of\ total\ protein = \frac{mass\ of\ loaded\ product \times yield \times (1 - purity)}{purity} \tag{B.25}$$

$$total\ protein\ concentration = \frac{mass\ of\ total\ protein}{cutting\ volume}; \tag{B.26}$$

Appendix C

CONFERENCE PROCEEDING

21st European Symposium on Computer Aided Process Engineering – ESCAPE 21
E.N. Pistikopoulos, M.C. Georgiadis and A.C. Kokossis (Editors)

Systematic Data and Knowledge Utilization to Speed up Bioprocess Design

Jun Zhang,[a] Anthony Hunter,[b] Yuhong Zhou[a]

[a]*Department of Biochemical Engineering,University College London,Torrington Place,London,WC1E 7JE,U.K.*
[b]*Department of Computer Science,University College London,Gower Street,London,WC1E 6BT,U.K.*

Abstract

Bioprocess design may require a substantial number of experiments to investigate the options for each bioprocess step. With the aim of reducing the amount of experimentation that is undertaken, we propose that data and knowledge about bioprocess design can be systematically exploited. We present a new general framework called the Bioprocess Data and Knowledge Framework (BDKF), for representing and reasoning with data and knowledge to produce the possible solutions for the bioprocess design. In the centrifugation case study, we established a database with 344 experiments described by 34 parameters, a knowledgebase with ontological, theoretical and empirical knowledge about centrifugation and showed how they were used for searching relevant process design information, predicting process performance and suggesting new experiments to be done. It demonstrates that BDKF is a promising approach for bioprocess data and knowledge utilization

Keywords: experimental data, ontological knowledge, theoretical knowledge, empirical knowledge, centrifugation

1. Introduction

Bioprocess Design involves identifying an optimal sequence of bioprocess steps to produce a targeted biomolecule. There are typically more than five steps, and each of them has complicated characteristics due to the complexity of the biomaterials and output requirements. The design of each step is achieved by extensive experimentation to explore the feasible space. This extensive experimentation may require substantial time and materials which are expensive to generate in the biopharmaceutical development phase [1].

A substantial amount of the data and knowledge has been generated by experimentation concerning the bioprocess steps. Yet, this data and knowledge are currently not being fully exploited. The use of it is limited because this data and knowledge is not stored in a form for re-use, and insufficient consideration has been given to ways that it could be harnessed for supporting bioprocess design [2]. Some specific computational approaches have been applied on the design issue, e.g. case based reasoning system which focused on the previous data reuse based on the similarity calculations [3], and simulations tools which calculated the possible results by the pre-defined equations and requiring the complete input [4], however these popular approaches are not able to systematically harness and reason with the complex data and knowledge that is available.

In this paper, we present a new general framework called Bioprocess Data and Knowledge Framework (BDKF), for representing data and knowledge on bioprocess

design and for reasoning with this data and knowledge in order to make predictions about output of bioprocess steps and to make suggestions about further experiments to be done in order to understand the landscape of the bioprocess step better.
For representing data and knowledge on bioprocess design in BDKF, we focus on four types of information, i.e. experimental data, ontological knowledge, theoretical knowledge and empirical knowledge. For reasoning with the data and knowledge, we focus on three functionalities, i.e. search, prediction and suggestion. A centrifugation prototype has been established in WinProlog 4800(LPA, UK), and in the following sections we use a centrifugation case study to illustrate how BDKF reasons with the four types of information.

2. Centrifugation Data and Knowledge

Experimental Data: Each experimental data point represented one centrifugation experiment fact which was represented by a subset of the 34 parameters in 3 categories, i.e. input, step and output. The parameters of input represented the biomaterials properties, such as the strain, etc., the step represented the centrifuge operating conditions, such as the flow rate, etc., and output indicated the centrifugation performance, such as the clarification efficiency (CE), etc. All data were kept in the form of a table, in which each column represented one parameter, and each row indicated one experimental data point. A total of 344 data points were collected from the centrifugation experiments done by the researchers at the Biochemical Engineering Department, University College London.
Ontological Knowledge: The ontological knowledge describes the terminologies of the specific domain and their relationships [5]. It can help to establish the logical link among the different bio-terms. The ontological knowledge concerning the strain has been explained as an example in our paper [6]. For instance, CHO and GS-NS0 are two types of mammalian cells, when the mammalian cell is queried, the CHO and GS-NS0 data would also be considered. It is helpful for the mammalian cell design because they have common physical properties which may perform the similar liquid-solid separation. A total of 22 pieces of ontological knowledge concerning the strain, feed, scale and centrifuge model were captured.
Theoretical Knowledge: The theoretical knowledge is the knowledge about bioprocess engineering that is represented by the formal definitions, such as the unit definitions, equations, etc. It can assist the bioprocess design or help to complete the inconsistent experimental data, for instance, the CE is calculated by optical density (OD) values of the feed [7]. If the experimental data only recorded the OD value, the corresponding CE value can be calculated by harnessing this knowledge.
Empirical Knowledge: The empirical knowledge is the knowledge about the bioprocess engineering that has been obtained from the empirical studies, and usually has been published in the scientific literature. For instance, it has been shown that for the CHO cell, how to use the specific shear device to make the performance of a small scale centrifuge can be used to simulate and thereby predict the performance of a specific pilot scale centrifuge [8].

3. Functionalities of Reasoning

Search: The search function requires the user to give the formalized design query which is comprised of the specifications of the centrifugation input, step and output, and then returns the data points from the experimental data base which have the similar input and step, and satisfy the output requirement. For instance, the design query

consisted of the input specification and output requirement, i.e. using centrifugation to harvest mammalian cell at 90% CE at least. For this query, the search function would return the data points that have the strain value as the mammalian cell, CHO or GS-NS0 (because the CHO or GS-NS0 is a type of the mammalian cell), as well as the CE value greater than 90%.

Prediction: Based on the results of the search functionality, the prediction returns a predicted features of the output of the bioprocess step. We propose to use the average to indicate the output feature that is likely to be achieved by implementing the user's specifications. Other options, such as the weighted average or a specific statistic algorithm may be applied to define the most similar result to the specified input, step and output.

Suggestion: Based on the prediction result, the suggestion returns suggestions for experiments that could be done in order to obtain the required output. For the given design query and the specific prediction, we propose to retrieve information from the data point, that has the similar output features with the prediction, as the experiment suggestions to help user to obtain the targeted output.

The aim of these functionalities is to help the user understand the possibilities for processing a specific bioprocess step input: the search enables the user to study the relevant previous experiments, the prediction enables the user to evaluate the utility of the particular bioprocess steps, and suggestion enables the user to better expand the understanding of the possible bioprocess steps.

4. Centrifugation Case Study

In this section, we will discuss the centrifugation case study to illustrate how the BDKF systematically harnesses the data and knowledge for the bioprocess design.

Design Query Formalization: In order to illustrate how we formalize a design query, suppose we propose to harvest the CHO cell by centrifugation, the solid density of culture is around 1.05 kg/m^{-3}, the liquid density is around 1 kg/m^{3}, and a pilot disk stack centrifuge is proposed to be used to achieve the 90% CE, and the question is "what flow rate should we use?" This information could be formalized in the design query shown by Fig.1.

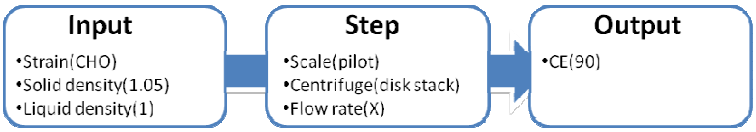


Fig.1 Formalized design query, where the queried flow rate is represented by 'X'

Given the formalized design query, the prototype processes the search, prediction and suggestion functionalities sequentially.

Reasoning of Search: Given the design query, the ontological, theoretical and empirical knowledge are harnessed in the search functionality.

- To harness the ontological knowledge, the prototype returns the data points whose terminology is the subset of the corresponding specification in the design query. E.g. for strain(CHO), the data points whose strain value is the subset concepts of CHO will be returned.
- To harness the theoretical knowledge, the prototype uses the theoretical knowledge to complete the inconsistent experimental data. E.g. to calculate the missing CE value by OD values.
- To harness the empirical knowledge, according to the empirical studies [9], the small scale (e.g. micro well scale) experimental data which were implemented

for simulating the specific pilot scale centrifuge performance is identical to the pilot scale data. By harnessing this knowledge, given the design query which specified the pilot and disc stack centrifuge, the corresponding ultra scale down (USD) data would also be searched.

By harnessing these 3 kinds of knowledge, all available design query-related data points would be returned by the search functionality.

Reasoning of Prediction: Based on the relevant data, the prototype would average the CE value as the predicted CE value (e.g. CE 96.4%). It indicates the possible CE value that may be achieved by implementing the specifications of the design query.

Reasoning of Suggestion: The prototype calculates the distance of the predicted CE and the CE value of the relevant data points, then retrieves the queried information from the data point that has the smallest distance (e.g. the flow rate 30L/h).

Besides, the empirical knowledge also can be harnessed here to generate suggestions for using USD approach to evaluate the results of the pilot scale, e.g. shear CHO cell by 6000 rpm at 20 seconds, the small scale centrifugation can be used to mimic the separation performance of the pilot disk stack [8].

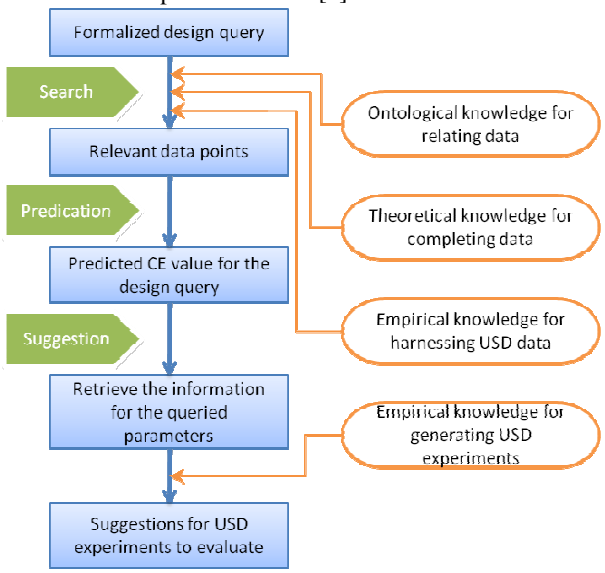


Fig 2 Process of reasoning with centrifugation data and knowledge

According to the Fig 2, the prototype suggested using 30 L/h as the answer of the flow rate which may lead to the 90% CE for mammalian cell harvest.

5. Evaluation

We undertook an evaluation to determine the quality of the results of BDKF regarding to the specific design query. A total of 170 CHO cell centrifugation data points were captured from published experimental data, each of them was named as an evaluation data point. The predicted CE value was evaluated by using real centrifugation experimental results (real CE) recorded in the evaluation data points. The design query only used the information of the type of strain, e.g. CHO, feed material, e.g. fermentation broth with whole cell, and the centrifuge separation capacity to produce the predicted CE, and then compared it with the real CE, the prediction error used to indicate the accuracy was defined by Equation 1.

Prediction error = |predicted CE – real CE| *Equation 1*

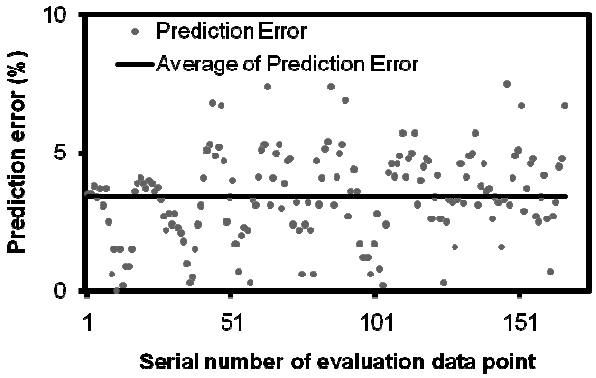


Fig.3 evaluation results of the predicted CE

From Fig.3, all of the 170 evaluation results were within 10% and the average error was only 3.4% which can be accepted by industry. This indicated the accuracy and reliability of the use of BDKF.

6. Conclusion

A centrifugation case study has been discussed to show how BDKF harnesses the experimental data, and ontological, theoretical and empirical knowledge concerning the bioprocess design. The evaluation results indicate that BDKF is an effective way to utilize the complex bioprocess data and knowledge, which may reduce the time and materials usage in biopharmaceutical development.

7. Acknowledgement

The advice and centrifugation data support of Dr. Andrew Tait, Dr. Jean Aucamp, Dr. Balasundaram Bangaru are gratefully acknowledged. Support for the IMRC for Bioprocessing housed in the Advanced Centre of Biochemical Engineering by the Engineering and Physical Sciences Research Council (EPSRC) under the Innovative Manufacturing Research initiative and UK Overseas Research Scheme (ORS) for J. Zhang are gratefully acknowledged.

References

[1] M.Rohner,H.-P.Meyer,1995,Bioprocess Eng,13,69-78
[2] S.Charaniya,WS.Hu,G.Karypis,2008,Trands Biotech,26,690-699
[3] N.Stéphane,LLJ.Marc,2008,Chem Eng Res Des,86,648-658
[4] SA.Rouf,PL.Doulas,M.Moo-Young,JM.Scharer,2001,Biochem Eng J,8,229-234
[5] Y.Ding,S.Foo,2002,J Inf Sci,28,123-136
[6] J.Zhang,A.Hunter,YH.Zhou,to be submitted
[7] H.Salte,JMP.King,F.Baganz,M.Hoare,NJ.Titchener-Hooker,2006,Biotechnol Bioeng,95,1218-1227
[8] AS.Tait,JP.Aucamp,A.Bugeon,M.Hoare,2009,Biotechnol Bioeng,104,321-331
[9] NJ.Titchener-Hooker,P.Dunnill,M Hoare,2008,Biotechnol Bioeng,100,473-487