THE DESIGN AND IMPLEMENTATION OF COMPARATIVE REASONING TOOLS FOR FERMENTATIONS

A thesis submitted to the University of London
for the degree of
Doctor of Philosophy

by
Carolynne Therese Marshall

Advanced Centre for Biochemical Engineering
University College London
Torrington Place
London WC1E 7JE

August 1992

To the memory of my dear friend Frances Elizabeth Jebson who kept me going first with letters, then with smiles, then, when she was gone, with inspiration, and always with thoughtfulness and love.

ACKNOWLEDGEMENTS

My thanks go to my supervisors at University College London, Dr Neville Fish and Professor Malcolm Lilly, for their advice and support. Special thanks to Mrs Nina Thornhill for all her encouragement. Thanks also go to my colleagues for their friendship and helpful advice.

My sincere gratitude is expressed to all the staff of the Biochemical Process R&D Group at Merck Sharp and Dohme Research Laboratories, especially Mr Leonard Lister for his assistance, Mr George Hunt for helpful discussions, Dr Kodzo Gbewonyo for his guidance and Dr Barry Buckland for providing the opportunity for me to work at Merck and for all his most valuable advice.

I acknowledge the financial support of the Commonwealth Scholarship Commission, the William Georgetti Scholarship Board and UCL.

Finally, a big thank you to my family and friends for all their encouragement and for getting me through the tough times.

ABSTRACT

The progress of a fermentation is usually assessed by visual comparison of the time profiles of the data with those from other batches or from a standard or model. In this work the comparative reasoning process was automated, thereby eliminating the problems caused by human inconsistencies and bias, and facilitating a more thorough usage of all available data. The comparative reasoning was extended to include non-numerical and single-value data.

A relational data base structure was designed to record all batch sheet and descriptive data from any fermentation and to enable the comparison of these data from one batch to another.

The quantitative time variant data from two fermentations may be dissimilar in a strict numerical sense but may exhibit similar patterns or trends. Conventional numerical techniques cannot be used to detect these similarities. A graphical analysis process was developed to enable detection of periods of approximate similarity in two time profiles: the data were simplified by segmentation into linear episodes, described qualitatively using the descriptive language of an expert, and algorithms were devised for the comparison of these data from batch to batch.

The results of comparing the data base and graphical information were used to identify discrepancies between fermentations and determine cause-effect relationships.

The comparative techniques were used to analyse the data from a set of recombinant, protein producing, laboratory scale fermentations and thus enable reasoning about the effects of sterilisation conditions and inoculum concentration on the progress of the fermentation. The results concurred with manual analysis of the data. The computerised tools improved understanding of the process because all available data could be analysed in a thorough and consistent manner.

The comparative reasoning tools have the potential to improve the on-line detection and diagnosis of faults in a production process and can provide a link between fermentation and downstream processing data analysis.

The comparative reasoning tools require no *a priori* knowledge of the process and can be applied to any type of data with no dependence on the magnitude of the values. The tools can therefore be used in both research and production environments and on any fermentation process.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ABBREVIATIONS

aFGF	acidic fibroblast growth factor
$aFGF_h$	aFGF at the recommended harvest point
$aFGF_m$	maximum aFGF
$aFGF_{sh}$	$aFGF_h$ in normalised units per gram of biomass
$aFGF_{sm}$	$aFGF_m$ in normalised units per gram of biomass
$aFGF_{vh}$	$aFGF_h$ in normalised units per litre
$aFGF_{vm}$	$aFGF_m$ in normalised units per litre
ASCII	American Standard Code for Information Interchange
CBR	case based reasoning
CER	carbon dioxide evolution rate (arbitrary units)
DCW	dry cell weight ($g.L^{-1}$)
DO(T)	dissolved oxygen (tension) (% air saturation)
DSIMP	data simplification (linearisation) routine
ECG	electrocardiogram
E. coli	*Escherichia coli*
g	grams
GCV	generalised cross validation
GMP	Good Manufacturing Practice
gof	goodness of fit
h	hours
HPLC	High Performance Liquid Chromatography
L	litres
lpm	litres per minute
M	molar
MATCHER	comparison routine
mins	minutes
mmol	millimoles
mL	millilitre
MSDRL	Merck Sharp and Dohme Research Laboratories
MSPLIN	cubic spline fitting routine
nm	nanometres
OD	optical density
OUR	oxygen uptake rate (arbitrary units)
PC	personal computer
psig	pounds per square inch gauge

QBE	Query By Example
QUAL	qualitative description routine
rpm	revolutions per minute
sat.	saturation
SQL	Structured Query Language
std dev	standard deviation
T	temperature (°C)
t_h	recommended harvest time (h)
t_m	time at which maximum aFGF concentration occurred
UCL	University College London
UV	ultraviolet
V	Volume (L or mL in 4.2.6.4)
v/v	volume per volume
x	abscissa variable
y	ordinate variable

1 INTRODUCTION

The interpretation of fermentation data generally involves a comparative assessment of the data with respect to other, historical or concurrent, fermentation data sets or a standard data set. This analysis process is usually carried out manually by a fermentation expert - a process operator, a plant manager, or a technologist. This thesis describes the development and implementation of techniques which automate this comparative analysis of fermentation data. The emphasis of the work was on the emulation of the techniques used by the fermentation experts as, thus far, these have proved the most suitable for application to a process that is best described using approximate quantitative values and qualitative statements. The idea was not to usurp the role of the expert in the comparative reasoning process but to improve on areas where the expert's abilities are limited.

In this chapter the data available from a fermentation facility are described and the techniques used for the analysis of these data are introduced. The current use of comparative analyses is highlighted along with a discussion of the limitations which result from the manual nature of these comparisons, in particular the lack of consistency, the existence of human bias and an inability to utilise all the available data. Computerised tools were developed to automate the comparative reasoning process and thereby eliminate the problems inherent in manual analysis. These tools are described in Chapters 2 and 3 and summarised in Table 1.1. The scope of this study is defined at the end of this chapter with a discussion of the extent of automation achieved (Section 1.2).

DATA TYPE	CURRENT COMPARATIVE ANALYSIS TOOLS	NEW COMPARATIVE ANALYSIS TOOLS
Time Invariant	Data are recorded manually in note books or on batch sheets. Numerical quantities are used for reporting (ie to summarise results) and to manually compare performance of batches. (Chapter 1).	Data are systematically recorded in a data base. Qualitative and quantitative information are used in automatic comparisons between batches. (Chapter 2).
Time Variant	Data are recorded automatically. Time profiles are used to visually compare data from batch to batch. (Chapter 1).	Data are recorded automatically. Three computer programs (DSIMP, QUAL, MATCHER) automatically compare all time profiles in a routine and consistent manner. (Chapter 3).

Table 1.1: The comparative reasoning tools developed in this work for the analysis of fermentation data and the analogous techniques in current use.

1.1 Fermentation Data

Two major categories of fermentation data were defined for this work: time invariant data and time variant data. The time invariant data category can be further subdivided into batch sheet data, descriptive data and expert comments. This section lists the components of each of these data categories and describes how they are used in the analysis of fermentations. Deficiencies in the analysis process are identified and the requirements for an improved analysis procedure are detailed.

1.1.1 Time Invariant Data

1.1.1.1 Batch Sheet Data

The batch sheet data are those pieces of information that describe the physical and chemical components of the fermentation, the operating conditions, and the control mechanisms used to achieve the desired operating conditions. A list of typical batch sheet data is given in Table 1.2.

Batch Number
Organism
Product(s)
Start date and time
End date and time
Seed or Production run
Medium details:
- composition
- component suppliers
- component lot numbers
- component grade

Details of medium make-up:
- mix tank batch number
- volume of mix tank
- operator

Vessel
Initial broth volume
Sterilisation details for vessel
Sterilisation details for medium
Inoculum details:
- source
- volume
- condition for addition, eg age
- operator

Initial operating conditions
Scheduled operating condition changes
Other operating condition changes
Details of feeds
Disposition of final broth
Equipment used, eg pumps

Table 1.2: Typical batch sheet data for a bacterial fermentation.

During developmental work the batch sheet data are recorded in notebooks so as to keep a record of the conditions of each experimental batch. It is very rare for the recording of the data to be formalised and even less common to use the data to its fullest potential during subsequent analyses.

The most obvious danger of an informal recording process at this level is that important information may be left out because the researcher may think it obvious or even unimportant to the investigations. As a simple example, consider a series of experiments designed to investigate the effect of temperature on a given fermentation: a number of fermentations would be carried out at different temperatures and the results analysed to find the temperature which resulted in the most productive fermentation. During the course of such a set of experiments all other environmental factors, such as pH and pressure, would be held at prespecified levels, however it is very easy to ignore the effect of, for example, different lot numbers of medium components or slightly different sterilisation regimes, especially when they have not been recorded. Similarly, when a large number of experimental fermentations have been carried out it becomes very difficult to include all the batch sheet information in the manual analysis process and the unintentional changes in conditions are often ignored. The comparison between fermentation batches thus only utilises part of the data available and may not elicit a true picture of the effects of various conditions.

Data recorded only in a researcher's notebook are not immediately available to other researchers. This becomes particularly important when downstream processing operations are developed separately from the fermentation process. The information passed on to the downstream operations research team, in many cases, would reflect the fermentation workers' interpretation of what is important and could well omit pertinent facts.

Pilot plant trials and production runs avoid some of these problems because the batch sheet data are usually recorded on standard forms, as required by regulatory bodies or in-house procedures. These forms are generally exhaustive (as long as recording is strictly enforced). However, the information available is usually only used for reporting purposes, little recourse is made to these data during data analysis.

During production it is feasible that the batch sheet data could provide important information for fault diagnosis. Again, it is often the 'not so obvious' changes in operating procedures, such as a new supplier of a medium component, that result in aberrant operation. Fault diagnosis procedures that do not include the batch sheet data cannot be fully effective. If the analysis process is manual, it may take some time to search through the plethora of data available from each batch to find the required information.

It is apparent that there is a need for a more sophisticated means of handling batch sheet data at both research and production levels. The requirements of such a tool would be:

1. to formalise the recording process at the research level;
2. to make all data readily available to all people involved in the work;
3. to facilitate the search for differences between fermentation batches.

Classical mathematical techniques cannot be used to improve the handling of this generally qualitative batch sheet information. However, the increasing sophistication of data base technologies offers a viable means of automating the storage of batch sheet data and facilitating comparisons between the data of different fermentations. The use of data bases in fermentation facilities is described in Chapter 2.

1.1.1.2 Descriptive Data

Descriptive data are single value quantities that are used to summarise various aspects of a fermentation and include various yields, rates, other model parameters and event times as listed in Table 1.3. These values are derived from directly or indirectly measured variables using defining relations, mass and energy balances, and process and mathematical models.

Initial biomass concentration
Initial substrate concentration
Maximum growth rate
Maximum specific growth rate
Maximum substrate utilisation rate
Maximum specific substrate utilisation rate
Maximum production rate
Maximum specific production rate
Yield of biomass on substrate
Yield of product on substrate
Maintenance coefficients
Final biomass concentration
Final product titre
Event times

Table 1.3: Examples of descriptive data for a batch bacterial fermentation.

The major benefit of these data is in reporting the results of experimental work or in summarising production runs. The facts represented by these data are actually displayed in the time profiles discussed in Section 1.1.2. As it is easier for the human mind to make comparisons between pictorial data rather than numerical data it is more common for the analysis of the descriptive data to take place along with the time variant data. In some situations however, these descriptive data are important in the characterisation of a fermentation and are used in comparisons between fermentations. In these cases, the descriptive data can be treated in a data base in a similar fashion to the batch sheet data as described in Chapter 2 and illustrated in Chapter 4.

1.1.1.3 Expert Comments

The data category 'Expert Comments' includes any observations made by anyone involved in the operation or analysis of a fermentation batch. Examples would include the colour of a broth after sterilisation, the failure of any sensors, and reasons for aborting a run. These comments are generally recorded as a postscript to the batch sheet information, either in the recording notebook or on the batch sheet record itself. Very little use is made of this information. It is possible that such comments, when linked with the quantitative data of the particular batch, could be useful in fault diagnosis procedures. For example, if a batch is found to be operating in a non-standard manner it may be possible to find a historical batch that behaved in a similar manner and use the expert comments of the previous batch to infer a diagnosis on the current batch and decide the type of action required based on the outcome of the historical batch.

The expert comments are qualitative information and can therefore be treated in the same way as the batch sheet data. As mentioned previously, a data base is an appropriate medium for the treatment of qualitative data and will be introduced in Chapter 2.

1.1.2 Time Variant Data

Any variable that is monitored over the course of a fermentation falls into the category of time variant data (Table 1.4). The controlled variables such as temperature, pressure, air flow rate, agitation rate, pH and dissolved oxygen are monitored to show that the desired set points are being achieved. The variables which reflect biological activity, eg biomass concentration, substrate levels, product concentration, carbon dioxide evolution rate, oxygen uptake rate, respiratory quotient, pH (or acid or alkali addition when pH is controlled), and dissolved oxygen, are used to assess the performance of the fermentation. This section considers the use and analysis of these data, firstly in a fermentation development environment and, secondly, in an established production process.

Temperature
Pressure
Air flow rate
Agitation rate
pH
Alkali addition volume
Acid addition volume
Broth volume
Dissolved oxygen concentration
Exit carbon dioxide concentration
Exit oxygen concentration
Carbon dioxide evolution rate
Oxygen uptake rate
Respiratory quotient
Biomass concentration
Substrate(s) concentration(s)
Product(s) concentration(s)
Growth rate
Product formation rate
Substrate utilisation rate

Table 1.4: Examples of time variant data for a batch bacterial fermentation.

During the development of a new fermentation the relative performance of the experimental batches is used to determine the optimal operating conditions for the fermentation. The most common means of assessing the relative performance of the fermentations is a comparison of the profiles of the time variant data from each batch. The type of information that is readily available from the time profiles includes trends, maximum and minimum values, slopes (or rates), abrupt changes, and event times. The combination of these features forms a pattern or picture which is readily amenable to comparison with other patterns. Although most of this information can be presented numerically, for example as the descriptive data described in Section 1.1.1.2, the human mind is better able to assess pictorial representations of data especially when it is necessary to compare one set of values with another: 'Graphics reveal data. Indeed graphics can be more precise and revealing than conventional statistical computations' (Tufte 1983). Thus it would be common for a researcher to overlay time profiles from different experimental batches, either using graphics packages or manually overlaying hard copies of the profiles, and assess whether or not the patterns were identical or showed significant differences.

For a complete picture of the relative performance of a fermentation it is not sufficient to consider the time profiles of only one or two variables. The pattern of any one fermentation actually consists of the information from all the monitored and calculated variables. It is the comparison of this overall pattern with that from other fermentations that gives a true picture of the relative performance of a fermentation. This is important in ascertaining the causes of any observed differences between fermentations. For example, if the pH profiles differed between two fermentations but no other profiles showed significantly different patterns, then it is likely that the pH measurement from one of the batches was faulty but the performance of the two fermentations would still be considered similar. This comparison of all the time variant variables from one fermentation with those from another fermentation can be quite an arduous task when a large number of fermentations have been performed. It is also very difficult to achieve consistency when comparing a large number of variables simultaneously: different signal and noise magnitudes must be dealt with, conflicting results may occur due to faulty measurements, and a lot of information must be retained and interpreted.

Consistency of comparison can be a major problem in any manual assessment of data especially when the data are contaminated with both sensor and process noise. In fermentation data analysis there is the additional problem of not being able to accurately monitor the true state of the biological system: a lack of suitable sensors prevents on-line measurement, and interference by media components, time delays and inadequate techniques impair off-line measurements. Because of the inability to adequately monitor the biological state variables understanding of the process is incomplete. Therefore it is usually assumed that small differences between the quantitative data of different fermentations do not indicate significant changes in the performance of the fermentation. These small differences may be a shifting in time of an event or a difference in the rate of change or level of a variable. However, judgement of what is a significant effect and what is not is purely subjective and may differ from expert to expert or even from day to day.

No automated techniques exist to facilitate this process of comparing data sets during developmental stages; it is left up to the researcher to be as thorough and as consistent as time constraints and human ability allow.

The data analysis requirements, personnel availability and time constraints in a production environment are quite different from those in a research situation. During a production run the time variant data are used for purposes of control, fault detection and fault diagnosis. Physical control of the environmental variables such as temperature and pressure is readily achieved using conventional control algorithms (Carleysmith and Fox 1984, Wang and

Stephanopoulos 1986, Fordyce *et al.* 1990). However, any automated control requiring some knowledge of the state of the process currently relies on simple empirical models of the process. The applicability of these models is limited to use over certain portions of the fermentation and can only be considered adequate if the environmental conditions are the same as those used in developing the model and if the process disturbances and noise characteristics can be considered invariant.

Despite progress in the area of modelling of fermentations (Fish *et al.* 1989, Montague *et al.* 1989, Dhurjati and Leipold 1990), still the most commonly used means for detecting faulty operation is the by-eye observations of the experts. On-line monitored and calculated variables are usually displayed on a computer screen and the plant managers or operators ensure the profiles follow the expected pattern. This expected pattern may be a mental picture of what the profiles usually look like, or a standard profile based on previous successful batches of the fermentation. The usefulness of a graphical representation for control purposes is identified by Hale and Sellars (1981): 'The human mind has a remarkable facility for extracting key observations out of the thousands of data points displayed [graphically]. Frequent use of this view of a process leads to mental models of these patterns ... Operators become conscious that things don't look right without going through a structured analysis.'

This manual comparison of on-line time profiles suffers from a lack of consistency between operators or from day to day because of the poorly characterised nature of the fermentation process: no matter how carefully the physical environment is controlled two batches of a fermentation will never be exactly the same, the assessment of what is similar and what is not is again a matter of opinion. The difference in processes from batch to batch is a direct result of the inability to monitor the biological state variables on-line. In the event of there not being an expert present, as may happen on shift work, the visual detection scheme is further hindered as an expert is generally better able to detect when something has gone wrong and is more likely to be able to track the cause through his/her experience with the process.

In an industrial environment it is essential to detect faults as soon as they occur, find the cause of the fault and take action to rectify the situation in as short a time as possible so as to prevent irreparable damage to the process or, where necessary, terminate the batch without wasting valuable time and resources. Manual comparison of fermentations is time consuming as it is necessary to look at a number of variables: a faulty sensor may result in abnormal behaviour for one variable while a fault with the biological system would normally be manifested in aberrant behaviour in more than one variable. The situation can be improved by automating the comparison of fermentation data.

In some industrial settings a 'band profile' is produced for each of the on-line variables: each time a successful fermentation is completed, the on-line data points from this run are added to the appropriate band profile, thus producing an envelope-shaped template which defines the outer limits for the values of the particular variable for subsequent fermentations (personal communication, A Stockett, Merck Sharp and Dohme, Woodbridge, NJ, USA). If the profile of a current fermentation deviates from the expected path or moves outside the envelope then it is possible that something has gone wrong with the batch. It is important to define how many erroneous data points can be tolerated before a fault is positively identified, this will vary with the frequency of data collection and the dynamics of the process. The definition of acceptable limits for each variable by this process alleviates the problem of inconsistency between comparisons and removes the need for an expert to be present. The process is still time consuming and relies on the operating staff performing the checks at regular intervals. It is also not possible to detect the causes of faults using this technique.

It is evident that the comparison of patterns in the data is the primary technique used in the interpretation of time variant fermentation data. Automation of this comparison process would lead to a more consistent interpretation of the data and should significantly reduce the manual work load required in the analysis. Techniques have been developed for the automation of the comparative analysis of time variant data. These employ ideas from the fields of pattern recognition, qualitative reasoning and fuzzy logic and are described in detail in Chapter 3.

1.2 Definition and Extent of 'Automation'

The aim of this work, as stated earlier, was to automate the comparative reasoning techniques used by an expert in the analysis of fermentation data. It is useful to define what is meant by the term 'automate' in this context and the extent of the automation.

In this work the term 'automate' is used to describe the process of employing computer packages or software to perform tasks usually carried out by humans. Therefore, automating the comparative reasoning techniques of fermentation experts required the development of computer programs to perform the task of comparing the fermentation data. A fully automated comparative reasoning process would also require that the results of the comparisons be interpreted by the computer, however this was not attempted in this work.

In developing the comparative reasoning tools it was acknowledged that the user is not incompetent and should be encouraged to apply his/her expertise to the reasoning process. Dreyfus and Dreyfus (1986) warn against attempting to provide computer-only technology as an alternative to human expertise: '[cognitive] support systems must be designed as vehicles through which the user can exercise his expertise more effectively.' The computerised tools were therefore designed to complement the skills of the expert. In the previous section some of the shortfalls of a manual analysis were identified: often not all data are recorded; qualitative data are commonly ignored in the analysis process; data are not readily available to other researchers; the comparison of time variant data lacks consistency from one expert to another and even from day to day; and the consideration of all time variant variables is time consuming. It is in these areas that the automation of the comparative analysis process would be beneficial. Computer programs and packages were therefore developed to:

1. formally record all fermentation data in a form that facilitates reporting of the data, makes the data available to other users and enables detection of differences between different batches with particular attention to qualitative data (Chapter 2);
2. compare time variant data between batches in a manner that is consistent from one comparison to the next and ensuring that all variables can be treated with the same techniques (Chapter 3).

The user is then able to interpret the results of the comparisons using his/her expertise. This is illustrated in Chapter 4 with the analysis of a set of experimental *Escherichia coli* fermentations.

An important feature of this work is that the computerised tools are generic to all batch and fed batch stirred tank fermentation processes, there is no system specificity built into the programs. This prevents loss of salience of the tools as new processes are developed and allows the tools to be applied to processes in the developmental stages.

It should be noted that the intention of the work presented here was not to produce an industrially useful fully integrated tool, but was rather to investigate the feasibility of automating comparative reasoning techniques and to demonstrate the efficacy of the individual tools developed. Industrial implementation requires a professional programmer who is able to optimise the programs developed in this work whilst tailoring the application to the specific needs of the user.

2 COMPARATIVE ANALYSIS OF TIME INVARIANT DATA

2.1 Introduction

The category 'time invariant data' consists of the batch sheet information, the descriptive data and expert comments. The majority of the data are qualitative and thus cannot be treated with classical mathematical techniques. This chapter outlines a detailed data base structure which enables the storage, manipulation and comparison of all time invariant fermentation data.

2.1.1 Data Bases

A data base is an organised collection of interrelated data. The data stored within the data base are totally independent of the programs that use or change them. A data base generally has two major functions: firstly, to store data in a logical, organised fashion and, secondly, to allow manipulation of the data for the purpose of extracting useful information usually by way of queries from the user.

There are three main types of data base architecture: hierarchical, network and relational.

In a hierarchical data base data are represented in a tree type structure, the hierarchy is chosen by the designer. Such an architecture is only useful if all queries are to be based on items in the top level of the hierarchy because items lower down are inaccessible other than by a path from the top. Fermentation data do not lend themselves to this type of structure because data are often accessed from different levels of the hierarchy. Two user queries exemplifying this are: 'access all *Penicillium chrysogenum* fermentations' and 'access all fermentations run at a temperature exceeding 40°C'; these queries could only be executed if the desired element were the top level of the hierarchy.

A network style data base is based on records and links. The developer defines relationships between data during the initial configuration of the data base. These predefined search paths allow rapid searching, however this is at the cost of flexibility as the user cannot alter relationships between data at a later stage.

Relational data bases are the most commonly used of the architectures. The relational data model was first proposed by Dr E F Codd of IBM in 1969 (Codd 1970). In this model all data are defined and accessed as simple tables made up of rows (*records*) and columns (*fields*). Different tables can be linked if they include common fields, for example a fermentation batch number could be used to link a table describing the broth components of a fermentation with a table describing the operating conditions of the same fermentation. Although searching through the data base is slower than in a network data base, the ability to define data relationships at the time of the search confers greater flexibility on the system.

A data base can be described as a server which hands out data to any client process that needs it (Lewis 1990). In small applications the data base and associated programs reside on a single machine whilst in larger applications the server is usually situated on a mainframe or mini or on a dedicated node of a Local Area Network (LAN) and the client programs reside on a separate system, often a personal computer (PC). When the data and programs are physically separated a bridge is required to link them. In February 1987 Structured Query Language (SQL) was made the official American National Standards Institute (ANSI) standard data base language and can be used, not only to provide the link between the client and the user, but also in stand-alone or single-user PCs to manipulate relational data (Pascal 1989). SQL uses simple COBOL-like English for data definition and manipulation. The major benefits of SQL are that, firstly, a single SQL statement can replace a dozen lines of code in an application program; secondly, the user can describe the data required and not be concerned with how to retrieve it; and, thirdly, SQL allows entire sets of data to be manipulated at once, instead of one record at a time as in traditional systems. The usual procedure is for SQL statements to be embedded in either third or fourth generation programming languages and then used to interact with the relational data base: the client systems frame their requests in SQL, the SQL server interprets those requests and chooses a reasonable strategy for implementing them.

Most data base packages allow an alternative to the use of programming languages to facilitate the query process: Query By Example (QBE). In QBE an image of the data base tables is filled in by the user stipulating the requirements of the query, SQL statements are then created by the system without the user needing to know any SQL syntax. This technique is very useful for simple or dedicated applications as very little programming knowledge is required.

There are many relational data bases on the market today, all of which are being constantly updated. The choice of an appropriate system depends on the size and complexity of the application. A number of articles appear in the various computing journals and magazines each year discussing the merits of different data bases (PC Magazine 7(9) 1988, Lewis 1990, Finkelstein 1990). Of the SQL-server relational data base management systems Ingres (Ingres Corp., Alameda, CA), Oracle (Oracle Corp., Belmont, CA) and Informix (Informix Software Inc., Menlo Park, CA) dominated the market in 1990 (Lewis 1990). The PC data base market is extremely competitive with all the major players such as Microsoft, Informix, Gupta Technologies and Borland International having very good products (PC Magazine 7(9) 1988, Lewis 1990).

2.1.2 Data Bases in Fermentation

In Section 1.1 the deficiencies of current data recording and analysis techniques in fermentation plants were detailed. The most conspicuous deficiency was the omission of qualitative data from the analysis process. This was because standard computing techniques could not readily manipulate this type of data. The rapid development of relational data base technology over the past decade has provided a platform for both the organised storage and effective analysis of qualitative data.

The descriptive data (Section 1.1.1.2) can, and should, be treated in the data base environment along with the batch sheet data and expert comments. Thus all information, other than the time profile data, can be readily stored and manipulated in one place, minimising the need for manual cross-referencing and significantly reducing search times. It is possible that the time profile data could also be stored within the data base, however this was not investigated here.

The availability of commercial relational data base packages has improved since Fox (1984) described the use of archived historical files to aid the analysis of fermentations. In his work each file required a user-specified descriptive header for the fermentation, giving batch details and subsequently providing a capability for the comparison of batch details in future search-and-analyse operations. With the availability of structured relational data bases under the control of efficient data base management systems, the ideas proposed by Fox (1984) can be put into practice more readily. Fastert (1990) observed that the use of data bases for the storage of fermentation data is in fact becoming more popular, citing the ease of retrieving data based on user-specified search parameters as the principal reason for the increasing usage.

Morris *et al.* (1991) have taken full advantage of emerging technologies in their proposed fermentation supervisory control system which employs both a relational data base (Ingres) and knowledge based systems. The data base records all on-line and off-line process data as well as set-up information and provides this information to the knowledge bases and other components of the system as required. The structure and function of the data base was not discussed in detail but its inclusion in the supervisory system is indicative of the benefits this technology can impart relative to the traditional data storage and retrieval methods.

2.2 Development of a Fermentation Data Base

This section describes the development of a data base for fermentation data. All data available from a fermentation were listed and grouped into table-like structures, emulating the format of a relational data base. The usage of the data base is also described: the user inputs required (Section 2.2.2), the outputs provided by the data base (Section 2.2.3) and the means by which the data base aids the desired comparative analysis of fermentation data (Section 2.2.4).

2.2.1 Data Base Structure

A relational architecture was chosen as the most appropriate for the development of a fermentation data base as it is important to be able to analyse data using any type of relationship and to change that relationship as required.

In any practical situation a server system would most likely be required because of the amount of data generated in a fermentation environment and to enable multiple users to gain access to the data. The emphasis of this work was on defining the structure of the data for the data base and not on creating a complete working system. The adaptation of the data structure to a working system will be specific to the particular data base chosen. Time constraints, software and hardware availability and the complexity of dealing with a large server system dictated the choice of a simple platform for development work: the personal computer (PC). SmartWare II (Informix Software Inc., Menlo Park, CA), was chosen for the development work. This is a stand-alone PC-based package which has its own programming language and allows all common data base transactions. The use of the SmartWare II Data Base enabled better understanding of the capabilities of relational data bases.

The planning stages of data base development are extremely important. In this work the required information, and its organisation, were carefully thought out prior to the building of the data base. The structure of the fermentation data base evolved by firstly listing all pieces of information available from a fermentation and then arranging the data into linked tables, emulating the format of a relational data base. This study was concerned only with batch and fed batch stirred tank fermentations, information specific to fermentations utilising immobilisation technologies, solid-state fermentations, continuous fermentations and biotransformations were not considered.

Two main types of tables were defined for this work: the *batch sheet tables* and the *inventory tables*.

The *batch sheet tables* contain all the information pertaining to the fermentation batches: each row, or *record*, represents a different fermentation batch and each column, or *field*, relates to a specific piece of information such as the vessel used. Some pieces of information cannot be described by single entries and thus require their own tables linked to the main table. For example a description of the medium composition of a batch requires a list of the medium components, their concentrations, and the grade, lot number and supplier of each component. This information can not reside in a single record thus a 'media' table is required. The main batch sheet table and its appendices are described in Tables 2.1 to 2.11.

The descriptive data are often peculiar to a given fermentation and thus are not included in the standard batch sheet table. For example, in one type of fermentation the maximum rate of substrate utilisation may be important whilst in another fermentation, which utilises a complex medium, substrate concentration may not be observable. The descriptive data are thus placed in *fermentation specific fields* in the data base. The user defines the individual fields as required for each process. Fermentation specific fields can be added at any time.

FIELDS	DETAILS
Batch Number	A string that uniquely identifies the batch
Organism	From the *Organism* inventory
Type of Organism	Automatic from *Organism* inventory
Strain	From the *Strain* inventory
Plasmid	From the *Plasmid* inventory
Purpose of Batch	Seed/growth/production
Oxygen Requirement	Aerobic/anaerobic/facultative
Number of Stages	Input by operator
Seed Type	From *Seed Type* inventory
Stage Number	Input by operator
Vessel Id	From *Vessel* inventory
Vessel Type	Automatic from *Vessel* inventory
Vessel Volume	Automatic from *Vessel* inventory
Mode of Operation	Batch/fed batch/continuous
Medium Name	From *Medium* inventory (if standard medium)
Mix Tank Batch Number	If medium mixed in a mix tank
Batched By	Operator responsible for media make up
Volume From Mix Tank	If medium mixed in a mix tank
Liquid Volume	Initial working volume in fermenter
Inoculum Volume	Could be volume/spore count/no. of loops etc
Inoculated By	Operator responsible for inoculation
Start Date	Input by operator
Start Time	Input by operator
End Date	Input at end of run
End Time	Input at end of run
Length of Run	Calculated
Comments	Observations made by operator

Table 2.1: A list of the column, or field, headings for the main data base table with details of the type of input required for the respective fields. (Id = identification number)

FIELDS	DETAILS
Batch Number	A link to the main data base
Product	From the *Product* Inventory
Type of Production	Primary/secondary/unknown
Comments	eg virus like particles, inclusion bodies

Table 2.2: The column, or field, headings for the product data base table. A new record is required for each different product expected.

FIELDS	DETAILS
Batch Number	A link to the main data base
Seed Id	Input by the operator

Table 2.3: The column, or field, headings for the seed data base table. (Id = identification number)

FIELDS	DETAILS
Batch Number	A link to the main data base table
Inoculum Source	Batch number of previous stage. If first stage this will be the seed identification number.

Table 2.4: The column, or field, headings for the inoculum data base table.

FIELDS	DETAILS
Batch Number	A link to the main data base table
Material	From *Materials* inventory
Concentration	Standardise units
Amount	Standardise units
Supplier	From *Supplier* inventory
Grade	Input by operator
Lot Number	Input by operator (if more than one lot number used then add another record, ie another row)
Sterilisation Group	Link to sterilisation data base

Table 2.5: The column, or field, headings for the medium data base table.

FIELDS	DETAILS
Batch Number	A link to the main data base table
Sterilisation Group	Each medium component (Table 2.5) is sterilised in a group, identified by this number
Method	From *Sterilisation* inventory
Equipment	From *Equipment* inventory
Temperature	Depending on the method, the appropriate fields are filled in
Pressure	
Length	
Volume	
Pre-sterile Volume	
Post-sterile Volume	
Pre-sterile pH	
Post-sterile pH	
Number of Loops	
Flow Rate	
Data File	The file containing the sterilisation data (if appropriate)

Table 2.6: The column, or field, headings for the sterilisation data base table.

FIELDS	DETAILS
Batch Number	A link to the main data base table
Variable	From *Variables* inventory
Set Point	Input by operator
Type of Control	From *Control* inventory
High Alarm	Input by operator
Low Alarm	Input by operator
Set By	Supervisor/operator
Data File	File containing data
Comments	Observations made by operator

Table 2.7: The column, or field, headings for the initial operating condition data base table.

FIELDS	DETAILS
Batch Number	A link to the main data base table
Variable	From *Variables* inventory
Age	Time of change
Set Point	Input by operator
Type of Control	From *Control* inventory
High Alarm	Input by operator
Low Alarm	Input by operator
Set By	Supervisor/operator
Basis	Scheduled/unscheduled
Comments	Observations made by operator

Table 2.8: The column, or field, headings for the operating condition changes data base table. A new record is required for each variable whose status changes and for each operating condition change.

FIELDS	DETAILS
Batch Number	A link to the main data base table
Material	From *Materials* inventory
Concentration	Standardise units
Supplier	From *Supplier* inventory
Grade	Input by operator
Lot Number	Input by operator
Method	Shotwise/continuous/on-demand
Basis	Age or condition eg pH<7.0
Equipment	From *Equipment* inventory
Amount/Feedrate	Input by operator
Sterilisation Group	Link to sterilisation data base
Data File	File containing data if appropriate

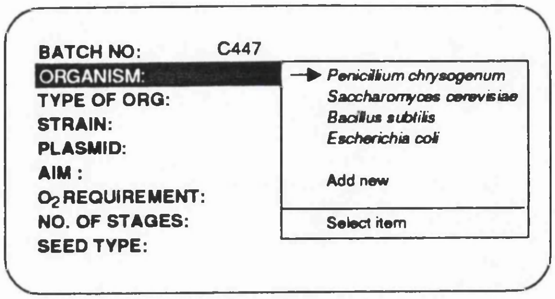
Table 2.9: The column, or field, headings for the feeds data base table. A new record is required for each feed and each change to the feed regime.

FIELDS	DETAILS
Batch Number	A link to the main data base table
Variable	From *Variables* inventory
Measurement Technique	Reference to technique
Measuring Instrument	From *Instruments* inventory
Measured By	Operator
Age	Input by operator
Value	Input by operator

Table 2.10: The column, or field, headings for the descriptive data table.

FIELDS	DETAILS
Batch Number	A link to the main data base table
Variable	From *Variables* inventory
Measurement Technique	Reference to technique
Measuring Instrument	From *Instruments* inventory
Measured By	Operator
Data File	File containing data

Table 2.11: The column, or field, headings for the time variant data table.



BATCH NO: C447
ORGANISM:
TYPE OF ORG:
STRAIN:
PLASMID:
AIM :
O_2 REQUIREMENT:
NO. OF STAGES:
SEED TYPE:

Figure 2.1: Example of the use of an inventory table to standardise data base input.

The second main type of table, the *inventory tables*, are like dictionaries as they provide an inventory of all previously, or commonly, used pieces of information such as types of organism and equipment available. An example of the use of an inventory table is given in Figure 2.1: a new batch is being added to the data base and the operator is about to input the name of the organism used, rather than type in the name, the operator selects it from the inventory table which is displayed when the *Organism* field is highlighted. If the required name were not present in the inventory the table would be updated by the inclusion of the new name and the data base management system would request other pertinent information such as the strain used. In this way the inventory tables grow with time. The purpose of the inventory tables is to standardise input: spelling mistakes are prevented and any one piece of information can have only one format so that data base searches and comparisons are facilitated. The different inventory tables are listed in Table 2.12.

The structure of the fermentation data base is summarised in diagrammatic form in Figure 2.2. The link between the main batch sheet table and the subsidiary batch sheet tables is the *Batch Number* which is the common field. Each record in the main table must have a unique batch number but the subsidiary tables may contain a number of records for each fermentation thus the batch number may be repeated, for example if there were two products expected from a particular fermentation then there would be two records with the same batch number in the *Product* data base table. The inventory tables are linked to the rest of the data base through the field names as described above and shown in Figure 2.2.

The fermentation data base was designed to fulfil two major tasks: firstly, the input and output of fermentation data for purposes of reporting and documentation and, secondly, comparative analyses of fermentations. These are discussed in the following sections.

DATA BASE	FIELDS	DETAILS
Organism	Organism Name	
	Type of Organism	bacteria/yeast/mould/...
	Oxygen Requirement	aerobic/anaerobic/facultative
Strain	Name of Strain	
	Organism Name	May have many different strains of one organism, but each strain is specific to only one organism
Plasmid	Name of Plasmid	
	Comments	eg stability
Product	Name of Product	
Seed Type	Seed Type	spores/lyophilised/loop/frozen suspension/...
Vessel	Vessel id	
	Vessel Type	Erlenmeyer/stirred tank/airlift/...
	Vessel Volume	
Medium	Medium Name	Some commomly used media are given names for ease of identification
Supplier	Supplier's Name	
Materials	Name of Material	
Sterilisation	Method	filtration/autoclave/continuous/batch - direct steam/batch - indirect steam
Equipment	Equipment id	eg type of filter, id of pump
Variables	Variable Name	
Control	Type of Control	automatic/cascade/...
Instruments	Instrument id	Instruments used for monitoring/measuring variables

Table 2.12: Details of the Inventory Data Bases used to standardise input to the Fermentation Batch Sheet Data Bases.
(id = identification number)

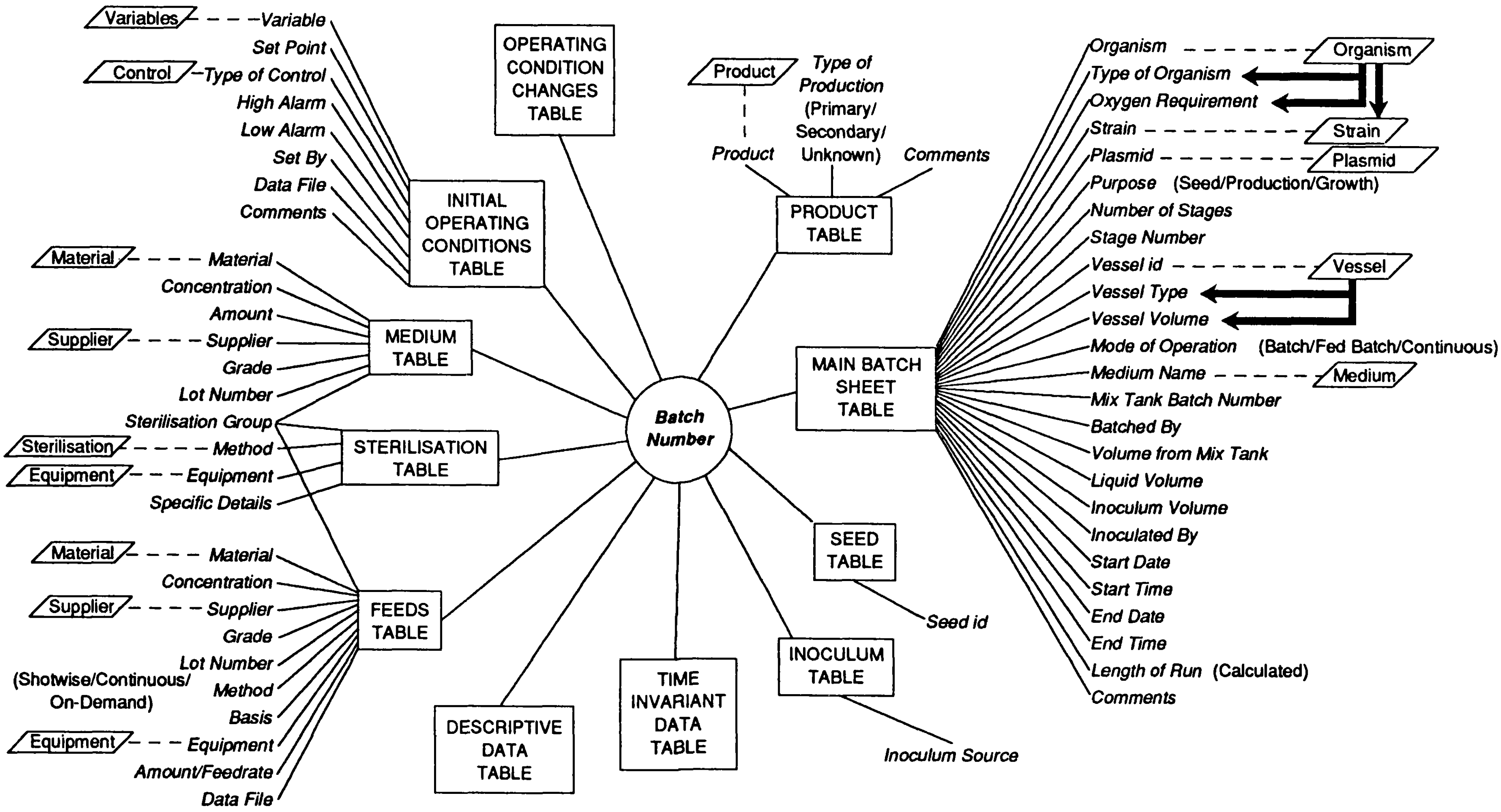


Figure 2.2: Summary of the data base structure. For any one fermentation batch there may be multiple entries in all of the tables except the main batch sheet table. The fields for each batch sheet table are shown in italics. (The fields are not shown for three of the tables). The connections between the batch sheet fields and the inventory data bases are shown by dashed lines. The inventory data bases are identified by slanted boxes. The thick arrows depict the automatic transfer of information from an inventory data base. More details of the data base tables are given in Tables 2.1 to 2.12.

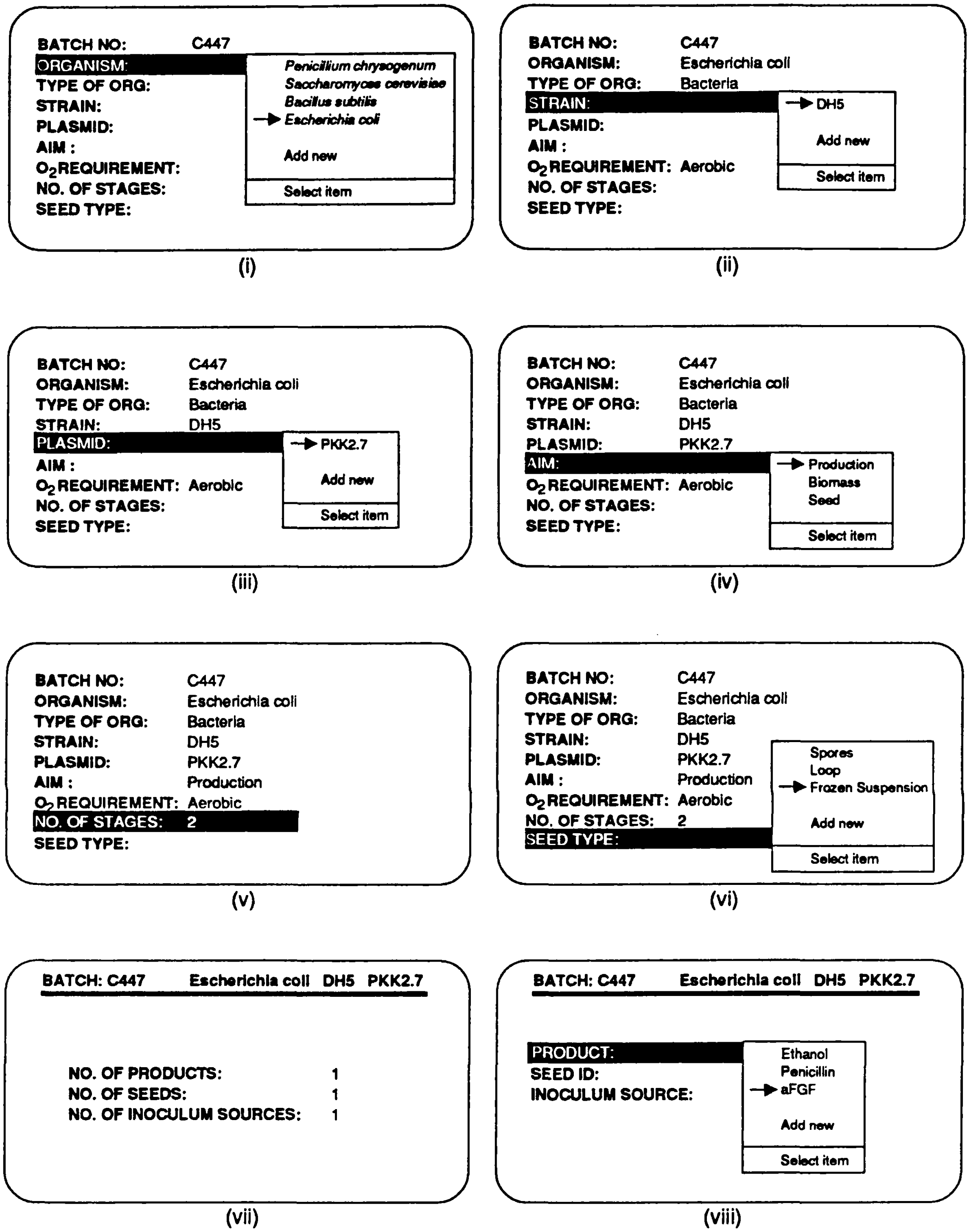
2.2.2 Data Input

Meticulous data recording is essential if the data base approach is to be successful: incomplete records would not only result in gaps in the information about the process but would also complicate the comparison of different data sets. Some fields, such as *Batch Number*, must be filled for every batch. However, it should be possible to leave fields empty when data are not available or are known to be incorrect, although undesirable this may be unavoidable. When gaps do occur in a data base record the comparison of that record with other records must indicate that data were unavailable otherwise false conclusions may be drawn (Section 2.2.4).

The data base management system should help the user during the data input process by specifying the order in which fields should be filled and by providing standardised inputs when required by way of the inventory tables. It is essential that any gaps in the data base can be filled after completion of the batch, for example at the time of data input it may not be known which pump will be used for a particular feed, this can be included in the data base when the information becomes available.

Some of the steps in the setting up of an *Escherichia coli* fermentation record (described in Chapter 4) are outlined in Figure 2.3. The figure shows a 'user friendly' interface which guides the user through the input procedure step by step. A format such as this ensures that no data are omitted in the recording process. The inventory data base tables are used to standardise and facilitate the input of data. It can be seen in Figure 2.3 that when the *Organism* field has been filled the *Type of Organism* and *Oxygen Requirement* fields are automatically filled from information in the *Organism* inventory table. The *Organism* inventory table is linked to the *Strain* inventory table so that only those strains that correspond to the organism being used appear in the pop-up menu for the *Strain* field. These links were summarised in Figure 2.2.

The above processes describe the input of qualitative data to the data base. The input of quantitative data is complicated by the need to include the uncertainty in the measurement. This is best achieved by recording the value, or mean value, in one field and then, in response to a request from the data base management system, providing the size of the uncertainty. From this the data base management system can record the data value as a range rather than an exact number. Where standard errors are involved, the data base management system can provide the required uncertainty data. The use of quantitative data in the data base is discussed in more detail in Section 2.2.4.



BATCH NO: C447
ORGANISM:
TYPE OF ORG:
STRAIN:
PLASMID:
AIM :
O_2REQUIREMENT:
NO. OF STAGES:
SEED TYPE:

BATCH NO: C447
ORGANISM: Escherichia coli
TYPE OF ORG: Bacteria
STRAIN:
PLASMID:
AIM :
O_2REQUIREMENT: Aerobic
NO. OF STAGES:
SEED TYPE:

BATCH NO: C447
ORGANISM: Escherichia coli
TYPE OF ORG: Bacteria
STRAIN: DH5
PLASMID:
AIM :
O_2REQUIREMENT: Aerobic
NO. OF STAGES:
SEED TYPE:

BATCH NO: C447
ORGANISM: Escherichia coli
TYPE OF ORG: Bacteria
STRAIN: DH5
PLASMID: PKK2.7
AIM:
O_2REQUIREMENT: Aerobic
NO. OF STAGES:
SEED TYPE:

BATCH NO: C447
ORGANISM: Escherichia coli
TYPE OF ORG: Bacteria
STRAIN: DH5
PLASMID: PKK2.7
AIM : Production
O_2REQUIREMENT: Aerobic
NO. OF STAGES: 2
SEED TYPE:

BATCH NO: C447
ORGANISM: Escherichia coli
TYPE OF ORG: Bacteria
STRAIN: DH5
PLASMID: PKK2.7
AIM : Production
O_2REQUIREMENT: Aerobic
NO. OF STAGES: 2
SEED TYPE:

BATCH: C447 Escherichia coli DH5 PKK2.7

NO. OF PRODUCTS: 1
NO. OF SEEDS: 1
NO. OF INOCULUM SOURCES: 1

BATCH: C447 Escherichia coli DH5 PKK2.7

PRODUCT:
SEED ID:
INOCULUM SOURCE:

Figure 2.3: The first stages in entering a new fermentation record to the data base. The data base management system guides the user through the input routine ensuring that all available information is entered. The pop-up menus provide a list of options for each entry.

2.2.3 Data Output

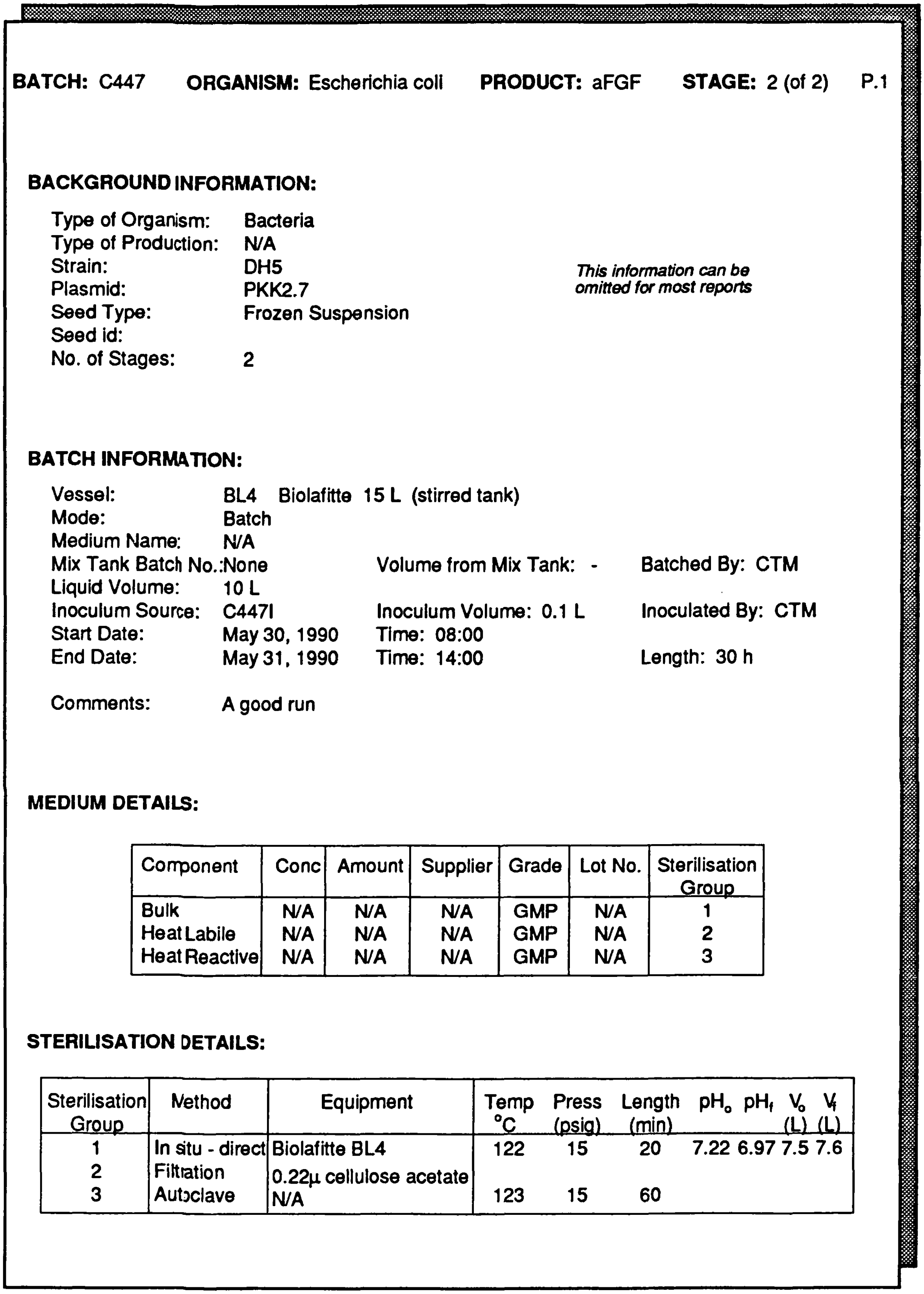
The output of batch information is important for reporting and for data analysis. Most data base packages have a reporting capability through which standardised output can be programmed according to the requirements of the user. A complete data base record for the production stage of a laboratory scale *Escherichia coli* acidic fibroblast growth factor (aFGF) fermentation (Chapter 4) is presented in Figure 2.4 (some details were omitted for proprietary reasons). It may not be necessary to include all information in some reports and the output can be tailored to meet the needs of the user.

2.2.4 Data Comparisons

The major purpose of this work was to develop techniques for comparative analysis of fermentation batches. A data base environment is ideal for the treatment of qualitative and single-value quantitative data. The data base tables described earlier (Tables 2.1 to 2.12) were designed specifically to facilitate the comparison of data from different fermentations.

It was not within the scope of this project to programme the comparison of the data base information as it is thought that individual users will tailor the system to suit their own needs. The programming will be dependent on the data base chosen. The envisaged process would be as follows.

The batches to be compared are specified directly by the user or are selected as a result of a search through the data base. The search for appropriate batches is easily carried out using Query By Example (QBE): the user enters information into various fields of an image of the data base tables specifying the requirements of the query. For example to retrieve all *E. coli* fermentations in which aFGF was the product and the inoculum concentration was 0.25%, '*Escherichia coli*' would be entered into the *Organism* field, 'aFGF' would be entered into the *Product* field and 0.25 would be entered into the *Inoculum Volume* field. The result of this query would be a list of all the records satisfying these criteria.



BATCH: C447 **ORGANISM:** Escherichia coli **PRODUCT:** aFGF **STAGE:** 2 (of 2) P.1

BACKGROUND INFORMATION:

Type of Organism:	Bacteria
Type of Production:	N/A
Strain:	DH5
Plasmid:	PKK2.7
Seed Type:	Frozen Suspension
Seed id:	
No. of Stages:	2

This information can be omitted for most reports

BATCH INFORMATION:

Vessel:	BL4 Biolafitte 15 L (stirred tank)		
Mode:	Batch		
Medium Name:	N/A		
Mix Tank Batch No.:	None	Volume from Mix Tank: -	Batched By: CTM
Liquid Volume:	10 L		
Inoculum Source:	C447I	Inoculum Volume: 0.1 L	Inoculated By: CTM
Start Date:	May 30, 1990	Time: 08:00	
End Date:	May 31, 1990	Time: 14:00	Length: 30 h

Comments: A good run

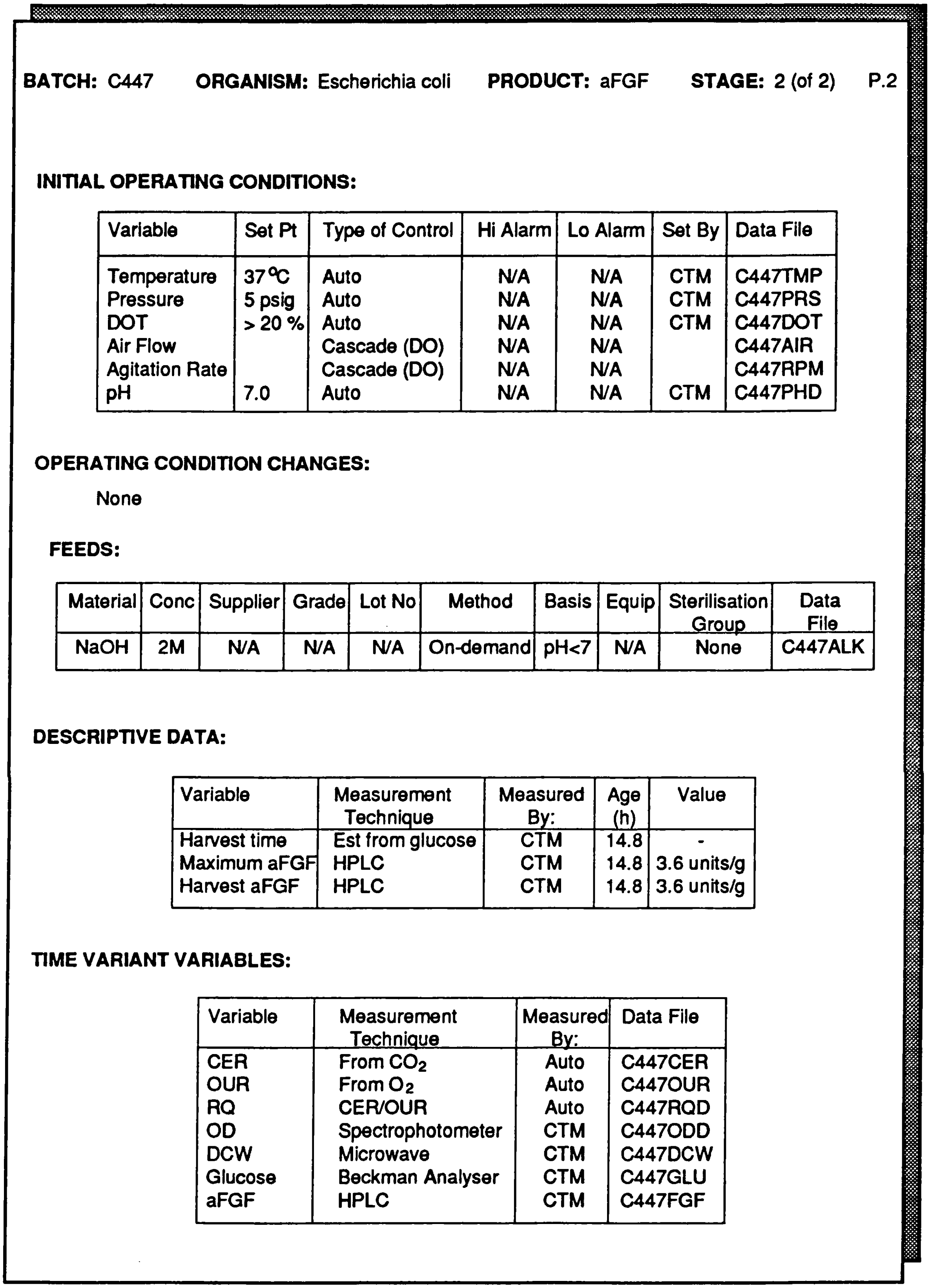
MEDIUM DETAILS:

Component	Conc	Amount	Supplier	Grade	Lot No.	Sterilisation Group
Bulk	N/A	N/A	N/A	GMP	N/A	1
Heat Labile	N/A	N/A	N/A	GMP	N/A	2
Heat Reactive	N/A	N/A	N/A	GMP	N/A	3

STERILISATION DETAILS:

Sterilisation Group	Method	Equipment	Temp °C	Press (psig)	Length (min)	pH_o	pH_f	V_o (L)	V_f (L)
1	In situ - direct	Biolafitte BL4	122	15	20	7.22	6.97	7.5	7.6
2	Filtration	0.22µ cellulose acetate							
3	Autoclave	N/A	123	15	60				

Figure 2.4 (a): A sample data base report. Some data are not presented for proprietary reasons, eg a full list of medium components. More detailed reports for each data base table can be requested. (N/A = unavailable information). Continued in Fig 2.4 (b).



BATCH: C447 **ORGANISM:** Escherichia coli **PRODUCT:** aFGF **STAGE:** 2 (of 2) P.2

INITIAL OPERATING CONDITIONS:

Variable	Set Pt	Type of Control	Hi Alarm	Lo Alarm	Set By	Data File
Temperature	37°C	Auto	N/A	N/A	CTM	C447TMP
Pressure	5 psig	Auto	N/A	N/A	CTM	C447PRS
DOT	> 20 %	Auto	N/A	N/A	CTM	C447DOT
Air Flow		Cascade (DO)	N/A	N/A		C447AIR
Agitation Rate		Cascade (DO)	N/A	N/A		C447RPM
pH	7.0	Auto	N/A	N/A	CTM	C447PHD

OPERATING CONDITION CHANGES:

None

FEEDS:

Material	Conc	Supplier	Grade	Lot No	Method	Basis	Equip	Sterilisation Group	Data File
NaOH	2M	N/A	N/A	N/A	On-demand	pH<7	N/A	None	C447ALK

DESCRIPTIVE DATA:

Variable	Measurement Technique	Measured By:	Age (h)	Value
Harvest time	Est from glucose	CTM	14.8	-
Maximum aFGF	HPLC	CTM	14.8	3.6 units/g
Harvest aFGF	HPLC	CTM	14.8	3.6 units/g

TIME VARIANT VARIABLES:

Variable	Measurement Technique	Measured By:	Data File
CER	From CO_2	Auto	C447CER
OUR	From O_2	Auto	C447OUR
RQ	CER/OUR	Auto	C447RQD
OD	Spectrophotometer	CTM	C447ODD
DCW	Microwave	CTM	C447DCW
Glucose	Beckman Analyser	CTM	C447GLU
aFGF	HPLC	CTM	C447FGF

Figure 2.4 (b): A sample data base report. Some data are not presented for proprietary reasons. More detailed reports for each data base table can be requested. (N/A = unavailable information). Continued from Fig 2.4 (a).

Once the appropriate batches have been selected the data base itself can only provide a listing of all the information about each batch. The actual comparison of this information is achieved through a programming language. The comparison program must compare each field in the batches being considered and record those fields that differ. As mentioned earlier the report of the comparison must also indicate those fields that are empty so as to prevent incorrect assumptions being made. Any 'expert comments' available for any of the batches being compared must also be included in the comparison summary as these may help the interpretation. The result of the comparison of the time invariant data is thus a *difference summary* containing:

1. a list of differences between the data contained in the respective data base records;
2. a list of all the missing data (empty fields) in the batches being compared;
3. any expert comment from the batches of interest.

This information is then combined with the outcome from the comparison of the time variant data (Chapter 3) and the analyst interprets the overall results.

The comparison of quantitative data is somewhat more complex than that of qualitative data. For example a final product titre of 1.5 $g.L^{-1}$ may not be significantly different from a titre of 1.7 $g.L^{-1}$, whereas two pH values that differ by 0.05 may be considered different. For this reason imprecise quantitative data must be recorded as an interval representing the range of uncertainty of the data. For some variables, such as pH, the uncertainty bounds will be a constant, eg +/- 0.02. This fact can be stored in the data base and linked to all *pH* fields (eg pH of a broth before and after sterilisation) so that when a pH value is read in it is automatically converted to a range encompassing 0.02 units either side of the reading. Other variables have non-constant uncertainty bounds in which case the operator must input the lower and upper limits of the data. The comparison of these data then requires investigation of the intervals: overlapping intervals indicate similarity whilst non-overlapping intervals indicate dissimilarity. Examples of the comparison of time invariant quantitative data are given in Chapter 4.

With sufficient foresight, the data base developer should be able to create a front-end process by which an inexperienced user can perform any desired query. Alternatively, an environment in which queries can easily be built would be desired.

2.3 Discussion

The amount of data available from a single fermentation is very large, as shown in the data base tables presented in this chapter. This is the main reason why information is often not recorded, especially during research work where only the 'essential' data are noted. By formalising the data recording process all data will be recorded and all relevant information will then be available for analysis. The data base also facilitates the dissemination of information as other personnel have access to the data in a standardised form.

The aim of this work was to investigate the feasibility of automating the comparative analysis of fermentation data. The data base described above allows data to be organised in such a way as to enable comparisons between data sets. The most important advantage of using a data base is that all the information is collected in one place in a structured manner. The inventory tables ensure that the input is standardised, eliminating potential grammatical problems.

A considerable amount of work is required in the development of a data base for a fermentation facility. The benefits achieved by formalising the recording of fermentation data alone justify the effort required. In Chapter 4 the data base tables are shown to be of great use in analysing a set of real fermentations. The comparative analysis of the data in these tables not only provided a list of differences between the fermentations, but also prompted the analyst to consider all the data before establishing the possible causes of differences in the performance of the fermentations. A much greater understanding of the fermentation was achieved. This further reinforces the usefulness of a data base in fermentation work.

2.4 Future Extensions

As mentioned earlier, it should be possible to include the time profile data in the data base, thus eliminating the need for different storage areas. With the possibility of embedding SQL statements in third generation languages the integration of the time profile comparisons and data base comparisons could be implemented thus reducing the amount of manual intervention.

The introduction of multi-media data bases (Informix OnLine, Informix Software Inc., Menlo Park, CA) brings with it the ability to store images, a feature which could be exploited to improve the reporting and analysis of fermentations. For example images from microscopic investigations or image analysers could be recorded and thus be available for analysis, comparative or otherwise, whenever required. It is also possible that the graphical comparison procedures outlined in Chapter 3 could be simplified using this medium but, at this stage, it is unclear how this could be achieved or whether the software systems are powerful enough to direct these analyses.

With the rapid advancement of computerised technologies it appears that there is never-ending potential for improving the way processes are monitored, controlled and analysed, both in research and production. Often it is not possible or feasible to investigate all these possibilities. Data bases are one technology that is worth adopting: the use of data bases for the storage of data greatly facilitates both reporting and retrieval of information for further analyses; and, if it can be used on a plant- or even company-wide scale, a data base can be even more powerful with data transfer between disciplines being greatly facilitated (making adequate allowance for security of sensitive information) resulting in a better awareness of the process as a whole (Svenson and McLean 1991).

3 COMPARATIVE ANALYSIS OF TIME VARIANT DATA

3.1 Introduction

The analysis of a fermentation involves comparison of the time variant data with data from other fermentations or with a model (Chapter 1). Computer routines were written to automate this comparative analysis: the routines simplify the data using a piecewise linear representation, describe the simplified data using approximate qualitative terms and compare the qualitative descriptions of one data set with another. The comparison process is summarised in Figure 3.1. This section reviews the relevant literature, explains the choice of a linear representation of the data and provides the rationale for the qualitative descriptors used in describing the data. Sections 3.2, 3.3 and 3.4 describe the computer routines for the simplification, qualitative description and comparison of the data respectively. The interpretation of the results is manual and is described in Section 3.5. The overall process for the analysis of time variant data is discussed in Section 3.6 and is applied to a set of experimental fermentations in the following chapter.

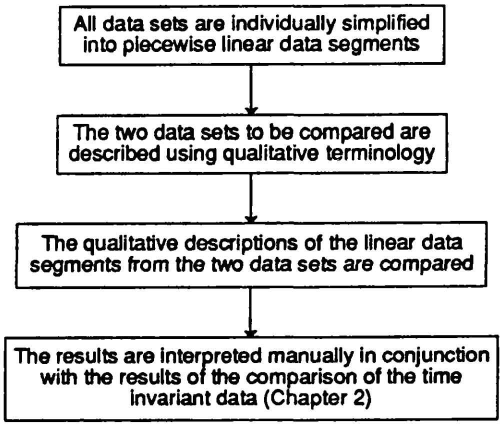

All data sets are individually simplified into piecewise linear data segments
The two data sets to be compared are described using qualitative terminology
The qualitative descriptions of the linear data segments from the two data sets are compared
The results are interpreted manually in conjunction with the results of the comparison of the time invariant data (Chapter 2)

Figure 3.1: Summary of the procedure for the comparative analysis of time variant fermentation data. All data sets are individually simplified by piecewise linearisation. The two data sets to be compared are then described using the approximate qualitative terminology of an expert. The data sets are then compared and the results interpreted manually. The first three steps are performed by computer routines.

3.1.1 Data Simplification

Many of the methods described in the literature for the analysis of graphical data consider groups of consecutive data points, rather than the data points themselves, as the primitive element of the graph. There are a number of reasons for taking this approach to simplify the data being analysed, those pertinent to fermentation data are as follows:

1. the effect of noise in the measurements is reduced;
2. when the data are to be compared with other data sets it is likely that the sampling points for the two sets do not coincide, thus a point by point comparison is not possible;
3. the manual analysis of fermentation time profiles would typically involve comparison of the characteristic microbial growth phases (Bailey and Ollis 1986) of the two batches being considered. The batch to batch variation commonly observed in fermentation processes is often manifested by slight variations in the lengths of these growth phases. This precludes the use of a point by point comparison of data sets: very little information can be obtained from comparing, say, a point in the lag phase with one in the exponential phase;
4. as alluded to in point 3, the grouping of data points emulates the way an expert views the data.

This simplification of the data is very much dependent on the type of information required from the data set and the form of the grouping is chosen to accentuate the underlying trends, or patterns, inherent in the data. The grouping of the data may be based solely on domain dependent knowledge, that is information specific to the process of interest, or it could be based on the geometric characteristics of the profiles themselves.

The grouping of fermentation data into distinct phases is an example of the use of domain dependent knowledge. Much of the recent work in automating fermentation data analysis for control purposes has concentrated on the development and use of process models which utilise domain dependent knowledge in the same way an expert would. The models created are based on an expert's view of the time variant data, ie the time profiles are segmented, by an expert, into intervals which reflect the physiological states of the organism (Halme 1989, Konstantinov and Yoshida 1989, 1990a, 1990b, Stephanopoulos and Tsiveriotis 1989, Locher *et al.* 1990, Morris *et al.* 1991). Process data are then compared with these models to enable identification of the current state of the bioreactor and thus trigger applications appropriate to this state. Where there are sufficient consistent data from a process, such models may prove valuable for state identification. However, in the developmental stages of a fermentation the trajectories of the process variables are not consistent because of the changes in operating conditions and therefore it is not feasible, at this stage, to define a model which describes the process. If the analysis of experimental data were to be based on the comparison of data grouped according to physiological states then an expert would be required to segment the data from each individual batch. Very little advantage could be achieved from automating the remainder of the analysis process.

In situations where domain dependent information is not useful in segmenting time profiles, the geometric characteristics of the profile may relay useful process information. Geometric characteristics describe the shape of a graph and do not require any process specific information for their definition. The geometric characteristics of interest are very much dependent on the type of process: linear features were considered important in the analysis of electrocardiograms (Udupa and Murthy 1980, Pietka 1991) and in the assessment of the performance of an aluminium rolling mill (Love and Simaan 1988) whilst curved data segments were used to identify subterranean geological formations in the analysis of pressure-time curves from well tests (McIlraith 1989) and in the extraction of trends from chemical engineering process data (Cheung and Stephanopoulos 1990a, 1990b). However, the geometric representation identified for a particular type of process is generic to all examples of that process.

A geometric representation of fermentation data was preferred to a domain dependent representation for two main reasons:

1. no process specific information is required for the simplification of the data thus the same tools can be used for any fermentation process;
2. no prior knowledge of the fermentation is required to perform the simplification and so the same tools can be used for the analysis of both developmental and industrial data.

The analysis of time variant fermentation data commonly involves consideration of such features as the occurrence, extent and level of any maxima or minima, the rate at which these extremes are approached, and the time and level at which rates change. An expert would thus segment a time profile into intervals of roughly constant slope and use this representation as the basis for interpretation. It would therefore seem reasonable to use a piecewise linearisation of the data as a starting point for the automatic analysis of time variant fermentation data. In this work techniques were developed to simplify fermentation data into piecewise linear segments.

The development of a piecewise linearisation scheme for noisy data is not trivial. The most important point to be considered is how the end points of each line are to be specified. If domain specific information is not to be used the end points must be found from the data themselves. The linearisation methods used by Pietka (1991) and Love and Simaan (1988) were investigated for their applicability to fermentation data.

Pietka (1991) used the 'fan method' (Blanchard and Barr 1985) to segment electrocardiograms (ECGs) into linear data pieces. In the fan method the first data point is taken as the anchor of the first line segment, straight lines are drawn from this anchor to successive data points until the maximum distance between the line segment and the points covered by the line exceeds a prespecified threshold value. At this stage the end point of the line segment is fixed and the process is repeated using this data point as the anchor for the next line segment. The fan method of linearisation is not appropriate to use in the presence of noisy data, as in fermentations, for two reasons. Firstly, it places considerable weight on the data points at which the lines begin and end. A technique which finds a 'best fit' line through the data, with each point having an equal weighting, would be more suitable. Secondly, if the data contained noise spikes (that is 'blips' in the data), the use of the maximum distance between the line and the raw data as the goodness of fit criterion would be inappropriate, noise spikes would not be removed from the data set without prior filtering. A measure of the average distance between the line segment and the raw data would be a better indicator for the goodness of fit in the presence of noisy data as this would enable smoothing of the noise spikes.

Love and Simaan (1988) employed a combination of three nonlinear filters to smooth data from an aluminium rolling mill and extract the linear features. The filters were variations on the general moving average filter and essentially removed noise from the data by replacing groups of consecutive points with the mean of those points. Piecewise linearisation was achieved by smoothing the slopes between consecutive data points in a similar manner. This linearisation process required the prior specification of four 'fitting variables': two filter lengths (the number of consecutive points that were averaged in the moving average filters) and two threshold values (values which distinguish noise from real process occurrences). Love and Simaan (1988) chose these fitting variables by trial and error but indicated that the values chosen have considerable effect on the resulting data structure. For a completely automatic system in which human bias is to be minimised the fitting variables should either be fixed for all data sets or should be automatically chosen as some function of the underlying noise of the signal.

A piecewise linearisation technique was developed for the simplification of fermentation data. The problems inherent in the simplification techniques described above were used as guidelines for the development of the novel linearisation technique:

1. the method must deal with noisy data that often contains noise spikes, thus an average distance, rather than a maximum distance, should be used as the measure of the acceptability of the fit of the linear data piece;
2. the weighting on all data points should be equal;
3. the number of fitting variables that require prior specification should be minimised and, if possible, the fitting variables should be related to the underlying noise of the signal.

Minimisation of the number of fitting variables also implied that the data should not require smoothing prior to the linearisation procedure as this would involve yet another fitting variable.

The piecewise linearisation procedure for the simplification of fermentation data uses a robust statistical fitting technique (M-estimates) to fit a line to a portion of a data set. The line is then extended or reduced until the average deviation between it and the raw data is within a limit specified by an estimate of the signal noise. Consecutive lines are extrapolated or interpolated so that they intersect but if in so doing the fitting requirement is violated the lines must be refitted over fewer data points. The process is repeated until the whole data set is covered. The piecewise linearisation technique is described in Section 3.2.

A number of fermentation researchers use curve fitting techniques to aid data analysis. Prior to commencing work on piecewise linearisation, a cubic spline fitting technique was investigated as a tool for the comparative analysis of fermentation data. The methods were complicated by the specification of smoothing parameters. Furthermore, it was found that simply smoothing the data, as is the general aim of fitting techniques, does not facilitate the comparison of fermentation data sets because of the batch to batch variations alluded to earlier: further tools must be employed to delineate which sections of two profiles are to be compared. The cubic spline technique is described in detail in Appendix 1.

3.1.2 Qualitative Representation of the Data

The expert analysis of a profile is not usually a purely quantitative procedure, it is common to consider the graph using a combination of qualitative and quantitative information. The linear segments described in the previous section may be viewed in terms of their relative slopes, for example 'steep' or 'shallow', and their relative lengths, for example 'long' or 'short'. The use of qualitative abstractions of numerical quantities facilitates the analysis process by providing a looser description of the data thus making allowance for ambiguities in the data set. This is of particular importance in fermentation data where both noise in the measurements and minor batch to batch variations would complicate a strict quantitative comparison of the data from different batches.

Numerical data provide the first and most complete description of the process: the quantitative description of an electrocardiogram (ECG) specified the length and angle of each linear data piece (Udupa and Murthy 1980, Pietka 1991) whilst the description of the linear portions of the aluminium rolling mill data consisted of the initial and final x and y coordinates of each segment (Love and Simaan 1988). These quantitative descriptions of the data were not used in subsequent analyses but were abstracted to qualitative descriptors.

Despite the approximate nature of the final description of the data, rules must be stipulated for the abstraction of the numerical quantities to the qualitative values, or labels. These rules define which aspects of the data are to be described qualitatively, the number of qualitative divisions required, and the actual mapping of the quantitative values to the qualitative regions.

The aspects of the data that are to be described qualitatively depend on the information required for the subsequent analyses and on the form of the data. For a curved data piece the qualitative description may include an indication of the length of the curve, eg 'long' or 'short', the rate of the curve, eg 'gradual' or 'rapid', and the shape of the curve, eg 'concave', 'convex' or 'linear' (McIlraith 1989). A number of representations have been used for linear data pieces. Love and Simaan (1988) described each linear segment of the aluminium rolling mill data as being a 'flat' or a 'ramp' and retained the quantitative information to provide a more complete picture. The flats and ramps description was useful in determining periods of constant and non-constant behaviour in the performance of the rolling mill. In the analysis of ECGs, Pietka (1991) divided the range of possible angles a line could make with the horizontal into nine qualitative regions, for example the qualitative region labelled 'h' described all lines whose angles with the horizontal were between 0° and 16°. Pietka's segmentation of the quantitative angle space was more sensitive than the seven qualitative regions used by Udupa and Murthy (1980) for the same type of data. Both Udupa and Murthy (1980) and Pietka (1991) retained a quantitative measure of the length of each line as this was considered important in the analysis of the ECGs. An inherent problem with using the angle a line makes with the horizontal is that the value of this angle is dependent on the scale of the two axes of the profile. ECGs are presumably drawn to a standardised scale, however, this is not the case for most engineering applications, including fermentation data. An alternative description of linear data pieces would include measures of the change in both the x-dimension and the y-dimension of the line. This provides the same amount of information as the methods of Udupa and Murthy (1980) and Pietka (1991) but may be more meaningful in the analysis of some processes. In the analysis of a fermentation time profile both duration and magnitude are important thus the latter description of the data is more applicable.

The number of qualitative regions and the boundaries of the regions are purely subjective choices and reflect the developer's interpretation of the system being studied. The adequacy of the choices can only be judged by applying the segmentation rules to the data and visually assessing the results. If the technique being developed is to be automatic then the segmentation rules must be applicable in any situation that may arise for the system in question.

The boundaries of the qualitative regions define the mapping of the quantitative values to the appropriate region. The definition of the boundaries can be crucial to the resulting representation of the data. Returning to Pietka's qualitative description of ECGs (Pietka 1991), the divisions between qualitative regions were fixed. This was possible because the limits of the angle space are well defined, that is the angles a line makes with the horizontal must fall within the range -90° to 90°. There are no such limits to the possible values of the change in magnitude or temporal extent of a linear segment in a fermentation time profile. Therefore the mapping of quantitative values to qualitative regions must be a relative function, dependent on the range of values present in the data being investigated.

A further consideration in the specification of segmentation rules is the sharpness of the boundaries between qualitative levels. Pietka's qualitative description of the angles of the lines present in an ECG (Pietka 1991), which was outlined above, implies that a line that makes an angle of 5° with the horizontal would be indistinguishable from one that makes an angle of 15° with the horizontal. However, a line whose angle with the horizontal was 15°

would be considered different from one whose angle with the horizontal was 17°. Such sharp distinctions between qualitative levels can lead to anomalies in the comparison of data pieces, it would be more desirable to allow some amount of fuzziness at the boundaries between qualitative values. This can be achieved by specifying a fuzzy area about each boundary, if a quantitative measure falls within this fuzzy area then it can belong to either of the two regions separated by the boundary. The size of the fuzzy area is again a matter of subjective choice for the developer of the system. An alternative would be to use a technique from the field of artificial intelligence known as 'fuzzy logic' (Zadeh 1965, Kaufmann and Gupta 1991).

Konstantinov and Yoshida (1989, 1990a, 1990b), Fu *et al.* (1989) and Postlethwaite (1989) have utilised fuzzy logic in the analysis of fermentation data. In these examples the qualitative regions were described as overlapping fuzzy sets and membership functions assigned to each variable according to the value of the variable, this membership function indicates the degree to which a variable is a member of a fuzzy set. Thus, at a particular value, a variable may belong to the 'high' set with a membership function of 0.7 and the 'medium' set with a membership function of 0.2. The technique of fuzzy logic exploits an expert's qualitative view of time variant data and eliminates the problem of sharp boundaries between qualitative regions. However, a number of drawbacks in the techniques precluded their use in the work reported here.

- The main drawback is the extensive amount of work required in the specification of the membership functions. In developmental work it is necessary to use all available variables in the relative assessment of performance of the experimental runs and membership functions would be required for each of these variables; the choice of these membership functions is entirely subjective and has considerable influence on the resulting data representation thus the developer must spend a lot of time achieving a suitable balance between the different membership functions.

- As mentioned previously, the analysis of time variant data during developmental work cannot proceed using a point by point comparison of the data because of the time shifts effected by the changes in operating conditions, thus it is not appropriate to describe the data points individually as done by Konstantinov and Yoshida (1989, 1990a, 1990b), Fu *et al.* (1989) and Postlethwaite (1989).

- If the data are to be represented by linear data segments, as indicated previously, both the temporal extent and the change in magnitude must be described and membership functions would be required for both these dimensions thus further increasing the work required in the development of the system.

- Furthermore, with the progress of the fermentations changing from batch to batch, the temporal extents and changes in magnitude for the various linear data pieces may vary considerably thus complicating the task of specifying membership functions that would be applicable in all situations.

A qualitative data structure was defined to describe linear data segments in fermentation data. The requirements of this data structure were:

1. both the duration and the change in magnitude of each linear segment must be described;
2. the number of qualitative regions must reflect an expert's interpretation of the data;
3. the specification of the boundaries, and therefore the mapping of the quantitative data to the qualitative regions, must cope with considerable variations in the changes in magnitude and temporal extent of the linear segments from batch to batch;
4. the boundaries between qualitative divisions must not be too severe.

The qualitative description of fermentation data proceeds by, firstly, defining the qualitative intervals and fuzzy boundaries for the two data sets being compared and, secondly, using these definitions to provide a qualitative representation of the data sets. The rules for defining the intervals and fuzzy boundaries are generic to all fermentation data sets but the specific values are dependent on the two data sets under consideration. The procedure is described in Section 3.3.

3.1.3 Comparison of the Time Variant Data

The previous section discussed the abstraction of the linearised time variant data to a qualitative representation. The next step is to undertake a meaningful comparison between the qualitative representation of one data set and that of another data set. In this work it was found that the most effective comparison method was simply to match the qualitative labels of each linear data piece with corresponding data pieces in another time profile (Section 3.4). The identification of congruous data pieces in different profiles was not straightforward and required a rather tortuous trial and error procedure.

Techniques that have been used for the comparative analysis of numeric data in the past, such as syntactic pattern recognition (Udupa and Murthy 1980, McIlraith 1989, Pietka 1991), rule based systems (Love and Simaan 1988, Chen *et al.* 1989, Halme 1989, Karim and Halme 1989, Stephanopoulos and Tsiveriotis 1989, Clapp and Ruel 1991, Cooney *et al.* 1991, Morris *et al.* 1991) and neural networks (Cooney *et al.* 1991, Morris *et al.* 1991) were not used for reasons described below.

Tou and Gonzalez (1974) defined syntactic pattern recognition as 'the characterisation of patterns by primitive elements and their relationships'. The relationships between patterns describe the various sequences of patterns, or grammars, that represent valid structures in the system of interest. For example, in the analysis of ECGs different sequences of straight lines, represented by lengths and angles, are indicative of different cardiac patterns (Pietka 1991). A number of training instances are required to define the grammar of the system, ie the set of all patterns that are recognised in the system. The comparison involves 'parsing' an unknown pattern, ie determining whether or not the new pattern represents a grammatically correct structure and, if so, labelling it according to the pattern class to which it belongs. Thus a successful match between an unknown pattern and a categorised pattern in ECG analysis would allow classification of the test signal and therefore a description of the cardiac cycle being studied.

Syntactic pattern recognition is successful in fields in which a strict grammar can be defined and a limited number of patterns will describe all possible scenarios. This is not the case during the developmental stages of a fermentation where changes in process conditions and unpredictable biological variations result in poorly characterised system performance. In an established process, a syntactic grammar could be devised for the ideal trajectories which would then allow the detection of out of control processes. However, the development of grammars for faulty operation, which would be needed for fault analysis, would require experience of all possible faults. Syntactic pattern recognition would not be suitable in the analysis of faults that had not been previously experienced. The process development work could be used to yield a number of 'fault grammars' and new faults encountered during production could be learnt by the pattern recognition system. However, each fault and each combination of faults would generally result in a different fermentation pattern making the set of possible patterns extremely large. It is unlikely that a comprehensive pattern recognition system could be developed to cover all possible scenarios required for the process of fault analysis. Syntactic pattern recognition in fermentation work is thus limited to fault detection in established processes and fault analysis only in situations that have been experienced previously.

Graphical reasoning may also be carried out using a rule based system in which previous experience is used to relate graphical observations or sequences of features to specific occurrences. In fermentation work a number of expert systems have been developed in which the time variant data at any point in the fermentation are compared with rule based models of the system and the current physiological state is identified (Chen *et al.* 1989, Halme 1989, Karim and Halme 1989, Stephanopoulos and Tsiveriotis 1989, Clapp and Ruel 1991, Cooney *et al.* 1991, Morris *et al.* 1991). The major disadvantage with the expert system approach is the extensive amount of work required to set up the system: a large number of test cases must be run and analysed by an expert before formulating the rules. An expert system can therefore only be used once the developmental work on a fermentation has been completed. Expert systems are expected to be as good as, or even better than, a human expert in assessing the current situation and suggesting appropriate actions. However, the performance of an expert system is totally unpredictable in situations for which rules are unavailable which can lead to unacceptable results. A human expert relies on intuition and years of experience which can never be encoded in series of rules; expert systems, as they are today, do not have the capacity to build up this type of knowledge. Furthermore, many of the rules in an expert system, especially those relating to metabolic events, are process specific thus the rule generation procedure must be repeated for every process to be analysed.

Neural network systems have been developed by Cooney *et al.* (1991) and Morris *et al.* (1991) as a means of recognising patterns in fermentation data and thus identifying the current state, detecting and analysing faults, and estimating variables. Neural networks are computing systems composed of a number of interconnected layers of simple neuron-like processing elements which process information by their dynamic response to external inputs (Rumelhart and McClelland 1986). They were designed to computationally emulate the cerebral neural structure and its behaviour. A neural net is trained by feeding it a large number of data sets, such as the raw sensor data and derived variables, along with the corresponding outputs, such as the state, estimated variables and information about faults. The network learns to recognise patterns in the data and, essentially creates an internal model relating the inputs to the outputs. After training is complete the network will determine the pattern of any input data and apply the appropriate 'model' to obtain the associated output. Neural networks are somewhat different from the other techniques presented as the raw data are treated directly rather than being simplified and described qualitatively. The concept of comparing the data with past cases is, however, similar. Like the other techniques presented, neural networks cannot be employed in the analysis of data from experimental fermentations because of the batch to batch variability and thus the lack of suitable training instances. Additionally, neural networks are 'black box' processes in that they are unable to provide the user with the line of reasoning followed in relating a set of inputs to the outputs. There is always some reluctance to employ analytical tools whose mechanism is not well understood, thus the introduction of neural networks to industrial processing staff would require a considerable amount of educational effort. Furthermore, a different neural network must be developed for each different process, which, in a large company, would result in a considerable amount of developmental effort and computing space.

The matching of qualitative labels between profiles has a number of advantages over the techniques described above:

1. the matching procedure can be used for both experimental and industrial data as no prior knowledge of the process is required;
2. no process specific information is required thus the matching procedure can be used on any fermentation process;
3. the line of reasoning in the matching procedure is easy to follow as it is based on a simple comparison of qualitative labels that have a clear interpretation.

The remainder of this chapter describes the techniques developed for the simplification, qualitative description and comparison of fermentation data. In Section 3.5 the interpretation of the results of the comparative process is described and the relative merits of the tools are discussed in Section 3.6. All the examples used in the following sections are from the fermentations described in Chapter 4, unless otherwise stated. The details of the fermentations are not required to understand the examples in this chapter as they are merely used to illustrate the techniques developed.

3.2 Data Simplification

3.2.1 Overview

The first step in the comparison of time invariant data is the simplification of the data into piecewise linear data segments. The objectives of the linearisation procedure were:

1. to remove noise and group the data to facilitate comparisons;
2. to summarise the data in a form that resembles an expert's view of the data.

It was important that the procedure did not require any *a priori* knowledge of the process and could be used to simplify both experimental and industrial fermentation data.

The simplification procedure can be broken down into three functional parts:

1. pretreatment and input of raw data;
2. simplification of data to give a piecewise linear approximation;
3. output of results.

The methods used to achieve each of these functions, and the theory behind some of the techniques used, are presented in the following sections.

The data simplification routines (DSIMP) were written in RM/FORTRAN v2.4 to run on a personal computer (Tandon PCA 40 AT) and later transferred (without alteration) to a Sun Sparc Station 1. Plotting was carried out using Simpleplot v2-10 (054) from Bradford University Software Services on the PC and Pro-Matlab from the Math Works Inc. on the Sun.

3.2.2 Raw Data

In some situations it may be necessary to 'clean up' the data prior to input to the simplification routines: any data logged prior to inoculation or subsequent to harvesting must be removed from the data files. These extraneous data should be eliminated at the source by ensuring the logging system is turned on and off at the appropriate times but in the less sophisticated logging systems this is often overlooked.

Where repeated measurements are available it is necessary to calculate mean values and measures of the spread of the data, such as the range and standard deviation. These calculations are readily carried out in a spreadsheet environment. The SmartWare II spreadsheet (Informix Software Inc., Menlo Park, CA) was used in this work, and the calculations are discussed in Section 3.2.3.2.1. A statistical data base (Ghosh 1989) may be a more appropriate medium for the storage and manipulation of the off-line data as it would eliminate the double handling of the data.

The raw on-line data are stored in standard ASCII files in two column format: (time,value). Off-line data, which are usually obtained by averaging a number of repeated measurements, are stored in ASCII files as (time,mean,minimum,maximum) as detailed in Section 3.2.3.2.1. The name of each data file identifies the batch number of the fermentation and the variable monitored for ease of identification, eg C447CER.DAT contains CER data from the fermentation with batch number C447. The input routine requires the user to identify the file containing the data to be simplified and places these data into two one dimensional arrays (TIME and VALUE).

3.2.3 Linearisation of the Data

3.2.3.1 The Best Fitting Line

Linear data pieces have been identified as the most appropriate data structure for the simplification of fermentation profiles (Section 3.1.1).

The choice of method for the linearisation of the data was dictated by the form of the data. Fermentation data are typically noisy containing both random fluctuations and occasional outliers. It is possible that the random element is normally distributed (ie a Gaussian model) in which case a piecewise linear least squares fit to the data would result in a reasonable approximation. However, in the presence of outliers a least squares fit may not be appropriate. In a Gaussian model the probability of occurrence of outliers is very small, thus a least squares fit of the data would be considerably distorted in order to bring any outliers into line. Robust statistics are used to deal with cases where the Gaussian model is a bad approximation or where outliers are evident in the data set (Press *et al.* 1986). A robust statistical estimator is insensitive to 'small departures' from the idealised assumptions for which the estimator is optimised. 'Small departures' may be either fractionally small departures from all data points or fractionally large departures for a small number of data points eg outliers. M-estimates are robust estimators that follow from maximum likelihood arguments, minimising some measure of the absolute deviations such as the mean or sum. The use of M-estimates means that no assumptions need to be made about the noise distribution of the data set and the prior removal of outliers is not required. It is assumed in the following theory of M-estimates that all data points are equally weighted.

Let y be the array of dependent variables and **x** the array of independent variables. The aim of the estimation procedure is to fit a straight line to this data:

$$y = a + bx \qquad (3.1)$$

where b is the slope of the line and a the point at which the line crosses the y-axis. The function to be minimised in finding the best line to represent this data is:

$$\sum_{i=1}^{N} |y_i - a - bx_i| \qquad (3.2)$$

where N is the number of raw data points over which the line is to be fitted.

The basis of this minimisation is that the median of a set of numbers is also that value which minimises the sum of the absolute deviations.

The method for minimising Equation 3.2 is as follows.

For fixed b the value of a which minimises Equation 3.2 is:

$$a = \text{median}\{y_i - bx_i\} \qquad (3.3)$$

The following must hold at the minimum:

$$0 = \sum_{i=1}^{N} x_i \, \text{sign}(y_i - a - bx_i) \qquad (3.4)$$

Substituting Equation 3.3 into Equation 3.4 results in an equation in a single variable (b) which can be solved by bracketing and bisection to give the 'best fitting' line to a series of raw data points.

This routine, coded in MEDFIT, ROFUNC and SORT in Press *et al.* (1986, pp 544-545) was used to fit straight lines to fermentation time profiles. (For the purposes of this discussion this routine is called PRESS). The initial guess to the best fitting line is the least squares solution, this is refined using bracketing and bisection until the line of least absolute deviation is found. It was observed that the PRESS routine occasionally did not find the line of minimum absolute deviation but that the bracketing and bisection methods used diverged from the optimal result. In order to rectify this the bracketing and bisection was carried out as described by Press *et al.* (1986) but at each step the minimum sum of absolute deviations of all steps thus far was recorded; if, at the termination of the PRESS routine, the sum of absolute deviations about the line was not smaller than the recorded minimum, the line that had resulted in the minimum was substituted as the best fitting line. The different fits obtained by the PRESS routine and the improved routines are shown in Figure 3.2.

3.2.3.2 The Goodness of Fit Criterion

One of the major problems encountered when fitting any form of line or curve to raw data is that of specifying how closely the line must fit the data: the fit must be close enough to permit the line to follow the trend of the data but not so close that the line is distorted by noise in the measurements. A suitable line is one in which some measure of the deviations about the line is less than a pre-specified goodness of fit criterion. Most curve fitting and linearisation techniques require manual specification of the goodness of fit criterion, there is no fundamental basis for the choices made other than the results are deemed adequate by the user. This is not suitable for a fully automated system which is intended to remove all sources of human bias and inconsistencies.

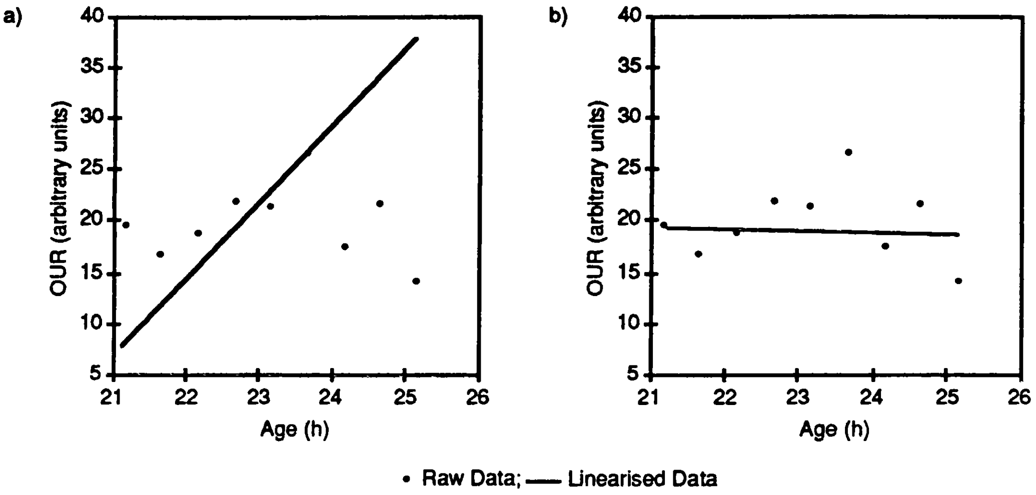


Figure 3.2: The robust estimation procedure of Press *et al.* (1986), used to fit the straight line in (a), does not always select the best fitting line of all the lines encountered in the iterative fitting procedure (as shown in this example). The fitting procedures were altered so that the best fitting line is always selected. The resulting line, for the same data set, is shown in (b). The best fitting line is that which has the smallest mean absolute deviation between the line and the raw data. The mean absolute deviation between the line in (a) and the raw data is 73.49 units while the mean absolute deviation between the line in (b) and the raw data is 24.71 units. (The oxygen uptake rate data are from batch C440 of the fermentations described in Chapter 4).

It is apparent that the goodness of fit criterion must in some way reflect the uncertainty or error in the measurement in question so that the resulting line is representative of the most plausible underlying trend in the data. The goodness of fit criteria used in the linearisation procedure are described below and summarised in Table 3.1.

DATA TYPE	GOODNESS OF FIT CRITERIA
Repeated measurements - retrospective analysis - a number of batches	The line must pass between the maximum and minimum values at each sample point. If the interval between the mean value and either the minimum or maximum value at any point is less than three average standard deviations, the line must pass within three average standard deviations of the mean value.
Repeated measurements - correlation	The line must pass between the maximum and minimum values at each sample point. If the interval between the mean value and either the minimum or maximum value at any point is less than the accuracy of the correlation, the fitting criterion at that point is overridden by the uncertainty in the correlation.
Repeated measurements - established process	The line must pass within three average standard deviations of the mean value at each sample point.
Single measurement from single instrument	The goodness of fit is specified by the instrument precision or a knowledge of the instrument's behaviour. The value, or proportionality constant, stays the same for all uses of a particular instrument.
Single measurement from multiple instruments	The goodness of fit is specified by the precision of the individual instruments or by experimental investigation. The value, or proportionality constant, stays the same for all uses of a particular combination of instruments.

Table 3.1: Summary of the fitting criteria for the piecewise linearisation of the different types of time variant fermentation data.

3.2.3.2.1 Repeated Measurements at Each Sample Point

Where repeated measurements are taken to obtain a single value, such as is usually the case for off-line data, the distribution of the repeats can be used to determine the goodness of fit criterion. The standard deviation is a measure of the spread of the readings for any one sample, however, unless a normal distribution can be assumed, interpretation of the standard deviation is difficult. It is rare to take more than four measurements at each sample point in fermentation work. With so few values there is no validity in assuming the errors in the readings would follow a normal distribution. The readings cannot be considered to be independent within each sample because the dilutions are usually made from a single sample tube (Esener *et al.* 1981), thus further invalidating the use of the standard deviation as a meaningful measure of the spread of the readings. For these reasons it was considered to be more valid to use the range of values as a measure of the accuracy of the readings, assuming that the true value would most likely fall somewhere between the minimum and maximum recorded values.

The analysis of experimental data usually occurs retrospectively with a number of batches being considered together. A large number of off-line data measurements are therefore available. It is thus possible to obtain an overall picture of the accuracy of the measurements of each variable by considering the average standard deviation of all samples from all available fermentations. From the Central Limit Theorem it is then reasonable to assume that the errors can be approximated by a normal distribution and, from the Empirical Rule, it can be expected that 99.73% of all readings would fall within three standard deviations of the mean value at each sample point.

For off-line data, therefore, the first fitting requirement is that the lines must pass within the range covered by the measured values at each point. Overriding this condition is a minimum goodness of fit criterion of three times the average standard deviation. Thus, if the interval between the mean value and either the minimum or maximum value at any one sample point is less than three times the average standard deviation the goodness of fit criterion at that point becomes three times the average standard deviation. This minimum fitting criterion was introduced to avoid the stringent fitting constraints arising when the range of repeats from a sample is very small. Although, taken on its own, this would seem to indicate an accurate measurement, the statistical evidence provided by the other samples refutes this possibility.

For some off-line measurements correlations are made between the physical measurement and a meaningful quantity, for example the optical density of a fermentation broth is usually converted to a measure of biomass concentration by way of an experimentally determined linear correlation. In this situation the measurement of biomass concentration cannot be considered to be any more accurate than the accuracy of the correlation. The minimum goodness of fit criterion is therefore stipulated from the accuracy of the regression equation parameters. An example of this is given in Chapter 4.

In the work reported in Chapter 4 these goodness of fit calculations were performed in the SmartWare II spreadsheet (Informix Software Inc., Menlo Park, CA) prior to the data being read in to the simplification routines. The outputs from the spreadsheet calculations were the average standard deviation over all the fermentations being considered and ASCII files containing the time, mean value, minimum value and maximum value at each sample point for each of the fermentations. In the simplification routine the minimum and maximum values were adjusted to reflect the goodness of fit requirement at each particular data point, that is when the standard deviation criterion or the correlation accuracy criterion overrode the range criterion the values were adjusted accordingly.

In an established process the average standard deviation of a particular type of off-line measurement is unlikely to change significantly over the course of time, unless measurement techniques or equipment are changed. In this situation, where there are a large number of examples from which to determine the appropriate statistics, it is possible to use three times the average standard deviation as the sole goodness of fit criterion.

3.2.3.2.2 Single Measurements at Each Sample Point

For on-line measurements that utilise a single instrument reading, such as dissolved oxygen and pH, the precision of the instrument can be used to specify the goodness of fit criterion. Approximate values for instrument precision are usually stipulated in the technical data sheet provided by the supplier. It is expected that the goodness of fit criteria selected for on-line measurements will be generic to all processes using the same equipment. Where technical data sheets are not available, the goodness of fit criteria can be chosen based on experience with the measurements being considered. Again, the criteria chosen in this way must be utilised for all fermentations so as to maintain consistency in the linearisation of the data. In Chapter 4 examples of appropriate goodness of fit values are given for on-line data from a set of experimental fermentations.

Some fermentation variables utilise measurements from a number of instruments. Carbon dioxide evolution rate (CER in (L of CO_2).(L of broth)$^{-1}$.h^{-1}) and oxygen uptake rate (OUR in (L of O_2).(L of broth)$^{-1}$.h^{-1}) are calculated from mass spectrometer, mass flow meter and volume readings as follows:

$$CER = (F_{in}/V)((CO_2)_{out}((N_2)_{in}/(N_2)_{out}) - (CO_2)_{in})$$

$$OUR = (F_{in}/V)((O_2)_{in} - (O_2)_{out}((N_2)_{in}/(N_2)_{out}))$$

(Omstead *et al.* 1990), where F_{in} is the inlet air flow rate (L of air.h^{-1}); V is the volume of broth in the fermenter (L); $(CO_2)_{in}$, $(O_2)_{in}$ and $(N_2)_{in}$ are the inlet volume fractions of carbon dioxide, oxygen and nitrogen respectively (L of gas.(L of air)$^{-1}$); $(CO_2)_{out}$, $(O_2)_{out}$ and $(N_2)_{out}$ are the outlet volume fractions of carbon dioxide, oxygen and nitrogen respectively (L of gas.(L of air)$^{-1}$). The values of CER and OUR are usually converted to mmol of gas per litre of broth per hour (mmol.L^{-1}.h^{-1}).

An estimation of the error in these CER and OUR measurements was given as four percent of the signal value (K.M. Stone, UCL, personal communication). This value was based on experimental work using a 7 L LH fermenter, a VG MM-80 magnetic sector mass spectrometer (VG Ltd, UK), a HI-TECH F100 thermal mass flow meter and controller (Bronkorst High Tech B.V., Netherlands) and manual volume measurements.

When the signal size was small it was found that this value underestimated the errors; it was also found that the implied proportionality of the error was only very approximate. To compensate for this the error in the CER and OUR measurements was calculated as follows:

1. the range of values is calculated for each batch:
   range = maximum value - minimum value;

2. if a series of batches is being analysed the average of the ranges is calculated;

3. the goodness of fit criterion (gof) is then calculated:
   gof = 0.04 * average range.

During the development of a new fermentation process it is feasible that the CER and OUR maximum values will change as the process is improved. The average range of CER and OUR values is updated each time a new batch or series of batches is available for analysis. However, the goodness of fit criterion used in previous batches is not altered - once a data set has been linearised it is stored as such and is not updated.

For an established process the average range of CER and OUR values is similarly updated as each batch is completed. If a run is to be analysed on-line the goodness of fit criterion calculated from all previous batches is used, the goodness of fit cannot be updated until the completion of a batch.

Despite the fact that the goodness of fit criteria for the CER and OUR were based on a particular experimental system the same values were found to give acceptable results for data obtained using completely different equipment (Chapter 4).

Other fermentation variables, such as the respiratory quotient and oxygen transfer coefficient, were not considered in this work. Determination of an appropriate goodness of fit for these variables would be required prior to their linearisation. The main requirement in determining a goodness of fit criterion is that the same criterion can be used for all processes using the same instruments. If different instruments are used for different processes it may be necessary to adjust the goodness of fit criterion.

3.2.3.2.3 Inputs Relating to Goodness of Fit

The linearisation routines are initiated by determining the type of data to be analysed, ie repeated readings at each point or individual readings. The appropriate data input routine is then employed. The user is asked to supply the goodness of fit criterion for the data: for off-line data this is the average standard deviation, for on-line data it is the precision of the instrument or an estimation thereof. At present this input process is manual and the calculations are carried out separately from the linearisation routines. The use of a statistical data base may facilitate automation of this procedure. Alternatively the goodness of fit information could be incorporated into the standard relational data base described in Chapter 2 using a separate 'Goodness of Fit' table with fields for the variable, the instrument and the goodness of fit.

3.2.3.3 Piecewise Linearisation Algorithm

The piecewise nature of the desired representation requires that there be some means of specifying the end points of each linear data segment. The constraints on each line are that they must intersect with the previous line and must satisfy the goodness of fit criterion. A recursive scheme was developed to find the best sequence of contiguous line segments as described below.

A line is initially fitted over one hundred raw data points using the procedure described in Section 3.2.3.1. The start point of the line is the first data point for the first line, and all subsequent lines start at the last point covered by the previously fitted line. The lines are not constrained to pass through these data points as the algorithm finds the best fit line through all the data points being covered. If there are less than one hundred points remaining in the profile, the line is fitted over all remaining data points. The choice of one hundred points was purely arbitrary. A number of initial step sizes were investigated but there was very little effect on the outcome of the simplified data.

The goodness of fit of the line is assessed by comparing the mean absolute deviation between the line and the raw data with the goodness of fit criterion (Section 3.2.3.2). If the mean absolute deviation is less than, or equal to, the goodness of fit criterion then the line is considered to be a good fit. An attempt is then made to extend the line to cover a further one hundred data points, the best fit line is obtained and the fit assessed. This is repeated until the whole data set is covered or a bad fit is encountered.

If, at any stage, the mean absolute deviation between the line and the raw data is greater than the goodness of fit criterion, the line is deemed to be a poor fit. Rather than reducing the number of points covered by the line, the line is extended over one more point (if not at the end of the data set), the best fit line is obtained and the fit reassessed. This is to ensure that the bad fit was not a result of a single outlying point. At this stage, if the fit is good, the line is extended over a further one hundred points and the process is repeated. However, if the fit is still poor or the line reached the end of the data set, it is necessary to fit the line over a reduced number of data points. The line is fitted over fifty less data points (or half the initial step size if this were less than one hundred). If the line is still not acceptable it is refitted to cover one more data point to ensure a 'bad' point had not been encountered. If the fit is still not good the line is then fitted over ten less data points and its acceptability reassessed. In the event that the fit is still poor the length of the line is increased by one again and the fit reassessed. This contraction by ten and expansion by one is repeated until within ten data points of the last acceptable fit, or within ten points of the previous line, and the line is then reduced one point at a time until a good fit is found and this line is then accepted. If, prior to this last set of reducing the line by single points, a good fit is encountered, the line is then increased by ten data points repeatedly until the fit is bad again and then reduced one data point at a time until the fit is good. The line is then accepted. The end point, in the time dimension, of this line becomes the starting point of the next line. The process is repeated until the whole data set is covered.

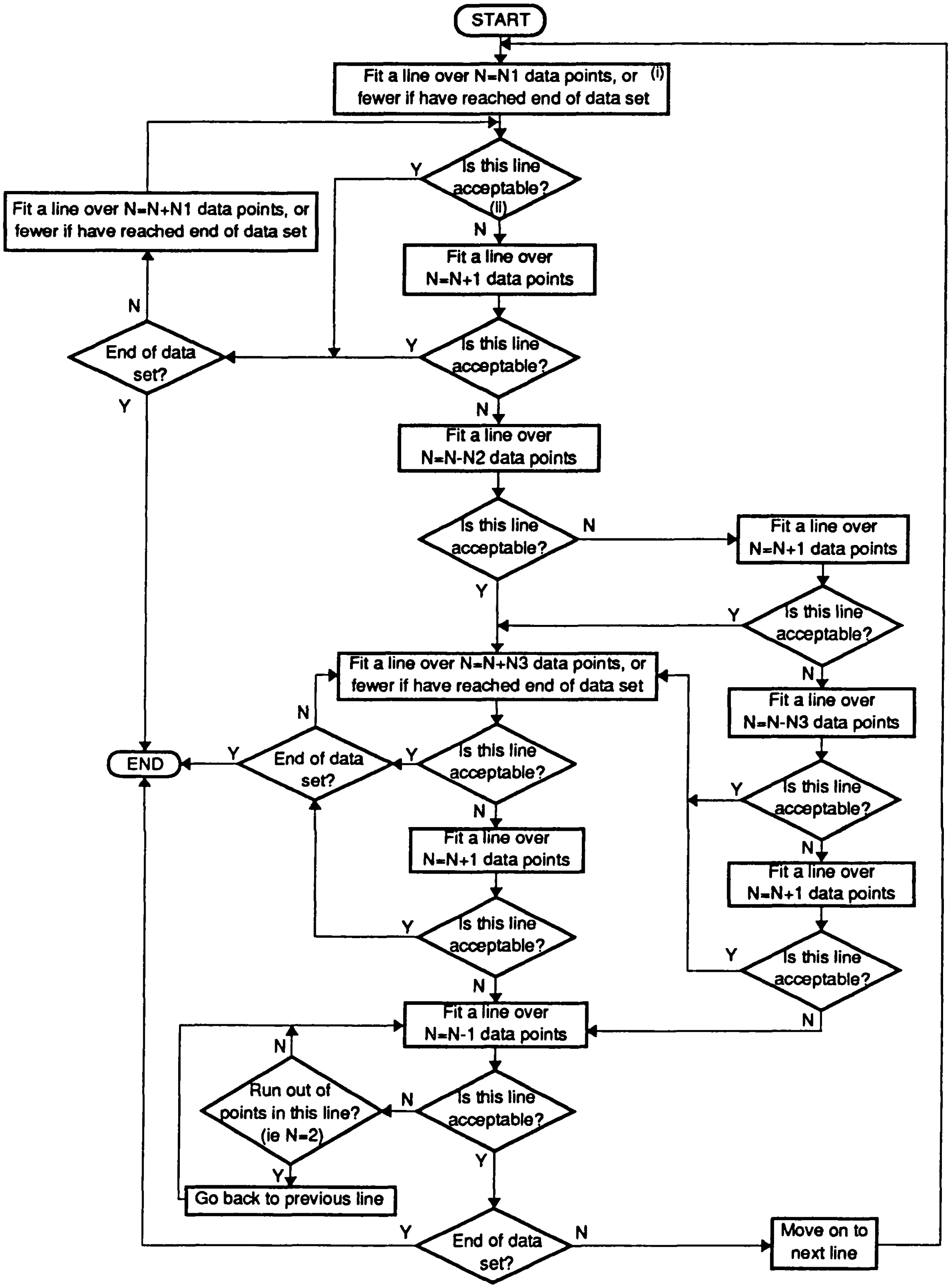
If the line being fitted is not the first line it must intersect with the previous line. Each time the length of a line is extended or reduced and the best fit line is obtained, the new line and the previous line are extrapolated or interpolated so that they intersect. If the goodness of fit of either line is poor then the new line must be extended or reduced (depending on the current position in the algorithm) as described above. In some situations an acceptable fit cannot be found for a sequence of data points and so it is necessary to return to the previous line and refit it over one less point (or more if this subsequently results in a bad fit).

The algorithm for extending and reducing the line segments is summarised in Figure 3.3.

An example of the piecewise linearisation of a carbon dioxide evolution rate profile from a recombinant, protein producing fermentation is shown in Figure 3.4(a). The piecewise linearisation of the oxygen uptake rate profile from a more complex mycelial secondary metabolite fermentation is shown in Figure 3.4(b). These two fermentations were carried out at Merck Sharp and Dohme Research Laboratories (Rahway, NJ, USA) using completely separate facilities, including fermenters and measuring instruments. However, the goodness of fit criterion, calculated as described in Section 3.2.3.2, was applicable in both cases. The linearisation of an off-line data profile is shown in Figure 3.4(c).

The examples in Figure 3.4 demonstrate the ability of the linearisation process to remove noise from the data. The simplified profiles in Figure 3.4 are also a good summary of how an expert views the data: variations in the metabolic activity in each of the data sets are clearly visible. A further important feature of the profiles in Figure 3.4 is that the simplification procedure adequately linearised two data sets with completely different process dynamics and was able to cope with variations in the range of magnitudes covered by the different data sets; no alterations in the fitting procedures or the goodness of fit criteria were required.

The piecewise linearisation can be carried out in real time with no alterations to the program. On the Sun Sparc Station the linearisation is almost instantaneous for a data set containing about seventy five data points, whilst for a data set containing 850 data points the linearisation occurs in approximately ten seconds. The linearisation may take up to two minutes on a PC AT which is not acceptable for real time analyses, but the faster Intel 80386 and 80486 based personal computers should linearise large data sets in a matter of seconds.


START
Fit a line over N=N1 data points, or fewer if have reached end of data set (i)
Is this line acceptable? (ii)
Y
N
Fit a line over N=N+N1 data points, or fewer if have reached end of data set
Fit a line over N=N+1 data points
Is this line acceptable?
Y
N
End of data set?
N
Y
Fit a line over N=N-N2 data points
Is this line acceptable?
N
Y
Fit a line over N=N+1 data points
Is this line acceptable?
Y
N
Fit a line over N=N+N3 data points, or fewer if have reached end of data set
Fit a line over N=N-N3 data points
END
End of data set?
Y
N
Is this line acceptable?
Y
N
Is this line acceptable?
Y
N
Fit a line over N=N+1 data points
Fit a line over N=N+1 data points
Is this line acceptable?
Y
N
Is this line acceptable?
Y
N
Fit a line over N=N-1 data points
N
Run out of points in this line? (ie N=2)
Y
Is this line acceptable?
N
Y
Go back to previous line
End of data set?
Y
N
Move on to next line

Notes: i) N1 = initial step size (100); N2 = N1/2 (minimum = 15); N3 = N1/10 (minimum = 10);
ii) Each time a line is fitted to N data points the line is extended or contracted to the point of intersection with the previous line; if the point of intersection lies outside the time range covered by the two adjacent lines or the absolute deviation over either line is too large then the line is not acceptable.

Figure 3.3: Algorithm for stepping through a data set to find the best sequence of piecewise contiguous straight lines

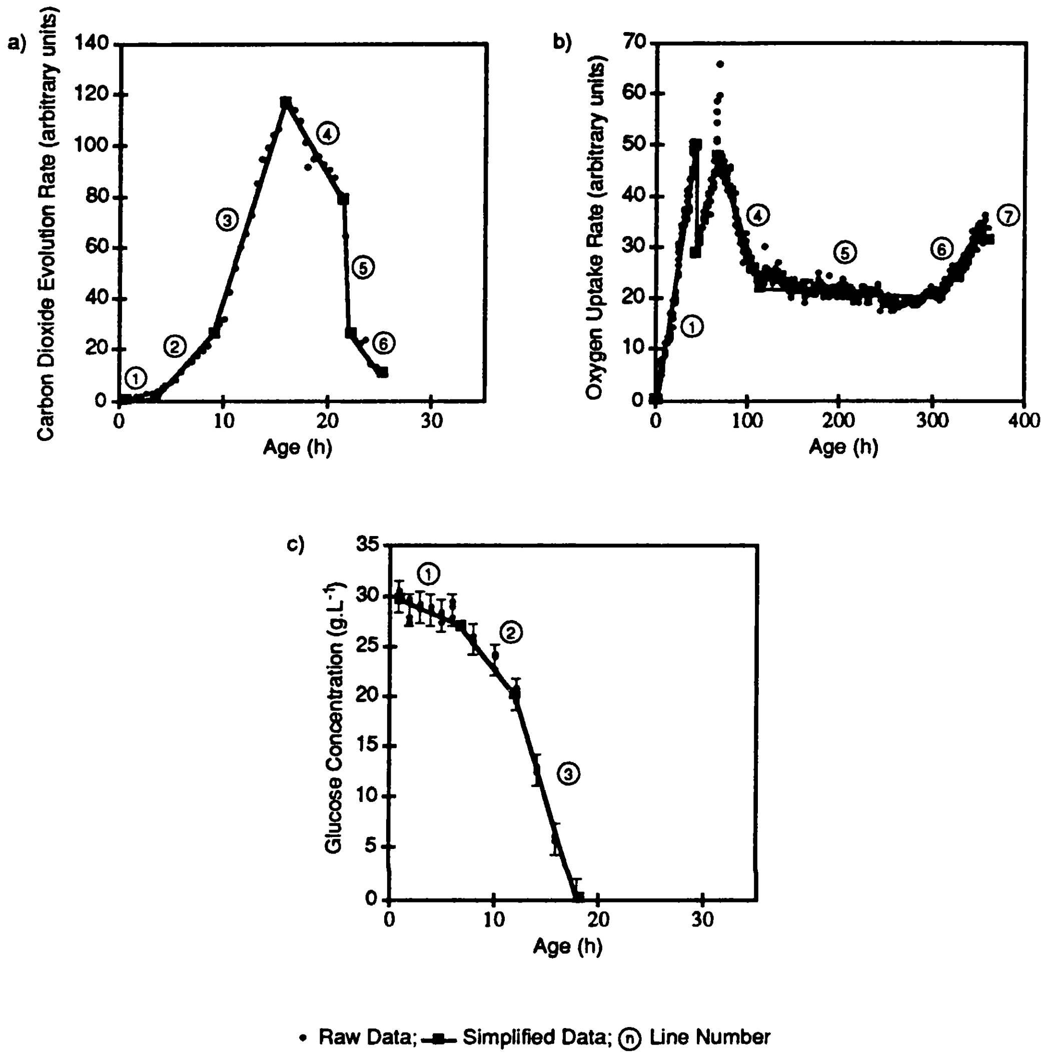


Figure 3.4: Examples of the piecewise linearisation of fermentation time variant data. a) on-line data from an *Escherichia coli* fermentation. b) on-line data from a mycelial secondary metabolite fermentation (courtesy of Merck Sharp and Dohme Research Laboratories, Rahway, NJ, USA). c) off-line data from an *Escherichia coli* fermentation, shown with error bars of three standard deviations.

3.2.3.4 Linearisation of Control Data

For set point controlled data it would be assumed that a single horizontal line through the data at the set point would adequately describe the data unless the control had been faulty or the set point had been changed intentionally. For the work presented in Chapter 4 the only information required from variables that were controlled at a constant level was whether or not the set point was adequately maintained. This information was obtained by the operator, from the control system, and was stored in the data base.

Some variables may be controlled at a constant level for only a portion of the fermentation, for example it is common to maintain the dissolved oxygen concentration above a minimum level, thus the dissolved oxygen concentration is only controlled during periods of high aerobic metabolic activity. In these situations the data prior and subsequent to the control periods provide important information about the process. The DSIMP routines are therefore applied to these data as described above.

3.2.4 Output

The output from the DSIMP routines is a two dimensional array containing the time and value of the endpoints of each linear data piece. These can be stored in ASCII files if required.

3.3 Qualitative Representation of the Data

3.3.1 Overview

The second step in the comparison of time invariant data produces an approximate qualitative representation of the linearised data.

A qualitative representation was chosen because it reflects an expert's view of the data and avoids the rigidity imposed by strict numerical quantities in the subsequent comparison process.

The minimum requirement of the qualitative representation was stated earlier to be the description of the magnitude and duration of each linear data segment. The rulers which map the quantitative data to the qualitative representation were required to be generic to all data sets but had to allow for the wide variations in both magnitude and duration that occur between different types of data and between different types of fermentation.

The qualitative description procedure can be broken down into three functional parts:

1. input of simplified data from DSIMP;
2. mapping of the quantitative data to qualitative intervals;
3. output of results.

The methods used to achieve each of these functions are described in the following sections.

The computer routines for the qualitative description of the time profile data (QUAL) were written in RM/FORTRAN v2.4 to run on a personal computer (Tandon PCA 40 AT) and later transferred (without alteration) to a Sun Sparc Station 1.

3.3.2 Input of Simplified Data

A data set is described qualitatively to facilitate comparison with another data set. The qualitative description of a data set is necessarily a relative phenomenon, for example the qualitative term 'short' implies that an entity is short relative to all other entities being considered. This relative nature of the qualitative description requires that the two data sets be considered together. The input to the qualitative description routines is therefore the simplified representation of the two data sets being compared, as generated by DSIMP. The data are stored in two two-dimensional arrays.

3.3.3 Mapping of Quantitative Data to Qualitative Intervals

The attributes chosen to provide the qualitative description of each linear segment were:

1. the *direction* of the line, which indicates whether the slope is positive or negative;
2. the *magnitude extent* of the line, which is defined as the change in y value from the beginning to the end of the line;
3. the *temporal extent*, or duration, of the line;
4. the *starting position* (y-value) of the line; and
5. the *slope* of the line.

The latter two descriptors are not necessary for the description of the important features of a fermentation profile (Section 3.1.2) but were included to facilitate the comparison procedure as described in Section 3.4.

Qualitative rulers were defined to dictate the mapping of quantitative values to qualitative intervals for each of the five attributes of a linear data piece.

The definition of the magnitude extent, temporal extent, starting position and slope rulers required reference values which were solely dependent on the range of values of each attribute in the two data sets being considered.

3.3.3.1 Magnitude Ruler

The magnitude ruler defines the relationship between the quantitative and the qualitative measures of the magnitude extent of a line.

The reference value for the magnitude extent ruler is the largest possible change in y-value for any one line, ie the difference between the largest and smallest signal values in the data set. This value is calculated for both profiles being considered and the average of the two becomes the reference value. The use of the average value is important when comparing two profiles with vastly different maximum values as it enables sufficient resolution between lines with small changes in magnitude.

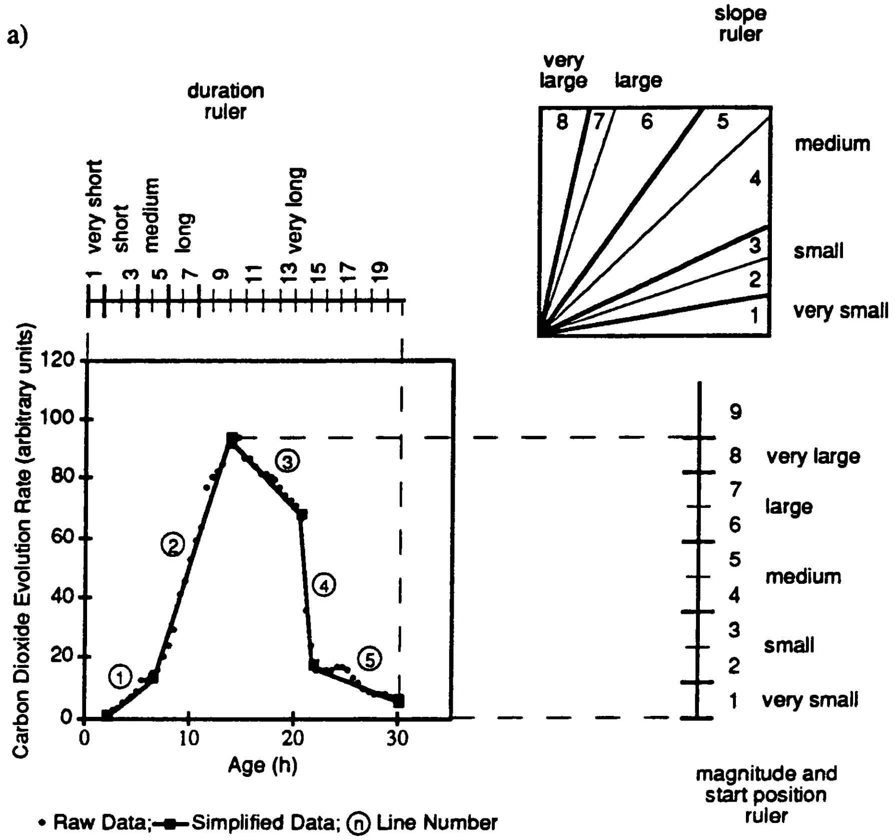
Five qualitative intervals for magnitude extent are calculated based on the reference value: the size of the basic interval of the ruler is one eighth of the reference value; the 'very small' and 'very large' intervals cover an area equivalent to the basic interval whilst the 'small', 'medium' and 'large' intervals cover an area twice the size of the basic interval, ie one quarter of the reference value (Figure 3.5(a)). The nine numerical labels depicted in Figure 3.5(a) enable interpretation by computer algorithms, each covers an area equivalent to the basic interval and their use is described in Section 3.3.3.6.

The magnitude ruler is a sliding ruler: the bottom of the ruler is placed at the beginning of the line being measured and the qualitative magnitude extent of the line is read from the ruler at the end point of the line. For lines with a negative slope the ruler is rotated through 180°.

3.3.3.2 Duration Ruler

The duration ruler defines the relationship between the quantitative and the qualitative measures of the temporal extent of a line.

a)



b)

Line	Direction	Magnitude	Duration	Slope	Start
1	1	2	4	4	1
2	1	7	6	7	2
3	-1	3	5	5	9
4	-1	5	1	8	6
5	-1	2	7	3	2

Figure 3.5: a) Qualitative rulers for magnitude extent, start position, duration and slope for a carbon dioxide evolution rate profile. It is assumed that the profile with which this is being compared has smaller dimensions and thus does not affect the ruler sizing.

b) Qualitative description of the profile in a). This description is dependent upon the profile with which it is being compared.

The qualitative intervals on the temporal extent ruler are delineated using the average duration of all the lines as a reference value. If each line were to cover the same temporal extent the average duration would be the overall duration of the fermentation divided by the number of linear data pieces in the time profile. This value is calculated for each profile being considered and the average of the two used as the reference value.

It was not possible to use a reference value similar to that adopted for the magnitude ruler because most of the lines would have temporal extents residing in the 'very small' or 'small' intervals, the resolution would be too coarse. This is most evident when a data set contains one line with a very long temporal extent, as is often encountered in data from fermentations whose product is formed subsequent to microbial growth (an example is displayed in Figure 3.4(b)).

Five major qualitative intervals for the temporal extent ruler are calculated based on the reference value: the size of the basic interval of the ruler is one quarter of the reference value; the 'very short' interval covers one basic interval, the 'short', 'medium' and 'long' intervals cover an area twice the size of the basic interval and the 'very long' interval covers the remainder of the duration ruler (Figure 3.5(a)). The number of numerical labels required depends on the duration of the fermentation and each one covers an area equivalent to the basic interval.

The duration ruler, like the magnitude ruler, is a sliding ruler: the left point of the ruler is positioned at the start point of the line being measured and the qualitative duration is read from the ruler at the end point of the line.

3.3.3.3 Starting Position Ruler

The starting position ruler assigns a qualitative label to the numerical value of the initial y-value of a line. It is essentially the same as the magnitude ruler however it is stationary: the bottom of the ruler remains fixed at the minimum point of the two profiles being compared, the qualitative starting point is read from the ruler at the start point of the line being measured.

3.3.3.4 Slope Ruler

The slope ruler defines the relationship between the quantitative and the qualitative measures of the slope of a line.

A human's perception of the slope of a line is heavily influenced by the scales of the x- and y-dimensions thus complicating the definition of the slope ruler and limiting its use in the comparison process. Nevertheless, the qualitative description of slope is used to identify horizontal lines (Section 3.3.3.5), to guide the combination of lines within a profile (Section 3.4.3), and to aid the decision of which lines should be compared (Section 3.4.4).

The average slope for each profile is determined as the ratio of the maximum possible change in magnitude for that profile to the overall duration of that fermentation. This value is calculated for both fermentations and the average used as the reference value for the slope ruler which is again divided into five qualitative regions (Figure 3.5(a)). The divisions between the qualitative intervals occur at one fifth, one half, one and a half times, and five times the reference value. The intermediate intervals, described by numerical labels, divide each interval in half again. The boundary points were chosen by trial and error and are purely the subjective choice of the developer.

For lines with a negative slope the mirror image (through the horizontal axis) of the slope ruler is used.

3.3.3.5 Direction Ruler

The qualitative description of direction distinguishes between increasing, decreasing and horizontal periods of a profile. It is important that increasing periods in one fermentation are not compared with decreasing periods in another fermentation as, even though the absolute changes in magnitude and time may be similar, the different direction of the lines points to obvious differences in metabolic action.

A horizontal line is defined as one whose change in magnitude is 'very small' and whose absolute slope is less than half of the 'very small' qualitative slope interval.

The ruler for direction was straightforward and followed the ideas of Bobrow (1985): the direction of a line with a positive slope is given a qualitative label of +1, a horizontal line has a direction of 0, and a line with a negative slope has a direction of -1.

3.3.3.6 Fuzzy Boundaries

Fuzzy intervals were required at the boundaries between qualitative regions for each ruler with the exception of the direction ruler. The use of fuzzy intervals ensures that a numerical quantity that falls at the very end of one qualitative interval on the ruler is not considered to be different from a quantity at the very start of the next qualitative interval.

Again it was necessary that the fuzzy intervals be defined relative to the values of the attributes of the linear data pieces in the data sets being investigated. Various sizes of fuzzy intervals were investigated. The applicability of the fuzzy intervals was judged visually: data sets were simplified, described qualitatively and compared, if the result of the comparison was considerably different from a manual comparison the description of the fuzzy interval was deemed unacceptable. The most successful fuzzy interval was that achieved by dividing each qualitative interval in half and stipulating that attributes whose values resided in adjacent half intervals could be considered similar. Thus a line whose temporal extent was in the upper half of the 'short' interval would be similar to a line whose temporal extent was in the lower half of the 'medium' interval. The numerical labels on each of the rulers in Figure 3.5(a) identify the fuzzy intervals. In the comparison process attributes are considered to be similar if their numerical labels describe the same or adjacent positions on the appropriate qualitative ruler.

Describing two profiles qualitatively by the above techniques is rapid and thus can be used for on-line applications.

3.3.4 Output

The output from the algorithm is two arrays containing the qualitative descriptions of the linear data pieces, one array for each profile. An example of one such array is given in Figure 3.5(b). Each array has six columns describing, from left to right, the line identifier, the direction of the line, the change in magnitude of the line, the duration of the line, the slope of the line and the starting position of the line. It is a very compact and clear representation of the fermentation profile. When each of the qualitative labels is translated into the corresponding qualitative term the language of an expert is evident: for example the carbon dioxide evolution rate of the fermentation described in Figure 3.5(a) initially increased from a *very small* value to a *small* value over a *medium* length of time at a *medium* rate.

The reference values which define the sizes of the intervals in each qualitative ruler are determined from the range of values in the two data sets being considered. The qualitative description of a profile may therefore vary depending on the data with which it is to be compared thus the output from the QUAL algorithms is not permanently stored. Figure 3.6 shows two different qualitative descriptions of a single carbon dioxide evolution rate profile resulting from comparison with two different fermentations. QUAL must therefore be performed on each pair of profiles that are to be compared.

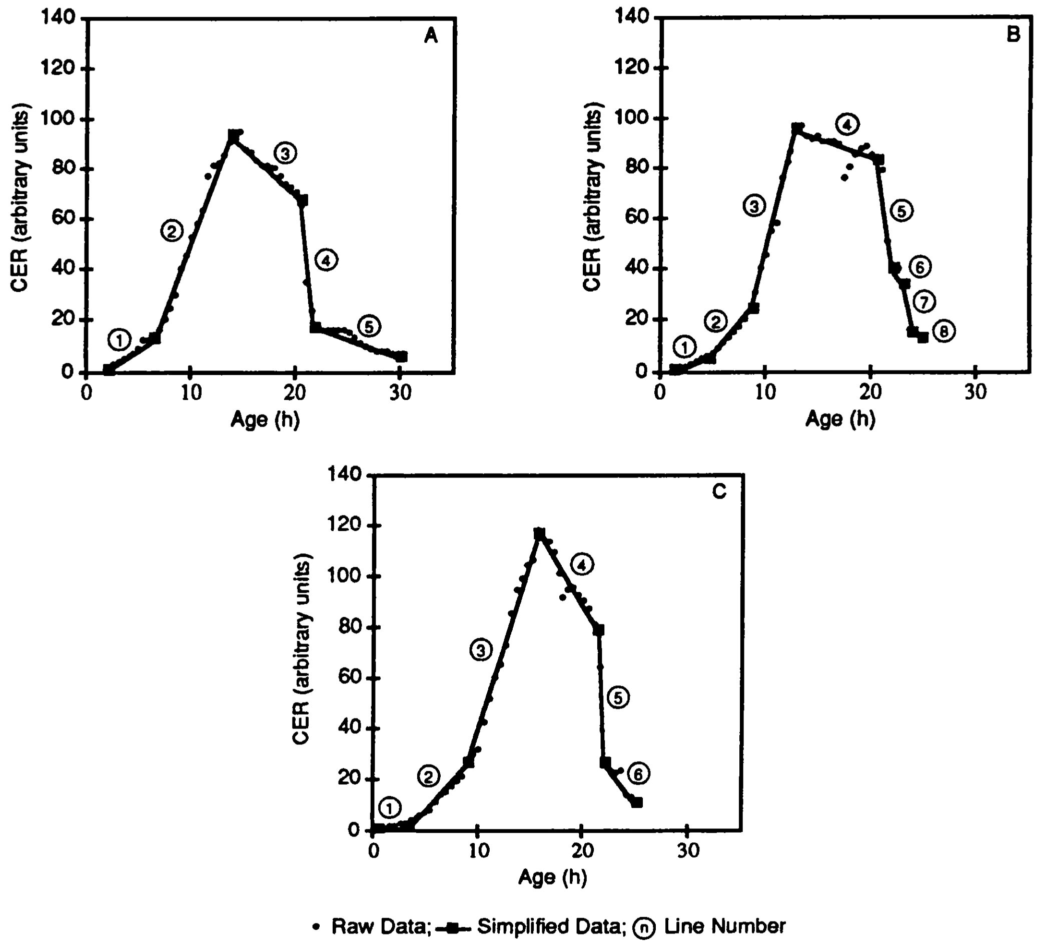
3.4 Comparison of Simplified Data

3.4.1 Overview

The comparison of two data sets occurs after they have been converted to a qualitative representation. It was important that the comparison procedure be generic to all fermentation data sets and not require any prior knowledge of the fermentation.

The data comparison routines can be broken down into four functional parts:

1. input of qualitative description of data from QUAL;
2. the combination of adjacent lines within a profile;



a)

Line	Dir.	Mag.	Dur.	Slope	Start
C1	0	1	3	1	1
C2	1	2	5	5	1
C3	1	7	6	7	2
C4	-1	3	5	6	9
C5	-1	4	1	8	6
C6	-1	2	3	5	3

b)

Line	Dir.	Mag.	Dur.	Slope	Start
C1	0	1	3	1	1
C2	1	2	6	5	1
C3	1	8	7	7	3
C4	-1	3	6	6	9
C5	-1	5	1	8	7
C6	-1	2	3	5	3

Figure 3.6: a) the qualitative description of profile C in relation to profile A. b) the qualitative description of profile C in relation to profile B. The differences between the two descriptions are highlighted in the tables. The two carbon dioxide evolution rate profiles are from the *E. coli* fermentations described in Chapter 4. Profile A is from batch C447, profile B from batch C444 and profile C from batch C441.

3. comparison of the linear data pieces from two profiles;
4. output of the results.

The methods used to achieve each of these functions are described in the following sections.

The profile comparison routines (MATCHER) were written in RM/FORTRAN v2.4 to run on a Tandon PCA 40 AT personal computer. These were later transferred (without alteration) to a Sun Sparc Station 1.

3.4.2 Input of Qualitative Description of Data

The qualitative descriptions of the profiles, as generated by QUAL, are transferred to the MATCHER algorithms in the form of two arrays which were described in Section 3.3.4. The order in which the data are input does not affect the result.

3.4.3 Combining Adjacent Lines

It would normally be expected that two profiles of the same variable from similar fermentation batches would exhibit the same number of linear data pieces upon linearisation. However, the imperfect nature of fermentation data sometimes precludes this. An example is shown in Figure 3.7. Visual inspection would suggest joining the second and third linear data pieces of profile A to enable an adequate comparison with the second line of profile B. As this occurred in a number of cases, a qualitative algebra for the combination of adjacent lines was devised as described below.

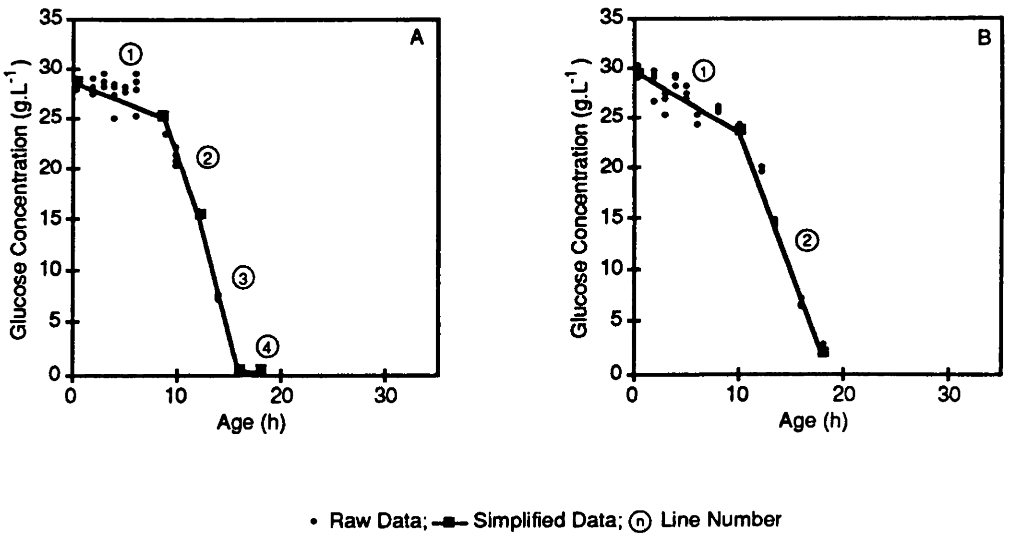


Figure 3.7: The linearisation of two glucose concentration profiles from the *E. coli* fermentations described in Chapter 4. Visual inspection would suggest joining line segments 2 and 3 in profile A. The comparison algorithm automatically joins these two lines so as to improve the match. Profile A is from batch C447 and profile B from batch C443.

1. If the qualitative direction of one line is +1 and the qualitative direction of the other line is -1 the lines cannot be combined.

2. If the slopes of the two lines differ by more than two qualitative units and neither of the lines is 'small' they cannot be combined because the resulting line would be significantly different from the raw data. A 'small' line is one whose magnitude extent and temporal extent are described as 'very small', ie have qualitative labels of 1.

3. When two lines are combined the resulting magnitude extent and temporal extent are found from adding the individual qualitative labels. If either of the lines being combined is horizontal (direction = 0) then the resulting magnitude extent is the same as that of the non-horizontal line. Table 3.2 shows the qualitative algebra devised for combining magnitude extents and Table 3.3 shows the rules for combining temporal extents.

+	1	2	3	4	5	6	7	8
1	2	3	4	5	6	7	8	9
2	3	4	5	6	7	8	9	9
3	4	5	6	7	8	9	9	9
4	5	6	7	8	9	9	9	9
5	6	7	8	9	9	9	9	9
6	7	8	9	9	9	9	9	9
7	8	9	9	9	9	9	9	9
8	9	9	9	9	9	9	9	9
direction= 0	1	2	3	4	5	6	7	8

Table 3.2: Qualitative algebra for the combination of magnitude extents of linear data pieces.

+	1	2	3	4	5	6	7 . . .
1	2	3	4	5	6	7	8
2	3	4	5	6	7	8	9
3	4	5	6	7	8	9	10
4	5	6	7	8	9	10	11
5	6	7	8	9	10	11	12
6	7	8	9	10	11	12	13
7	8	9	10	11	12	13	14
:							

Table 3.3: Qualitative algebra for the combination of temporal extents of linear data pieces.

+	1	2	3	4	5	6	7	8	9
1	1	2	2	3	3	4	4	5	5
2	2	2	3	3	4	4	5	5	6
3	2	3	3	4	4	5	5	6	6
4	3	3	4	4	5	5	6	6	7
5	3	4	4	5	5	6	6	7	7
6	4	4	5	5	6	6	7	7	8
7	4	5	5	6	6	7	7	8	8
8	5	5	6	6	7	7	8	8	9
9	5	6	6	7	7	8	8	9	9
mag. = 0 dur. = 0	1	2	3	4	5	6	7	8	9

Table 3.4: Qualitative algebra for the combination of slopes of linear data pieces. The shaded region marks those combinations that would result in unacceptable lines. (mag. = magnitude extent, dur. = duration or temporal extent).

4. The starting position of the combined lines remains the same as the starting position of the first line.

5. When two lines are combined the resulting qualitative slope is the average of the two qualitative slope intervals as shown in Table 3.4.

It was found that the automatic combination of adjacent lines whose qualitative slopes were identical would help the comparison process. This is performed prior to comparison of the data set.

3.4.4 Comparison of Linear Data Pieces

The comparison algorithm compares two time variant fermentation data sets which have already been linearised and described in qualitative terms. The algorithm performs like a human expert by selecting those lines from each profile which correspond. The two profiles are assumed to be similar until proven otherwise, thus the algorithm investigates a number of combinations of lines and determines which gives the most satisfactory comparison between the two profiles.

Some terms will be defined prior to describing the comparison algorithm. A number of worked examples are provided at the end of this section to clarify the different aspects of the routines.

The comparison routine compares two profiles, one is labelled 'A', the other 'B'. The lines in each profile are labelled, for example, 'A1' for the first line in profile A, 'B4' for the fourth line in profile B.

A *directional episode* is made up of a series of consecutive line segments with the same direction, ie the slope of each line in a directional episode has the same sign. Any line that is deemed to be horizontal (qualitative value of direction = 0) is initially included in the directional episode of the previous line. In the case of fermentation time profiles a change in direction from one linear data piece to the next is indicative of changes in metabolic action thus, when comparing profiles, lines of different direction cannot be considered similar.

Dividing the profiles into directional episodes introduces boundaries across which comparisons cannot be made.

A *match* is the result of comparing lines from two different profiles. There are four types of matches:

1. a *one-to-one match*, eg line A1 is compared with line B1;
2. a *two-to-one match*, eg line A1 is combined with line A2 and compared with line B1;
3. a *one-to-two match*, eg line A1 is compared with the combination of lines B1 and B2;
4. a *null match* where there are no lines in one of the profiles to compare with a line, or lines, in the other profile.

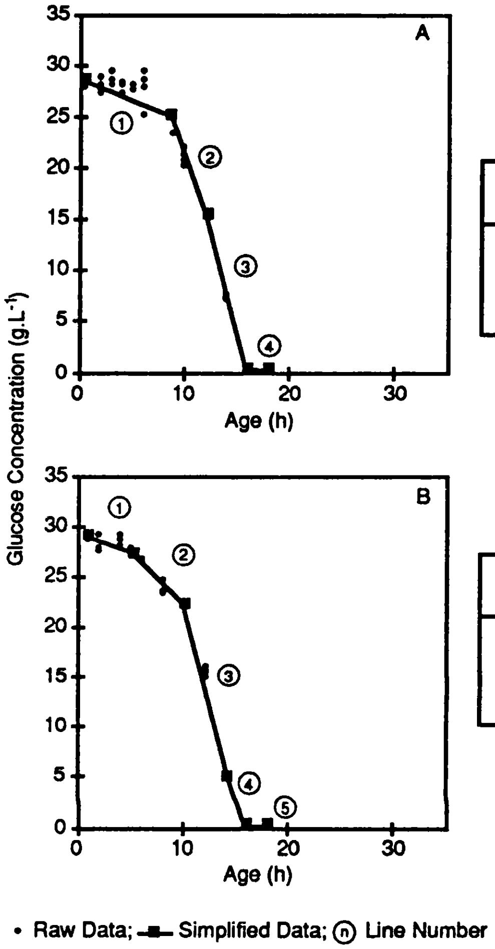
For linear data pieces to be considered *similar* they must have the same qualitative descriptions, that is the same direction, magnitude extent, temporal extent and starting position. For each of the qualitative descriptors similarity is denoted by congruent or adjacent numerical labels. Thus a horizontal line (direction = 0) can be similar to a line with positive slope (direction = +1) if all other attributes are the same; a line with a 'medium' change in magnitude (magnitude = 5) is similar to a line with a 'large' change in magnitude (magnitude = 6) if all other attributes are the same.

Often not all of the four qualitative descriptors are the same for the two components of a match. A measure of the degree of similarity was therefore introduced in the form of a *match score.* At each comparison the similarities of the magnitude extents, temporal extents, starting positions and slopes are determined. Qualitative descriptors that are similar are given a *match value* of one, whilst those that are dissimilar are given a *match value* of zero. The *match score* is the sum of the match values over all the qualitative descriptors of a match excluding direction and starting position. Similarity of direction is a necessity for all lines that are compared and is thus not included in the match score. The starting position descriptor was introduced late in the work to aid recovering the match after a bad fit had been encountered (described later), it should be included in the match score in the future.

The algorithm is initiated by matching line A1 with line B1, lines A1 and A2 with line B1, and line A1 with lines B1 and B2. The best of these matches is the one with the highest match score. Figure 3.8 demonstrates this portion of the routine. If it is not clear which match is best it is then necessary to assess the results obtained by comparing the next available line in each profile for each of the three scenarios. For example, in the one-to-two match, line A2 would be compared with line B3, as shown in the hypothetical example in Figure 3.9. The match score for this match would then be added to that of the previous

SCENARIO: Currently at start of both profiles. Attempt to find the best match between profiles A and B.

STEPS:
i) Compare A1 with B1, A1+A2 with B1, A1 with B1+B2.
ii) The best of these matches is recorded.



Line	Direction	Magnitude	Duration	Slope	Start
A1	-1	2	7	3	9
A2	-1	7	5	6	7
A1+A2	-1	9	12	[5]	9

Line	Direction	Magnitude	Duration	Slope	Start
B1	-1	1	3	2	9
B2	-1	2	4	4	8
B1+B2	-1	3	7	3	9

Trial Matches

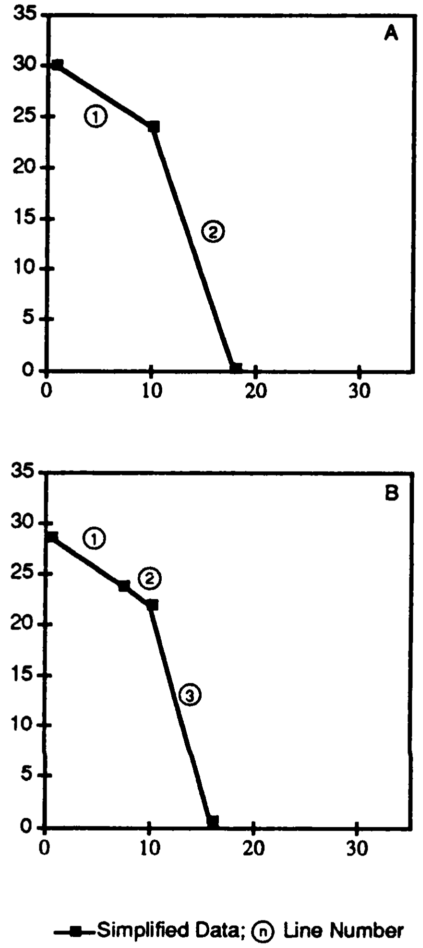
Lines	Magnitude	Duration	Slope	Start	Match Score
A1 → B1	1	0	1	1	2
A1+A2 → B1	0	0	0	1	0
A1 → B1+B2	1	1	1	1	3

Best Match: A1 → B1+B2

Figure 3.8: Example showing the first step in the comparison of glucose concentration profiles from two *E. coli* fermentations. The fermentations are described in Chapter 4. Profile A is from batch C443 and profile B from batch C442.

SCENARIO: Currently at start of both profiles. Attempt to find the best match between profiles A and B.

STEPS:
i) Compare A1 with B1, A1+A2 with B1, A1 with B1+B2.
ii) A1 matched to B1 and A1 matched to B1+B2 have the same match score, therefore examine the next matches;
iii) A2 is compared with B2, and A2 is compared with B3, the match scores from these matches are added to those of the respective previous matches;
iv) the best match is recorded.



Line	Direction	Magnitude	Duration	Slope	Start
A1	-1	2	6	3	9
A2	-1	7	5	6	7

Line	Direction	Magnitude	Duration	Slope	Start
B1	-1	1	5	3	8
B2	-1	1	1	4	8
B3	-1	7	5	6	7
B1+B2	-1	2	6	4	8

Trial Matches

Lines	Magnitude	Duration	Slope	Start	Match Score
A1 → B1	1	1	1	1	3
A1 → B1+B2	1	1	1	1	3
A2 → B2	0	0	0	1	0
A2 → B3	1	1	1	1	3

Sums of Match Scores: (A1→B1) + (A2→B2) = 3
(A1→B1+B2) + (A2→B3) = 6

Best Match: A1 → B1+B2

Figure 3.9: Example showing how the comparison algorithm chooses between matches with identical match scores. The profiles are hypothetical examples.

match, in this example the one-to-two match. The largest of the combined match scores indicates which of the initial matches is best. When the best possible match for the lines being investigated has been found it is recorded in the *match record* (Section 3.4.5).

If the last entry in the match record does not include the first lines of a directional episode and the previous entry involves a combination of lines for one of the profiles, the algorithm then attempts to improve the match even further: the separation and subsequent rematching of the combined lines is investigated as follows. By way of an example (Figure 3.10), lines A1 and A2 have been matched with line B1 and line A3 has been matched with line B2, the combined match score for these two matches is MS1. Separation and rematching results in line A1 being matched with line B1 and line A2 being matched with line B2, the combined match score for these two matches is MS2. The higher of MS1 and MS2 indicates the best two matches and these are thus recorded in the match record.

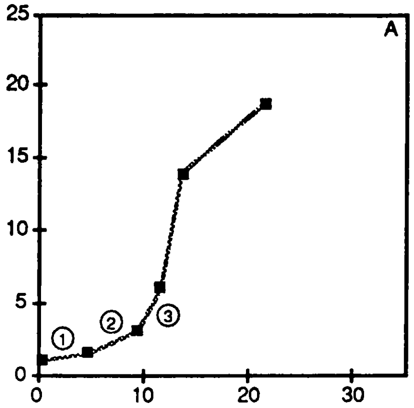
If the latest entry in the match record is bad, ie the match score is less than two, and one of the lines in the match is horizontal, an attempt is made to move the horizontal line over to the next directional episode. If the horizontal line compares well with the first line of the next directional episode, ie the match score resulting from the comparison is greater than, or equal to, two, then the horizontal line becomes the first line of the next directional episode and the match record is updated. The rationale behind this follows from the definition of a horizontal line. Lines with extremely small slopes are described as 'horizontal' because the inaccuracies in the data may mask the 'true' direction, ie a line with an extremely small positive slope is indistinguishable from, and could equally have been described by, a line with an extremely small negative slope. Thus the presence of a horizontal line will complicate the process of distinguishing the end points of directional episodes as the line could belong to either the increasing or the decreasing episode. If a horizontal line disturbs the matching process the possibility of an incorrectly partitioned time profile is investigated by moving the horizontal line to the next directional episode to see if this improves the comparison.

The algorithm now moves on to the next available line in each profile. If there are lines left in the current directional episode of both profiles the above steps are repeated. However, if only one of the profiles has unmatched lines in the current directional episode, there are three possible courses of action, as depicted in Figure 3.11.

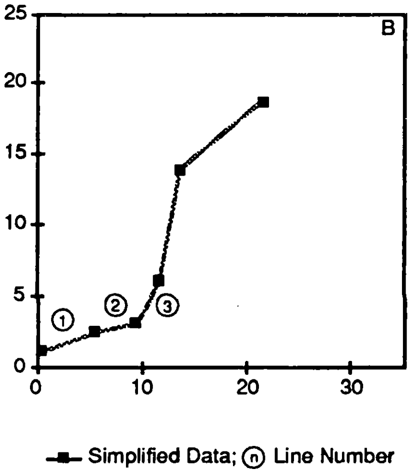
1. If any of the unmatched lines are horizontal an attempt is made to move the line over to the next directional episode as described above.

SCENARIO: Currently have A1+A2 matched to B1. Would separating A1 and A2 improve the subsequent match?

STEPS:
i) Compare A3 with B2, A1 with B1, A2 with B2;
ii) which pair of matches gives the best combined match score?
iii) this match is recorded.



Line	Direction	Magnitude	Duration	Slope	Start
A1	1	1	2	1	1
A2	1	2	2	2	2
A3	1	4	1	6	3
A1+A2	1	3	4	1	1



Line	Direction	Magnitude	Duration	Slope	Start
B1	1	3	3	2	1
B2	1	1	2	1	3
B3	1	4	1	6	3

Trial Matches

Lines	Magnitude	Duration	Slope	Start	Match Score
A1 → B1	0	1	1	1	2
A2 → B2	1	1	1	1	3
A1+A2→ B1	1	1	1	1	3
A3 → B2	0	1	0	1	1

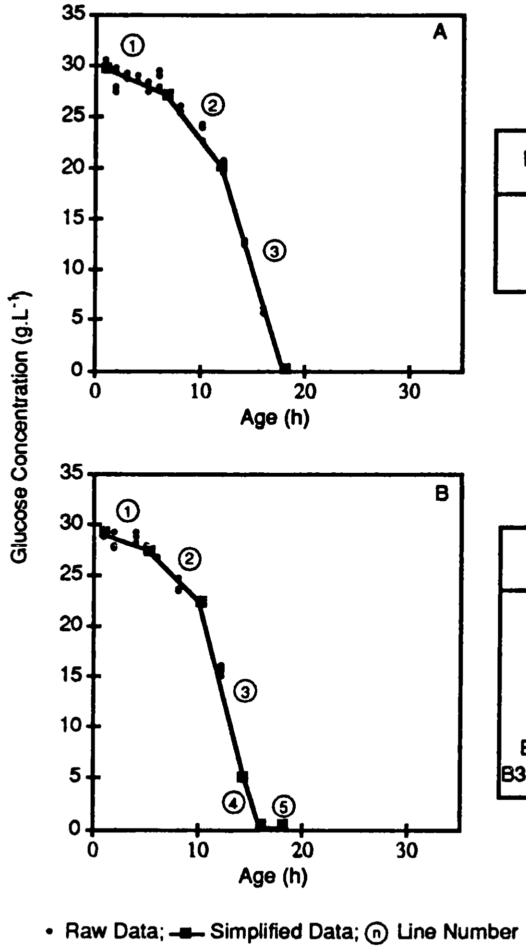
Sums of Match Scores: (A1→B1) + (A2→B2) = 5
(A1+A2→B1) + (A3→B2) = 4

Best Match: A1→B1 / A2→B2

Figure 3.10: Example showing the separation and rematching of lines in an attempt to improve a match record where a previous match had contained combined lines. The profiles are hypothetical examples.

SCENARIO: Currently have A1 matched with B1, A2 matched with B2 and A3 matched with B3+B4. There are no more lines in profile A to match with B5.

STEPS:
i) B5 is horizontal but there are no more directional episodes to move it to;
ii) the previous match does not contain combined lines (profile A) therefore cannot attempt to separate and rematch;
iii) attempt to combine B5 with B3+B4 and assess the resulting match score.



Line	Direction	Magnitude	Duration	Slope	Start
A1	-1	1	5	2	9
A2	-1	2	5	4	8
A3	-1	6	6	6	6

Line	Direction	Magnitude	Duration	Slope	Start
B1	-1	1	4	2	8
B2	-1	2	4	4	8
B3	-1	5	4	6	7
B4	-1	2	2	6	2
B5	0	1	2	1	1
B3+B4	-1	7	6	6	7
B3+B4+B5	-1	8	8	[4]	7

Trial Matches

Lines	Magnitude	Duration	Slope	Start	Match Score
A2 → B2	1	1	1	1	3
A3 → B3+B4	1	1	1	1	3
A3 → B3+B4+B5	0	0	0	1	0

Best Match: A2→ B2 / A3 →B3+B4

Resulting Match Record

Lines	Magnitude	Duration	Slope	Start
A1 → B1	1	1	1	1
A2 → B2	1	1	1	1
A3 → B3+B4	1	1	1	1
- → B5	-	-	-	-

Figure 3.11: Example demonstrating the options available when there are extra lines in either of the current directional episodes. The two glucose concentration profiles are from the *E. coli* fermentations described in Chapter 4. Profile A is from batch C441 and profile B from batch C442.

2. If the previous entry in the match record contained combined lines, the separation and rematching of these lines is investigated as described previously.
3. The possibility of combining the left over lines into the last match is investigated. For example, if the last entry in the match record was line A3 matched to line B2 and the left over line is line A4, then line A4 is combined with line A3 and compared with line B2. If the resulting match score is greater than, or equal to, that of line A3 matched to line B2, then the combined lines are accepted as a good match. More than two lines can be combined in this situation.

If all three of these attempts to deal with the left over lines are unsuccessful then these lines are recorded in the match record as a null match.

When the end of a directional episode is reached the algorithm moves onto the next directional episode and the above routines are repeated. However, if the last entry in the match record was bad as shown by a match score of zero or one, an attempt is made to rematch the lines. The main benefit of this process is seen when one of the profiles suffers some small disturbance but then recovers to its normal path; without the rematching routine the profiles would be considered dissimilar from the point of the disturbance, whereas with the ability to rematch the lines it is possible to record the areas of the disturbance as a null match and then continue matching the remaining lines. The rematching routine is as follows and is shown in the example in Figure 3.12.

- The first step is to find the line in profile A from which to begin the rematching: the match record is searched backwards until either a match is found in which all four match values are equal to one, in which case the next line in profile A is the starting point, or the beginning of the current episode is reached and this line is then the starting point.

- The next step is to find the line in profile B from which to start the rematching: if the starting point in profile A belongs to a non-null match then the starting point in profile B is the line after that which is currently matched to the starting line of profile A, otherwise the starting line in profile B is the line after that which belongs to the last non-null match prior to the match containing the starting line of profile A. These two starting lines must be of the same directions and must have the same qualitative description for their starting positions. The starting line in profile B is increased by one until these criteria are met or there are no more lines present in profile B, in which case the starting line in profile A is increased by one and the search process repeated.

- Having found the appropriate starting lines these lines are then compared.

- If the resulting match score is greater than, or equal to, two then the match is considered adequate, the match record is updated and the algorithm continues, as before, to compare subsequent lines.

- If the match is poor then the algorithm steps through all remaining lines in profile B attempting to match them to the starting line in profile A using the same procedure. If the end of profile B is reached and a suitable match has not been found, the algorithm returns to the starting line in profile B and attempts to match this with the next line in profile A.

In this manner, all the lines in profile B are compared with all the lines in profile A until either a suitable match is found or all the lines have been exhausted.

The rematching routine also makes it possible to skip the first linear data piece of either profile in the comparison process. If the first directional episode ended in a bad match then the rematching routine starts at lines A1 and B2, or the next line in B that has the same direction and starting position as line A1. There are two situations where the skipping of the first line may be important. Firstly, it may happen that the recording instrument is not turned on at the start of the process resulting in missing data which would disrupt the comparative analysis. Secondly, the initial, or lag, phase of a fermentation is notoriously variable as it is dependent on a large number of factors (Bailey and Ollis 1986), thus it is often desirable to relax the comparison over the initial stages. An example is shown in Figure 3.13.

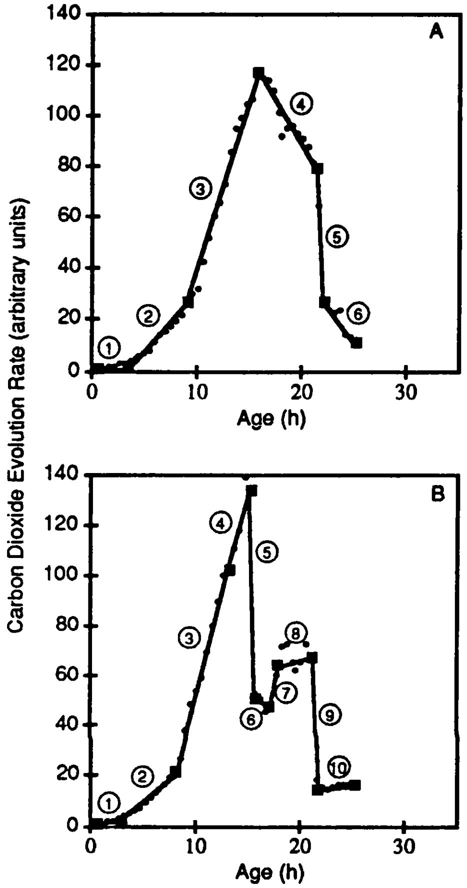
Finally, when the end of one of the profiles is encountered, any remaining lines in the other profile are recorded as null matches after ensuring that they cannot be combined with the lines in the previous match.

The general flow of the comparison algorithm is summarised in Figure 3.14.

The comparison of two profiles using the above algorithm is a rapid process and is thus suitable for on-line applications.

SCENARIO: Currently, in the second directional episode, have A4 matched with B5, A5 matched with B6 and A6 is recorded as a null match. Thus have a bad match at the end of a directional episode, attempt to rematch the lines.

STEPS:
i) the initial starting points for rematching are lines A4, the beginning of the directional episode, and B6, the line after that which is currently matched to A4, but the starting points differ so this match is not allowed;
ii) lines B7 and B8 cannot be matched to A4 because they are of a different direction;
iii) lines B9 and B10 have different starting points from A4 and thus cannot be matched;
iv) line A5 becomes the starting point in profile A for rematching and, following the same reasoning as i), ii) and iii), line B9 is the starting point in profile B for rematching;
v) the resulting match is good thus the match record is updated and the comparison continues.



Line	Direction	Magnitude	Duration	Slope	Start
A4	-1	3	7	5	8
A5	-1	4	1	8	6
A6	-1	2	4	5	2

Line	Direction	Magnitude	Duration	Slope	Start
B5	-1	6	1	8	9
B6	-1	1	2	3	4
B7	1	2	1	7	4
B8	1	1	4	2	5
B9	-1	4	1	8	5
B10	0	1	5	1	1

• Raw Data; -■- Simplified Data; ⓝ Line Number

Match Record Prior to Rematching

Lines	Magnitude	Duration	Slope	Start
A4 → B5	0	0	0	1
A5 → B6	0	1	0	0
A6 → -	-	-	-	-

Match Record After Rematching

Lines	Magnitude	Duration	Slope	Start
A4 → B5	0	0	0	1
- → B6+B7+B8	-	-	-	-
A5 → B9	1	1	1	1
A6 → B10	1	1	0	1

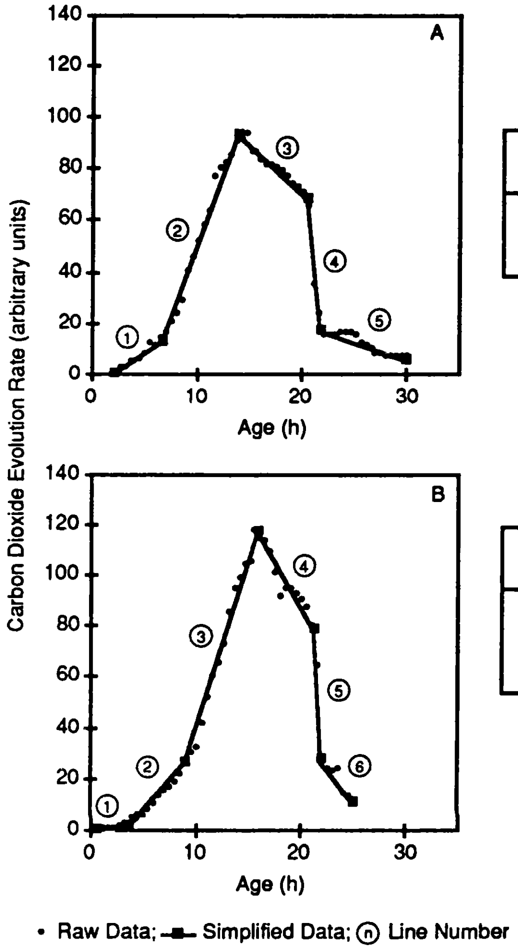
Figure 3.12: Example demonstrating the rematching of lines after a directional episode ends with a bad match. The two carbon dioxide evolution rate profiles are from the *E. coli* fermentations described in Chapter 4. Profile A is from batch C441 and profile B from batch C442.

SCENARIO: Currently, in the first directional episode, have A1 matched with B1, A2 matched with B2 and B3 is recorded as a null match. Thus have a bad match at the end of a directional episode, attempt to rematch the lines.

STEPS:

i) because this is the first directional episode the rematching starts with lines A1 and B2, note that the starting positions of these two lines are similar;

ii) the resulting match is good thus the match record is updated and the comparison continues.



Line	Direction	Magnitude	Duration	Slope	Start
A1	1	1	4	4	1
A2	1	7	6	6	1

Line	Direction	Magnitude	Duration	Slope	Start
B1	0	1	3	1	1
B2	1	2	5	5	1
B3	1	7	6	7	2

Match Record Prior to Rematching

Lines	Magnitude	Duration	Slope	Start
A1 → B1	1	1	0	1
A2 → B2	0	1	1	1
- → B3	-	-	-	-

Match Record After Rematching

Lines	Magnitude	Duration	Slope	Start
- → B1	-	-	-	-
A1 → B2	1	1	1	1
A2 → B3	1	1	1	1

Figure 3.13: Example demonstrating the rematching of lines after the first directional episode ends with a bad match. The two carbon dioxide evolution rate profiles are from the *E. coli* fermentations described in Chapter 4. Profile A is from batch C447 and profile B from batch C441.

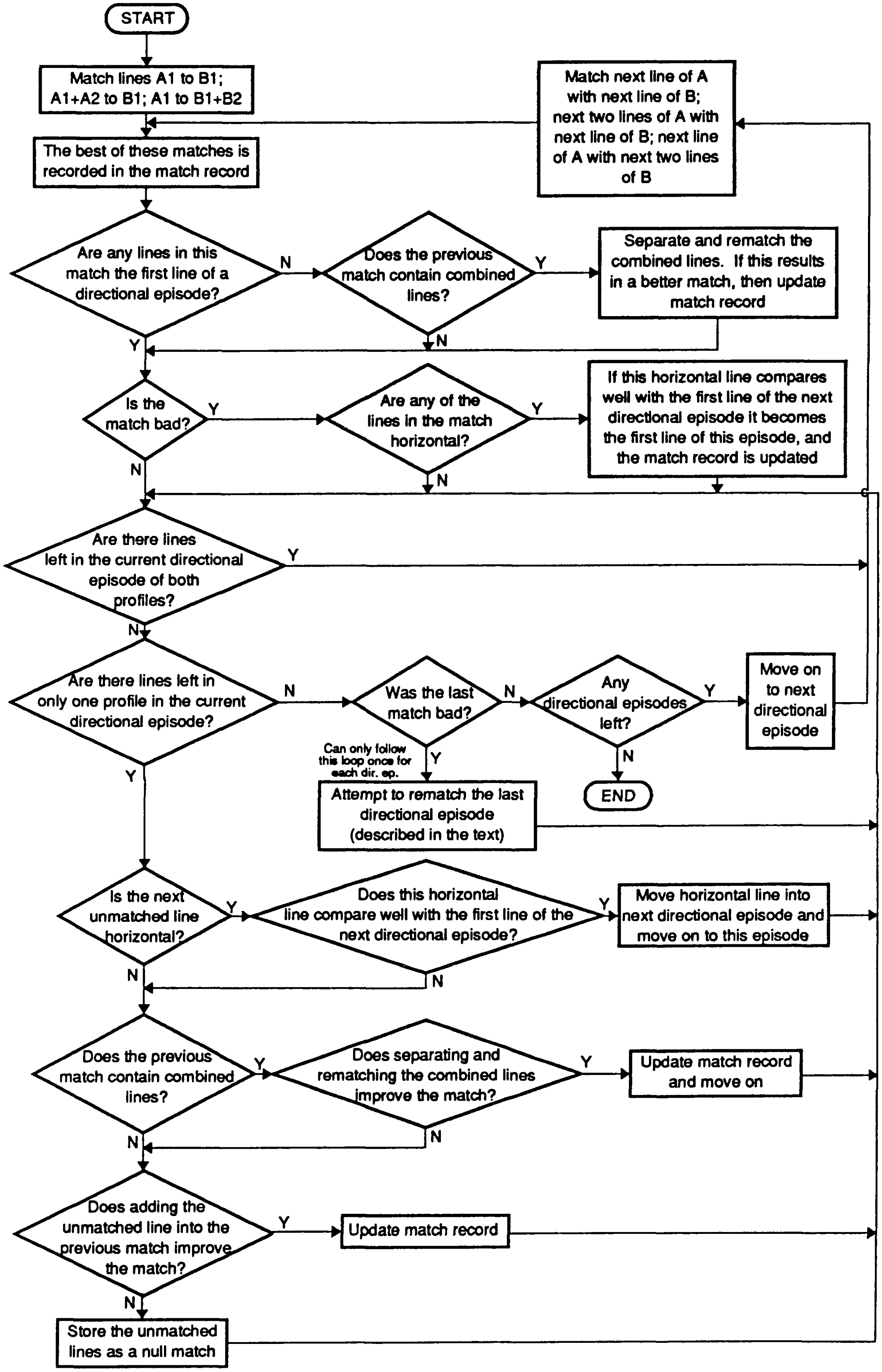

START
Match lines A1 to B1; A1+A2 to B1; A1 to B1+B2
Match next line of A with next line of B; next two lines of A with next line of B; next line of A with next two lines of B
The best of these matches is recorded in the match record
Are any lines in this match the first line of a directional episode?
N
Does the previous match contain combined lines?
Y
Separate and rematch the combined lines. If this results in a better match, then update match record
N
Y
Is the match bad?
Y
Are any of the lines in the match horizontal?
Y
If this horizontal line compares well with the first line of the next directional episode it becomes the first line of this episode, and the match record is updated
N
N
Are there lines left in the current directional episode of both profiles?
Y
N
Are there lines left in only one profile in the current directional episode?
N
Was the last match bad?
N
Any directional episodes left?
Y
Move on to next directional episode
N
END
Can only follow this loop once for each dir. ep.
Y
Attempt to rematch the last directional episode (described in the text)
Y
Is the next unmatched line horizontal?
Y
Does this horizontal line compare well with the first line of the next directional episode?
Y
Move horizontal line into next directional episode and move on to this episode
N
N
Does the previous match contain combined lines?
Y
Does separating and rematching the combined lines improve the match?
Y
Update match record and move on
N
N
Does adding the unmatched line into the previous match improve the match?
Y
Update match record
N
Store the unmatched lines as a null match

Figure 3.14: Simplified flow of MATCHER software for the comparison of two profiles (particulars are described in the text)

3.4.5 Output

The result of the comparison process is an array called the match record, an example of which is given in Figure 3.15. Each row of the array describes the outcome of a comparison and is interpreted as follows:

column 1: first line of profile one in this match
column 2: last line of profile one in this match
column 3: first line of profile two in this match
column 4: last line of profile two in this match
column 5: match value resulting from comparing the magnitude extent of these lines
column 6: match value resulting from comparing the temporal extent of these lines
column 7: match value resulting from comparing the slopes of these lines
column 8: match value resulting from comparing the starting positions of these lines

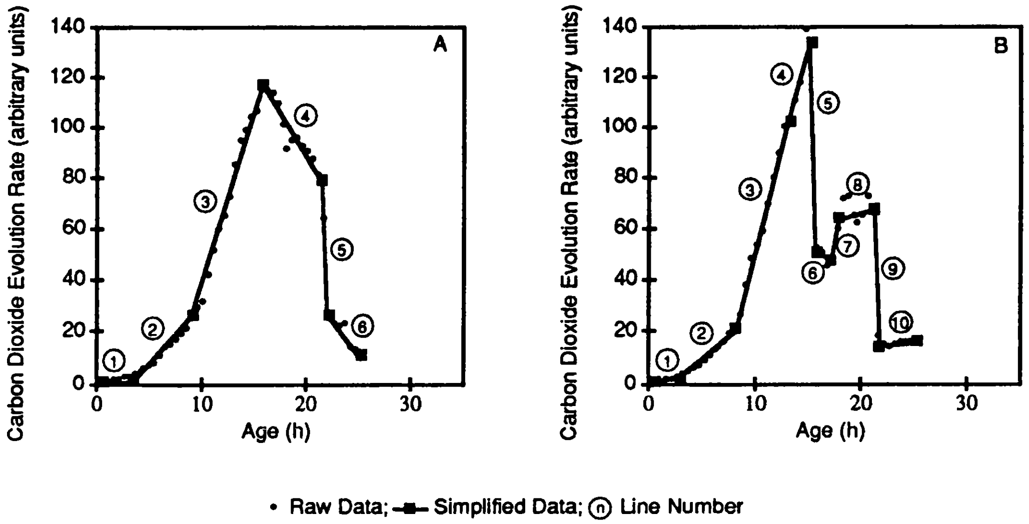
A match value of one indicates that the qualitative descriptors of the lines being compared are similar, whilst a value of zero indicates that the qualitative descriptors are dissimilar. The sum of columns five, six and seven is the match score.

3.5 Interpretation of the Comparison

The match record, obtained from the comparison of time variant data, is interpreted manually. It is envisaged that the initial steps of the interpretation will be automated using a small number of coded rules. The final interpretation is to be carried out by an expert.

The aim of the comparative analysis is, initially, to determine whether or not two profiles are similar. Similarity of two profiles is denoted by a match record in which all the match values are equal to one. A number of other results also indicate similarity:

1. As mentioned earlier the lag phase of a fermentation is notoriously variable, thus dissimilarity of the first lines of the profiles is not usually considered significant although it should still be noted. It is essential that, unless a null match has been recorded, the qualitative descriptor of the starting position of the first lines is the same.



Match Record

Lines				Magnitude	Duration	Slope	Start
A1	A1	B1	B1	1	1	1	1
A2	A2	B2	B2	1	1	1	1
A3	A3	B3	B4	0	1	1	1
A4	A4	B5	B5	0	0	0	1
-		B6	B8	-	-	-	-
A5	A5	B9	B9	1	1	1	1
A6	A6	B10	B10	1	1	0	1

Figure 3.15: The match record summarising the comparison between profiles A and B. Columns one to four, under the heading 'Lines', are the lines compared in a match. In other examples of the match records in this chapter the second and fourth columns have been omitted unless lines have been combined. The results of the comparison are in columns five to eight: a value of 1 indicates similarity, a value of 0 indicates dissimilarity and a '-' represents a null match. The two carbon dioxide evolution rate profiles are from the *E. coli* fermentations described in Chapter 4. Profile A is from batch C441 and profile B from batch C442.

2. In a number of situations it was observed that match values of one were obtained for the magnitude extent and temporal extent of two lines being compared while the match value for slope was zero. Quantitatively, this is nonsense as the definition of slope is the ratio of the change in magnitude to the change in duration of a line, thus if the magnitude and duration of two lines are the same then the slope should also be the same. The discrepancy arises from the difficulty of describing the slope qualitatively. In such a situation the zero match value for slope is ignored. However, when two adjacent lines are combined to form a match with another line, a zero match value indicates that the slopes of the combined lines were significantly different and the combination was not desirable, thus this zero match value cannot be ignored.

3. At the end of the match record a null match, or null matches, for only one of the profiles implies that this fermentation was carried on longer than the one it is being compared with. The null match is noted but does not influence the general similarity of the two profiles.

4. When performing an on-line comparison, the magnitude extent or temporal extent of the latest linear data piece may not compare well with those of the appropriate line in the historical or standard profile. If these qualitative descriptors are smaller than those of the historical or standard batch, but the slope descriptors are similar, then it is assumed that the in-progress batch will eventually match that of the historical or standard batch. No action is taken to alter the course of the in-progress run.

These rules could feasibly be formalised to form a small rule-based interpretation of the match record. At present the rules are applied to the match records manually. The result of applying the rules to all the time variant variables of a pair of fermentations is a list summarising those variables that differed between the two fermentations. This list is combined with that produced by the comparison of the time invariant data to give a concise *difference summary* of the two fermentations.

The analyst must then interpret these results to establish which factors caused the deviations in fermentation performance. The analyst must return to the qualitative descriptions and match records to determine what the differences in time variant variables were. This process could easily be automated and the results provided in the difference summary. The analyst then uses his/her expertise to relate causes to effects. The process is greatly facilitated by having all the information summarised in one place. If conclusive evidence is not available then further investigations may be required; the information in the difference summary gives an indication of what factors need to be looked at in these investigations.

Examples of the use of these tools in a research environment, and the interpretation of the results, is presented in Chapter 4. The use of the tools in an on-line environment is discussed in Chapter 5.

3.6 Discussion

The procedure developed for comparing time variant fermentation data consists of four main steps:

1. simplification of the data into piecewise linear segments;
2. abstraction of the data to a qualitative description;
3. comparison of the qualitatively described linear data pieces from two data sets;
4. interpretation of the results.

FORTRAN computer routines have been written for the first three steps and the final interpretation is carried out manually. The process is summarised in Figure 3.16.

These tools for the comparison of time variant fermentation data require no prior knowledge of the fermentation process being analysed and can therefore be applied to both developmental and industrial data. The tools do not rely on process specific information either so they can be used to analyse data from any type of fermentation without adaptation. The only user-specific information required is the goodness of fit criterion used to guide the linearisation of each data type. However, this fitness criterion is defined for each type of data and each instrument used and does not alter from one batch to another or from one process to another.

There were two objectives to be achieved by the linearisation of the time variant data: firstly, to facilitate comparison of the data by removing noise and grouping data so that non-corresponding time points from two data sets did not hinder the comparison, and, secondly, to summarise the data in a form that resembles an expert's view of the data. This latter objective was included to ensure that the manipulation of the data was meaningful to the user thus encouraging use of the tools. The results obtained by the linearisation routines

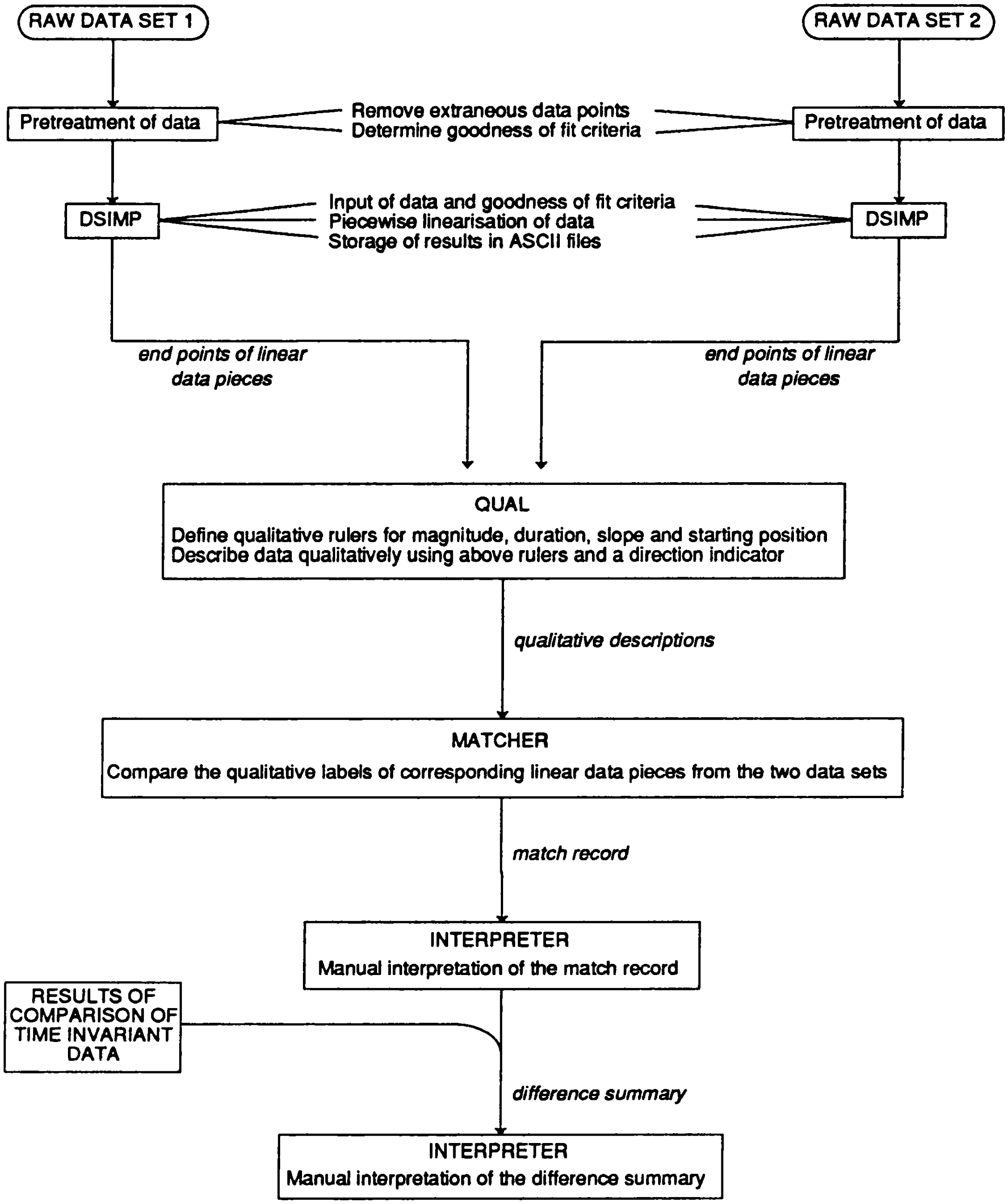

RAW DATA SET 1
RAW DATA SET 2
Pretreatment of data
Remove extraneous data points
Determine goodness of fit criteria
Pretreatment of data
DSIMP
Input of data and goodness of fit criteria
Piecewise linearisation of data
Storage of results in ASCII files
DSIMP
end points of linear data pieces
end points of linear data pieces
QUAL
Define qualitative rulers for magnitude, duration, slope and starting position
Describe data qualitatively using above rulers and a direction indicator
qualitative descriptions
MATCHER
Compare the qualitative labels of corresponding linear data pieces from the two data sets
match record
INTERPRETER
Manual interpretation of the match record
RESULTS OF COMPARISON OF TIME INVARIANT DATA
difference summary
INTERPRETER
Manual interpretation of the difference summary

Figure 3.16: Summary of the tools developed for the comparative analysis of time variant fermentation data. All data sets are individually simplified by piecewise linearisation. The two data sets to be compared are then described using the approximate qualitative terminology of an expert. The qualitative descriptions of the linear data segments from the two data sets are then compared. A difference summary is produced by combining the results of comparing both the time variant data and the time invariant data (Chapter 2) and is interpreted by the human analyst. DSIMP, QUAL and MATCHER are FORTRAN computer routines that were developed to perform the individual steps of the comparison procedure. The terms in italics are the outputs from each step.

were amenable to comparison as shown by the examples in this chapter, thus the first objective was achieved. The time profiles used to illustrate the workings of the comparative analysis tools also show that the second objective was met: the linearisation of the time profiles very approximately segmented the fermentation into its growth phases, ie lag phase, fast growth phase, stationary phase and decline phase (Bailey and Ollis 1986). In some situations more than one line segment was used to describe the data in each phase of the fermentation but it was still possible to apply some physiological meaning to the simplified data sets. This was achieved without providing the simplification routines with a template of what the data were expected to look like. No prior knowledge of the fermentation was required as all the information came from the data sets themselves.

A goodness of fit value dictates how close the line segments must be to the raw data in the simplification routines. Guidelines for the specification of goodness of fit criteria were developed in this work. The means of determining appropriate goodness of fit values was achieved by trial and error. It was found that the use of slightly different values did not actually affect the final comparison results so long as the same technique was used for finding the goodness of fit in both data sets being considered and the fits were reasonably close to the raw data. However, the resulting simplified data may not resemble an expert's view of the process and thus acceptability of the tools would be limited. In Chapter 4 it is shown that the linearisation of the data highlights correlations between various events in the variables of a fermentation, for example the maximum value in one variable may always occur at the same time as the minimum value in another variable. If an 'incorrect' goodness of fit value were applied to the data, the highlighted 'events' may not have any physiological meaning and thus would not be of any use in improving understanding of the process. The choice of an appropriate goodness of fit value is therefore very important. The guidelines presented in this chapter result in goodness of fit values that dictate an adequate fit in all situations examined.

The simplified data were described qualitatively so as to enable 'imprecise' comparisons. The qualitative rulers were not based upon any fundamental principles but were chosen by trial and error. The results appear to be adequate and it will be seen in the next chapter that comparisons based on these qualitative descriptions do concur with visual comparison of the data. The slope ruler was the only one that had questionable results and may need further refinement. The importance of the slope descriptor in the final comparisons was reduced as a result.

The definition of the qualitative rulers remains the same for every variable and for every type of fermentation, no process specific information is required.

The use of the qualitative rulers not only provides a mechanism for finding approximate similarities between data sets but does so in a consistent manner. This is what makes these techniques superior to manual analysis of the data.

The comparison routines work on the principle of 'similar until proven otherwise' and recognise that some portions of two data sets may be similar whilst others differ substantially: lines are combined, separated and skipped over until almost every combination has been investigated in an attempt to find the best possible match between two data sets. Humans have a remarkable ability to perform such manipulations without conscious effort. However, the benefits of consistency and the ability to consider all variables rather than a select few, justify the effort required in automating the comparison process.

The comparison tools can be applied to any data set that reflects the dynamics of a batch or fed batch process but cannot be used for data whose values do not change considerably over the course of the fermentation. For example, the qualitative description routine (QUAL) and the comparison routine (MATCHER) cannot be applied to temperature or pressure data which have been controlled at a constant level throughout a fermentation. The piecewise linearisation routine (DSIMP) could be applied to these data if required. The application of QUAL to set point data gives unacceptable results: in data sets that have been controlled at a single set point the maximum possible change in magnitude is very small thus the divisions of the magnitude ruler would be minuscule and comparison of the magnitude descriptor would be meaningless. The same is true for the start position ruler and the slope ruler. The definition of a horizontal line is also affected by the small overall change in magnitude of set point controlled data, lines that would be considered horizontal by visual inspection would not be defined as such by the QUAL routines. For this reason the set point controlled data are recorded with the time invariant data in the data base, with the appropriate set point being noted. If a deviation from the set point occurs this would normally be detected by the control system and the operator would be alerted by an alarm. This fault would be recorded in the data base in the *comments* field. It is possible to use QUAL and MATCHER on faulty set point data. This would be useful for a fault analysis system where it is required to utilise information from previous fermentations in which similar faults had occurred; the detection of similar faults would be facilitated by QUAL and MATCHER. This is discussed further in Chapter 5.

The final step in the comparative analysis of fermentation data is the interpretation of the difference summary, ie determining why the differences between two fermentations occurred and what implications this has to the process as a whole. This final interpretation has not been automated but left to the analyst. It is important that both the time variant and time invariant data are considered at this stage as both provide vital information about the process. This is demonstrated in the examples in the next chapter.

4 COMPARATIVE REASONING IN A DEVELOPMENTAL ENVIRONMENT - AN EXAMPLE

In this chapter the comparative reasoning tools developed in the previous two chapters are used to analyse a set of experimental fermentations. The conclusions of the analyses were the same as those obtained by a previous manual analysis of the data but significantly more insight into the process was obtained using the computerised techniques. The analysis demonstrates the efficacy of the individual comparative reasoning tools and highlights the benefits of automating the comparison of fermentation data.

4.1 Introduction to the Experimental System

During the course of this study a series of laboratory scale *Escherichia coli* fermentations were carried out at Merck Sharp and Dohme Research Laboratories (MSDRL, Rahway, New Jersey, USA). The objectives of the experiments were:

1. to determine the effects of sterilisation conditions on fermentation performance with a view to scale up;

2. to determine the effects of inoculum concentration on fermentation performance with a view to scale up;

3. to provide a comprehensive data set with which to develop and test the techniques for the automated comparative analysis of fermentation data.

The product of the experimental fermentations was Acidic Fibroblast Growth Factor (aFGF), described in the next section.

4.1.1 Acidic Fibroblast Growth Factor

Acidic Fibroblast Growth Factor (aFGF) is a protein produced in the brain, heart and other organs. The protein is a mitogen, a substance which induces cellular division. This mitogenic activity has been observed on a wide variety of cells in culture, including fibroblasts, vascular and corneal endothelial cells, chondrocytes, osteoblasts, myoblasts and glial cells (Thomas and Gimenez-Gallego 1986). Little is known about the normal activities of aFGF *in vivo* although it is known to induce blood vessel capillary growth. Thomas and Gimenez-Gallego (1986) hypothesised that, because of the broad spectrum of target cells in culture, it is likely that fibroblast growth factors are general tissue growth factors that stimulate and coordinate mitogenesis in many cell types during animal growth, maintenance and tissue repair. aFGF is thus thought to have potential therapeutic use as a topical wound healing agent for burns, bedsores and ulcers, and as a healing agent in corneal transplants.

Bovine and human aFGFs have been purified and amino acid sequencing has shown that both have 140 amino acid residues and are very similar in construct (Gimenez-Gallego *et al.* 1985, Gimenez-Gallego *et al.* 1986).

The bovine aFGF gene has been synthesised (Linemeyer *et al.* 1987) and subsequently converted to a human synthetic aFGF gene by point mutations (Linemeyer *et al.* 1987). The synthetic gene was cloned into a pKK2.7 plasmid vector through an EcoRI-SalI linkage, downstream to a tac promoter (Linemeyer *et al.* 1987). The bacterial host chosen by Linemeyer *et al.* (1987) for expression of the recombinant aFGF was *E. coli* DH5. This transformed strain was used in these studies.

4.1.2 The Effect of Sterilisation Conditions on Fermentations

One of the objectives of the experimental work was to assess the effect of differing sterilisation conditions on the aFGF fermentation with a view to scale up. The effects of sterilisation on a fermentation process are dependent on the medium components and, in a batch process, differ with the size of the operation. A new technique for assessing the effect sterilisation has on a fermentation medium is introduced in Section 4.1.2.1 and was used successfully in these investigations.

The most common method of sterilising fermentation media and equipment is steam heating under pressure (Corbett 1985). Despite the many advantages of using continuous sterilisation (Corbett 1985), a batch process is still frequently used on both laboratory and industrial scales.

The primary deleterious effect of heat sterilisation is the destruction of medium components, either by thermal degradation or by unwanted chemical reactions, and this in turn affects the performance of the microbial culture growing in that medium. The most commonly observed effect is the destruction of sugars by the Maillard or browning reaction: reducing sugars react with the amino groups of proteins when heated resulting in a decrease in the amount of sugars available, the destruction of amino acids and the formation of toxic substances which may interfere with the subsequent fermentation (Gottschalk 1972). The solubility of various substances in the medium is also affected by sterilisation conditions: different combinations of media components, pre-sterile medium pH and length of heating have been shown to have different effects on the solubility of certain media constituents such as the nitrogen source, altering the availability to the microorganism and thus having considerable influence on the subsequent fermentation pattern (Corbett 1985). Heat labile media components, such as many vitamins, are degraded during the sterilisation process (Benterud 1977) and, where possible, are usually separately filter sterilised.

Mathematical and experimental investigations of batch sterilisation processes (Deindoerfer and Humphrey 1959, Singh *et al.* 1989) show that increasing the heat stress on the medium, ie subjecting the medium to longer or harsher sterilisation, increases the extent of nutrient degradation. The effect on process performance depends on the nutrients present and on the nutritional requirements of the microorganism.

A typical batch sterilisation cycle is characterised by a heat up phase, followed by a holding period, and finally a cool down phase. The slope and duration of the heating and cooling temperature profiles are very dependent on equipment heat transfer capabilities and are therefore a function of scale. In large fermenters the medium is subjected to increased heat stress because of the longer sterilisation times required to achieve sterility and longer heat up and cool down periods resulting from the smaller heat exchange surface per unit volume (Buckland 1984). Prior to scale up of a fermentation it is therefore desirable to examine what effect increasing the heat stress has on the medium and on the resulting performance of the fermentation. In the work reported here large scale sterilisation conditions were simulated on a smaller scale by altering the hold time and cooling water flow rate during sterilisation. The effect these changes had on the progress of the aFGF fermentation gave an indication of what would be expected on scale up.

4.1.2.1 Detecting Media Changes Caused By Sterilisation

The effects of differing sterilisation regimes are process dependent. It is difficult to determine the extent of degradation that has occurred in a medium during sterilisation without carrying out a complete chemical analysis on the broth which is an extremely time consuming task and does not give immediate results. A qualitative assessment of the relative amount of degradation occurring may be obtained from the change in pH of the medium during sterilisation, however, this is not a very sensitive indicator. An alternative technique, using the absorption spectrum of the medium, was investigated in this work as a novel method for assessing the amount of nutrient degradation that occurred during sterilisation. The rationale for introducing this new technique is described below.

The absorbance of proteins in the UV range of the spectrum is caused by the peptide groups, the aromatic amino acids and the disulphide bonds. Aromatic amino acids also give rise to fluorescence emission. Schmid (1989) explains the phenomena of absorption and emission in proteins and describes in detail the experimental protocols for measuring these occurrences. Differing concentrations and conformations of proteins and protein components give rise to different spectral characteristics, thus measurement of the difference in spectral properties of a protein-containing medium before and after sterilisation should give an indication of the extent of changes that occurred in the protein component of the medium.

Peptide bonds absorb strongly below 230 nm, whilst the aromatic side chains of tyrosine, tryptophan and phenylalanine absorb in the 230-300 nm range, and disulphide bonds show weak absorbance at about 250 nm (Schmid 1989). The position of the absorption peaks depends on the nature of the molecular neighbourhood of the chromophores and the intensity and shape of the peaks depend on the number and kind of chromophores and their position in the protein molecule, that is the degree of burial of the respective side-chains in the interior of the protein. For these reasons it is difficult to interpret the absorption spectrum of an unknown complex mixture of proteins and protein components.

The applicability of absorption spectroscopy to fermentations was investigated using broth samples from some of the fermentations carried out in this work. The changes in the spectra over sterilisation were found to provide useful information, but should be used in conjunction with pH measurements, not as an alternative. Although not investigated here, it is also possible that the absorption spectrum of a medium prior to sterilisation could be used to indicate non-standard broth compositions resulting from incorrect batching or a change in the quality of the broth components.

A single beam spectrophotometer was used in this work to obtain the absorption spectra of pre- and post-sterile fermentation broths (Section 4.2.6.5). Schmid (1989) describes the use of double-beam spectrophotometers, or microprocessor-controlled single-beam spectrophotometers, for accurately measuring difference spectra which are a useful means of monitoring conformational changes in a protein as the difference in the absorption spectra between the native and the unfolded protein is generally small. These techniques would obviously improve the analysis of changes in fermentation broths.

Measurements of fluorescence could also be used to measure the effects of sterilisation on a protein-containing fermentation medium and, because of the large changes in fluorescence emission as a result of conformational changes in the proteins, would be more sensitive than absorption measurements (Schmid 1989, Cantor and Timasheff 1982).

4.1.3 The Effect of Inoculum Concentration on Fermentations

The second objective of these experiments was to determine the effect of inoculum concentration on the aFGF fermentation.

In order to provide the same inoculum concentration on a large scale as on the laboratory scale a number of seed stages may be required. Several problems may occur as a result of this:

1. the likelihood of contamination increases;

2. a further source of variation is introduced especially when the progress of the seed trains is not monitored;

3. if a mutant organism is being used the risk of reversion to wild type increases as the number of generations increases;

4. for recombinant organisms increasing the number of generations may result in the loss of an unstable plasmid;

5. each time the culture is transferred to a new stage it is subjected to a change in environment; as the number of transfers increases the organism may become less capable of adjusting to these changes and a loss in viability or slow down in metabolism may result;

6. for each extra seed stage another fermenter is required thus another series of issues relating to GMP (Good Manufacturing Practice) operation is introduced.

For these reasons it is often worth reducing the number of seed stages used in a process. This may be achieved either by using a lower inoculum concentration or by combining a number of first stage seeds to provide the required inoculum concentration.

The use of a lower inoculum concentration may result in a longer lag phase thus increasing the overall length of the fermentation. The implications of this must be assessed prior to scale up. Experiments were carried out to determine the effect of a low inoculum concentration on the aFGF fermentation.

4.2 Materials and Methods

The materials and methods used in these fermentation experiments followed the MSDRL Standard Operating Procedures (SOPs) for laboratory scale aFGF fermentations (with the exception of the scanning spectroscopy work and the data analysis procedures which were new). Some details have been omitted from the descriptions given below for proprietary reasons.

4.2.1 Organism

The organism used for the aFGF fermentations was *Escherichia coli* strain DH5 containing the pKK-aFGF plasmid vector. The cultures were provided by Merck Sharp and Dohme, Rahway, New Jersey, USA.

4.2.2 Seed Preparation

The seed medium consisted of glucose, salts and complex protein sources. All medium components were GMP grade and the same lot number of each component was used for all fermentations in this work. The heat sensitive medium components were filter sterilised through 0.22 μm cellulose acetate membrane units. Medium components that react when heated were dissolved in deionised water and separately sterilised at 123°C for 90 minutes. The remaining components were dissolved in deionised water in a 2 L Erlenmeyer flask, pH adjusted with 50% sodium hydroxide to 7.0 and sterilised at 123°C for 90 minutes. The seed media for this work were prepared at the outset and stored at 4°C until required.

Prior to inoculation, the sterilised medium components were combined in the 2 L Erlenmeyer flask to give a volume of approximately 300 mL. The seed medium was inoculated with 0.6 mL of a frozen suspension of *E. coli* DH5 pKK-aFGF. The seed flask was incubated for 12 hours on a shaker at 220 rpm and 37°C.

4.2.3 Fermentations

The aFGF fermentation experiments were carried out in two 15 L Biolafitte fermenters (BL3 and BL4, Biolafitte, Poissy, France).

The fermentation medium was the same as the seed medium containing glucose, salts and complex protein sources. Lactose was included as the inducer of aFGF synthesis. All medium components were GMP grade.

The preparation of the heat sensitive medium components and those that react on heating was the same as for the seed medium described above. The remaining components, including antifoam, were dissolved in deionised water, pH adjusted with 50% sodium hydroxide to 7.2 and sterilised *in situ* at 122°C using direct steam injection. In some experiments glucose was added to the bulk broth prior to sterilisation so as to determine the effect this had on the subsequent fermentation (Table 4.1). The sterilisation conditions often found in larger vessels were simulated over the course of the experiments by subjecting the media to longer sterilisation holding times and by turning off the cooling water to extend the cool down phase. Details of the sterilisation conditions of each fermentation are given in Table 4.1. The sterilisation sequence was computer controlled and is described in Section 4.2.4.4.

BATCH NUMBER	STERILISATION HOLDING TIME (h)	GLUCOSE STERILISED IN SITU	OTHER CONDITIONS
C439	1.00	no	
C440	1.00	no	
C441	0.33	no	
C442	0.33	no	
C443	1.50	no	
C444	1.50	no	
C446	0.33	yes	
C447	0.33	no	
C449	1.00	yes	
C450	0.33	no	slow cooling after sterilisation
C451	0.33	no	
C452	0.33	no	0.25% v/v inoculum

Table 4.1: Details of aFGF fermentation experiments.

During sterilisation the condensate from steam injection increased the broth volume to no more than 8.5 L so that, with the addition of the remaining medium components, the final working volume was 10 L. The ‘open position’ of the valve controlling the flow of steam to the broth was adjusted manually to compensate for fluctuations in the steam supply. However, in some batches the valve was not adjusted correctly and the condensate pick-up resulted in a broth volume exceeding 8.5 L. Broth was removed from these fermenters prior to the addition of the remaining broth components thus giving a final working volume of 10 L but with a lower concentration of some components.

After sterilisation, the remaining medium components were added to the fermenter and sterile deionised water was added when necessary to give a final broth volume of 10 L. The pH was adjusted to 7.0 with 2M sodium hydroxide. The control system actually maintained the pH at about 6.8 rather than the desired 7.0, however this was consistent throughout the study.

100 mL (25 mL in C452) of seed culture were inoculated into the sterilised growth medium in the fermenters. Whenever two fermentations were run concurrently the seed was taken from the same inoculum vessel. The culture was grown for about 24 hours at 37°C and 5 psig. Dissolved oxygen tension in the broth was controlled above 20% of air saturation with the agitation rate and air flow rate under cascade control. During growth 2M sodium hydroxide was added as required to maintain a neutral pH.

4.2.4 Computer Process Control and Data Acquisition System

The laboratory scale fermenters were under the control of a Fisher Provox Process Control System enhanced with a Hewlett Packard A series supervisory computer. The coupling of these two systems enabled the following range of facilities for all fermentations:

1. monitoring directly measured variables;
2. mathematical manipulation of acquired data;
3. control loop implementation;
4. on-line display of process data;
5. sequential (batch) control;
6. archival of batch data.

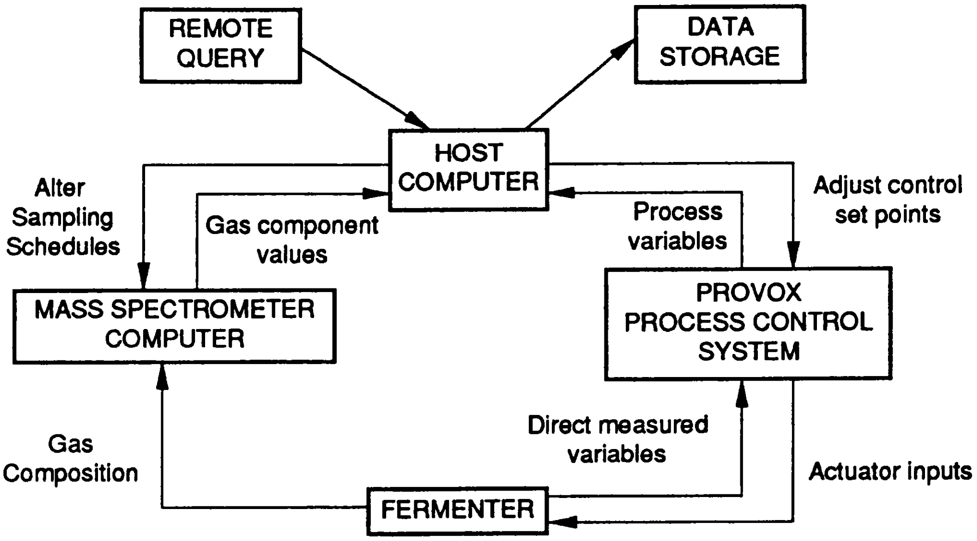


Figure 4.1: Process control and data acquisition system (adapted from D.R. Omstead, K.D. Reda, J.M. Maglaty, T.D. Harrington, 'Continuous and Batch Control of Fermentors', MSDRL report).

Examples of these capabilities are described below. The complete process control and data acquisition system is summarised schematically in Figure 4.1.

4.2.4.1 Directly Measured Variables

The variables that were monitored directly, via analog to digital conversion, are listed in Table 4.2.

4.2.4.2 Derived Variables

Most data obtained from a fermentation require some form of mathematical manipulation, whether it be a simple linear scaling to produce engineering units or a more complicated combination of variables to provide more useful information. The variables calculated by the supervisory computer from the fermentation data are listed in Table 4.2.

DIRECTLY MEASURED VARIABLES	DERIVED VARIABLES
Temperature	Carbon Dioxide Evolution Rate
Dissolved Oxygen Tension	Oxygen Uptake Rate
Pressure	Respiratory Quotient
pH	
Air Flow Rate	
Agitation Rate	
Vent Gas Composition	
Volume	
Cumulative Alkali Addition	

Table 4.2: Directly measured and derived variables monitored by the Provox and host computer.

4.2.4.3 Controlled Variables

Temperature, pressure, pH, dissolved oxygen, air flow rate and agitation rate were controlled by the Fisher Provox Process Control System. Most of the parameters were under set point control using the standard proportional, integral, derivative (PID) algorithm.

Dissolved oxygen was under cascade control with, firstly, the air flow rate and, secondly, the agitation rate. In this the output from a standard PID dissolved oxygen control loop is cascaded to control the set point for the air flow rate. When the air flow rate reaches a pre-specified maximum value the dissolved oxygen control loop output then dictates the set point for the agitation rate (Buckland 1990). The relationships between the dissolved oxygen value and the air flow rate and agitation rate set points were previously coded into the Process Control System.

4.2.4.4 Sequential Control

Automatic sequential control was used to provide a reproducible sterilisation sequence for all fermentation batches. This process is essentially the opening, closing and adjustment of utility valves such that the temperature in a fermenter rises to 122°C for a predetermined time and is then systematically lowered to the required operating temperature. As described earlier, the valve controlling the steam flow to the broth was adjusted manually to compensate for fluctuations in the steam supply. A simplified description of the sterilisation sequence is given in Table 4.3.

STEP LABEL	BASIS FOR ACTION	CONTROL ACTION
Heat	Start	Open jacket steam valve Close air inlet
Heat filter	When T=95°C	Open filter steam valve
Superheat	When T=105°C	Set vent orifice
Sterilise	When T=122°C	Control temperature Start timer
Cool	When time = hold time	Close steam valves Open city water valves
Hold	When T=50°C	Close city water valves Open chilled water valves

Table 4.3: Steps in the sterilisation sequence for the Biolafitte fermenters. (Adapted from D.R. Omstead, 'Computer Applications in Fermentation Processes', MSDRL report).

4.2.4.5 Data Archival

All monitored variables from the fermenters were stored in two column ASCII format (time,value) on the host computer. Each different variable was stored in a separate file which was labelled by batch number and variable type for ease of recognition.

4.2.5 Sampling

Samples of 20 to 40 mL were aseptically removed from the fermenters at one to two hourly intervals for off-line analysis of biomass concentration by optical density and dry weight, glucose concentration and aFGF concentration. Samples were put on ice as soon as they were removed from the fermenter so as to slow down metabolic activity as quickly as possible and thus obtain a more accurate picture of the state of the culture at the time of sampling. Broth samples were stored at -18°C prior to aFGF analysis.

A small amount of inoculum from each batch was reserved for determination of the optical density and glucose concentration of the seed culture at the time it was introduced to the fermenters. Again this was stored on ice until the measurements were taken.

4.2.6 Analytical Techniques

In each of the techniques described below the preparation and measurement process was performed three or four times on each sample. The SmartWare II spreadsheet (version 1.02, Informix Software Inc., Menlo Park, CA) was used to obtain average values and standard deviations for each sample. Average standard deviations for each off-line variable were calculated for use in the simplification routines: this required averaging the standard deviation of all samples from all the experimental fermentations. The results were stored in four columns (time, average value, minimum value, maximum value) in files for input to the simplification routines (DSIMP, described in Chapter 3). For each set of readings Chauvenet's criterion (Holman and Gajda 1978) was used to test for outliers which were discarded when found (Appendix 2).

4.2.6.1 Biomass Measurement by Optical Density

Whole broth was accurately diluted with deionised water to obtain an optical density reading of between 0.2 and 0.5 units at 550 nm. A Bausch and Lomb Spec 20 spectrophotometer was used. The biomass concentration was expressed in grams of biomass per litre of broth by correlating the optical density readings with dry cell weight measurements (Section 4.3.2.1).

4.2.6.2 Biomass Measurement by Dry Weight

10 mL of whole broth were vacuum filtered through a pre-weighed Millipore HA 0.45 µm filter using a Gelman Sciences Vacuum Filtration Manifold. The filter was washed with the same volume of deionised water and microwaved in a Micro-Mite Conair Cuisine microwave for 10 minutes, allowed to cool and reweighed to give the dry cell weight of the broth.

4.2.6.3 Glucose Assay

A Beckman Glucose Analyzer was used to measure the glucose concentration in the fermentation broth. Deionised water was used for dilution as required.

4.2.6.4 aFGF Crude HPLC Assay

Broth samples were prepared for aFGF analysis as follows. A known volume of broth (V) was spun for 15 minutes at 5600 rpm in a Beckman TJ-6 centrifuge. 0.5*V mL of 6M guanidine hydrochloride were added to the pellet and mixed at room temperature for 5 to 15 minutes. 0.5*V mL of 0.1% trifluoroacetic acid were added to the suspension and microcentrifuged in eppendorf tubes for 30 seconds. The supernatant was filtered through a 0.45 µm CR PTFE Gelman Acrodisc for HPLC analysis.

A Spectra-Physics HPLC system was used for the detection of aFGF in the prepared samples. The HPLC system consisted of a Spectra-Physics SP8800 ternary HPLC pump, SP8880 autosampler and SP4400 integrator, and a LDC/Milton Roy Spectro Monitor III detector. 20 µL of prepared sample were injected onto a Polymer Laboratories PLRP-S 8 300 A column (# 1512-1801) operated at 60°C. A gradient solvent system was employed: solvent A was 0.05% trifluoroacetic acid in HPLC grade water, solvent B was 60% acetonitrile and 0.05% trifluoroacetic acid in HPLC grade water. With a flow rate of

1 mL/min the mobile phase was initially 70% A and 30% B, after 0.3 minutes equal amounts of A and B were pumped through the column, at 9 minutes the proportion was 40:60 A:B, and at 9.1 minutes the original proportions were resumed, the elution was complete after 15 minutes. UV absorbance of the column effluent was monitored at 220 nm. An aFGF standard, supplied by MSDRL, was used to provide a calibration curve for conversion of the absorbance values of the fermentation samples to aFGF concentrations.

4.2.6.5 Spectroscopic Analysis

The absorption spectra of pre- and post-sterilisation fermentation broths were used to investigate the effects of sterilisation on the protein component of the broth (Section 4.1.2.1). The absorption spectra were obtained by scanning spectroscopy using a Hewlett Packard 8451A Diode Array Spectrophotometer. The broth samples were diluted 1:10 with deionised water and scanned at wavelengths from 190 nm to 820 nm.

4.2.7 Data Analysis Tools

Tools for the comparative analysis of fermentation data were described in the previous two chapters. These tools were used for the analysis of the twelve aFGF fermentations described above. The application of the comparative analysis tools was divided into two parts: preparation of the data and comparative reasoning as summarised in Figure 4.2 and described below.

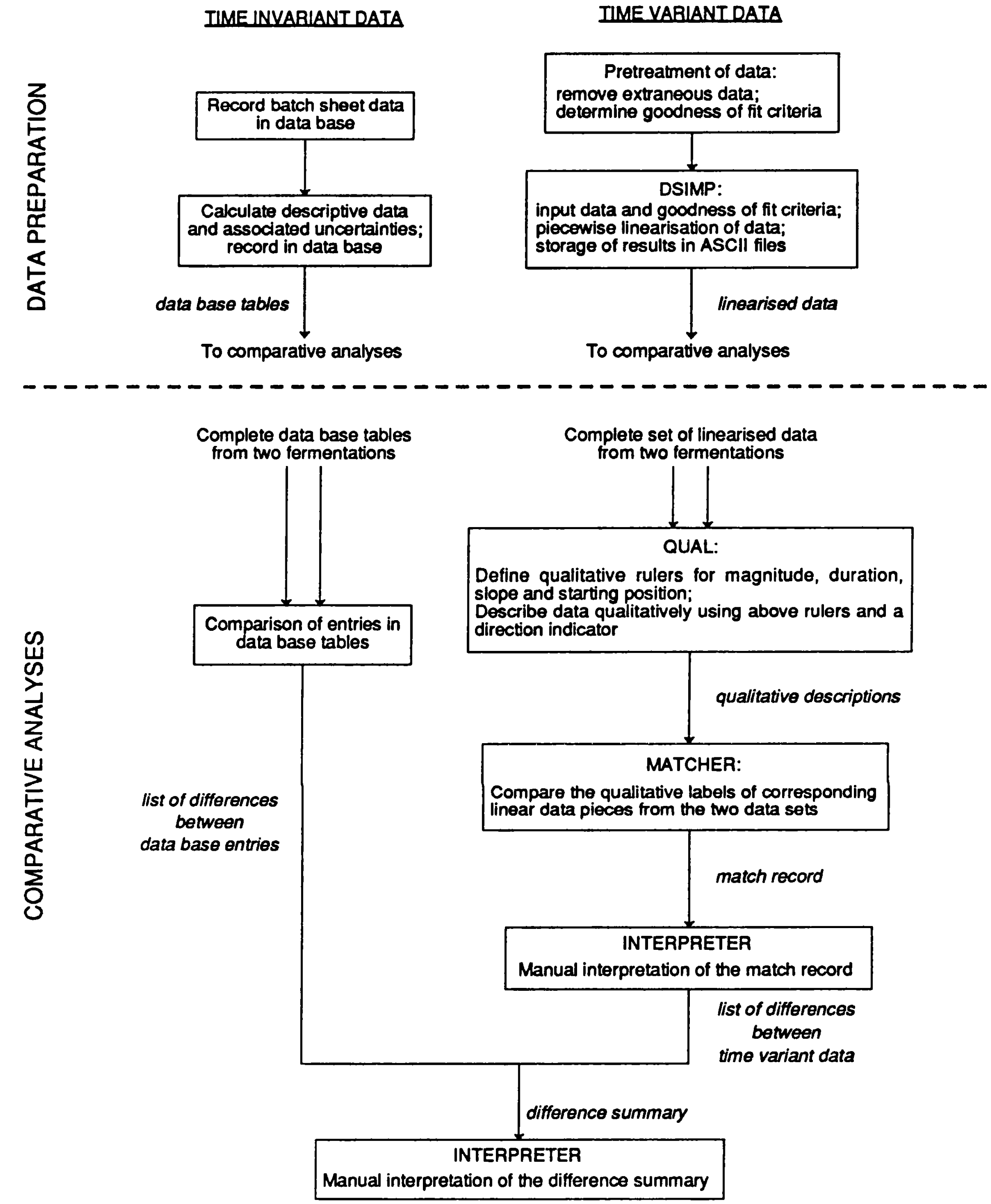

TIME INVARIANT DATA
TIME VARIANT DATA
DATA PREPARATION
Record batch sheet data in data base
Calculate descriptive data and associated uncertainties; record in data base
data base tables
To comparative analyses
Pretreatment of data: remove extraneous data; determine goodness of fit criteria
DSIMP: input data and goodness of fit criteria; piecewise linearisation of data; storage of results in ASCII files
linearised data
To comparative analyses
COMPARATIVE ANALYSES
Complete data base tables from two fermentations
Comparison of entries in data base tables
list of differences between data base entries
Complete set of linearised data from two fermentations
QUAL: Define qualitative rulers for magnitude, duration, slope and starting position; Describe data qualitatively using above rulers and a direction indicator
qualitative descriptions
MATCHER: Compare the qualitative labels of corresponding linear data pieces from the two data sets
match record
INTERPRETER Manual interpretation of the match record
list of differences between time variant data
difference summary
INTERPRETER Manual interpretation of the difference summary

Figure 4.2: Summary of the process by which the comparative analysis tools were applied to the aFGF data. The data from all twelve fermentations were prepared as described in the upper part of the diagram ('DATA PREPARATION'). The analysis of any two fermentations then proceeded through the 'COMPARATIVE ANALYSES' steps. DSIMP, QUAL and MATCHER are FORTRAN computer routines that were described in Chapter 3. The data base was described in Chapter 2. The terms in italics are the outputs from each step.

4.2.7.1 Data Preparation

The data from all twelve aFGF fermentations were treated in the same manner prior to the comparative analysis.

1. The time invariant batch sheet data and any comments on the experiments were recorded in the data base tables described in Chapter 2.

2. Descriptive data were calculated using the SmartWare II spreadsheet (version 1.02, Informix Software Inc., Menlo Park, CA) and included in the data base tables (Section 4.3.1.2).

3. On-line time variant data were 'cleaned up' by removing all data recorded prior to inoculation and subsequent to harvesting. This was carried out in the SmartWare II spreadsheet environment.

4. The first major step in the comparative analysis of time variant fermentation data was the simplification of each data set into piecewise linear segments using a FORTRAN program, DSIMP, which was described in Chapter 3. The inputs required for this routine were:

    - time, mean value, minimum value and maximum value at each sample point for each off-line data set;
    - the average standard deviation over all data sets for each off-line variable;
    - time and value at each sample point for each on-line data set;
    - the goodness of fit value for each on-line variable.

    Where necessary, the SmartWare II spreadsheet was used to calculate these values.

Each of these processes is described in detail in Section 4.3.

4.2.7.2 Comparative Reasoning

For any two fermentations being compared in this work the comparative analysis process was as shown in the second part of Figure 4.2. The simplified time variant data were described in qualitative terms and compared using two FORTRAN programs, QUAL and MATCHER respectively, described in Chapter 3. The results of this comparison and the comparison of the data base information were then combined in a *difference summary* which was interpreted manually.

A difference summary lists all the differences between the two fermentations being compared. The interpretation of each difference summary required the distinction between *causes* and *effects*. The analyst had to establish which data were indicative of performance changes within the fermenter, these are the *effects* which necessarily must have been a result of a perturbation to the system, ie a *cause*. The perturbation to the system may be obvious, such as a different fermenter, it may be intentional, such as a change in the length of sterilisation, or it may be unintentional, such as a faulty control variable. Links between causes and effects were made by careful analysis of the difference summaries. The data which were indicative of performance changes in the aFGF fermentations are listed in Table 4.4. A difference in any of these variables between two fermentations was indicative of a perturbation to the system.

TIME INVARIANT DATA INDICATIVE OF FERMENTATION PERFORMANCE	TIME VARIANT DATA INDICATIVE OF FERMENTATION PERFORMANCE
Harvest time	Biomass
aFGF titres	Glucose
	Carbon dioxide evolution rate
	Dissolved oxygen tension
	pH
	Agitation rate
	Air flow rate
	Alkali addition

Table 4.4: The data reflecting the dynamics of the aFGF fermentations. Any differences detected in these variables during the comparative analyses were indicative of variations in the performance of the fermentation.

4.2.8 Data Analysis Protocol

The analysis of the twelve aFGF fermentations was carried out after completion of the experiments. The analysis protocol was as follows:

1. the reproducibility of the aFGF fermentation was examined;
2. features in the time variant data were identified;
3. the specific research objectives were investigated.

4.2.8.1 Reproducibility

The aim of most fermentation research work is to investigate the behaviour of a process in different environments. As described in Chapter 1 this investigation involves a comparative analysis of data from experiments in which the microorganism is exposed to the different environments. For the comparative analysis of the data to be valid it is essential that the process is approximately reproducible: if fermentations run under identical conditions do not exhibit similar behaviour it is not possible to divorce the effects of enforced operating condition changes from those of unknown factors.

A particular fermentation process is considered to be reproducible if two fermentations, run under identical conditions, show no difference in performance when compared using the comparative analysis tools described in the previous section. The indicators of performance for the aFGF fermentations were listed in Table 4.4. The reproducibility of a particular fermentation process needs to be demonstrated only once.

The steps involved in examining the reproducibility of the fermentations were:

1. pairs of fermentations were selected, from the data base, such that the only intentional differences were the fermenters used and/or the inoculum and/or the date of operation;
2. the time invariant data and linearised time variant data from each pair of fermentations were compared using the comparative analysis tools as described in the lower part of Figure 4.2;
3. the results, ie difference summaries, were searched for evidence of reproducibility;

4. where differences in performance were found in these supposedly identical fermentations, attempts were made to explain the differences using the information in the difference summaries; conclusions were recorded in the data base comments fields and were utilised in the subsequent analyses.

The search for pairs of 'identical' fermentations must be described in more detail. In this work the pairs of fermentations that had been operated under supposedly identical conditions were identified by the analyst. These batches were then extracted from the data base tables using the batch number as the identifier. This was possible because of the small number of fermentations and because the experimental programme had been designed to provide repeated batches. In other situations it may be necessary to compare the batch sheet tables of the data base to find suitable fermentations. However, a well designed experimental programme will always provide repeats and pairs could be identified in the comments section of the appropriate records in the data base.

4.2.8.2 Features of the Data

The linearised time profiles produced by DSIMP, the linearisation routine, were arranged in two ways for analysis:

1. all profiles of one variable were grouped together and compared from batch to batch using the comparative analysis tools QUAL and MATCHER as described in the lower part of Figure 4.2;
2. all profiles of one fermentation were grouped together to identify features in the data: when the end points of the linear segments from different variables occurred at approximately the same time a correlation between the variables was noted thus improving understanding of the fermentation.

The fermentations that were used to identify features of the aFGF data were those that had been designated as 'identical' fermentations in the investigation of reproducibility. The identification of features of the data sets is described in more detail in Section 4.4.2.

4.2.8.3 The Research Objectives

The effects of the different sterilisation regimes and inoculum concentrations were assessed by comparing the data from the appropriate fermentations using the comparative analysis procedure described in Figure 4.2. The identified data features and information from the reproducibility investigations were used in the analysis.

4.3 Data Preparation

The first step in the analysis of the aFGF fermentations was the preparation of the data for comparative analysis. The batch sheet data and experimental observations were recorded in data base tables, descriptive data quantities were calculated and recorded in the data base tables, and the time variant data were simplified by piecewise linearisation after determining appropriate fitting criteria. These processes are described below.

4.3.1 Time Invariant Data

4.3.1.1 Batch Sheet Data

The batch sheet data, recorded for each batch, are summarised in the example in Table 4.5. Most of this information was provided prior to start-up of each batch. However, some data were dependent on occurrences during the run, thus the batch sheets required updating. For example, post-sterile volume and pH were added after sterilisation, while the duration of the fermentation and any changes to operating conditions were added after each run was completed.

Quantitative data generally have some inherent uncertainty in the measurement. For comparative purposes it is therefore not feasible to consider a single value as being representative of the measurement. As mentioned in Chapter 2, numerical data must be

FERMENTATION DETAILS

Batch Number:	C447
Organism:	*Escherichia coli*
Type:	Bacteria
Strain:	DH5
Plasmid:	PKK2.7
Organism id:	***
O_2 Requirement:	Aerobic
Mode of Operation:	Batch
Aim:	Production
Product:	aFGF
Number of Stages:	2
Seed Type:	Frozen Suspension

STAGE DETAILS

Stage:	1	2
Vessel:	2 L Erlenmeyer Flask	15 L Biolafitte BL4
Liquid Volume:	0.3 L	10 L
Start Date:	May 30, 1990	May 31, 1990
Start Time:	20:00 h	8:00 h
Length:	12 h	30 h
Inoculum Volume:	0.6 mL	100 mL
Medium:	MedM1	MedM2

MEDIUM DETAILS (MedM1 and MedM2)

MEDIUM COMPONENT	STERILISATION GROUP
Glucose	1
Heat Labile Components	2
Heat Reactive Components	1
Other	3

Note: MedM1 contains no inducer (lactose)

Table 4.5(a): Batch sheet information for aFGF fermentation (Batch C447) (continued in Table 4.5(b)).

STERILISATION DETAILS

GROUP	METHOD	DETAILS
1	Autoclave	123°C; 15 psig; 90 mins
2	Filter	0.22µm cellulose acetate
3	In situ steam sterilisation	122°C; 15 psig; 20 mins; V_o=7.5 L; V_f=7.6 L; pH_o=7.2; pH_f=7.0

INITIAL CONDITIONS

VARIABLE	TYPE OF CONTROL	SET POINT
Temperature	Auto	37°C
Pressure	Auto	5 psig
Air Flow	Cascade from DO	4.9 lpm
Agitation	Cascade from DO	375 rpm
pH	Auto	7.0
DO	Auto	> 20 % sat.

FEEDS

MATERIAL	CONC	METHOD	BASIS
NaOH	2M	On demand	pH < 7.0

Table 4.5(b): Batch sheet information for aFGF fermentation (Batch C447) (continued from Table 4.5(a)).

recorded in the data base as an interval covering the range of possible values for that particular quantity. When comparing the quantity with one from another data base record overlapping intervals indicate similarity. In this work uncertainty intervals were required for the broth volumes before and after sterilisation, the pH values of the broth before and after sterilisation, and the optical densities and glucose concentrations of the inocula prior to inoculation.

Broth volumes were read from a scale marked in 0.5 L increments on the sight glass of the fermenters. The measurements were assumed to be within 0.25 L of the true value, that is half of the smallest interval on the scale, as per engineering convention, and were recorded as an interval from the smallest possible value to the largest possible value. This can be done automatically by programming the data base management system to convert the single value input to a range according to a pre-recorded error convention.

Individual pH values before and after sterilisation were also recorded as a range, encompassing 0.05 pH units either side of the recorded value. The accuracy of a pH meter would normally be 0.01 to 0.02 pH units, however, after sterilisation the pH meter must be recalibrated which involves aseptically removing a sample of sterile broth, measuring the pH of the broth on a separate pH meter and adjusting the fermenter pH meter as appropriate; the resulting calibration, and any subsequent recalibrations, can only be approximate.

The inoculum data (optical density and glucose concentration) were recorded as a range from the minimum to the maximum of the measured values for each inoculum. This can be misleading as one extremely large or small measurement could distort the recorded value. However, the application of Chauvenet's criterion (Appendix 2) did not detect any outliers in the readings so this was not of great concern. These measurements were only very approximate: the measurements were taken up to an hour after inoculation and, even though stored on ice in the meantime, it is conceivable that further growth had taken place; also the inoculation flask may not have been well mixed during inoculation thus resulting in an unrepresentative sample.

4.3.1.1.1 Media Changes Caused by Sterilisation

As the aim of the experiments was initially to determine the effect of sterilisation conditions on the performance of the fermentation, the batch sheet data that show changes in the fermentation broth during sterilisation are highlighted here.

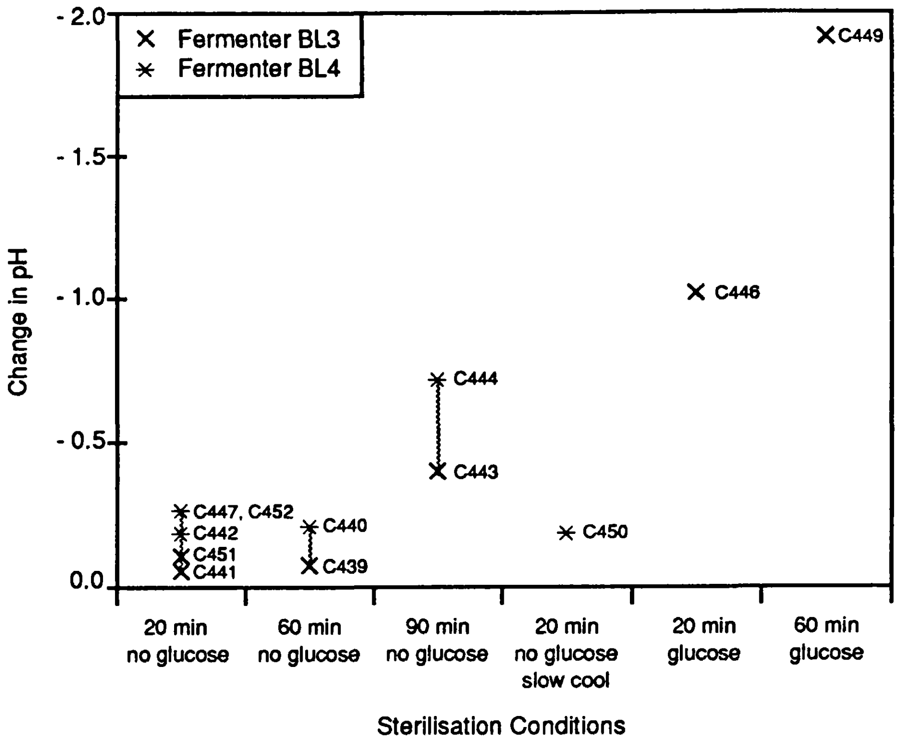


Figure 4.3: Changes in pH of aFGF fermentation broths over sterilisation. Sterilising for 90 minutes with no glucose present and sterilising with glucose in the bulk medium resulted in significant changes in the pH of the broth.

The most common indicator of medium changes during sterilisation is the broth pH. A summary of these results is presented in Figure 4.3. These data were recorded in the batch sheet with error bounds of 0.1 pH units either side of the calculated value (when adding or subtracting values their absolute errors are summed). It is evident from Figure 4.3 that, during sterilisation, the broths in fermenter BL3 consistently had less of a decrease in pH than the broths in BL4. Significant changes in broth pH were seen when the medium was sterilised for 90 minutes with no glucose present and when the medium was sterilised for 20 or 60 minutes with glucose *in situ*. These are discussed in more detail in Section 4.4.

The absorption spectra of pre- and post-sterile fermentation broths were investigated as an alternative indicator of the effects of sterilisation on the media (Section 4.1.2.1). These measurements were taken after completion of all the fermentations, using broth samples that had been stored at -18°C. Not all broths were analysed. The broths measured showed strong absorbance at wavelengths of approximately 210 nm, most likely as a result of the peptide bonds in the protein source, and 255 nm, as a result of the presence of either aromatic acids or disulphide bonds (Schmid 1989). Both the wavelength and absorbance of each absorption peak were recorded in the batch sheet; the accuracy of the readings would be greatly improved if they could be taken directly from a microprocessor attached to the spectrophotometer rather than reading the values from the hard copies of the spectra.

The absorption spectra of pre- and post-sterilisation broths of batch C447 (sterilised for 20 minutes with no glucose in the medium) are shown as an example in Figure 4.4. There was no change in the absorption spectrum of the broth during sterilisation. The results of all the absorption readings are summarised in Table 4.6 and the changes in absorption over sterilisation are summarised in Figure 4.5. Significant changes in absorption at both 210 nm and 255 nm were observed in the two batches sterilised with glucose in the bulk medium, these changes increased with increasing length of the heating process. These effects are discussed in more detail in Section 4.4.

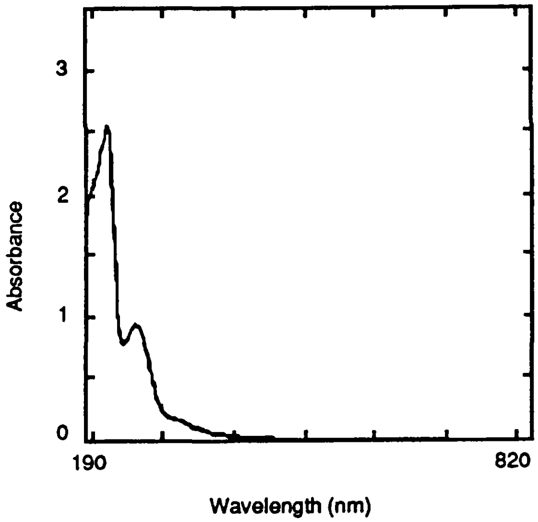
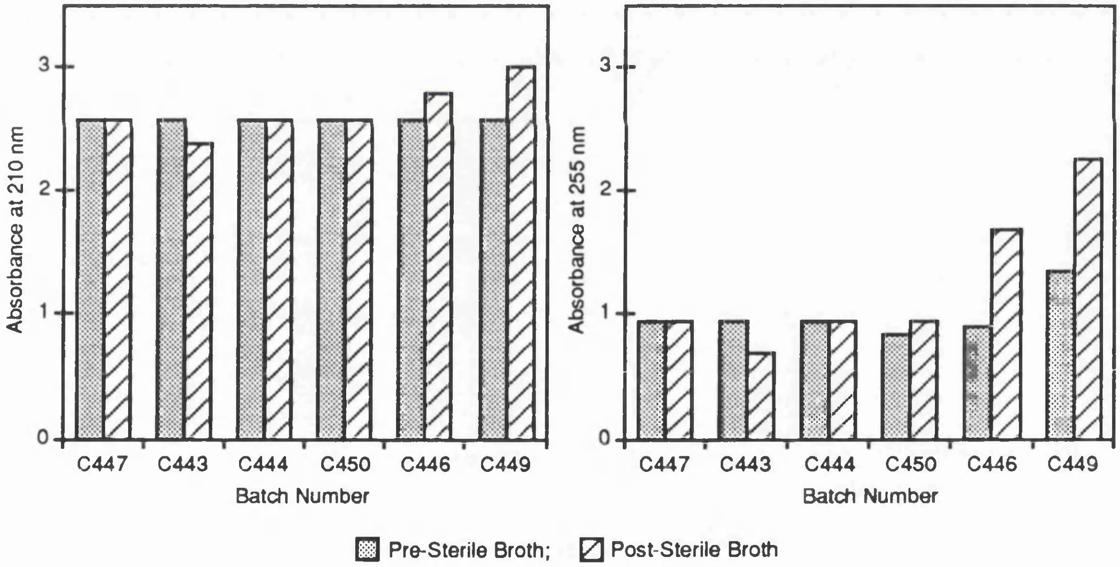


Figure 4.4: Absorption spectra of pre- and post-sterile fermentation broths of batch C447. The absorption spectrum of the broth did not change during sterilisation. A number of the pre-sterilisation broths exhibited the same spectrum, thus this was defined as the standard.

	PRE-STERILE ABSORBANCE		POST-STERILE ABSORBANCE	
BATCH NUMBER	at 210 nm	at 255 nm	at 210 nm	at 255 nm
C439	N/A	N/A	2.39	0.75
C440	N/A	N/A	2.39	0.75
C441	N/A	N/A	N/A	N/A
C442	N/A	N/A	2.42	0.61
C443	2.57	0.95	2.39	0.69
C444	2.57	0.95	2.57	0.95
C446	2.57	0.89	2.79	1.68
C447	2.57	0.95	2.57	0.95
C449	2.57	1.35	3.00	2.25
C450	2.57	0.84	2.57	0.95
C451	N/A	N/A	2.42	0.78
C452	N/A	N/A	2.57	0.89

Table 4.6: Summary of absorbance spectra for aFGF fermentation broths. The absorbance maxima occurred at wavelengths of approximately 210 nm and 255 nm. The changes that occurred during sterilisation are indicative of changes in the protein components of the media. (N/A = data not available).



Description of fermentations: C447: 20 minute sterilisation, no glucose present
C443: 90 minute sterilisation, no glucose present
C444: 90 minute sterilisation, no glucose present
C450: 20 minute sterilisation, slow cool, no glucose present
C446: 20 minute sterilisation, glucose present
C449: 60 minute sterilisation, glucose present

Figure 4.5: Changes in the absorbance of some of the fermentation broths over sterilisation. The wavelengths at which maximum absorption occurred were approximately 210 nm and 255 nm. The most significant changes occurred when glucose was present in the medium during sterilisation (C446 and C449). The protein component of batch C443 was diluted due to poor control of steam injection and this resulted in a decrease in the absorption levels at both wavelengths.

A number of the pre-sterile broths had identical absorption characteristics, this spectrum was thought to be most representative of the fermentation media prior to sterilisation and thus was identified as the standard (Figure 4.4).

The measurement of absorption spectra was found to be rapid and the results were very reproducible and extremely useful, as will be demonstrated later in this chapter.

4.3.1.2 Descriptive Data

The descriptive data consisted of the 'recommended harvest time' (t_h), the aFGF concentration at the recommended harvest time ($aFGF_h$), the maximum aFGF concentration ($aFGF_m$), the time at which this maximum aFGF concentration occurred (t_m). These values were defined and calculated as described in the following sections and were recorded in the data base tables. The correlation between optical density and dry cell weight was also considered to be a descriptive entity but is described in Section 4.3.2.1 with the optical density and dry cell weight data.

4.3.1.2.1 The Recommended Harvest Time

During aFGF production the broth is usually harvested before the glucose concentration has fallen to 5 $g.L^{-1}$, so as to avoid the problems that may occur when the microbial metabolism switches from utilisation of glucose to utilisation of other carbon and nitrogen sources such as amino acids. This point is termed the 'recommended harvest time' (t_h) and typically occurs about 15 hours after inoculation. Most of the fermentations were run for approximately 25 hours, allowing them to progress past the recommended harvest point. This was done to obtain as much information as possible from each fermentation but, as the metabolism of the organism changes after glucose exhaustion, data obtained after t_h have no direct bearing on the industrial aFGF production process.

During production the harvest point is predicted by extrapolating the glucose concentration profile, which is usually approximately linear approaching 5 $g.L^{-1}$. The linearisation of the glucose profiles using DSIMP roughly simulates the on-line extrapolation procedure and so was used to find the approximate harvest points by linear interpolation. The results are presented in Table 4.7 and were included in the data base tables.

The accuracy of the recommended harvest point could not be determined. It was thus recorded in the data base without the error bounds required for comparisons. However, the similarity of the recommended harvest points from separate batches could be obtained from comparison of the respective glucose profiles as described in Section 4.4.

4.3.1.2.2 Characterisation of aFGF Production

All aFGF titres were converted to normalised units.L^{-1} and normalised units per gram of dry cells (Appendix 3) for proprietary reasons. The results are presented as time profiles in Appendix 3. The crude HPLC analysis of aFGF was found to be very unreliable with relative errors of up to 60% of the mean aFGF value calculated. The HPLC technique only gave an approximation of the relative levels of aFGF.

The relatively infrequent aFGF measurements and the unreliability of the results precluded the use of the linearisation routine, DSIMP, on the product profiles. The remaining comparative analysis tools, QUAL and MATCHER, could also not be used on the aFGF profiles. aFGF production was thus described by way of the aFGF concentration at the recommended harvest point (aFGF$_h$) and the maximum aFGF concentration achieved during the fermentation (aFGF$_m$). The calculations are presented in Appendix 3. At each sample point the aFGF concentration was recorded as an interval incorporating an appropriate measure of the spread of the repeats (Appendix 3). The maximum aFGF value, for each batch, was taken as the aFGF range which contained the largest of the maximum values. The true maximum aFGF value could not be accurately determined because of the low sampling frequency. The harvest aFGF value was estimated at the recommended harvest point, ie the time at which glucose reached 5 g.L^{-1}. In these calculations it was assumed that the true aFGF profile, between any two sample points, was linear, making this an extremely approximate determination. The aFGF harvest values were calculated by linear interpolation to give an expected value and its associated uncertainty and then converted to an interval so as to facilitate comparisons (Appendix 3). The results of the aFGF calculations are summarised in Table 4.7 and were recorded in the data base tables.

a)

BATCH NUMBER	HARVEST TIME (t_h) (h)	aFGF @ t_h (aFGF$_{sh}$) (units.g^{-1})	aFGF @ t_h (aFGF$_{vh}$) (units.L^{-1})
C439	17.5	2.9 - 3.4	42.1 - 53.2
C440	14.9	1.7 - 2.5	22.8 - 31.7
C441	16.4	4.0 - 4.8	65.4 - 74.4
C442	14.7	4.1 - 4.9	61.1 - 72.3
C443	16.8	2.2 - 2.6	28.8 - 38.7
C444	15.4	2.5 - 3.2	37.9 - 48.4
C446	13.6	1.9 - 2.6	13.3 - 24.8
C447	14.8	3.3 - 3.8	48.9 - 58.5
C449	12.4	0.2 - 0.6	00.0 - 6.96
C450	14.7	1.4 - 1.6	17.9 - 27.2
C451	15.1	1.5 - 1.7	25.1 - 34.0
C452	16.0	2.4 - 3.0	34.8 - 47.4

b)

BATCH NUMBER	TIME OF MAX aFGF (h)	MAX aFGF (aFGF$_{sm}$) (units.g^{-1})	MAX aFGF (aFGF$_{vm}$) (units.L^{-1})
C439	25.0	3.1 - 3.7	54.6 - 67.2
C440	14.0	1.7 - 2.8	23.0 - 35.6
C441	18.0	5.0 - 6.1	90.6 - 103.2
C442	14.2	4.2 - 5.1	61.7 - 74.3
C443	10.0	2.2 - 2.8	29.7 - 42.3
C444	18.0	2.7 - 3.3	44.5 - 57.1
C446	18.0	2.1 - 2.6	22.3 - 34.9
C447	16.0	3.2 - 4.1	60.8 - 73.4
C449	19.9	1.0 - 1.6	9.7 - 22.3
C450	17.9	2.7 - 3.6	47.3 - 59.9
C451	18.0	2.2 - 2.8	44.6 - 57.2
C452	18.0	2.4 - 3.0	44.3 - 56.9

Table 4.7: aFGF harvest values (a) and maximum values (b) for each aFGF fermentation. The aFGF values are given in specific terms, ie normalised units per gram dry weight of cells, and in volumetric terms, ie normalised units per litre of broth.

4.3.1.3 Expert Comments

Comments made about the fermentations included such observations as the colour of the broth subsequent to sterilisation and the occurrence of any operational faults. These will be reported as necessary in the following discussion.

4.3.2 Off-Line Time Variant Data

The off-line measurements made during the aFGF experiments consisted of the optical density of the broth, the dry cell weight, the glucose concentration and the aFGF concentration (discussed in Section 4.3.1.2). These off-line measurements were investigated for the presence of outliers, their accuracies were determined, conversions to meaningful data were made and the goodness of fit criteria, for use in the linearisation routine (DSIMP), were calculated.

Chauvenet's criterion (Holman and Gajda 1978) did not detect any outliers in the off-line data from the laboratory scale experiments (Appendix 2). This does not indicate that the data were 'good' but rather that there was no evidence to assume that any of the readings, for any one sample, were significantly different.

A summary of the accuracy of each monitored off-line variable is given in Table 4.8 and the goodness of fit criteria used in the simplification routines are listed in Table 4.9.

4.3.2.1 Biomass Concentration

The biomass concentration was initially determined by measuring the dry cell weight of the broth samples. The relative error in these measurements was approximately 10% on values ranging from 0.4 $g.L^{-1}$ to 17.8 $g.L^{-1}$. Dry cell weight measurements were made less frequently than optical density readings because they were considerably more time consuming and generally less reproducible (the relative error in the optical density readings was 2%). The dry cell weight measurements were used to find the relationship between the

OFF-LINE VARIABLE	MIN VALUE	MAX VALUE	AVE STD DEV	AVE ABS ERROR	AVE REL ERROR	MAX REL ERROR
Optical Density	0.7	42.3	0.4	0.3	2%	8%
DCW (1) ($g.L^{-1}$)	0.4	17.8	0.6	0.4	10%	60%
DCW (2) ($g.L^{-1}$)	0.7	20.0	0.2	0.1	2%	20%
Glucose ($g.L^{-1}$)	0.2	34.1	0.5	0.4	2%	10%
aFGF ($units.L^{-1}$)	1.2	96.9	2.1	1.5	10%	60%
aFGF ($units.g^{-1}$)	0.3	5.5	N/A	0.3	20%	70%

Table 4.8: Accuracy of off-line monitored and derived variables.
(DCW (1): measured dry cell weight; DCW (2): dry cell weight from regression; std dev: standard deviation; rel error: relative error; abs error: absolute error)

VARIABLE	GOODNESS OF FIT (GOF) CRITERIA	GOF (MIN GOF)
Glucose	within bounds set by range (minimum = 3 * average std dev)	(1.5 g.L^{-1})
Dry Cell Weight	within bounds set by range (minimum = 3 * average std dev)	(0.5 g.L^{-1})
CER	4% of average maximum value	3.85 arb. units
pH	estimated measurement precision	0.05
DOT	estimated measurement precision	2 % air sat.
Air Flow Rate	estimated measurement precision	0.5 L.min^{-1}
Agitation Rate	estimated measurement precision	10 rpm
Alkali Addition	estimated measurement precision	0.5
Temperature (sterilisation)	estimated measurement precision	1 ^{o}C

Table 4.9: Goodness of fit criteria for the fermentation variables simplified using DSIMP.

optical density and dry cell weight so that optical density measurements alone could be used to determine the approximate biomass concentration in the broth.

Optical density readings are affected by the presence of coloured substances in the broth which cause background absorbance. It would be expected that, if the colour of the media changed, the optical density of the broth, for a given cell concentration, would also change. In the fermentations where glucose was sterilised *in situ* considerable colouring of the medium was observed and the longer the sterilisation the darker the colour. A rough plot of dry cell weight versus optical density for all the fermentations suggested the possibility of a difference in the correlation between optical density and dry cell weight for the batch that had undergone a sixty minute sterilisation with glucose *in situ* (C449). A statistical test was devised to investigate the possibility of a difference in correlations (Mendenhall and Sincich 1988) and is presented in detail in Appendix 4.

The conclusion of the statistical analysis was that the correlation between optical density and dry cell weight for batch C449 was significantly different from that of all the other aFGF fermentations. The equations describing the relationship between optical density and dry cell weight were found to be:

$$\hat{y} = -1.76 + 0.47x_1 \qquad \text{for C449} \qquad (4.1)$$

$$\hat{y} = 0.55 + 0.46x_1 \qquad \text{for the other fermentations} \qquad (4.2)$$

where $\hat{y}$ is the predicted value of the dry cell weight and x_1 is the optical density.

The correlations are shown in Figure 4.6 and were stored in the data base tables in the appropriate field.

Each optical density reading was converted to a dry cell weight measurement using the correlation equations. These dry cell weight values were averaged for each sample and used as the biomass concentration profiles. Determination of the accuracy of this calculated dry cell weight is described in Appendix 4 and summarised in Table 4.8.

Two factors were taken into consideration when specifying the goodness of fit criterion for the linearisation of the biomass concentration profiles. The lines should pass within the range covered by the minimum and maximum measured value at each sample point, unless this spread is smaller than three times the average standard deviation of all samples (Section 3.2.3.2.1). The determination of the range of values must take into consideration the uncertainties in the regression equation and the optical density readings; the calculations are presented in Appendix 4.

The biomass concentration data used in the linearisation routine, DSIMP, were thus the mean, minimum and maximum at each sample point and the average standard deviation.

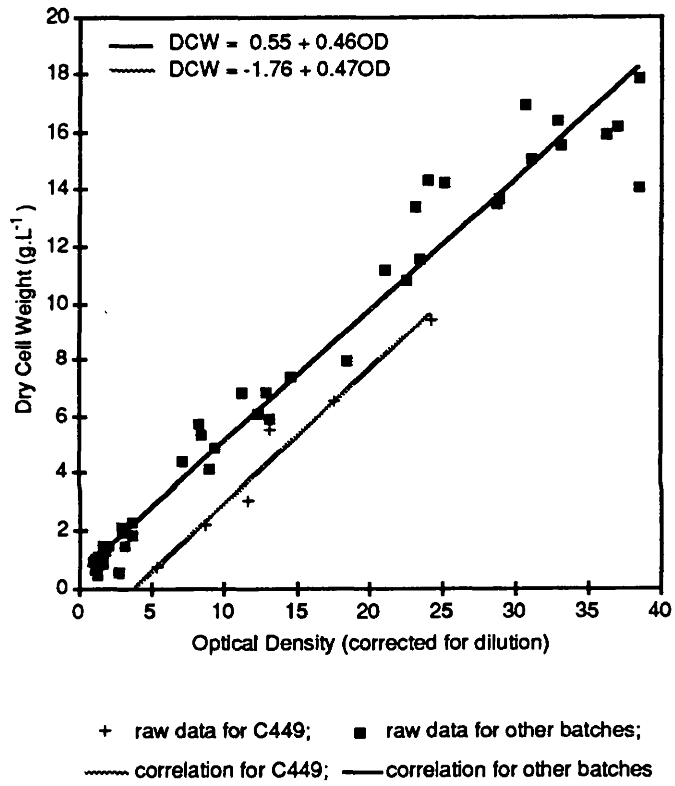


Figure 4.6: Linear correlations between optical density (OD) and dry cell weight (DCW) for aFGF fermentations. The medium in batch C449 was coloured due to a long period of sterilisation with glucose in the medium and thus the correlation was affected.

4.3.2.2 Glucose Concentration

The first fitting requirement in the simplification of glucose concentrations was that the lines should pass within the range covered by the measured values at each point. However, as described in Section 3.2.3.2.1, a minimum goodness of fit criterion was specified as three times the average standard deviation of the glucose measurements (Table 4.8).

The glucose concentration data used in the linearisation routine, DSIMP, were the mean, minimum and maximum at each sample point and the average standard deviation.

4.3.3 On-Line Time Variant Data

The monitored and derived on-line variables were listed previously in Table 4.2.

Before the on-line data could be used in the simplification routines it was necessary to remove any data that had been logged prior to inoculation, as indicated by the inoculation time in the batch sheets, or subsequent to harvesting, as indicated by the rapid changes in data values. In some fermentations the control signals from the previous fermentation were inadvertently not cleared from the computer prior to inoculation; any data recorded prior to the clearing of these signals were removed from the data file.

A number of comments about the on-line data were recorded both during operation and in the retrospective analysis as described below.

The data from batches C439 and C440 may not be very accurate as all instruments were recalibrated after these runs.

After completion of the work it was found that there was an error in the CER and OUR equations on the host computer. This error affected the scale of the values but not the shape of the time trajectories. The CER and OUR are thus presented in arbitrary units.

The oxygen uptake rate (OUR) data were found to be unsuitable for linearisation. OUR data, like carbon dioxide evolution rate (CER) data, are generally useful because they can be correlated with the physiology and metabolism of the microorganism (Omstead *et al.* 1990). However, when the sampling frequency is low relative to the dynamics of the process, as in these experiments, the relatively large noise in the OUR measurements masks the true dynamics and precludes the use of the OUR data in further analyses. Ongoing research (P.N.C. Royce, UCL, personal communication) has indicated that noise in mass spectrometry data is approximately proportional to the size of the measurement signal. The maximum exit gas concentration of carbon dioxide is approximately 3% whilst the inlet concentration is close to 0%. These two values are subtracted from each other in the carbon dioxide evolution rate (CER) calculations and the resulting noise is essentially that of the 3% signal reading. The inlet gas oxygen concentration is approximately 21% and the minimum outlet concentration is about 18%. The error associated with each of these values is approximately six to seven times that of the maximum carbon dioxide reading and, when the two oxygen values are subtracted from each other to give OUR, the uncertainties are additive giving an error of the order of fourteen times larger than the error in the CER values. In this work, even though the general shape of the oxygen uptake rate profiles was discernible to the human eye (Appendix 5, Figure A5.9), the data simplification routines were too sensitive to the large fluctuations that occurred and it was not possible to specify an adequate goodness of fit parameter for the linearisation routines. The goodness of fit suggested in Chapter 3, ie 4% of the maximum value, resulted in a 'spiky' linearisation, the lines followed the data too closely. The goodness of fit was increased to find a suitable value. A linearisation with no 'spikes' was obtained with a goodness of fit of nearly twice the initial value however the resulting fits were poor with large deviations between the raw data and the linear segments. Thus the simplification of the OUR data was deemed unacceptable. By increasing the sampling rate, relative to the dynamics of the process, it would be easier to extract the true trends of the process from the noise. This is shown in the OUR data from a mycelial secondary metabolite fermentation obtained from Merck Sharp and Dohme Research Laboratories (Figure 3.4(b)) where the sampling rate for mass spectrometry data was the same as in these experiments but the dynamics of the mycelial process were much slower, consequently it would be possible to simplify the OUR data and use it in further analyses.

As a result of the poor quality of the OUR data, respiratory quotient data (CER/OUR) were also unavailable.

The alkali addition data were also treated as suspect. During the initial comparative analyses (Section 4.4) using MATCHER (the computer routine which compares time variant data) the alkali addition data were often found to differ from batch to batch. On closer examination of the match records (produced by MATCHER) and the difference summaries it was found that the amount of alkali added to fermenter BL3 was consistently less than that added to fermenter BL4 (Appendix 5, Figure A5.6). The reasoning behind this was that the alkali addition pumps had not been calibrated prior to use and the conversion to a volume, utilised by the host computer, relied on this calibration. It was also not known whether the same pumps were used throughout the study or if the settings had been changed. This further justifies the need for strict recording of all information including pumps and pump settings. In the comparative analyses presented in Section 4.4, alkali addition was only recorded in the difference summaries if the match record showed a difference in the temporal extents of the linear data pieces; differences in the magnitudes, slopes and starting positions were ignored.

This work was carried out away from the site of the experimental research thus the instrument precision data, required for the specification of the goodness of fit values, were not available for dissolved oxygen, pH, alkali addition, air flow rate and agitation rate. Values were chosen based on a general knowledge of the approximate accuracies that should be expected from these instruments when operating in a fermentation environment. The goodness of fit values are listed in Table 4.9.

Temperature and pressure were controlled at a constant level throughout the fermentations, and thus the only information required about these data was whether or not the control was adequate. This information was recorded by the operator for inclusion in the data base, based on observations from the control system during the process.

4.3.4 Linearisation of Time Variant Data

The FORTRAN routine DSIMP, described in Chapter 3, was used to linearise the time variant data. The results are shown in Appendix 5 and are reproduced in the text where necessary to clarify discussions.

4.4 Comparative Reasoning (Analysis and Discussion of Results)

In the previous section the aFGF data were prepared for comparative analysis. The results of these preparations were data base tables containing all the time invariant data and linearised time profiles of the time variant data. This section describes the analysis of these data using the protocol described in Section 4.2.8: the reproducibility of the aFGF fermentation was examined, features of the data were identified and the effects of sterilisation conditions and inoculum concentration were investigated. The analyses used the tools described in Chapters 2 and 3: the data base tables were used to compare time invariant data; the FORTRAN routines QUAL and MATCHER were used to compare time variant data; the results of the comparisons were summarised in difference summaries (Section 4.2.7.2) and interpreted manually.

As mentioned in Section 4.3.3, some data were not available for comparative analysis: oxygen uptake rate data, respiratory quotient data, and the magnitudes, slopes and starting positions of the alkali addition data.

The comparison of the recommended harvest times was not straight forward and was carried out manually. The comparison of quantitative data in the data base requires the consideration of error bounds (Section 2.2.4); however, the recommended harvest point, calculated from the glucose concentration profiles (Section 4.3.1.2.1), was recorded as a single value without error bounds. Information on the similarity of the harvest points was extracted from the comparison of the glucose time profiles: if two glucose profiles were found to be similar, the recommended harvest point of the two batches was considered to be similar; if two glucose profiles were different the qualitative values of the magnitudes and durations were summed separately, up to and including the line covering the harvest point, if the resulting sums for both profiles were equal then the harvest points were considered to be the same otherwise it was concluded that the two batches would have been harvested at different times. This procedure was carried out manually. Summing the qualitative values for magnitude and duration is, in effect, the same as overlaying the profiles, ignoring what happens either side of the point of interest, and determining if the glucose concentrations approach the harvest level at approximately similar times. The data base management system must contain an option to exclude fields, such as the recommended harvest time, from the standard data base comparisons.

The aFGF values were not very accurate and thus differences detected when comparing these values, whilst investigated, were treated with caution.

In the following sections the difference summaries for all the comparative analyses are presented and discussed. The difference summaries do not provide 'answers' but indicate which fields in the data base tables to investigate further and which match records contain important information. The match records and qualitative descriptions have been presented along with the simplified time profiles to illustrate some of the results obtained from the comparative analyses of the time variant data. The profiles were not required for the actual analysis as all the necessary information was contained in the qualitative descriptions.

4.4.1 Reproducibility of the Fermentations

A fermentation must be shown to be approximately reproducible prior to an investigation into the effects of enforced operating condition changes: if a fermentation does not perform consistently under identical conditions it is then not possible to analyse the true effects of intentional changes in the fermentation environment. The information contained in this section, whilst not directly relevant to the investigation of the effect of sterilisation conditions and inoculum concentration on the aFGF fermentation, provides the justification for eliminating some of the fermentations from the study and for treating the data of other fermentations with caution. The investigation is reported in detail to demonstrate the utilisation of the computerised comparative analysis tools. A summary of the findings is presented in Section 4.4.1.6.

The fermentations used in the reproducibility investigation were:

- C439/C440 sterilised for sixty minutes with glucose sterilised separately; the media were prepared simultaneously and the two batches were seeded from the same inoculum; the fermentations were run concurrently under the same operating conditions; different vessels were used.

- C441/C442/C447/C451 sterilised for twenty minutes with glucose sterilised separately; the fermentations were run under the same operating conditions; the media for C441 and C442 were prepared simultaneously, these two batches were seeded from the same inoculum and were run concurrently; batches C442 and C447 were fermented in the same vessel (BL4), C441 and C451 were fermented in the other vessel (BL3).

- C443/C444 sterilised for ninety minutes with glucose sterilised separately; the media were prepared simultaneously and the two batches were seeded from the same inoculum; the fermentations were run concurrently under the same operating conditions; different vessels were used.

The results of comparing the data base tables and the simplified time variant data of each pair of fermentations are summarised in the difference summary in Table 4.10 (the key to this table, and all subsequent difference summaries, is given in Table 4.11) and the interpretation of these results is presented below.

The criterion for reproducibility is that the comparative analysis tools detect no difference in performance between two fermentations operated under identical conditions. The reproducibility of a fermentation process needs to be demonstrated only once. The comparison of batches C447 and C451 provided evidence of the reproducibility of the aFGF fermentation (Table 4.10): the MATCHER routine detected no differences in the time variant data and the only performance differences reported from the data base comparison were the aFGF values which were earlier stated to be very approximate and thus of little value in assessing the performance of the fermentations. (The differences in the aFGF values are discussed later). The conclusions of the comparative analysis tools were in agreement with visual inspection of the time variant data from these two fermentations (Figure 4.7) thus demonstrating the efficacy of the automated techniques.

Eight pairs of fermentations were expected to behave similarly but only C447/C451 demonstrated this reproducibility. Visual comparison of the time profiles of C443/C444 (Figure 4.8), C441/C447 and C441/C451 indicated approximate similarity, however, the comparative analysis tools detected a number of significant differences between these batches (Table 4.10). During visual inspection of the time profiles it was assumed that the uncertainties in the data were sufficiently large to account for the slight differences in the profiles. The computerised tools were superior to visual analysis in these situations because the software made use of all available quantitative data, ensuring that the variation in each

BATCH NUMBERS	DIFFERENCES IN TIME INVARIANT DATA	DIFFERENCES IN TIME VARIANT DATA	OBSERVATIONS
C439/C440	Tanks Post-sterile pH $\Delta V_{sterilisation}$ Harvest Time $aFGF_{vh}$, $aFGF_{vm}$ $aFGF_{sh}$, $aFGF_{sm}$	Sterilisation Biomass Glucose CER DOT pH Agitation	Medium dilution in both No pre-sterile abs. spec. in C439 and C440 Temperature fault C440 Foaming C440
C441/C442	Tanks $aFGF_{vm}$	Sterilisation CER DOT pH Agitation	Temperature fault C442 DOT fault in both Foaming C442 No abs. spec. C441
C443/C444	Tanks Post-sterile pH $\Delta pH_{sterilisation}$ $\Delta V_{sterilisation}$ Post-sterile abs. spec. $aFGF_{vm}$	Biomass DOT pH Agitation	Medium dilution C443
C441/C447	Tanks Inoculum Post-sterile pH $\Delta V_{sterilisation}$ $aFGF_{vh}$, $aFGF_{vm}$ $aFGF_{sh}$, $aFGF_{sm}$	Sterilisation Glucose CER DOT pH Agitation	DOT fault C441 No abs. spec. C441
C442/C447	Inoculum $\Delta V_{sterilisation}$ Post-sterile abs. spec. $aFGF_{vh}$ $aFGF_{sh}$, $aFGF_{sm}$	Sterilisation Glucose CER DOT pH Agitation	No pre-sterile abs.spec. in C442 Temperature fault C442 DOT fault C442 Foaming C442

Table 4.10(a): Difference summary of fermentations intended to be operated under identical conditions (continued in Table 4.10 (b)). The abbreviations are described in Table 4.11.

BATCH NUMBERS	DIFFERENCES IN TIME INVARIANT DATA	DIFFERENCES IN TIME VARIANT DATA	OBSERVATIONS
C447/C451	Tanks Inoculum $\Delta V_{sterilisation}$ Post-sterile abs. spec. $aFGF_{sh}$, $aFGF_{sm}$ $aFGF_{vh}$, $aFGF_{vm}$		No pre-sterile abs.spec. in C451 C451 shorter No off-line data first 10 h of C451
C441/C451	Inoculum $\Delta V_{sterilisation}$ $aFGF_{sh}$, $aFGF_{sm}$ $aFGF_{vh}$, $aFGF_{vm}$	Sterilisation CER DOT Agitation	No abs. spec. C441 DOT fault C441 C451 shorter No off-line data first 10 h of C451
C442/C451	Tanks Inoculum $\Delta V_{sterilisation}$ Post-sterile abs. spec. $aFGF_{sh}$, $aFGF_{sm}$ $aFGF_{vh}$, $aFGF_{vm}$	Sterilisation CER DOT pH Agitation	No pre-sterile abs.spec. in C442 and C451 Temperature fault C442 DOT fault C442 C451 shorter No off-line data first 10 h of C451

Table 4.10(b): Difference summary of fermentations intended to be operated under identical conditions (continued from Table 4.10 (a)). C447/C451 demonstrates the reproducibility of the aFGF fermentation. The abbreviations are described in Table 4.11.

TABLE ENTRY	EXPLANATION
abs.spec.	Absorption spectrum
$aFGF_{sh}$	The specific aFGF concentration at the harvest time ($units.g^{-1}$)
$aFGF_{sm}$	The maximum specific aFGF concentration ($units.g^{-1}$)
$aFGF_{vh}$	The volumetric aFGF concentration at the harvest time ($units.L^{-1}$)
$aFGF_{vm}$	The maximum volumetric aFGF concentration ($units.L^{-1}$)
Batched medium	The media components sterilised in the fermenter by direct steam injection
Harvest Time	Time at which glucose reached 5 $g.L^{-1}$, the recommended harvest time
Inoculum	Different source vials of inoculum were used for fermentations carried out on different days
Medium dilution	The bulk medium components were diluted by excessive condensate accumulation during sterilisation
$\Delta pH_{sterilisation}$	Change in broth pH over sterilisation
Tanks	Two fermenters were used in the experiments: BL3 and BL4
Variable (eg pH)	The match record indicated differences in the time profiles of this variable
$\Delta V_{sterilisation}$	Change in broth volume over sterilisation
$V_{0,sterilisation}$	Broth volume prior to sterilisation

Table 4.11: Key to entries and abbreviations in difference summaries (Tables 4.10 and 4.13 to 4.17). Other explanations are given in the text.

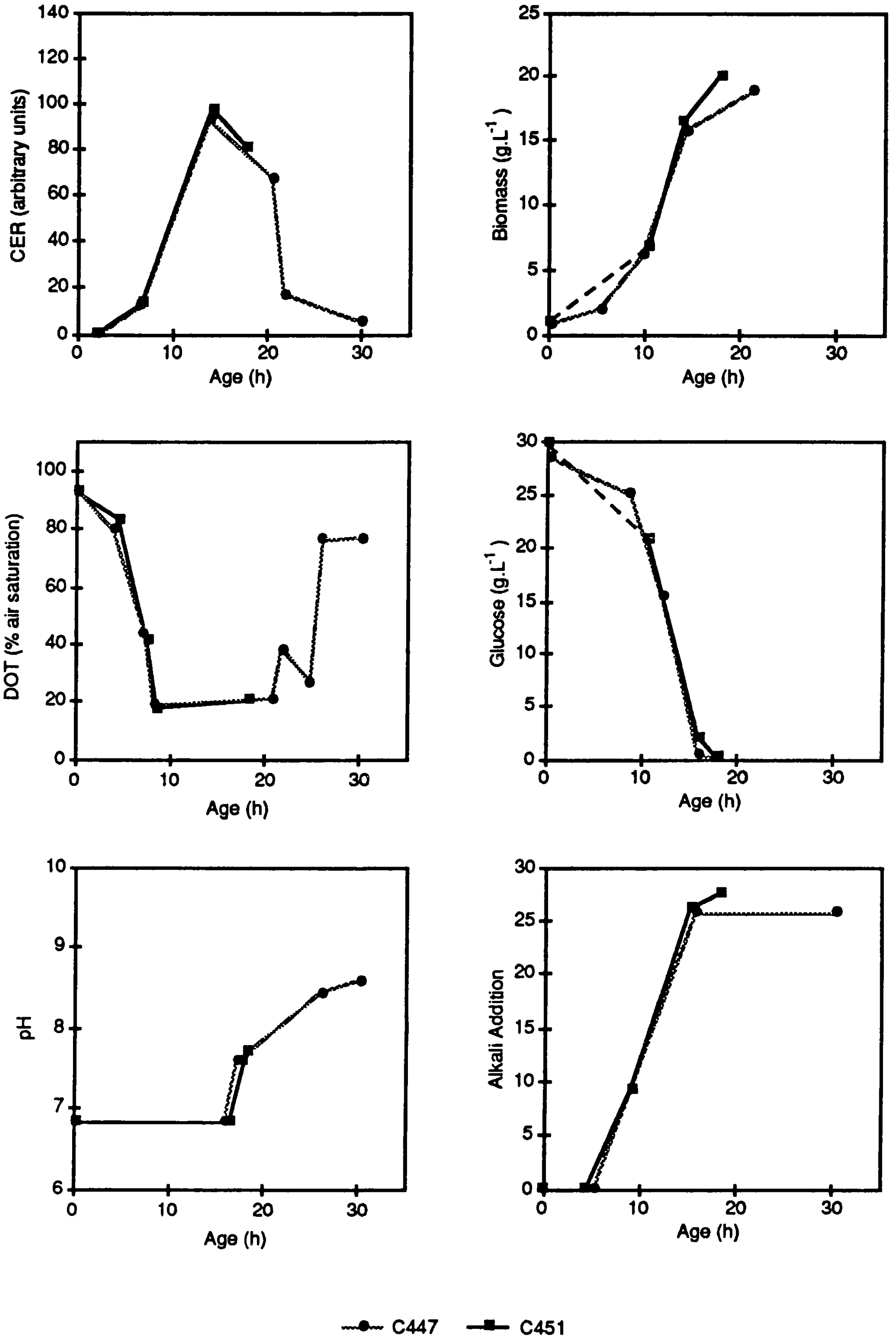


Figure 4.7: Simplified time variant data from batches C447 and C451 illustrating the reproducibility of the aFGF fermentations. (There were no data available for the first 10 h of C451 glucose and biomass profiles, the dashed lines are an extrapolation back to the time zero concentrations).

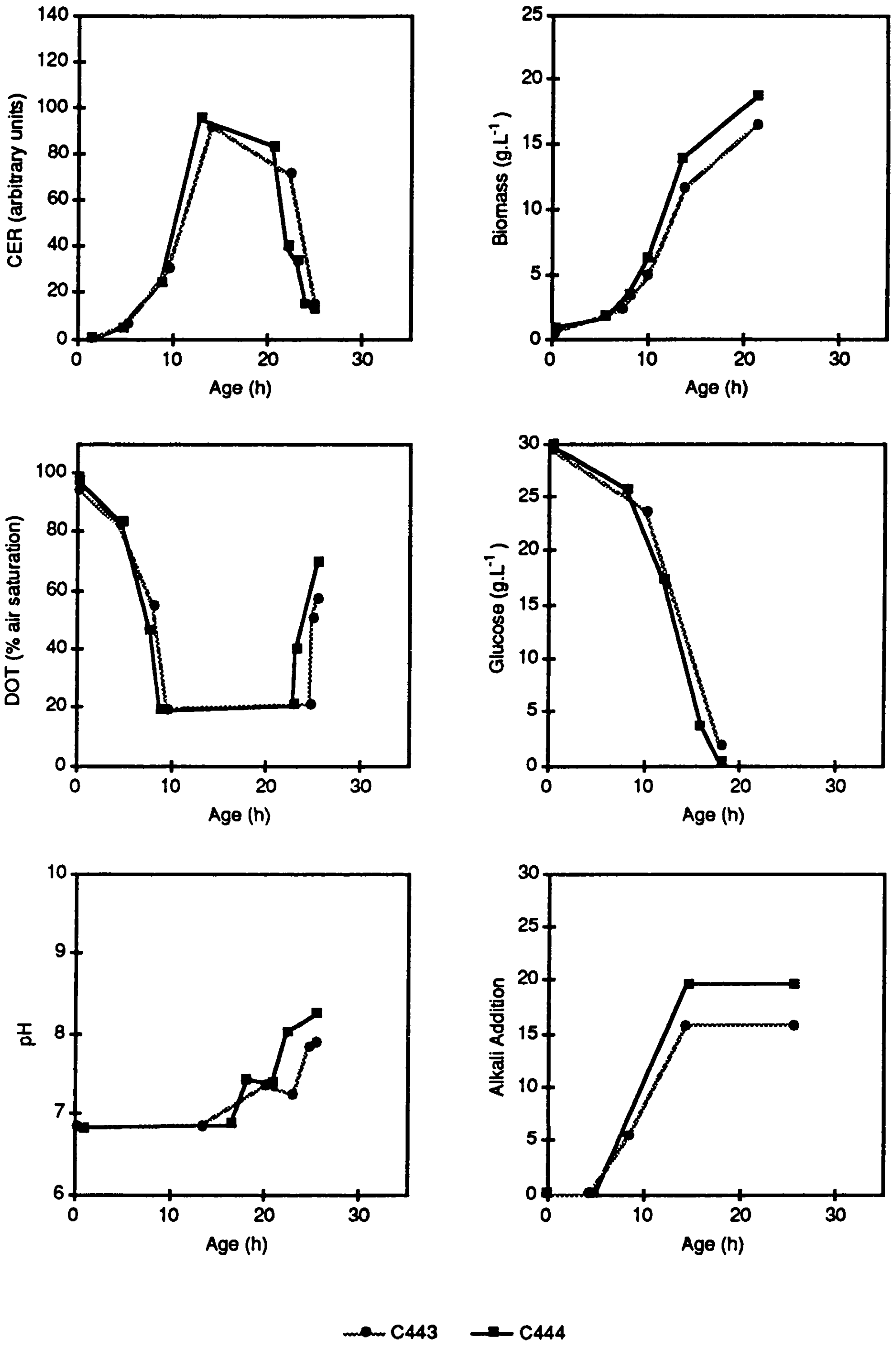


Figure 4.8: Simplified time variant data from batches C443 and C444. Visual comparison of these profiles had assumed that there was very little difference between the two fermentations but the comparative analysis tools detected significant differences which were explained by the dilution of medium components that occurred in batch C443.

measurement was taken into account and thereby concluded that the differences in the time profiles were significant. Additionally, the results of the computerised analysis were not prejudiced by prior expectations: even though these pairs of fermentations were expected to be the same, small variations were recognised and recorded. During the interpretation of the results the researcher may discount these variations as being unimportant but, in this case, if these variations had been detected whilst the experiments were still in progress further investigations would have been warranted.

Comparison of batches C447 and C451 demonstrated the reproducibility of the aFGF fermentation. The variations in the other fermentations that were expected to demonstrate reproducibility had to be explained before the data from these fermentations could be used in any further analyses. The effects of the perturbations observed during the reproducibility investigations are discussed in the following sections.

The data from batches C439 and C440 were treated with caution in the following analyses. The comparison of these two batches highlighted a number of differences in performance (Table 4.10); both media dilution (Section 4.4.1.2) and faulty temperature control (Section 4.4.1.4) were possible causes. However, the accuracy of the data from these two batches was questionable as the fermenter instruments were not calibrated until after their completion.

4.4.1.1 Fermenters BL3 and BL4

The similarity of fermentations C447/C451 show that the two fermenters, BL3 and BL4, performed comparably during the course of the experiments.

The media sterilised in fermenter BL4 consistently underwent a larger change in pH during sterilisation than the media sterilised in BL3 (Figure 4.3). The reason for this is unknown but there was no apparent effect on the fermentations as shown by the similarity of C447/C451.

4.4.1.2 Media Changes During Sterilisation

When direct steam injection is used for sterilisation it is inevitable that there is an increase in volume during the sterilisation process due to condensation of the steam. In these experiments the broth volume, prior to sterilisation, was set so as to allow for this volume increase and not dilute out the medium components.

Differences in the amount of condensate accumulated ($\Delta V_{sterilisation}$) were detected in the comparison of the batch sheet data from a number of the fermentations and recorded in the reproducibility difference summary (Table 4.10). This batch to batch variation was due to poor adjustment of the valve controlling the flow of steam to the broth (Section 4.2.3). No analysis of the incoming steam was undertaken so it was not known whether any impurities were being added during sterilisation. It is possible that the large amount of condensate accumulated in batch C451 (1.15 L), relative to that accumulated in batch C447 (0.1 L), was responsible for the difference in the absorption spectra of the post-sterile media of these two batches (Figure 4.9). However, without the analysis of the pre-sterile medium of C451 the possibility of an altered medium composition cannot be eliminated. The comparison of batches C447 and C451 (Table 4.10) showed that the difference in the post-sterile media had

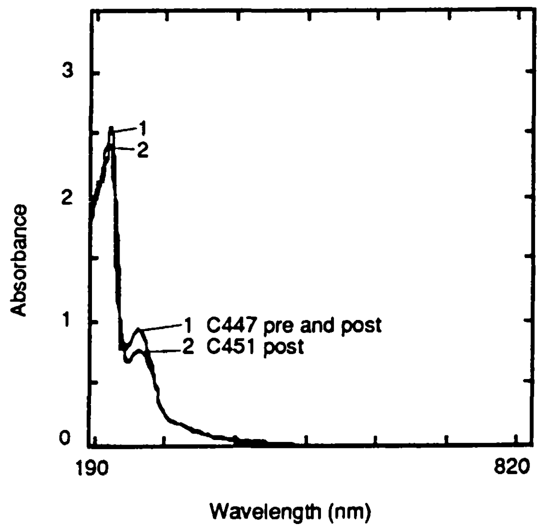


Figure 4.9: Absorption spectra of pre- and post-sterile fermentation broths of batch C447 and post-sterile broth of C451. There was no pre-sterile spectrum for C451. The lower absorption maxima in C451 may have been the result of an incorrect broth composition or of changes occurring during sterilisation.

very little effect on the performance of the fermentations. It is possible that the recorded difference in the product titres of the two batches was a result of the different post-sterile broth compositions but, with the poor aFGF data, this could not be verified. If it were found that a medium exhibiting a low absorbance resulted in low productivity then, based on this evidence, a production batch could be abandoned, or the required adjustments to the media made, prior to inoculation. This demonstrates the potential of scanning spectroscopy as a tool for fault detection in a fermentation facility.

It proved difficult to control the amount of steam introduced to the vessel and, in three fermentations, the volume increase was greater than anticipated and some media had to be removed after sterilisation to achieve the desired working volume. As a result the media components present during sterilisation were diluted: C439 suffered a 5% dilution, C440 an 8% dilution and C443 a 15% dilution. The post-sterile additions to the media, such as glucose and lactose, were not diluted. The extent of the dilution is illustrated in the respective absorption spectra of the post-sterile broths (Figure 4.10). It is not known whether the change in the absorption spectra was due solely to the diluted medium components or if impurities in the condensate contributed.

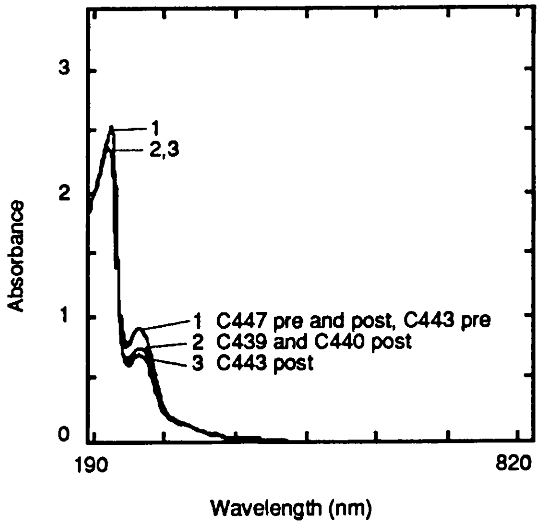
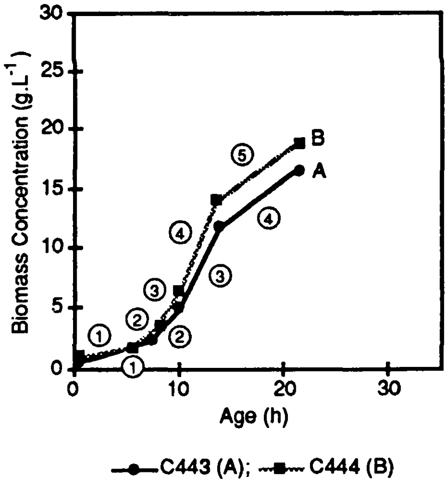


Figure 4.10: Absorption spectra of pre- and post-sterile fermentation broths of batches C447 and C443 and post-sterile broths of C439 and C440. The media in C439, C440 and C443 had been diluted during sterilisation, this is reflected in the lower absorption maxima.

In the reproducibility investigations dilution of the broth affected the comparisons of batches C439/C440 and C443/C444. The effects of dilution on batches C439 and C440 could not be isolated from the effect of the temperature fault in C440: either of these perturbations could have resulted in any of the performance differences in these two batches. However, as fermentation C443 suffered the largest amount of dilution, any effects resulting from dilution would be more evident in this fermentation than in the others. The differences between C443 and C444 were the changes in the medium as a result of dilution during sterilisation, the biomass, DOT, pH and agitation rate profiles and the maximum volumetric aFGF concentrations (Table 4.10). These differences can be explained with reference to the dilution of the broth in C443.

The effect of dilution on the absorption spectrum of the broth was described above. The change in pH of the broth during sterilisation was significantly less in C443 than in C444, however, it was noted earlier that the different fermenters had some influence on this.

The comparison of the biomass data of batches C443 and C444 is shown in Figure 4.11.



Qualitative Descriptions

Line	Dir	Mag	Dur	Slope	Start
A1	1	1	6	2	1
A2	1	2	3	5	1
A3	1	4	4	6	3
A4	1	3	7	4	6
B1	1	1	5	1	1
B2	1	1	3	4	1
B3	1	2	2	6	2
B4	1	4	3	6	3
B5	1	3	7	4	7
B3+B4	1	6	5	6	2

Match Record

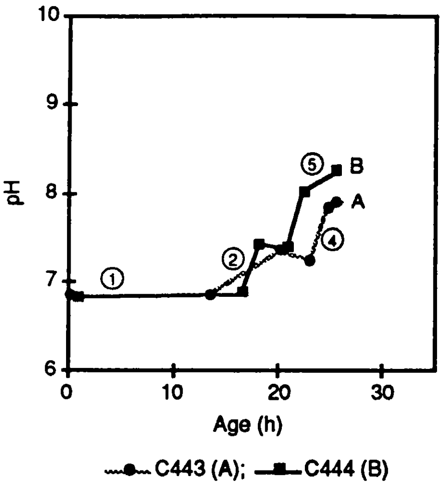
Lines	Magnitude	Duration	Slope	Start
A1 → B1	1	1	1	1
A2 → B2	1	1	1	1
A3 → B3+B4	0	1	1	1
A4 → B5	1	1	1	1

Figure 4.11 : Comparison of biomass data from batches C443 and C444. Dilution of the medium components in C443 resulted in a lower biomass yield as shown by the different magnitudes in the match record.

The change in magnitude of the lines representing the period of fastest growth was smaller in C443 than in C444. It was concluded that the loss of complex broth components caused by dilution of the broth in C443 was detrimental to biomass production. The growth rates of the two fermentations were the same, as shown by the similarity in the slopes of all the linear data pieces in the qualitative descriptions of the profiles (Figure 4.11). There was no evidence that the specific production rate had been affected by the dilution.

The comparison of the pH profiles of batches C443 and C444 detected a difference in the temporal extent of the lines representing the control period: the time between inoculation and the onset of increasing pH was shorter in C443 than C444 (Figure 4.12). During the interval that pH is held at a constant level the organism is metabolising the glucose resulting in acidic products being released from the cells, the addition of alkali maintains a neutral pH in the broth. It would normally be expected that the end of glucose metabolism would correspond with the end of pH control, this is shown to be the norm in Section 4.4.2. However, in both C443 and C444 pH control ended before the exhaustion of glucose (Figure 4.13) indicating that some other nutrient was limiting the growth (this is discussed further in Section 4.4.3.1). This growth limiting substrate was in shorter supply in C443 than in C444, as indicated by the earlier end of pH control in C443, implying that the substrate was present in the bulk media during sterilisation and was diluted in C443. The difference in the pH control periods was not detected in the comparison of the alkali addition profiles but was still thought to be a real effect: the greater number of lines in the pH profiles means that the duration descriptor is more sensitive than in the alkali addition profiles. The difference is obviously small but the detection of it was important to the analysis of the effect of dilution in the media. Visual inspection of the pH profiles, prior to the availability of the comparative analysis tools, had not detected the difference in the length of pH control in these two fermentations and had not observed the lack of correlation between the end of pH control and the depletion of glucose. Understanding of the process was enhanced by the use of the comparative analysis tools.

The differing post-control pH behaviours in C443 and C444 (Figure 4.12) concur with the expected behaviour in a medium in which the levels of secondary nutrients have been altered. Upon depletion of the simple carbohydrates the organism's metabolism switches to utilise other carbon and nitrogen sources such as organic acids or the amino acids of any proteins in the medium, resulting in an increase in pH and cessation of pH control. In batch C443 there was a lower concentration of secondary nutrients because of the dilution that occurred during sterilisation and it would therefore be expected that the behaviour after glucose exhaustion would be different.



Qualitative Descriptions

Line	Dir	Mag	Dur	Slope	Start
A1	0	1	11	1	1
A2	1	4	6	6	1
A3	-1	1	3	4	4
A4	1	4	2	8	3
A5	1	1	1	5	7
B1	0	1	13	1	1
B2	1	4	2	8	1
B4	-1	1	3	3	4
B5	1	5	2	8	4
B6	1	2	3	6	8

Match Record

Lines	Magnitude	Duration	Slope	Start
A1 → B1	1	0	1	1
A2 → B2	1	0	0	1
A3 → B3	1	1	0	1
A4 → B4	1	1	1	1
A5 → B5	1	0	1	1

Figure 4.12 : Comparison of pH data from batches C443 and C444. The control period was longer in C444 (duration of first lines). The post-control profiles were initially different as shown by the poor matching in the match record. These differences were a result of the diluted medium in batch C443.

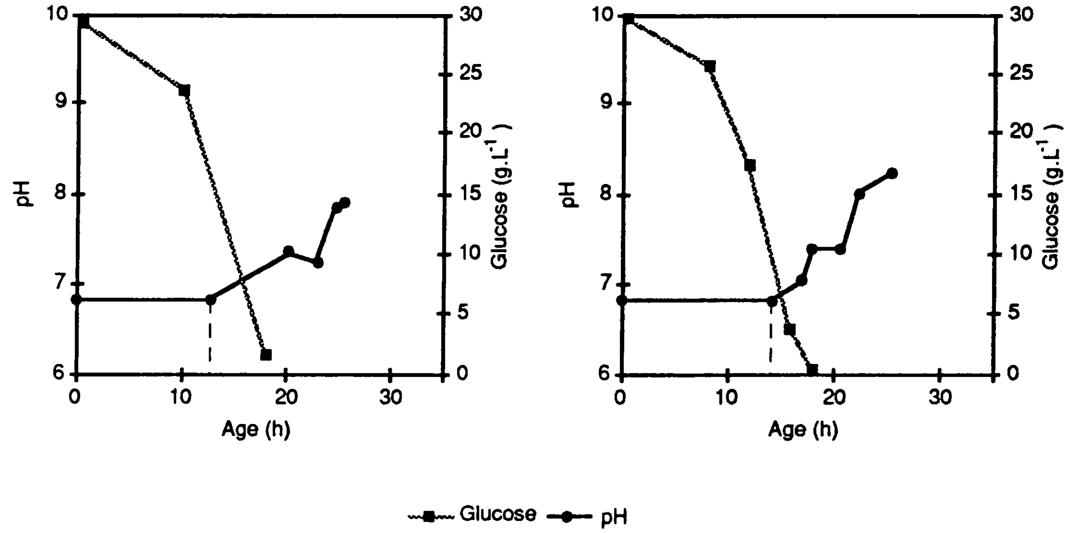
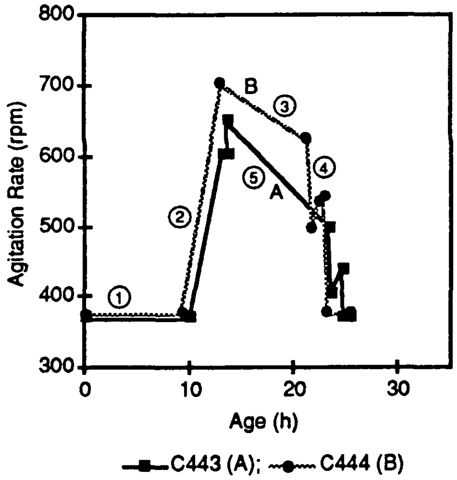


Figure 4.13: The end of pH control did not correspond with glucose exhaustion in batches C443 and C444 which underwent prolonged sterilisations. This implied that a nutrient, other than glucose, was limiting the growth. The availability of this essential nutrient had been reduced by the excessive heat stress.

In the aFGF experiments a high agitation rate was employed when required to maintain the dissolved oxygen level above 20% of air saturation; as the metabolic activity of the organism decreases, due to exhaustion of readily assimilated nutrients, the demand for oxygen decreases and the agitation rate can be lowered. A number of differences in the time courses of the agitation rate data of batches C443 and C444 were indicated in the match record of these profiles (Figure 4.14). In C443 the agitation rate was lowered more rapidly than in C444 implying that the supply of secondary nutrients was lower in C443 and thus the metabolic activity ceased more abruptly after glucose exhaustion. This would not have an effect on the production process as it occurred after the usual harvest time.

The reported differences in the DOT profiles occurred very late in the fermentation and would not have an effect on the production process but are, again, indicative of changes in the availability of the nutrients utilised after glucose exhaustion.



Qualitative Descriptions

Line	Dir	Mag	Dur	Slope	Start
A1	0	1	14	1	1
A2	1	7	5	8	1
A3	-1	1	1	2	7
A4	1	2	1	8	7
A5	-1	5	13	6	8
B1	0	1	13	1	1
B2	1	8	6	8	1
B3	-1	3	11	4	8

Match Record

Lines	Magnitude	Duration	Slope	Start
A1 →B1	1	1	1	1
A2 →B2	1	1	1	1
A3+A4 → -	-	-	-	-
A5 →B3	0	0	0	1

Figure 4.14: Comparison of the beginning of the agitation rate data from batches C443 and C444. The maximum agitation rate, used to maintain dissolved oxygen in the broth, was lowered more rapidly in C443 as shown by the comparisons of lines A5 and B3.

It was concluded that reducing the concentrations of the pre-sterile media components by up to 15% has a slight deleterious effect on biomass production in the aFGF fermentation as a result of decreasing the supply of essential nutrients. The other effects were small and generally occurred after the recommended harvest time (around 16 h, Table 4.7) thus would not be detrimental to the production process.

The result of this comparison must be heeded when using the biomass and pH profiles of C443 in subsequent analyses; a note to this effect was made in the data base 'observations' field of batch C443. However, as all the differences between batches C443 and C444 were readily explained as being effects of the dilution of media components in C443, the data from these two batches was considered to be suitable for use in any further analyses.

This example demonstrates the role of the difference summary in tracing cause-effect relationships and stresses the importance of recording all information in the data base. An observation such as accidental medium dilution may easily be overlooked in manual analysis but has been shown to be important in interpreting the different behaviour of two fermentations.

4.4.1.3 Variations in Inocula

Age was used as the basis for the time of transferral of the inoculum to the fermenters. This is not considered to be a suitable practice as the resulting variability in inoculum condition and size can have deleterious effects on the ensuing fermentation (Buckland 1984). However, on a small scale, where the inoculum vessel is a shake flask, it is difficult to monitor the growth of the culture and often the only indicator available is culture age. The optical densities and glucose concentrations of each inoculum prior to inoculation are listed in Table 4.12. It is evident from the similarity of batches C447 and C451 that inocula with a wide range of optical densities (4.8 to 6.6) and glucose concentrations (7.4 to 8.6 g.L^{-1}) had very little effect on the progress of the aFGF fermentation.

However, there is evidence that the inocula used in batches C441 and C442, which came from the same source vial, performed differently from the other inocula used. The difference summary reported variations in the CER and glucose data of batches C441 and C442 when compared with C447 and C451; these differences occurred in the first linear data piece, ie the lag phase, which is usually evidence of variations in the inoculum. It will also be shown later (Section 4.4.1.5) that the metabolic activity of C441 and C442 was slightly higher than in the other batches and more aFGF was produced. These effects are indicative of a highly active inoculum which may have been a result of the growth state of the bacteria at the time of inoculation but may also be a result of a spontaneous mutation in the recombinant organism. Changes in the genetic make-up of an organism cannot be routinely monitored and thus present a barrier to any comprehensive comparative analysis programme.

BATCH NUMBERS	OPTICAL DENSITY	GLUCOSE ($g.L^{-1}$)
C439 & C440	5.8	8.4
C441 & C442	6.4	7.3
C443 & C444	5.7	6.9
C446 & C447	4.8	7.4
C449 & C450	4.9	7.5
C451 & C452	6.6	8.6

Table 4.12: Inoculum data for aFGF fermentations.

Inoculum variation may have been responsible for the reported low aFGF titre in batch C451 (Table 4.10). This was investigated further by comparing the aFGF titre of C447 with that of C452, which had been seeded with a smaller volume of the same inoculum as that used in C451. The appropriate batch to use in this investigation was found by searching the data base tables for a fermentation with the same inoculum identification number as C451. The aFGF titre of C452, 41.1 normalised units.L^{-1}, was less than that of C447, 53.7 normalised units.L^{-1}, but more than that of C451, 29.6 normalised units.L^{-1}. A difference in the activity of the inocula of batches C447 and C451 cannot be ruled out; however, it is also possible that the difference in the post-sterile broths, indicated by the different absorption spectra values, was responsible for the lower aFGF production in C451.

The biomass and glucose profiles for C451 and C452 contained very little data in the early stages of the fermentations. The shapes of the biomass and glucose profiles during the initial growth phases were therefore not available and the effect of the inocula on biomass production and glucose consumption for these two batches could not be determined. Although this suggests a fundamental deficiency in the comparative analysis tools it should be noted that this lack of data would also hinder manual analysis.

4.4.1.4 Temperature Control

A problem with the service module attached to fermenter BL4 resulted in a fault in the temperature control of batches C440 and C442 as noted in the batch sheet tables. The results of comparing these batches with other batches gave an indication of the effects a temperature fault has on the aFGF fermentation. The CER, DOT, pH and agitation rate profiles of batches C440 and C442 differed from those of the fermentations they were compared with (Table 4.10). The temperature excursion in C442 had no effect on the cell growth or glucose utilisation as evidenced by the comparisons with batches C441, C447 and C451 (the reported difference in the glucose profiles of C442 and C447 occurred in the lag phase, prior to the temperature fault). The poor temperature control in C440 had no effect on glucose utilisation (the differences in the glucose profiles of C439 and C440 occurred prior to the temperature problem). The difference in biomass between C440 and C439 may be a result of the difference in temperature but seems unlikely based on the evidence of C442.

The effects of the temperature fault on batch C442 were complicated by the occurrence of a fault in the control of dissolved oxygen which is discussed in the next section.

The CER (Figure 4.15), agitation rate and post-control pH profiles of C440 and C442 were very different from those of the other batches but the effects were not consistent. This was attributed to the differences in dissolved oxygen in the broths of C440 and C442, as discussed in Section 4.4.1.5, and also the fact that the fermenter instruments were recalibrated after batch C440 which may have caused some discrepancies in the data. Despite these differences, the lack of effect of temperature on the biomass and glucose data indicates that the aFGF fermentation is fairly resistant to brief temperature excursions.

The aFGF yield in C442, at the recommended harvest point, was higher than that in both C447 and C451, but comparable to the yield in C441 (Table 4.7). It was thought that the high yield was a result of a highly active inoculum, as the same inoculum was used to seed C441 and C442. The similarity of the aFGF yield at the recommended harvest point with that of C441 suggests that the temperature fault in C442 did not adversely affect the specific production rate of the organism. The high temperature in C442 probably prevented batch C442 from reaching the same maximum aFGF yield as C441 but the anaerobic conditions experienced would also have contributed to this effect, the faults could not be treated separately.

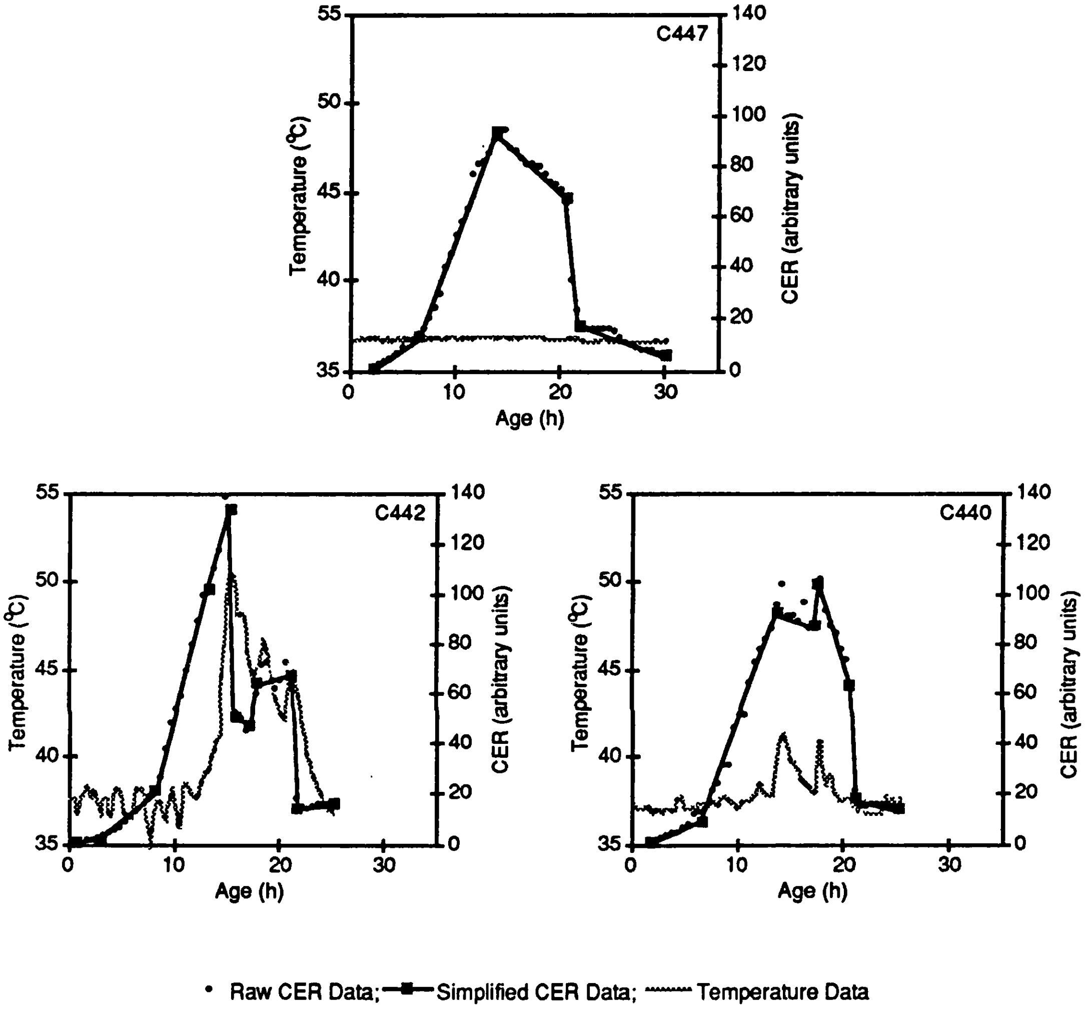


Figure 4.15: 'Normal' temperature and carbon dioxide evolution rate (CER) profiles were seen in batch C447. Batches C442 and C440 had poor temperature control and their CER profiles did not follow the expected pattern.

The differences between C439/C440 and C441/C442 could not be fully explained by the temperature fault because of the inconsistencies observed between the two comparisons. As stated earlier the data from batches C439 and C440 were suspect and should not be used in further analyses. The comparison of C441 and C442 is further discussed in the next section.

4.4.1.5 Dissolved Oxygen Concentration Control

The control settings did not allow a quick response when the DOT initially fell below the specified minimum level of 20% of air saturation. As a result the dissolved oxygen concentration in the broth fell below the control level for a brief period about 10 hours after inoculation in all the fermentations. This is seen in the raw data in Appendix 5 (Figure A5.4) but does not show up in the linearised data. As this occurred in all the fermentations its effect could not be determined.

A problem with the control of dissolved oxygen was noted in batches C441 and C442: the dissolved oxygen levels fell below the control level a second time, at about 13 hours (Figure 4.16). In both fermentations the point at which dissolved oxygen was lower than the specified minimum corresponded to the peak respiration rates indicating high levels of metabolic activity (Figure 4.16). The effect of the low dissolved oxygen concentration in batch C442 was complicated by the occurrence of the temperature fault.

The linearised time profiles (Figure 4.16) showed the course of events to be as follows: when the CER exceeded 100 arbitrary units (approximately) the dissolved oxygen decreased below the minimum value of 20% of air saturation; in batch C442 the increase in temperature, which occurred at about the same time as the dissolved oxygen fault, initially led to a further increase in metabolic activity, as illustrated by the increasing CER, and consequently a further decrease in dissolved oxygen followed by a rapid decrease in respiration; when the CER fell below 100 units the dissolved oxygen concentration returned to the control limit of 20% of saturation and the agitation rate was decreased as the high level was no longer required to maintain the dissolved oxygen level. No other aFGF batches had carbon dioxide evolutions rates exceeding this level of 100 units. The agitation rate and air flow rate upper limits were not high enough to introduce sufficient oxygen into the system to maintain an acceptable dissolved oxygen level during high levels of metabolic activity.

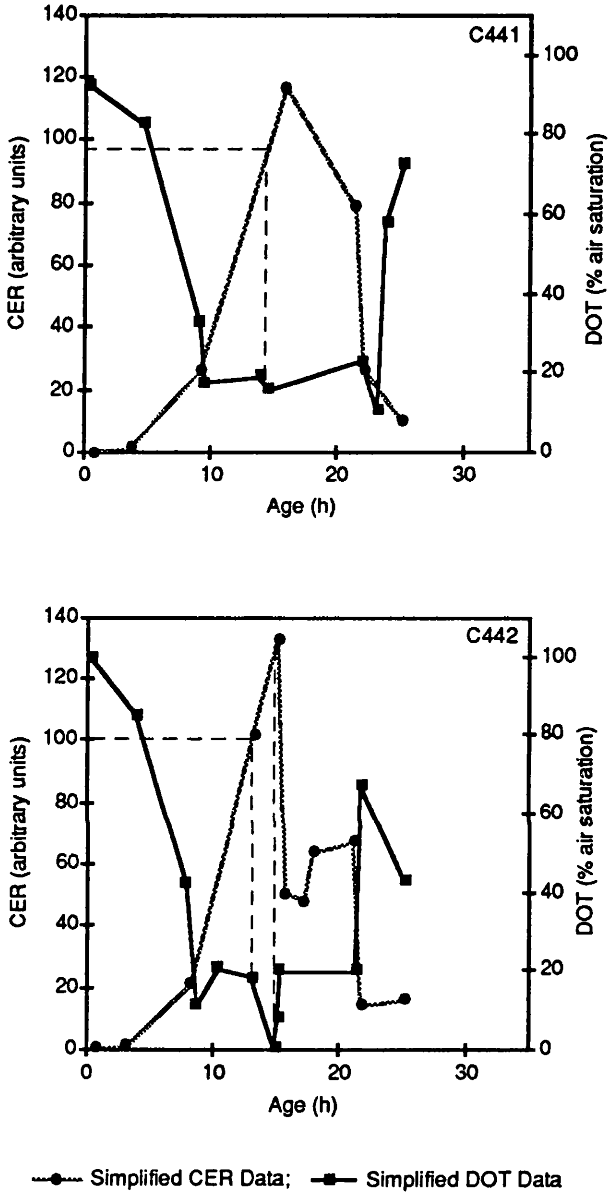


Figure 4.16: Poor control of the dissolved oxygen tension (DOT) was observed in batches C441 and C442. The DOT fell below the lower limit of 20% when the carbon dioxide evolution rate (CER) reached approximately 100 units. When the DOT reached 0% in C442 the CER decreased rapidly.

In C442 the dissolved oxygen was reduced to 0% of air saturation. The lack of oxygen in the broth necessarily affected the metabolic activity of the cells. This was observed in the rapid decline in carbon dioxide evolution rate (which may also have been influenced by the high temperature). The decrease in the respiration rate of the organism allowed the dissolved oxygen to increase again and the cells continued to metabolise the substrates available. When the respiration rate of the organism slowed down the agitation rate decreased rapidly as the demand for oxygen lowered, this accounts for the observed differences in the agitation rate profiles. The lack of oxygen in the broth may also have been responsible for the aFGF concentration in C442 not reaching the same high level as in C441.

The DOT in C441 fell to approximately 10% of air saturation with only a brief fluctuation in respiration which did not significantly affect the CER, biomass or glucose time profiles as shown by comparison of C441 with C447 and C451 (the noted differences in CER and glucose occurred in the lag phase, prior to the DOT problem). The alteration in metabolic activity in C441 was reflected in the differences between the agitation rate profiles. The low dissolved oxygen concentration caused the metabolic activity of the organism to slow down, the oxygen requirement was therefore lower and the agitation rate decreased. The effect on pH was not clear as the observed differences may also have been a result of the high metabolic activity of the organism or the medium composition (discussed below).

In conclusion it is reasonable to assume that a reduction of DOT to 10% of air saturation for a brief period of time is of no consequence to the fermentation whilst DOT levels below 10% of air saturation alter the pattern of the respiration data but have little effect on biomass or glucose.

It is interesting to note that the high aFGF values recorded in both C441 and C442 occurred after the DOT fell below the control level.

During manual analysis of the data, prior to the availability of the comparative analysis tools, it had been assumed that the unusual time profiles of batch C442 had resulted from the temperature fault and that there was no reason to assume that C441 was any different from C447 and C451. No connection was made between the dissolved oxygen fault and high metabolic activity. The computerised comparisons did not detect any differences between the CER profiles of C441/C447 and C441/C451 (other than in the lag phase, prior to the DOT problem), they were not considered significant. The connection was made obvious when looking at the linearised time variant data: the simplified DOT data highlighted where the fault occurred and the coincidence with the high metabolic rate was easily seen (Figure

4.16). This was not detected whilst the experimental system was still available: the high metabolic activity in C441 and C442 resulted in high aFGF titres (69.9 normalised units.L^{-1} and 66.7 normalised units.L^{-1} respectively at the recommended harvest point, Table 4.7) thus, if the cause could be found and implemented, an improvement in aFGF production could be achieved.

Comparisons of batches C441 and C442 with C447 and C451 (Table 4.10) indicated the possible causes of the high levels of metabolic activity in C441 and C442: these include the sterilisation temperature profile, the inoculum and the medium composition, as indicated by the different absorption spectrum of C442. The sterilisation temperature profiles were very slightly different in the heating up and cooling down phases. It was thought that there would not be any significant effect on the medium as a result of these differences. The inoculum used in C441 and C442 was from the same seed vessel. Although there was no evidence to show that the inoculum was significantly different from those used in the other fermentations (Table 4.12) no analysis of the original frozen suspension was provided and thus this cannot be discounted as a possible cause of the high metabolic activity. The possibility of an incorrect medium composition was implied by the lower absorption maxima in the post-sterile broth of C442 (Figure 4.17). There were no pre-sterile absorption data for C442 and no absorption data for C441. The media for C441 and C442 were prepared at the same time and any mistakes may have been duplicated. In conclusion it is likely that the high metabolic activity observed in C441 and C442 was a result of either the inoculum or the medium composition.

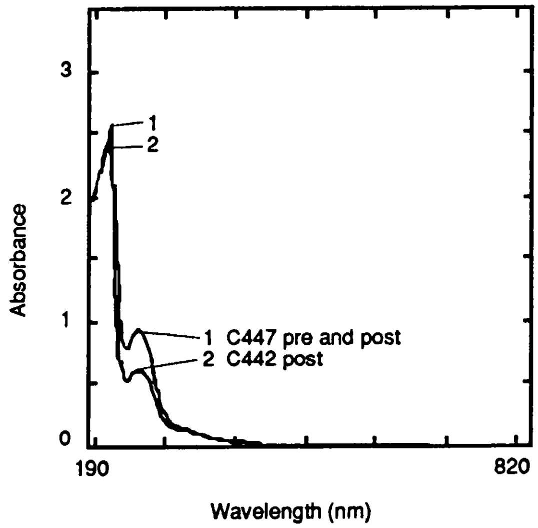


Figure 4.17: Absorption spectra of pre- and post-sterile fermentation broths of batch C447 and post-sterile fermentation broth of C442. The difference in the absorption spectrum of C442 was probably due to an incorrect medium composition.

The analyses presented in this section and the previous one discussed the possible causes of the differences observed when batches C441 and C442 were compared with each other and with C447 and C451. The inoculum or medium composition of C441 and C442 resulted in high metabolic activities, and temperature and DOT faults affected the pH and agitation rate profiles of C441 and C442 and the CER profile of C442. The data from C442 were not used in subsequent analyses because of the extremely different behaviour of the CER profiles. C441 data were used but with caution.

This example again shows the role of the difference summary in tracing cause-effect relationships. The list of differences presented in the difference summary indicates where the analyst must look to reason about what occurred in the fermentations.

4.4.1.6 Summary of Reproducibility Investigations

Few conclusive cause-effect relationships were found in this analysis. However, the process illustrates the important role of the comparative analysis in pointing where to direct any further investigations. If the above analyses had been carried out when the experimental system was still available a number of the indicated cause-effect relationships could have been investigated.

Although this analysis was time consuming it was important to ensure that the aFGF fermentation is reproducible; there would be little point in attempting to compare data from fermentations where, even under identical conditions, the performance varies significantly. In situations where fermentations were expected to behave similarly and did not, it was necessary to investigate why the variations had occurred, ie to determine that the fermentations had not in fact been operated under the intended identical conditions. This led to an improved understanding of the process as the effects of these unintentional perturbations were examined. It was also important to determine if any of the fermentations were faulty before using them in the planned comparisons.

The conclusions from the reproducibility investigations were as follows:

1. comparison of batches C447 and C451 demonstrated the reproducibility of the aFGF fermentations;

2. the data from batches C439 and C440 should only be used with caution in subsequent analyses;
3. a problem with DOT control and the possibility of either an incorrect medium composition or an unusually active inoculum suggest caution in the use of data from batch C441;
4. batch C442 should not be used in subsequent analyses because of poor temperature control and temporary anaerobic operation;
5. the dilution of batch C443 resulted in a reduced biomass yield and a shorter period of pH control;
6. glucose was not the limiting substrate in batches C443 and C444;
7. the aFGF yield differed in batches that were otherwise reproducible, C451 and C447, it was not possible to determine what a typical value was;
8. batches C443, C444, C447 and C451 may be used with confidence in subsequent analyses (keeping in mind points 5, 6 and 7);
9. the transfer of inoculum on the basis of age may not be satisfactory based on the evidence of batches C441 and C442, this requires further investigation;
10. a 15% reduction in the concentration of the bulk media components may be possible as it has very little effect on the resulting fermentation (up to the recommended harvest point) as shown by the comparison of C443 and C444;
11. the effect of temperature excursions was not conclusive but appeared to be small;
12. a reduction of dissolved oxygen tension in the broth to 10% of air saturation has little effect on the fermentation (C441) but a reduction to 0% of air saturation causes an immediate precipitous decline in respiration (C442).

These observations were utilised in the analysis of the effects of sterilisation conditions and inoculum volume on the aFGF fermentation in Sections 4.4.3 and 4.4.4.

Prior to the availability of the comparative analysis tools a manual analysis of the aFGF data had been carried out. This analysis had concluded that there were no significant differences between batches C443/C444, C441/C447, C441/C451 and C447/C451. Use of the comparative analysis tools detected significant differences in all these pairs of fermentations except C447/C451 thus prompting a detailed investigation of why these differences had occurred which resulted in an improved understanding of the process. The comparative analysis tools were more consistent than visual analysis in the detection of differences between the data sets; manual analysis was prejudiced by prior expectations of which batches should have behaved similarly.

Perhaps the most notable benefit of the comparative analysis tools was that the effort involved in comparing the time variant data was significantly reduced as there was no longer a need to manually overlay all the different profiles to perform the relevant comparisons. The computer routines performed the comparisons automatically and summarised the results in match records thus allowing more time to be spent on determining why differences occurred rather than on detecting the differences. The inclusion of the data base comparisons also significantly improved the analysis by ensuring all information was summarised in one place allowing a more comprehensive reasoning process.

4.4.2 Features of the aFGF Data

The linearised time profiles were very useful in highlighting correlations between different variables. The points at which the linear segments join identify *events* in the fermentation; events identified in one variable often coincide with events in another variable; these correlations are *features* of a particular fermentation. Correlations between different variables are useful in fault detection and diagnosis: when an expected correlation is not apparent in a data set there is evidence of aberrant behaviour. Correlations can also be used to predict when an event is likely to occur in on-line operation.

The desired correlations are found in the time variant data of fermentations operated under standard conditions in which no faults are apparent.

Fermentations C441, C442, C447 and C451 were intended to be operated under standard conditions, that is, under the conditions that would be used in an aFGF production run. The data from C447 and C451 were very similar, C441 deviated slightly from these 'standard' batches and C442 was shown to be faulty. Features common to batches C447, C451 and C441 were identified to aid subsequent analyses.

Acidic products are usually produced as a result of simple carbohydrate metabolism, this is borne out by the fact that the point at which glucose was exhausted in the aFGF fermentations corresponded to the point at which pH control ended as shown by the cessation of alkali addition to the broth (Figure 4.18).

The growth rate of the organism (slope of the biomass profile) began to decrease just prior to glucose exhaustion (Figure 4.18). This decrease in growth rate is a result of a reduced

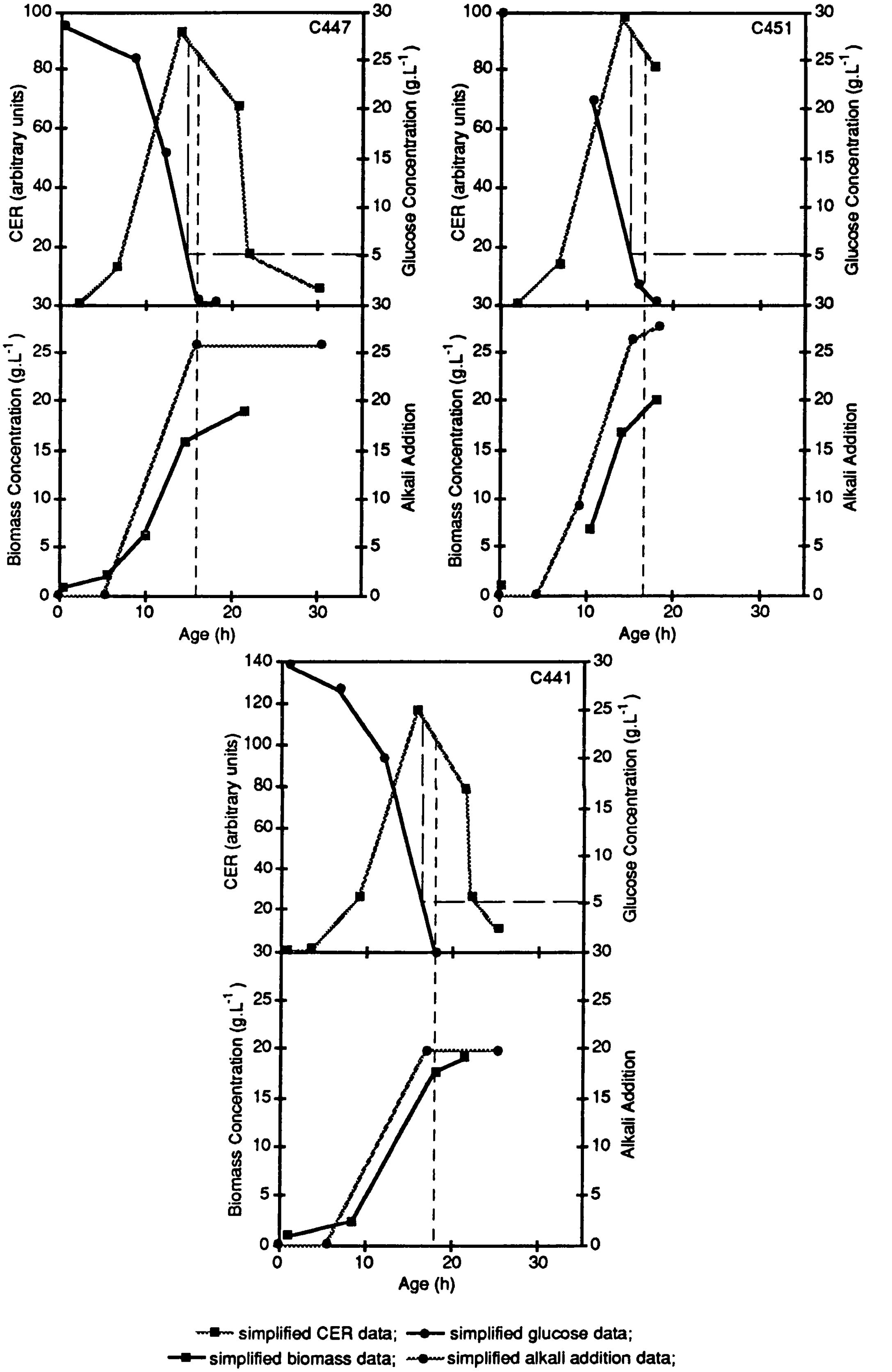


Figure 4.18: Correlations between variables from 'standard' aFGF fermentations. The exhaustion of glucose corresponds with the end of alkali addition and occurs about the same time as the growth rate slows down. The recommended harvest time (glucose = 5 g.L^{-1}) occurs just after the peak CER.

driving force for transport of glucose across the cell membrane caused by the low sugar concentration.

The point at which the carbon dioxide evolution rate began to decrease occurred just prior to the recommended harvest point, ie a glucose concentration of 5 g.L^{-1} (Figure 4.18). During operation of a production scale process the harvest point is determined by extrapolating the off-line glucose measurements to predict the time at which 5 g.L^{-1} will be reached. However, as is usually the case, it would be beneficial to be able to use an on-line measurement for prediction of this point and the carbon dioxide evolution rate profile could be used for this.

4.4.3 The Effect of Sterilisation Conditions

One of the aims of this work was to determine the effects of different sterilisation regimes on the aFGF fermentations. The data from these experiments are analysed here using the computerised comparative techniques.

A typical sterilisation temperature profile for the 15 L Biolafitte fermenters is shown in Figure 4.19. During the reproducibility investigations small differences were observed in

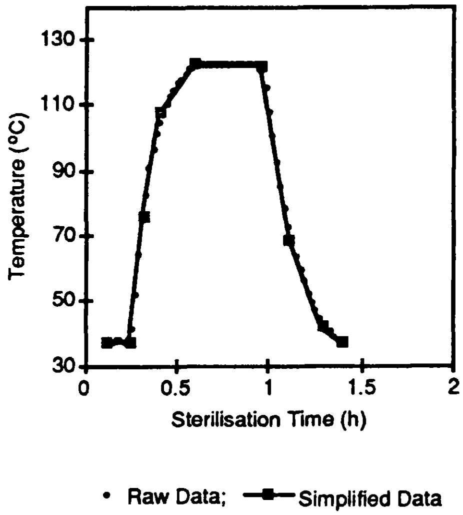


Figure 4.19: Temperature profile for the sterilisation of batch C447. The shape of this profile is typical of the 15 L Biolafitte sterilisation process.

the heating up and cooling down phases of the sterilisation temperature profiles. These were a result of attempting to control the amount of steam introduced to the vessel. It was thought that these small differences would have had no effect on the fermentation broths. The sterilisation pressure profiles were not available for analysis.

4.4.3.1 Increased Length of Sterilisation - No Glucose Present

The data from the fermentations in which glucose was sterilised separately from the bulk medium (C441, C443, C444, C447 and C451) were compared, using the comparative analysis tools, to assess the effects of lengthening the sterilisation holding time. The results of the comparisons are summarised in Table 4.13. Batches C439, C440 and C442 were not included in the comparative analyses because of observed anomalies (Section 4.4.1). The interpretation of the comparisons also acknowledged the facts that C441 exhibited a relatively high metabolic activity and that the biomass yield in batch C443 had suffered because of dilution of the medium components (Section 4.4.1).

The effects on the media are shown in the absorption spectra in Figure 4.20 and the changes

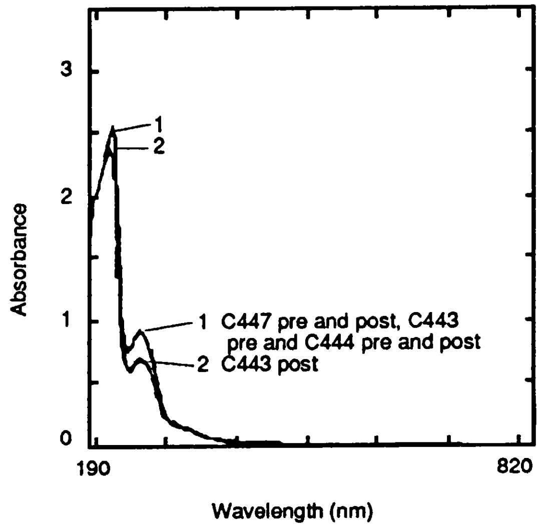


Figure 4.20: Absorption spectra of pre- and post-sterile fermentation broths of batches C443, C444 and the standard (C447). Sterilising for ninety minutes did not affect the absorption characteristics of the broth (C444). The lower absorption of C443 was due to a diluted medium composition.

BATCH NUMBERS	DIFFERENCES IN TIME INVARIANT DATA	DIFFERENCES IN TIME VARIANT DATA	OBSERVATIONS
C441/C444	Tanks Inoculum Post-sterile pH $\Delta pH_{sterilisation}$ $aFGF_{sh}$, $aFGF_{sm}$ $aFGF_{vh}$, $aFGF_{vm}$	Sterilisation Biomass CER DOT pH Agitation	DOT fault C441 No abs. spec. C441
C447/C444	Inoculum Post-sterile pH $\Delta pH_{sterilisation}$ $aFGF_{vh}$,$aFGF_{vm}$ $aFGF_{sh}$	Sterilisation CER DOT pH Agitation	
C451/C444	Tanks Inoculum Post-sterile pH $\Delta pH_{sterilisation}$ $\Delta V_{sterilisation}$ $aFGF_{vh}$,$aFGF_{sh}$	Sterilisation CER pH Agitation	No pre-sterile abs.spec. in C451 C451 shorter No off-line data first 10 h of C451
C441/C443	Inoculum Post-sterile pH $\Delta pH_{sterilisation}$ $\Delta V_{sterilisation}$ $aFGF_{sh}$, $aFGF_{sm}$ $aFGF_{vh}$, $aFGF_{vm}$	Sterilisation Biomass CER DOT pH Agitation	Medium dilution C443 No abs. spec. C441 DOT fault C441
C447/C443	Tanks Inoculum Post-sterile pH $\Delta V_{sterilisation}$ Post-sterile abs. spec. $aFGF_{vh}$, $aFGF_{vm}$ $aFGF_{sh}$, $aFGF_{sm}$	Sterilisation Biomass CER DOT pH Agitation	Medium dilution C443
C451/C443	Inoculum Post-sterile pH $\Delta pH_{sterilisation}$ Post-sterile abs. spec. $aFGF_{vm}$, $aFGF_{sh}$	Sterilisation Biomass CER pH Agitation	Medium dilution No pre-sterile abs.spec. in C451 C451 shorter No off-line data first 10 h of C451

Table 4.13: Difference summary for fermentations sterilised over differing lengths of time with glucose not present in the bulk medium during sterilisation. The abbreviations are described in Table 4.11.

in pH in Figure 4.3. The broths of batches C447 and C444, sterilised for twenty minutes and ninety minutes respectively, had identical absorption characteristics both before and after sterilisation, ie sterilisation did not alter the protein components of the broths. Lower absorption maxima were observed in the post-sterile broth of batch C443 as a result of the dilution of the medium components. There were no absorption data for C441. The broths sterilised for twenty minutes underwent very small changes in pH: from 0.05 to 0.25 pH units. Sterilisation for sixty minutes did not result in a larger change in pH, however a sterilisation holding period of ninety minutes resulted in a significantly larger decrease in pH (in the range of 0.4 to 0.7 pH units). In many fermentations the hydrolysis of proteins would constitute the major change in the medium during sterilisation. However, in the medium used for the aFGF fermentations the major protein source consisted of partially hydrolysed proteins and it would be expected that the effect of heat stress on this medium would be considerably less than on a medium containing a more complex protein source. The similarity of the absorption spectra before and after sterilisation demonstrated this. The large changes in pH of batches C443 and C444 indicated that sterilising the broth for extended periods of time (ninety minutes) did have some influence on the media. This was most likely a result of changes in the solubility of some media components (Corbett 1985) and indicates that if, on scale up, the sterilisation holding time were to be increased substantially, there would be some effect on the fermentation broth.

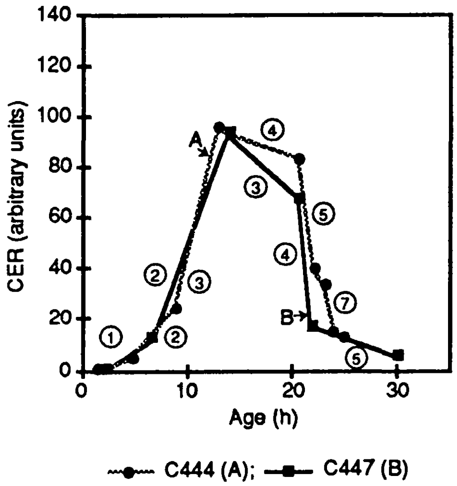
The changes in broth pH and absorption spectra as a result of sterilisation reflected different occurrences in the media and thus should be used in conjunction to determine the effects of sterilisation.

The difference summary (Table 4.13) lists the effects of the increased sterilisation hold times on the performance of the fermentations. These include the biomass concentrations, the carbon dioxide evolution rates, the dissolved oxygen, the pH, the agitation rates, and the aFGF titres.

There was no effect on the biomass production in C444 (except when compared with C441) or on glucose utilisation. The discrepancy in biomass production between C441 and C444 was further evidence of the high metabolic activity in C441 and did not imply that the biomass yield in C444 was any lower than in a typical fermentation. The differences in biomass concentrations between C443 and the other batches were a result of the dilution of the media (Section 4.4.1.2) and were not relevant to this part of the investigation.

The comparison of the CER profiles of batches C444 and C447 is shown in Figure 4.21.

The line representing the most rapid increase in CER, ie the fast growth period, had a shorter temporal extent in C444 than in C447. Upon closer inspection this was seen to be a result of the fast growth period starting later in C444 than in C447. After the fast growth period the CER begins to decrease but the match record in Figure 4.21 shows that this decrease was more rapid in C447 than in C444. Further differences, towards the end of the fermentations, were also noted in the match record. The comparison of the CER profile of C444 with C451 and C441 and the CER profile of C443 with C447, C451 and C441 gave similar results.



Qualitative Descriptions

Line	Dir	Mag	Dur	Slope	Start
A1	1	1	3	3	1
A2	1	2	4	5	1
A3	1	7	4	7	3
A4	-1	2	7	3	9
A5	-1	4	2	8	8
A6	-1	1	1	6	4
A7	-1	2	1	8	3
A8	-1	1	1	4	2
B1	1	2	4	4	1
B2	1	7	7	7	2
B3	-1	3	6	5	8
B4	-1	5	2	8	6
B5	-1	2	8	3	2

Match Record for C444/C447

Lines	Magnitude	Duration	Slope	Start
A1 → -	-	-	-	-
A2 → B1	1	1	1	1
A3 → B2	1	0	1	1
A4 → B3	1	1	0	1
A5+A6 → B4	1	1	1	0
A7 → -	-	-	-	-
A8 → B5	1	0	1	1

Figure 4.21: Comparison of carbon dioxide evolution rate profiles from batches C444 and C447. The poor matches in the match record show that sterilising for 90 minutes (C444) resulted in a slightly different fermentation from one in which the broth was sterilised for 20 minutes (C447).

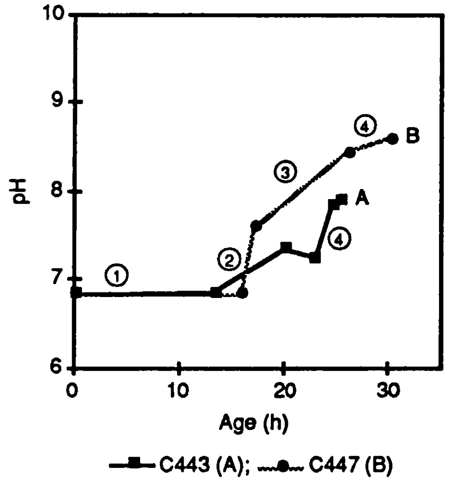
Comparison of the pH profile of C443 with those of C447 (Figure 4.22), C451 and C441 showed that the length of the first linear data piece, ie the control period, was shorter in C443. This was a result of the dilution of secondary nutrients during sterilisation of batch C443 (Section 4.4.1.2).

The comparison routine, MATCHER, concluded that the lines representing the post-control pH data of the batches that had been sterilised for ninety minutes (C443 and C444) were significantly different from the corresponding lines in the batches sterilised for twenty minutes (C447, C451 and C441). The comparisons with C447 are shown in Figure 4.22.

Glucose was not the limiting substrate in batches C443 and C444 (Section 4.4.1.2). In the fermentations operated under standard conditions the end point of pH control coincided with glucose exhaustion and occurred just after the growth rate and carbon dioxide evolution rate slowed down (Section 4.4.2). In batches C443 and C444 the correlation between carbon dioxide evolution rate, growth rate and pH control was as expected but glucose exhaustion occurred nearly four hours after the expected time (Figure 4.23). The comparison of C443 and C444 in the reproducibility investigations (Section 4.4.1.2) had concluded that the limiting substrate had been present in the bulk medium during sterilisation as it was present in smaller amounts in C443 which had been diluted during sterilisation. It is therefore apparent that sterilising the broth for ninety minutes results in the destruction of an essential nutrient. This would account for some of the differences observed in the CER and post-control pH profiles. In Figure 4.23 it is seen that, in a 'standard' fermentation (C447), the decrease in carbon dioxide evolution rate corresponded with the switch from glucose metabolism to the metabolism of secondary nutrients. It was stated earlier that this initial period of decreasing CER was slower in C443 and C444 than in C447, C441 and C451. In C443 and C444 the rate of metabolic activity had started to decrease because an essential nutrient (not glucose) had run out, but as glucose was still present the metabolic rate did not decrease as rapidly as in C447 where glucose had been exhausted by this stage. Similar reasoning can be applied to explain the differences in the post-control pH profiles.

The differences in dissolved oxygen and the agitation rates observed between C447 and the two fermentations sterilised for 90 minutes (C451 did not proceed this far) were most likely a result of different availabilities of minor substrates in the medium.

The effect on the aFGF production was not conclusive. The aFGF titres at the recommended harvest points were lower than those in C441 and C447 but higher than those in C451 (Table 4.7).

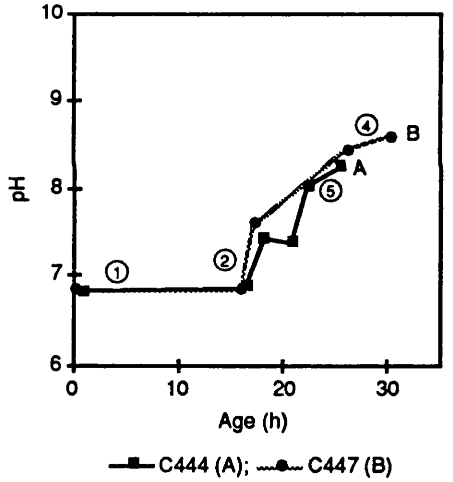


Qualitative Descriptions

Line	Dir	Mag	Dur	Slope	Start
A1	0	1	9	1	1
A2	1	3	5	6	1
A3	-1	1	2	4	3
A4	1	4	2	8	3
A5	1	1	1	5	6
B1	0	1	11	1	1
B2	1	5	1	8	1
B3	1	5	6	6	5
B4	1	1	3	4	8

Match Record C443/C447

Lines	Magnitude	Duration	Slope	Start
A1 → B1	1	0	1	1
A2 → B2	0	0	0	1
- → B3,B4	-	-	-	-
A3,A5 → -	-	-	-	-



Qualitative Descriptions

Line	Dir	Mag	Dur	Slope	Start
A1	0	1	10	1	1
A2	1	3	1	8	1
A3	-1	1	2	2	4
A4	1	4	1	8	3
A5	1	2	2	5	7
B1	0	1	11	1	1
B2	1	4	1	8	1
B3	1	5	6	6	4
B4	1	1	3	4	8

Match Record C444/C447

Lines	Magnitude	Duration	Slope	Start
A1 → B1	1	1	1	1
A2 → B2	1	1	1	1
- → B3	-	-	-	-
A3,A4 → -	-	-	-	-
A5 → B4	1	1	1	1

Figure 4.22: Comparison of pH data from batches C443, C444 and C447. The comparisons of C443 and C444 with each of C451 and C441 were similar. The pH control of C443 was over a shorter period (durations of first lines) and the post-control profiles for both C443 and C444 were different from those of the fermentations run under standard conditions.

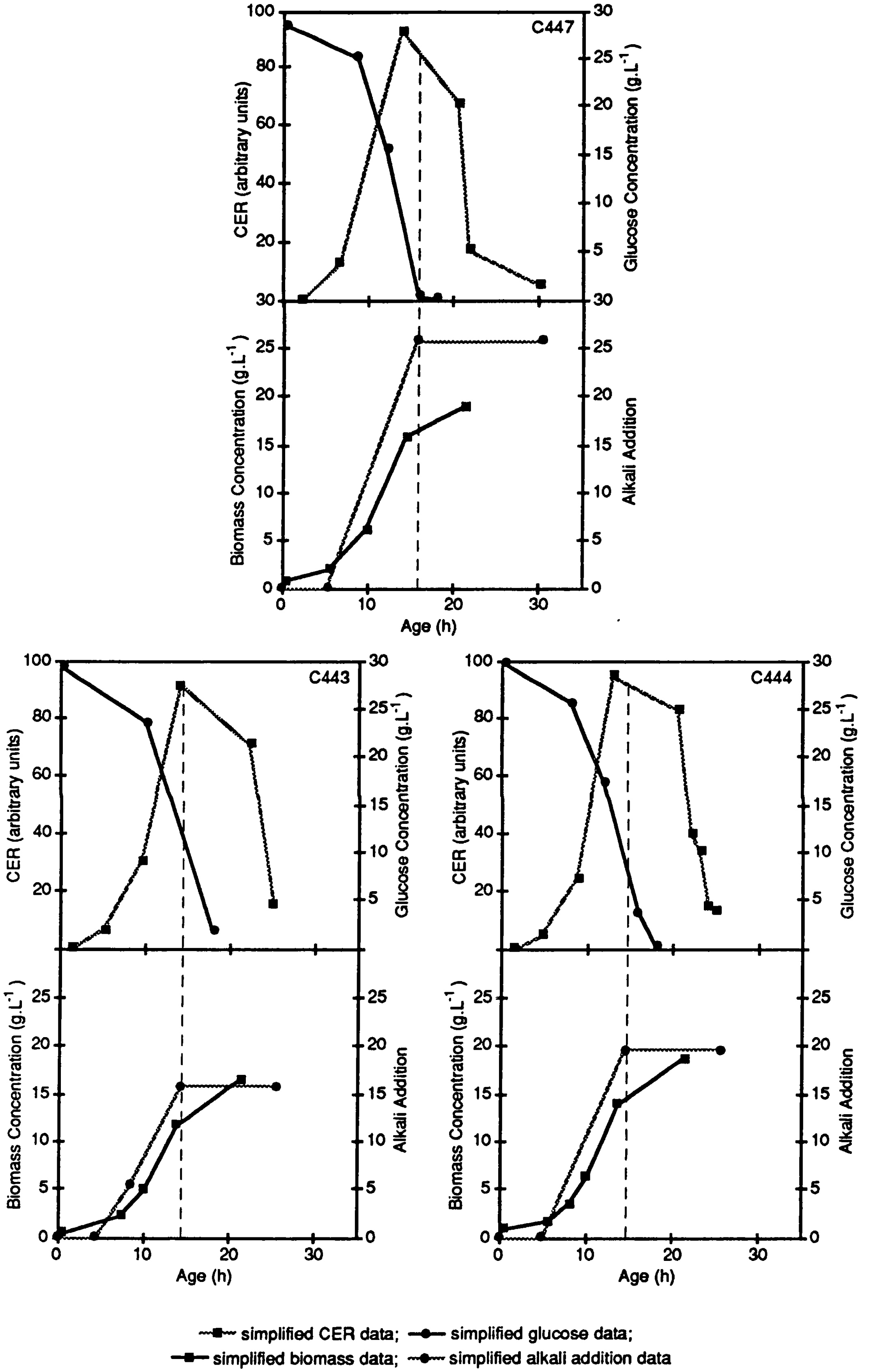


Figure 4.23: In C447, the 'standard' fermentation, glucose exhaustion corresponded with the cessation of alkali addition and the slowing down of growth rate and CER. In C443 and C444 glucose exhaustion occurred about 4 hours later than was expected from the patterns exhibited in C447.

These effects that occurred as a result of longer sterilisation times indicate that, even though the heat reactive and heat labile medium components were sterilised separately, there is the potential for some changes in the performance of the fermentation on scale up. As mentioned earlier these changes are a result of the longer sterilisation periods having an effect on the solubility of some of the medium components, especially the nitrogen source (Corbett 1985). Although the performance of the fermentation was obviously affected, albeit very slightly, by lengthening the sterilisation hold time, the minimal effects on biomass production and glucose utilisation indicate that the overall effect on the fermentation is not detrimental. However, the true effect on the aFGF yield needs further investigation. The ninety minute sterilisation period experienced in C443 and C444 is more excessive than would normally be seen on scale up, thus it is believed that the effect of the increased heat stress on scale up would be minimal to the aFGF fermentation.

Prior to the availability of the computerised tools this comparative analysis had been carried out manually. The manual analysis had concluded that there was no difference between the batches sterilised for twenty minutes and the batches sterilised for ninety minutes and therefore the increased heat stress on scale up would not affect the aFGF fermentation. The structured analysis carried out using the computerised tools not only detected differences in the data but enabled the analyst to reason about the causes of the differences and, in so doing, learn more about the process.

4.4.3.2 A Longer Sterilisation Cool Down Period

A slower cool down period subsequent to sterilisation was used to simulate the conditions of a larger fermenter in batch C450. The sterilisation profile is shown in Figure 4.24. The effect this had on the progress of the fermentation was assessed by comparing the data from this fermentation with data from fermentations in which the standard sterilisation sequence had been used. Again the comparisons with C447 and C451 were thought to be most indicative of the effects the extended cooling period had on the fermentation. The results of the comparative analysis are summarised in Table 4.14.

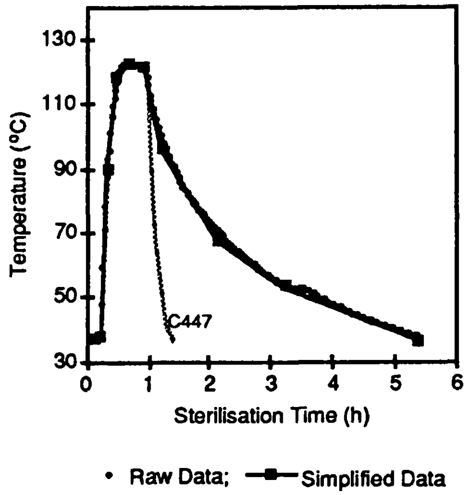


Figure 4.24: Temperature profile for the sterilisation of batch C450 showing the long cooling time. The profile from batch C447 is shown for comparison.

BATCH NUMBERS	DIFFERENCES IN TIME INVARIANT DATA	DIFFERENCES IN TIME VARIANT DATA	OBSERVATIONS
C441/C450	Tanks Inoculum Post-sterile pH $\Delta V_{sterilisation}$ aFGF$_{sh}$, aFGF$_{sm}$ aFGF$_{vh}$, aFGF$_{vm}$	Sterilisation Glucose CER DOT pH Agitation	No abs. spec. C441 DOT fault C441
C447/C450	Inoculum Pre-sterile abs. spec. aFGF$_{sh}$ aFGF$_{vh}$, aFGF$_{vm}$	Sterilisation Glucose CER Agitation	
C451/C450	Tanks Inoculum $\Delta V_{sterilisation}$ Post-sterile abs. spec.	Sterilisation CER Agitation	No pre-sterile abs.spec. in C451 C451 shorter No off-line data first 10 h of C451

Table 4.14: Difference summary for fermentations with different lengths of cooling after sterilisation. The abbreviations are described in Table 4.11.

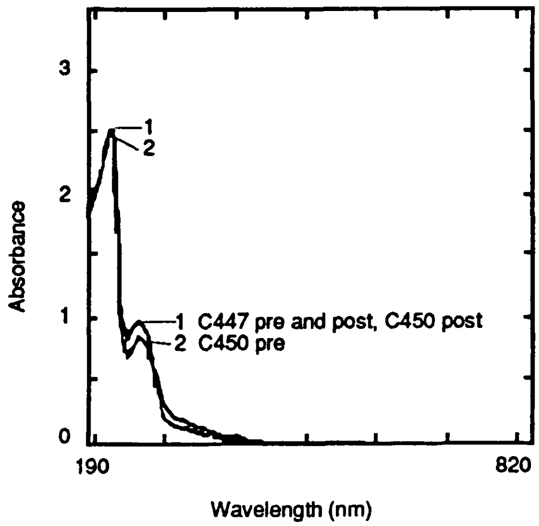
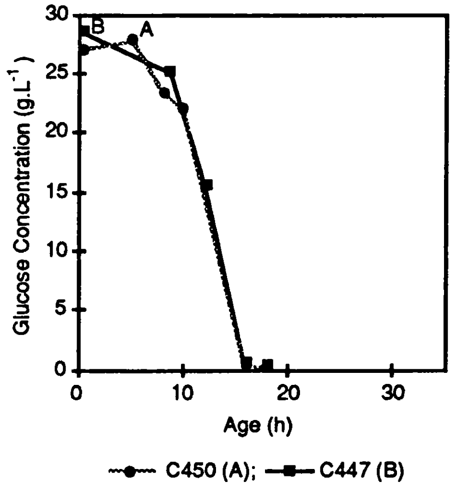


Figure 4.25: Absorption spectra of pre- and post-sterile fermentation broths of batch C450 and the standard (C447). The initial medium compositions differed and the longer cool down period after sterilisation in C450 increased the absorbance of the broth.

The effect on the media was a small increase in the broth's absorbance at 255 nm showing the potential for media degradation on scale up where harsher sterilisation conditions generally occur (Figure 4.25). The post-sterile broth's absorbance characteristics were similar to those of the standard broth. The change in the broth pH with the extended cooling period, 0.18 pH units, was similar to that observed for the standard cooling period (Figure 4.3).

The comparison of the glucose data from batches C450 and C447 highlighted a number of dissimilarities (Figure 4.26). This was a result of C450 showing an initial increase in glucose concentration. As no glucose was added to the medium during this period it was thought that this was a result of poor measurement and was ignored. In this example the conclusions of the comparative analysis tools were overridden by expert interpretation of the results. However, this was done with careful consideration of the linearised profiles and the facts presented in the difference summary.

The comparison of the carbon dioxide evolution rate profiles is shown in Figure 4.27. The match record shows a difference in the initial lines of the CER profiles, ie the lag phase, and a difference in the change in magnitude of the lines representing the fast growth period. Despite the difference in the fast growth period the maximum CER levels were similar (the



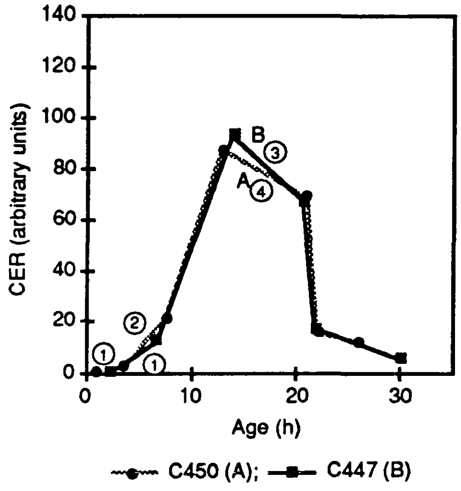
Qualitative Descriptions

Line	Dir	Mag	Dur	Slope	Start
A1	1	1	5	1	8
A2	-1	2	3	4	8
A3	-1	1	2	3	7
A4	-1	7	6	6	7
A5	0	1	2	1	1
B1	-1	1	9	2	9
B2	-1	3	4	6	8
B3	-1	5	4	6	5
B4	0	1	2	1	1

Match Record for C450/C447

Lines	Magnitude	Duration	Slope	Start
A1 → -	-	-	-	-
A2+A3 → B1	1	1	0	1
A4 → B2+B3	1	0	1	1
A5 → B4	1	1	1	1

Figure 4.26: Comparison of glucose concentration profiles from batches C450 and C447. The linearisation of the data showed an initial increase in glucose concentration in C450 but no glucose was added during the fermentation. This effect was ignored, ie the user overrode the conclusions of the comparative analysis tools.



Qualitative Descriptions

Line	Dir	Mag	Dur	Slope	Start
A1	1	1	2	2	1
A2	1	2	4	5	1
A3	1	6	5	7	2
A4	-1	2	7	4	8
A5	-1	5	1	8	7
A6	-1	1	4	2	2
B1	1	2	4	4	1
B2	1	8	6	7	2
B3	-1	3	6	5	9
B4	-1	5	1	8	6
B5	-1	2	7	3	2

Match Record for C450/C447

Lines	Magnitude	Duration	Slope	Start
A1 → -	-	-	-	-
A2 → B1	1	1	1	1
A3 → B2	0	1	1	1
A4 → B3	1	1	1	1
A5 → B4	1	1	1	1
A6 → B5	1	0	1	1

Figure 4.27: Comparison of carbon dioxide evolution rate profiles from batches C450 and C447. The comparison of C450 with C451 was similar. The null match indicates a difference in the lag phase of the fermentations. The smaller change in magnitude of line A3, when compared with B2, did not indicate a lower peak CER in C450 because the starting positions of the next lines were similar; it was in fact an effect of the differences in the lag phase.

match record showed no difference in the starting positions of the lines following the fast growth period). It was concluded that the recorded difference in the fast growth lines was a result of the differences in the lag phase. It is likely that the lag phase differences were a result of the different inocula used in the two fermentations. Similar effects were seen in the comparison of C450 with C451.

The aFGF concentration in C450 at the recommended harvest time, 22.6 normalised units.L^{-1}, was similar to that of C451 but lower than the titres in C447 and C441. No reduction in biomass production was detected by the comparison routine.

It was not possible to conclude whether these differences in the performance of batch C450 were results of the longer sterilisation hold time or if the inoculum had had some effect. Batch C449 was seeded from the same inoculum source as C450 but was run under completely different conditions and thus could not be used to check the performance of the inoculum. The possibility that the inoculum could have been the cause of the variations in batch C450 had not been considered during manual analysis; the computerised tools highlighted a fact that had been previously overlooked.

It is apparent that there is the potential for changes in the medium to occur when subjected to increased heat stress. The effects on the subsequent fermentation were small and could not conclusively be attributed to the different sterilisation conditions. It was believed that the scale up of the aFGF fermentation should not be hindered by the effects of sterilisation.

4.4.3.3 Sterilising with Glucose Present

Comparing the fermentations where glucose was sterilised in the bulk medium (C446 and C449) with those in which glucose was sterilised separately (C439, C440, C441, C447 and C451) showed many differences in the progress of the fermentations (Table 4.15). Batch C446 had been sterilised for twenty minutes and was therefore compared with fermentations that had also been sterilised for twenty minutes but in the absence of glucose. The same criteria were used for choosing the batches to compare with C449 but the sterilisation period was sixty minutes. Again C442 was not used in the comparisons and the results of comparisons with C439 and C440 were treated with caution (Section 4.4.1).

BATCH NUMBERS	DIFFERENCES IN TIME INVARIANT DATA	DIFFERENCES IN TIME VARIANT DATA	OBSERVATIONS
C441/C446	Inoculum Post-sterile pH $\Delta pH_{sterilisation}$ Harvest Time $aFGF_{sh}$, $aFGF_{sm}$ $aFGF_{vh}$, $aFGF_{vm}$	Sterilisation Glucose CER DOT pH Agitation	Batched medium differs Broth darker in C446 No abs. spec. C441 DOT fault C441
C447/C446	Tanks Post-sterile pH $\Delta pH_{sterilisation}$ $\Delta V_{sterilisation}$ Pre-sterile abs. spec. Post-sterile abs. spec. Harvest Time $aFGF_{sh}$, $aFGF_{sm}$ $aFGF_{vh}$, $aFGF_{vm}$	Biomass Glucose CER DOT Agitation	Broth darker in C446 Batched medium differs
C451/C446	Inoculum Post-sterile pH $\Delta pH_{sterilisation}$ $\Delta V_{sterilisation}$ Post-sterile abs. spec. Harvest Time $aFGF_{sh}$ $aFGF_{vh}$, $aFGF_{vm}$	Glucose CER DOT Agitation	Broth darker in C446 Batched medium differs No pre-sterile abs.spec. in C451 C451 shorter No off-line data first 10 h of C451
C439/C449	Inoculum $V_{0,sterilisation}$ Post-sterile pH $\Delta pH_{sterilisation}$ $\Delta V_{sterilisation}$ Post-sterile abs. spec. Harvest Time $aFGF_{sh}$, $aFGF_{sm}$ $aFGF_{vh}$, $aFGF_{vm}$	Sterilisation Biomass Glucose CER DOT pH Alkali addition Agitation Air Flow	Broth darker in C449 Batched medium differs Medium dilution C439 No pre-sterile abs.spec. in C439 OD-DCW correlation
C440/C449	Tanks Inoculum $V_{0,sterilisation}$ Post-sterile pH $\Delta pH_{sterilisation}$ $\Delta V_{sterilisation}$ Post-sterile abs. spec. Harvest Time $aFGF_{sh}$, $aFGF_{sm}$ $aFGF_{vh}$, $aFGF_{vm}$	Sterilisation Biomass Glucose CER DOT pH Alkali addition Agitation Air Flow	Broth darker in C449 Batched medium differs Medium dilution C440 Foaming C440 No pre-sterile abs.spec. in C440 OD-DCW correlation Temperature fault C440

Table 4.15: Difference summary of fermentations sterilised with and without glucose in the bulk medium. The abbreviations are summarised in Table 4.11.

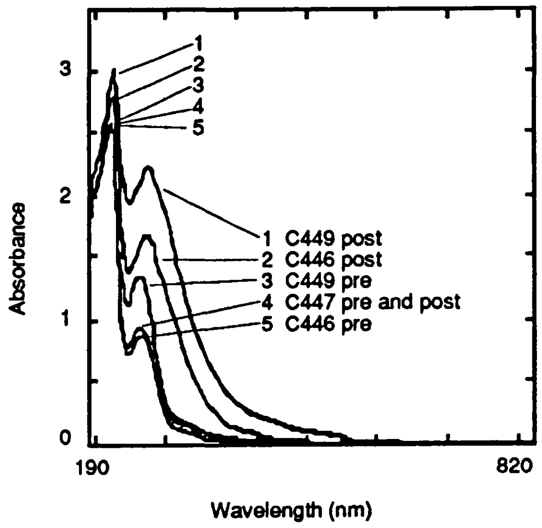
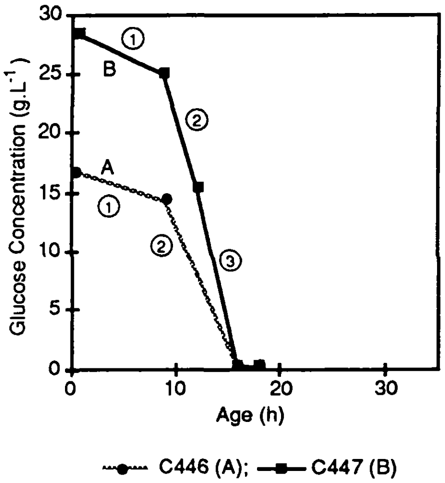


Figure 4.28: Absorption spectra of pre- and post-sterile fermentation broths of batches C446, C449 and the standard (C447). The glucose in the medium of C446 and C449 reacted with the other broth components before and during sterilisation resulting in different absorption spectra. Sterilising for sixty minutes (C449) had a greater effect than sterilising for twenty minutes (C446).

The pre-sterile fermentation broths of C446 and C449 had different absorption characteristics from the standard broths but the results were not consistent (Figure 4.28). This was most likely a result of the components of the media reacting prior to the measurement. Sterilisation with glucose *in situ* resulted in significant changes in the post-sterile absorption spectra with increased absorbance at both 210 nm and, more noticeably, at 255 nm, accompanied by a slight red shift, that is a slight increase in the wavelength of the absorption maxima, and a broadening of the absorption bands. Large changes in pH of the media were also noted (Figure 4.3). These changes in the media were a result of Maillard reactions occurring during the heating process (Section 4.1.2).

In the comments section of the data base tables it was noted that the broths of both C446 and C449 were somewhat darker after sterilisation. This was a result of caramelisation of the sugars during the excessive heating of the sterilisation process. The effect on C449 was very pronounced and altered the optical density of the broth resulting in a different correlation between optical density and dry cell weight as discussed in Section 4.3.2.1.

The comparison of the glucose profiles of C446 and C447 is shown in Figure 4.29. The starting positions of the first linear data pieces, ie the initial glucose concentrations, differed. During sterilisation of batch C446 the glucose reacted with the amino groups of the proteins thus reducing the glucose concentration in the broth. The low initial glucose concentration in batch C446 resulted in a different pattern of glucose metabolism as shown by the dissimilarities in the match record of the comparison (Figure 4.29). Similar effects were noted in all the comparisons in Table 4.15. The actual rates of glucose utilisation, that is the slopes of the linear data pieces in the glucose profiles, were not affected and thus the recommended harvest time was earlier in those fermentations that had started with lower glucose concentrations.



Qualitative Descriptions

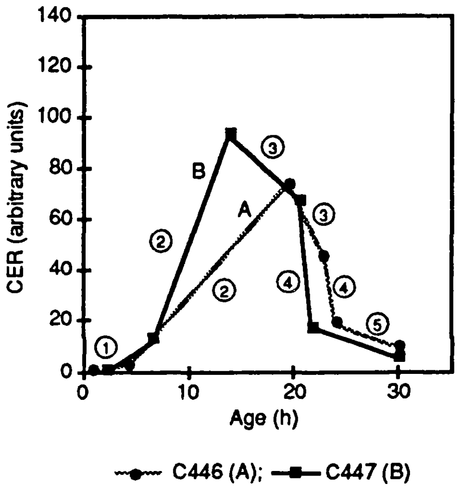
Line	Dir	Mag	Dur	Slope	Start
A1	-1	1	7	2	6
A2	-1	6	6	6	6
A3	0	1	2	1	1
B1	-1	2	7	2	9
B2	-1	4	3	6	9
B3	-1	6	3	6	6
B4	0	1	2	1	1

Match Record for C446/C447

Lines	Magnitude	Duration	Slope	Start
A1 → B1	1	1	1	0
A2 → B2+B3	0	1	1	0
A3 → B4	1	1	1	1

Figure 4.29: Comparison of glucose concentration profiles from batches C446 and C447. Sterilising with glucose in the bulk medium (C446) resulted in a lower initial glucose concentration (start positions of first lines). The maximum glucose utilisation rate (slope of line 2 in C446 and lines 2 and 3 in C447) was not affected.

An example of the differences in the CER profiles resulting from sterilising glucose *in situ* is given in Figure 4.30. The line representing the fast growth period covered a longer temporal extent in batch C446, ie it took longer to reach the maximum carbon dioxide evolution rate. The other comparisons summarised in Table 4.15 also showed this effect and indicated that the maximum CER value was generally lower when glucose had been sterilised *in situ*. These effects are a result of the lower availability of substrate.



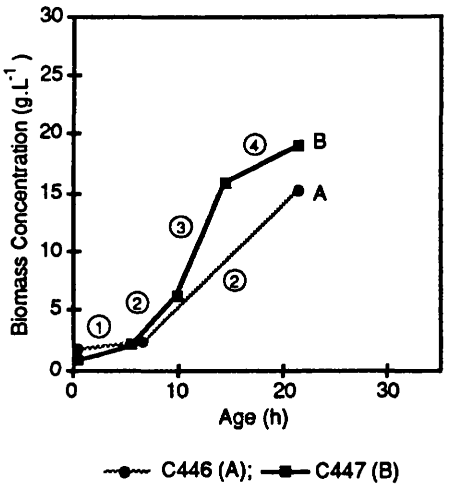
Qualitative Descriptions

Line	Dir	Mag	Dur	Slope	Start
A1	1	1	3	2	1
A2	1	7	11	6	1
A3	-1	3	3	6	8
A4	-1	3	1	8	5
A5	-1	1	5	4	2
B1	1	2	4	4	1
B2	1	8	5	7	2
B3	-1	3	5	5	9
B4	-1	5	1	8	7
B5	-1	2	6	4	2

Match Record for C446/C447

Lines	Magnitude	Duration	Slope	Start
A1 → B1	1	1	0	1
A2 → B2	1	0	1	1
A3 → B3	1	0	1	1
A4 → B4	0	1	1	0
A5 → B5	1	1	1	1

Figure 4.30: Comparison of carbon dioxide evolution rate (CER) profiles from batches C446 and C447. C446, the batch sterilised with glucose in the bulk medium, took longer to reach the peak CER as shown by comparison of the durations of the second lines.



Qualitative Descriptions

Line	Dir	Mag	Dur	Slope	Start
A1	1	1	4	1	1
A2	1	7	8	5	1
B1	1	1	3	2	1
B2	1	3	3	5	1
B3	1	5	3	6	3
B4	1	2	4	4	8

Match Record for C446/C447

Lines	Magnitude	Duration	Slope	Start
A1 → B1	1	1	1	1
A2 → B2+B3	1	0	1	1
- → B4	-	-	-	-

Figure 4.31: Comparison of biomass concentration profiles from batches C446 and C447. The duration of the last line in C446 was longer than the corresponding lines in C447, but the increase in biomass concentration (magnitude) was the same. C447 went on to achieve a higher biomass yield than C446.

The low glucose concentration also led to a reduced biomass yield in fermentations C446 and C449 (Figure 4.31).

The air flow rate was the first level of control used to counteract the lowering of the dissolved oxygen in the broth as the cells consumed oxygen. The difference in the air flow rate of C449 when compared with C439 and C440 was that the time at which the air flow began to increase was delayed in C449. The lower amount of glucose available in the broth resulted in a slower metabolism. The same effect was observed in the agitation rate profiles of C449.

As mentioned earlier a high agitation rate was employed during intervals of high metabolic activity so as to maintain the dissolved oxygen above a minimum level of 20% of air saturation. The agitation rate profiles are therefore indicative of the amount of effort required to counteract the consumption of oxygen by the organism. Unlike the other fermentations, the rate of agitation in C446 and C449 was not increased to the maximum level of 700 rpm, indicating a lower level of metabolic activity in these fermentations. This was again due to the lower initial glucose concentration.

The differences in the pH and DOT profiles in the latter parts of some of these fermentations were indicative of differences in the availability of minor substrates which was expected as many of them would have been involved in the Maillard reactions during sterilisation.

Comparison of the pH profile of C449 with that of C439 and C440 had shown a discrepancy in the temporal extent of the first lines, ie the 'control' portion of the profile. This was investigated using the data features identified in Section 4.4.2. Alkali addition to the broth stopped about six hours before the pH began to increase (Figure 4.32) thus it was not the length of pH control that was unusual but the fact that the broth remained neutral without the addition of alkali. Contrary to the previously described relationships between variables (Section 4.4.2) the maximum growth rate was maintained and the carbon dioxide evolution rate continued to increase after alkali addition had ceased (Figure 4.32) and glucose had been exhausted (not shown in Figure 4.32). This indicates that the organism was able to utilise the products of the Maillard reactions as substrates to maintain cell production. The by-products were not acidic like those of glucose metabolism. This effect had not been detected during manual analysis of the data prior to the availability of the comparative analysis tools. The linearisation of the profiles and subsequent identification of features in the data (Section 4.4.2) are very powerful aids to the understanding of the process.

The aFGF concentration at the recommended harvest time in C446, 19.0 normalised units.L^{-1}, was less than that in both C447 and C451. It was not clear if the productivity of the cells in C446 had been affected, as the amount of aFGF produced per gram of biomass was between that of C447 and C451. The lower biomass concentration in C446 had resulted in the lower volumetric aFGF yield. Sterilisation for sixty minutes with glucose *in situ* (C449) resulted in a significant decrease in both specific and volumetric aFGF yields.

These large discrepancies identified in this comparative analysis justify the extra work involved in sterilising glucose separately from the bulk medium.

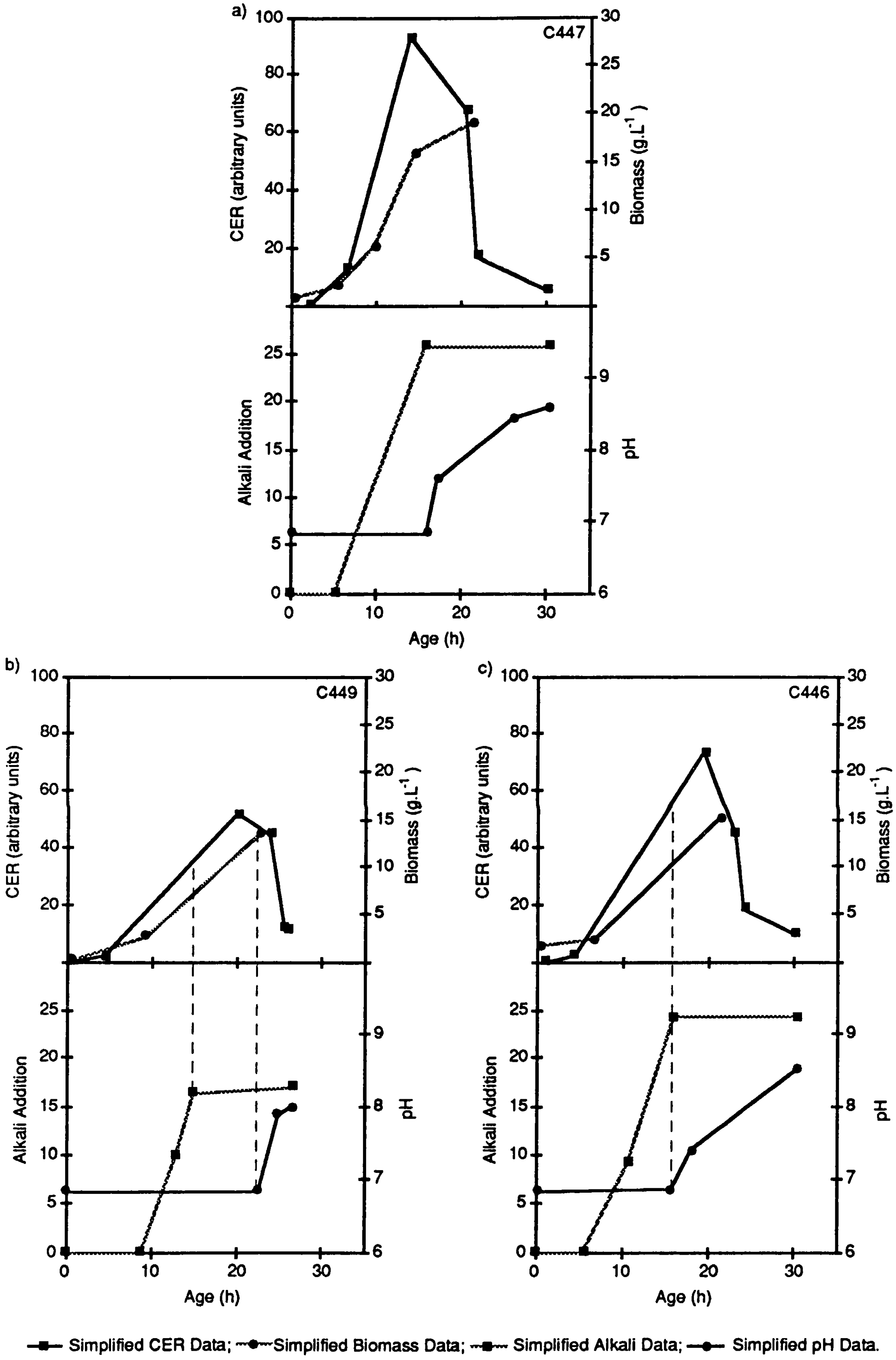


Figure 4.32: The broth in C449 remained neutral for approximately six hours after alkali addition ceased (b). In both C446 (c) and C449 the maximum growth rate and increase in CER continued after cessation of alkali addition to the broth. These effects were a result of sterilising broths C446 and C449 with glucose *in situ*. Batch C447 had glucose sterilised separately and is provided for comparison (a).

4.4.3.4 Increased Length of Sterilisation - Glucose Present

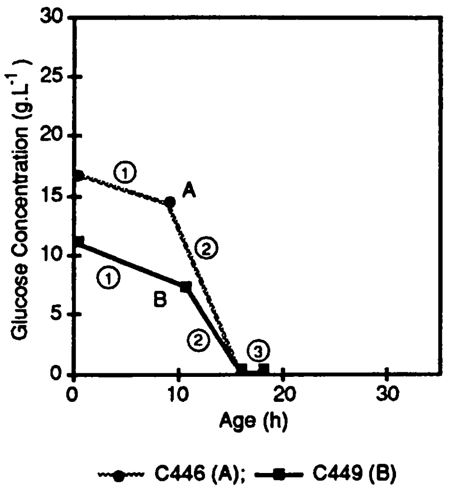
The effects of increasing the sterilisation period with glucose present in the bulk medium were more pronounced than when glucose had been sterilised separately. The differences between batches C446 and C449, sterilised for twenty and sixty minutes respectively, are summarised in Table 4.16.

BATCH NUMBERS	DIFFERENCES IN TIME INVARIANT DATA	DIFFERENCES IN TIME VARIANT DATA	OBSERVATIONS
C446/C449	Inoculum $V_{0,sterilisation}$ Post-sterile pH $\Delta pH_{sterilisation}$ $\Delta V_{sterilisation}$ Pre-sterile abs. spec. Post-sterile abs. spec. Harvest Time $aFGF_{sh}$, $aFGF_{sm}$ $aFGF_{vh}$	Sterilisation Glucose CER DOT pH Agitation	Broth dark in both Broth darker in C449 OD-DCW correlation

Table 4.16: Difference summary for fermentations sterilised over differing lengths of time with glucose present in the bulk medium during sterilisation. The abbreviations are described in Table 4.11.

The longer sterilisation time of sixty minutes (C449) resulted in larger increases in the absorbance of the broth at both 210 nm and 255 nm and a more pronounced broadening of the absorption peaks (Figure 4.28). A much larger change in the pH of the medium was also observed after sixty minutes sterilisation than occurred after twenty minutes sterilisation (1.91 and 1.01 respectively, Figure 4.3). These observations verified the fact that the amount of glucose that reacts with the amino groups of the proteins in the Maillard reactions is dependent on the extent of the heat stress.

The broth in batch C449 was considerably darker than that in C446 as a result of the caramelisation of the sugars during prolonged sterilisation. This affected the optical density of the broth and resulted in a different correlation between optical density and dry cell weight as discussed in Section 4.3.2.1.



Qualitative Descriptions

Line	Dir	Mag	Dur	Slope	Start
A1	-1	2	6	3	9
A2	-1	8	5	6	9
A3	0	1	2	1	1
B1	-1	3	7	3	7
B2	-1	5	4	6	5
B3	0	1	2	1	1

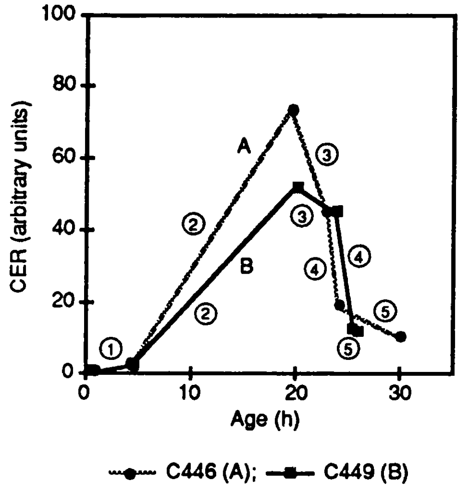
Match Record for C446/C449

Lines	Magnitude	Duration	Slope	Start
A1 → B1	1	1	1	0
A2 → B2	0	1	1	0
A3 → B3	1	1	1	1

Figure 4.33: Comparison of glucose concentration profiles from batches C446 and C449 both of which had glucose sterilised *in situ*. Sterilising for 60 minutes (C449) resulted in a lower initial concentration than sterilising for 20 minutes (C446) as shown by the start positions of the first lines.

The difference in the glucose profiles is also evidence of the longer sterilisation time allowing more glucose to react with the protein components: the glucose concentration in the broth after sterilising for sixty minutes was significantly less than that after twenty minutes (Figure 4.33). This resulted in a different pattern of glucose metabolism. The rates of glucose consumption did not alter as a result of the lower initial glucose concentration, as indicated by the similarity of the slopes of the lines in the glucose profiles (Figure 4.33), and so the recommended harvest time was earlier when the initial glucose concentration was lower.

The smaller amount of glucose available in batch C449 also resulted in a lower peak carbon dioxide evolution rate and a slower initial decline in the carbon dioxide evolution rate after the fast growth period (Figure 4.34).



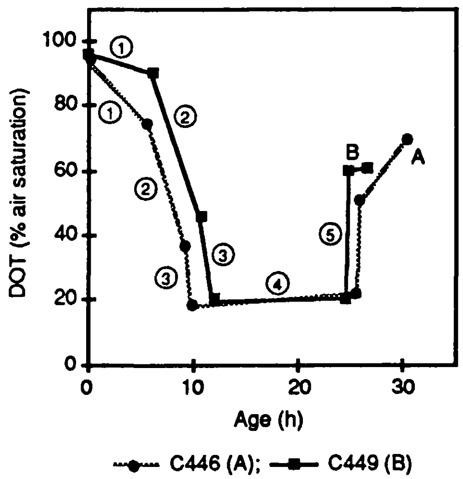
Qualitative Descriptions

Line	Dir	Mag	Dur	Slope	Start
A1	1	1	3	2	1
A2	1	8	11	6	1
A3	-1	4	3	7	9
A4	-1	4	1	8	6
A5	-1	2	5	4	3
B1	1	1	3	2	1
B2	1	7	12	5	1
B3	-1	1	3	4	7
B4	-1	5	2	8	6
B5	-1	1	1	3	2

Match Record for C446/C449

Lines	Magnitude	Duration	Slope	Start
A1 → B1	1	1	1	1
A2 → B2	1	1	1	1
A3 → B3	0	1	0	0
A4 → B4	1	1	1	1
A5 → B5	1	0	1	1

Figure 4.34: Comparison of carbon dioxide evolution rate profiles from batches C446 and C449. Glucose was sterilised with the bulk medium. Sterilising for 60 minutes (C449) resulted in a lower peak CER than a sterilisation period of 20 minutes (C446) as shown by the different start positions of the third lines.



Qualitative Descriptions

Line	Dir	Mag	Dur	Slope	Start
A1	-1	3	5	5	8
A2	-1	4	3	7	6
A3	-1	2	1	8	2
A4	0	1	14	1	1
B1	-1	1	6	2	8
B2	-1	5	4	7	8
B3	-1	3	2	8	3
B4	0	1	11	1	1

Match Record for C446/C449

Lines	Magnitude	Duration	Slope	Start
A1 → B1	0	1	0	1
A2 → B2	1	1	1	0
A3 → B3	1	1	1	1
A4 → B4	1	0	1	1

Figure 4.35: Comparison of beginning of dissolved oxygen tension (DOT) profiles from batches C446 and C449. The initial difference in magnitude describes a slower metabolic rate in C449 which resulted from a lower initial glucose concentration. The length of the control region (lines 4) was shorter in C449.

The initial decrease in dissolved oxygen concentration was also much slower in batch C449 and the length of the control period for DOT was shorter in C449 (Figure 4.35) which is again indicative of a slower metabolic rate as a result of the lower substrate availability.

The agitation rate profiles of C446 and C449 were very different. The most notable effect was that the amount of agitation required to maintain the dissolved oxygen above the required level was significantly less in C449 than in C446. The demand for oxygen in C449 was obviously much lower than in C446 because of the smaller amount of carbon source available.

Similar to the cases where glucose was sterilised separately, the lengthening of the sterilisation had no observable effect on the biomass profiles. A marked difference in the aFGF concentration at the harvest point was noted, with less aFGF being produced in the batch that had been sterilised longer (C449).

The difference in pH indicated in the difference summary (Table 4.16) was a shorter 'control' length in C446. In the previous section the neutrality of the broth in C449 after cessation of alkali addition was described as being a result of the organism consuming products of the Maillard reactions without the production of acidic by-products. It is clear from the comparison with C446 that the longer sterilisation period was required to synthesise these alternative substrates. However, a look at the biomass and carbon dioxide evolution rate profiles of C446 in relation to the cessation of glucose metabolism (Figure 4.32) shows that alternative substrates were also available in this fermentation as the metabolic activity did not slow down as was expected (Section 4.4.2).

The inocula used in batches C446 and C449 came from different sources and thus could have been responsible for some of the observed variations between the two fermentations. However, all the variations could be explained with reference to the changes in the media occurring during sterilisation thus any effect of the inocula was thought to be minor.

4.4.3.5 Summary of the Effects of Sterilisation on the aFGF Fermentation

The presence of partially hydrolysed proteins and the absence of glucose in the bulk media greatly reduced the effects of increasing heat stress during sterilisation on the subsequent fermentation. The medium did alter during prolonged sterilisation even when glucose was not present: essential nutrients were lost and replaced glucose as the limiting substrate thus affecting the pattern of growth. The yield of biomass was not significantly altered and the effect on productivity was inconclusive. The results implied that the effect of sterilisation on scale up would be minimal.

Glucose should not be sterilised with the bulk medium components as this results in a significantly inferior fermentation. Increasing the length of sterilisation when glucose is present in the bulk medium reduces the quality of the fermentation even further.

Absorption spectroscopy was shown to be a useful technique for analysing the effect of sterilisation conditions on the protein component of the fermentation broth. It was evident that the information from pH changes and absorbance changes as a result of sterilisation should be used in conjunction to ascertain the effects of the sterilisation on the medium. The changes in absorbance were indicative of both concentration and conformational changes in the protein components of the media such as those caused by the Maillard reactions and other thermal degradation products, whilst the changes in pH reflect this information to some extent but also include the effect of changes in solubility of various medium components (Corbett 1985). Further chemical analyses are required to determine the true meaning of these indicators.

4.4.4 The Effect of Inoculum Concentration

The differences between a fermentation with a 1% v/v inoculum (C451) and one with a 0.25% v/v inoculum (C452) are summarised in Table 4.17. The post-sterile absorption spectrum of C451 had lower maxima than that of C452 (Figure 4.36). This could have been a result of the larger amount of steam condensate that accumulated in C451 during sterilisation: 1.15 L in C451 compared with 0.75 L in C452. The first lines of the alkali addition rate profiles differed (Figure 4.37) indicating that C452 had a slightly longer lag phase than C451 but this was not seen in any of the other data. The harvest aFGF concentration in C452 was greater than that in C451 but the discrepancies in this measurement emphasised throughout this discussion preclude drawing any conclusions from this. No other differences were observed between these two fermentations.

BATCH NUMBERS	DIFFERENCES IN TIME INVARIANT DATA	DIFFERENCES IN TIME VARIANT DATA	OBSERVATIONS
C451/C452	Tanks Inoculum Size Post-sterile pH $\Delta V_{sterilisation}$ Post-sterile abs. spec. $aFGF_{sh}$, $aFGF_{vh}$	Alkali Addition	No pre-sterile abs.spec. in C451 and C452 No off-line data first 10 h of C451 and C452

Table 4.17: Difference summary for fermentations with different inoculum sizes. The abbreviations are described in Table 4.11.

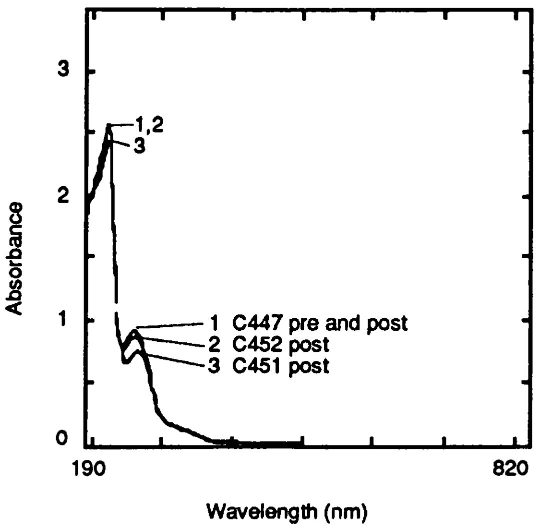
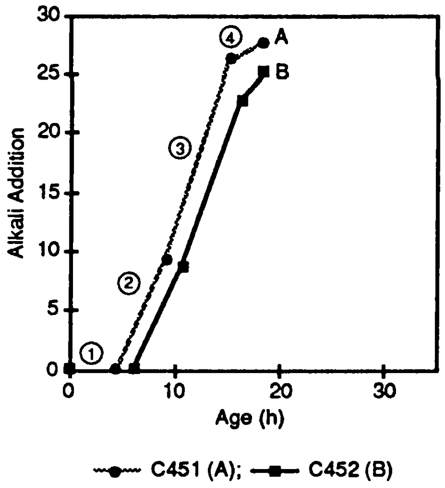


Figure 4.36: Absorption spectra of post-sterile fermentation broths of batches C451 and C452 and pre- and post-sterile standard broth (C447).



Qualitative Descriptions

Line	Dir	Mag	Dur	Slope	Start
A1	0	1	4	1	1
A2	1	3	5	5	1
A3	1	6	6	6	3
A4	1	1	3	2	8
B1	0	1	6	1	1
B2	1	3	5	5	1
B3	1	5	5	6	3
B4	1	1	2	4	7

Match Record for C451/C452

Lines	Magnitude	Duration	Slope	Start
A1 → B1	1	0	1	1
A2 → B2	1	1	1	1
A3 → B3	1	1	1	1
A4 → B4	1	1	0	1

Figure 4.37: Comparison of alkali addition profiles from batches C451 and C452. Alkali addition started later in C452 as shown by the different durations of the first lines.

The production vessel for the aFGF fermentation is not likely to be very large thus the problems of providing a large inoculum are not particularly relevant. It should also be noted that the aFGF plasmid is reportedly relatively stable (personal communication, K Gbewonyo, Merck Sharp and Dohme) thus again the provision of a smaller inoculum is not of great import. The major reason for investigating the possibility of using a smaller inoculum is that it reduces the number of seed vessels required, thus reducing the lead up time and the number of issues relating to GMP (Good Manufacturing Practice) operation. If these latter issues are of great concern, the results show that reducing the size of the inoculum is a viable option.

4.5 Scale Up

The aim of the laboratory scale aFGF experiments reported in the previous sections was to evaluate some of the effects scale up would have on the aFGF fermentations. A pilot plant scale aFGF fermentation (1900 L, working volume of 840 L) was carried out in conjunction with MSDRL personnel and provided verification of the predictions relating to sterilisation conditions. The results are not reported here.

Aside from the scale of operation there were two major differences between this run and the laboratory scale experiments: two seed stages were used, the second being run in the Biolafitte fermenter (BL4), so as to provide a 1% v/v inoculum; the medium components were of a lower quality with non-GMP grade materials being used and cerelose replacing glucose. Minor changes to the operating conditions were also made with the pressure being held at 10 psig and the dissolved oxygen being controlled above 30% of air saturation using the agitation rate, air flow rate was constant at 0.5 vvm.

The pilot scale fermentation was subjected to a longer sterilisation than expected and the heat up and cool down periods were considerably longer than in the 15 L fermentations. The complex medium components batched prior to sterilisation were of a lower quality in the large scale fermentation than in the 15 L experiments and would have contained higher levels of carbohydrates and non-hydrolysed proteins, thus it was expected that there would be slightly more nutrient degradation as a result of sterilisation. This was in fact observed as the post-sterilisation medium was slightly coloured. However, by separately sterilising all heat sensitive and heat reactive components the effect of the increased heat stress on the medium components was minimal and there was little apparent difference in the kinetics of the fermentation as predicted from the smaller scale experiments: the biomass production was very similar to the 15 L experiments and the aFGF harvest concentration was comparable to the smaller scale results.

4.6 Assessment of the Computerised Comparative Techniques as a Tool for Data Analysis in a Developmental Environment

The aims of this chapter were threefold: to determine the effects of the sterilisation regime and the inoculum concentration on the aFGF fermentation and to demonstrate the utilisation and efficacy of the computerised comparative analysis techniques presented in Chapters 2 and 3. The preceding sections used the computerised comparative techniques to fulfil the first two objectives. The performance of these tools is assessed in this section.

A manual analysis of these fermentations had been carried out prior to the availability of the computerised tools. The major conclusions reached by the two methods were the same. However, considerably more information about the process was obtained using the computerised techniques.

The procedure for applying the comparative analysis tools to experimental fermentation data was summarised in Figure 4.2: the data were prepared for comparative analysis by recording the time invariant data in the data base and linearising the time variant data using a FORTRAN program called DSIMP; the linearised data sets were then described qualitatively and compared with other data sets using two more FORTRAN programs, QUAL and MATCHER; the results from this comparison were combined with the results from comparing the data base information in a difference summary which was interpreted manually.

The data preparation steps were time consuming but were essential to the effective comparison of the data. The framework of the data base tables had been prepared prior to the experimental work and was an excellent guide to structured recording of process information. The calculations for the uncertainty levels in the descriptive data, and goodness of fit values for the time variant data, were stipulated at the beginning of the analysis and were consistent throughout. This was important as it maintained a consistent basis for the comparison of the data. It was a simple task to program a spreadsheet to perform the desired calculations on the raw data and output the required pieces of information in a form suitable for the simplification routines.

The final step in the preparation of the time variant data was the simplification of the data into piecewise linear segments. The user must exercise discretion in deciding when the linearisation of a particular variable is not practical. In these analyses the product titre and the OUR were excluded from the linearisation and comparison processes. The low frequency of product samples and the large amount of variability in each set of repeats precluded the use of the linearisation process on these data. An expert may have been able to visualise 'best fit' lines through these data for each fermentation but these would be purely speculative and would probably differ from one expert to another. The OUR data were very noisy and it was not possible to use the specified goodness of fit on these profiles. The main problem with this is that the resulting linear fits would have very little physical meaning: the linearisation of, for example, the carbon dioxide evolution rate profiles approximately divided the time course into the typically recognised growth phases of the fermentation; attempts at linearising the oxygen uptake rate profiles did not achieve this as the fluctuations in the data interfered with the fits. The user must use his/her expertise to make judgements on situations like these, the computer does not totally usurp the human's role in the analysis process.

The piecewise linearisation of the remaining variables reflected how an expert would view the data. It is important to recognise that the algorithm for the linearisation of the time variant data is generic to all fermentation processes and can be applied to any fermentation variable without the need for *a priori* knowledge of what the profiles 'should look like'. The same techniques could be used on a completely different process with the only alterations being the goodness of fit values for on-line variables monitored using different pieces of equipment. This is necessary for a process in the developmental stages as the relative positions of event times (the join points of adjacent lines), and even the number of events, may change with the different conditions being investigated.

The linearisation of the time variant data highlighted relationships between variables of the same fermentation. In some fermentations these relationships were violated and detection of these discrepancies gave further insight into the effects of some perturbations to the system. An example of this was the effect of sterilising for prolonged periods with glucose excluded from the bulk medium (Section 4.4.3.1). It was concluded that prolonged sterilisation had altered the availability of an essential nutrient, other than glucose, and thus affected the fermentation pattern. It was the linearisation of the data that highlighted these relationships thus initiating this line of reasoning. An additional computer algorithm could be developed to extract these correlations from the simplified data: each event time in each variable would be systematically compared with the event times in all other variables and a list of correlations provided. The qualitative duration descriptor could be used for this with the event times being depicted by the sum of the qualitative durations of all earlier linear data pieces. Two event times would coincide if their qualitative descriptions were within one unit of each other. During developmental work these relationships may change as the fermentation environment is changed, their inclusion in the comparison process would provide useful information.

The reproducibility of the aFGF fermentation was examined prior to investigation of the research objectives. During manual analysis of the fermentations a cursory assessment of reproducibility had been made and concluded that batches C443/C444, C441/C447, C441/C451 and C447/C451 demonstrated the reproducibility of the aFGF fermentation. A detailed examination had not been attempted because of the excessive time involved in visually comparing all the various time profiles. The computerised comparative analysis tools automatically compare the time profiles and summarise the results in match records thus reducing the time involved in determining which data differ. The analyst is then able to investigate why variations occur and is aided by the difference summary which lists all possible (identifiable) causes. Comparative analysis using the computerised tools concluded that the reproducibility of the aFGF fermentation was demonstrated by batches C447 and C451. Significant differences were found in the other fermentations that had previously been identified as similar. These differences were identified as being significant because they exceeded the uncertainty limits of the comparisons. The automated tools were more consistent than visual comparisons because the uncertainties in the data were taken into account in the comparison, manual analysis had assumed that the uncertainties were large enough to account for the variations in the data sets. One of the biggest advantages of the automated techniques is that expectations do not affect the outcome of the comparisons, all differences are recorded whether they are considered to be important or not. It is only in the interpretation of the difference summaries that expert opinion is introduced. However, because all differences between two fermentations are recorded in a single place, the interpreter would tend to consider all possibilities before making decisions on cause-effect relationships.

The reproducibility investigations were extremely important in justifying the subsequent comparative analyses between the fermentations which provided the desired information regarding the sterilisation conditions and inoculum concentrations. There would be little point in comparing fermentations which did not behave consistently under identical operating conditions; the true effects of the operating condition changes could not be determined. The additional information obtained during the examination of reproducibility improved understanding of how the process responded to various perturbations in the system, provided plausible explanations for a number of observations and identified some factors as having no effect on the outcome of the fermentation. Very little of this information had been obtained during the manual analysis of the data. It is common for a researcher to look only for the information that is being sought, in this case, the effect of the sterilisation conditions and inoculum concentration. A lot of other information is often available as was shown here. The combination of the data base and MATCHER routines facilitated the extraction of information from the fermentations that were expected to behave similarly. The summaries of the results, ie the difference summaries, enabled the tracing of cause-effect relationships.

The final step in the analysis of the aFGF fermentations was the investigation of the research objectives, ie the effects of sterilisation conditions and inoculum concentration on the fermentations. The major conclusions reached by the automated analysis were the same as those reached by manual analysis of the data, however, considerably more information was gained by using the computerised comparative analysis tools. The knowledge gained by applying the computerised tools during the reproducibility investigations helped to explain a number of the effects observed when examining the research objectives. The features of the data, highlighted by the linearisation of the profiles, were also instrumental in explaining variations in the fermentation patterns.

It is expected that even greater benefits will be achieved by applying the automated tools to the on-line analysis of fermentation data. The comparisons involving batches C441 and C442 showed the potential of the comparison tools for detecting faults in a process. The fault in the dissolved oxygen control of these two batches had been noted during operation and was thus recorded in the 'Expert Comments' but the fault was also identified by MATCHER when comparing these DOT profiles with those from other batches. This implies that the computerised comparative analysis tools could feasibly be used for the on-line detection of faults. This is discussed further in Chapter 5.

The overall aim of this work was to demonstrate the feasibility of automating the comparative analysis of fermentation data. The results of this chapter have shown that the analysis of data in a developmental environment was greatly improved by automation. The analysis was:

- thorough because all information was routinely recorded in a structured fashion;
- consistent because the uncertainty criteria were not altered between batches;
- not influenced by prior expectations of how the data should behave;
- not dependent on prior knowledge of the process and could therefore be used on completely unseen fermentations;
- not dependent on the particular data being analysed or the magnitudes of the data;
- not totally under the control of the computer, the analyst performs the final interpretation and can override any of the conclusions of the computerised tools.

A much greater understanding of the aFGF fermentation was gained by application of the computerised analysis tools. It is possible that the same information could have been obtained from a thorough, and time consuming, manual analysis but it is unlikely that the same consistency and unbiased results would have been achieved.

The data base tables and FORTRAN routines (DSIMP, QUAL and MATCHER) developed in this work are the main elements of an automated comparative reasoning programme. The intention of this work was not to provide a complete working system but to investigate the competence and benefits of the individual tools and thus demonstrate the feasibility of automating the comparative analysis process. The individual components of the comparative analysis programme were shown to perform the desired functions and they greatly facilitated the analysis of a set of twelve experimental fermentations. Professional computer systems personnel are required to fully automate the tools and to link them for optimal performance and it is likely that individual implementations of the tools will be tailored to the specific needs of the users. The obvious improvements in the analysis of experimental data justify the work involved in developing the complete system.

4.7 Conclusions

1. The effects of changes in heat stress during sterilisation on the kinetics of the aFGF fermentation were minimal when the heat labile and heat reactive medium components, including glucose, were sterilised separately from the bulk medium. Some effect on the medium was observed: the availability of an essential nutrient was

reduced and subsequently became the limiting substrate altering the pattern of the fermentation slightly.

2. Sterilising glucose *in situ* resulted in a decrease in performance of the fermentation as increases in the heat stress further decreased the performance.

3. Because of the minimal effect of heat stress the kinetics of a large scale aFGF fermentation can be approximately predicted from small scale experiments.

4. The absorption spectra of pre- and post-sterile broths, in conjunction with pH measurements, give valuable information regarding the extent of medium degradation occurring as a result of sterilisation.

5. The absorption spectra of a medium differs when components are diluted and thus could be used to detect changes in the composition of the medium.

6. Lowering the inoculum level from 1% v/v to 0.25% v/v resulted in a slightly longer lag phase but otherwise had no effect on the growth or production kinetics of the fermentation.

7. Analysis by way of the computerised comparative reasoning tools reached the same overall conclusions as manual analysis. However, understanding of the process was greatly improved by a more thorough analysis.

5 FURTHER APPLICATIONS OF THE COMPARATIVE REASONING TOOLS

The capabilities and benefits of the individual components of the automated comparative reasoning programme have been demonstrated in the retrospective analysis of twelve experimental fermentations. With this foundation it is possible to consider the application of the comparative analysis tools to other areas of the fermentation industry. This chapter outlines the extension of these tools to:

1. the detection of faults during a production run;
2. the diagnosis of faults on-line;
3. the integration of information from the fermentation and information from the downstream processing operations.

These applications were outside the scope of this study but are included here to show the potential of automated comparative reasoning as a generic and universal instrument for fermentation data analysis. A brief description of each application is given below.

5.1 On-Line Fault Detection

5.1.1 Fault Detection Techniques

In any production process aberrant operation must be detected rapidly so that action can be taken to rectify the problem. A number of fault detection techniques have been employed in industry and many more have been proposed in the literature. These will not be reviewed here but three are worth mentioning because of their wide use.

Perhaps the most widely utilised technique in a processing environment is a visual assessment of the data. In a typical fermentation plant on-line evaluation of the process is carried out by the plant manager or operators who ensure the time profiles of the monitored and calculated variables follow an expected pattern. The expected pattern may be a mental picture of what the profiles usually look like, or a standard profile based on previous successful runs. If the profile of a current fermentation deviates from the expected path then it is possible that something has gone wrong with the batch and some action may be required to rectify the problem. The problems with this manual approach were described in Chapter 1: the comparison lacks consistency from batch to batch, operator to operator and day to day, and the consideration of all variables is time consuming.

A technique commonly used in the fermentation industry for the detection of faults was introduced in Chapter 1: a 'band profile' is produced for each of the on-line variables; if the data from a current fermentation move outside this band then faulty operation is indicated. This process is usually performed manually and is dependent on the operator being vigilant.

Statistical process control is a fault detection technique used in other engineering fields (Oakland 1986, Keats and Hubele 1989) which could possibly be applied to fermentations. Essentially, statistical process control procedures differentiate between chance or random variations in a process and 'assignable causes', ie large variations that are attributable to some cause. The concept is similar to that of the band profile but utilises a statistical basis for the analysis of the data. The assumptions used in statistical process control methods are (Montgomery and Friedman 1989):

1. the data are obtained from the process via periodic samples;
2. observations are statistically independent, both between and within samples;
3. rational subgrouping is used in the selection of samples and sample sizes are larger than one;
4. the data follow some particular probability distribution; for variables data it is usual to assume normality.

Control charts, the most common technique in statistical process control, are not robust to departures from the independent or uncorrelated data assumption. The assumption of normality is of somewhat less concern. In fermentations, data from one sample to the next are correlated, for example the biomass concentration at one time is dependent on how the biomass has developed throughout the fermentation thus far. The assumption of within sample independence in fermentations is also violated: all samples are drawn from one fermenter and thus must be related. Furthermore, for fermentation on-line data only one measurement is taken at each time point thus the third assumption is not valid.

One method of dealing with serial correlation is to model the process data with an empirical stochastic model (Montgomery and Friedman 1989). The residuals from such a model would then be uncorrelated if the process is in statistical control. The usual control charting methods of statistical process control could then be applied to the residuals. The determination of models for fermentation data is not always possible and could frustrate attempts to employ statistical process control.

However, the number of assumption violations would indicate that statistical process control techniques would not be appropriate for fault detection in fermentation systems.

The comparative analysis tools provide an alternative technique for fault detection. The adaptation of the tools to a processing environment is discussed in the next section.

5.1.2 Adaptation of the Comparative Analysis Tools to Fault Detection

The use of a computerised comparative analysis process for on-line fault detection requires the definition of an average, or standard, data set as a baseline for the comparison. A standard data base record and standard time profiles are required.

The standard data base record consists of all information that is generic to the process, such as the usual operating conditions, medium components, suppliers of the ingredients, equipment used, and average values of any descriptive variables (based on all successful historical batches). The record can be updated after completion of any successful batch or after adjustments to the process.

The standard time profiles are also based on all previous successful batches. For each variable the data from the historical fermentations are combined to create one profile which is then simplified into linear segments using the algorithm coded in DSIMP (Chapter 3). The goodness of fit values described in the previous chapters can be used here but it may not be necessary to do so. When all the data are plotted on a single graph a band or envelope is formed and the outer limits of this band could be used to constrain the piecewise linearisation of the data, that is the lines must remain inside the envelope. The standard profiles can be updated after each completed run or after alterations to the process.

The fault detection process then proceeds as described in Figure 5.1. The comparison procedure is repeated at regular intervals during a fermentation and all data up to the current point are included in the analysis.

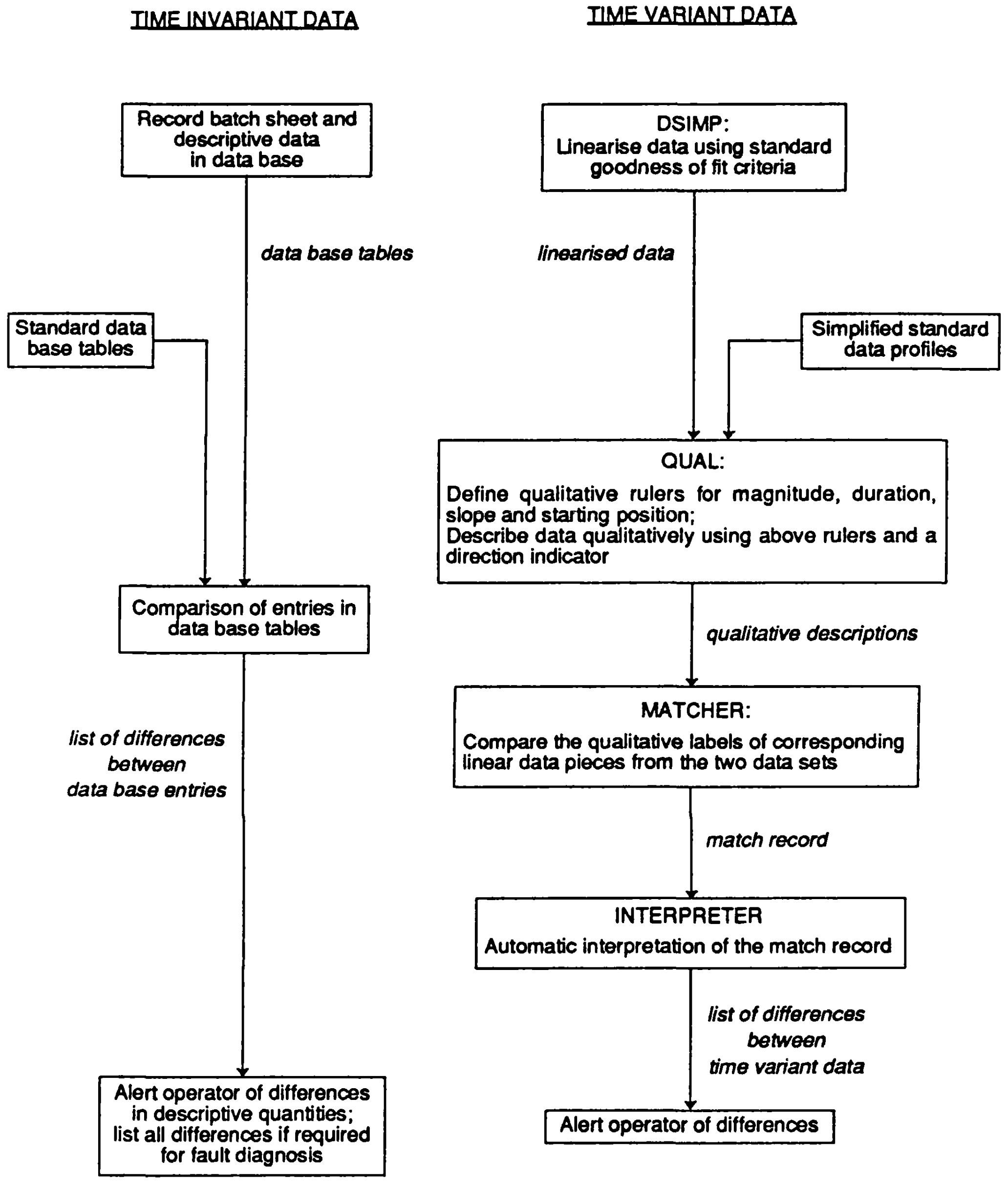

TIME INVARIANT DATA
TIME VARIANT DATA
Record batch sheet and descriptive data in data base
data base tables
Standard data base tables
Comparison of entries in data base tables
list of differences between data base entries
Alert operator of differences in descriptive quantities; list all differences if required for fault diagnosis
DSIMP: Linearise data using standard goodness of fit criteria
linearised data
Simplified standard data profiles
QUAL:
Define qualitative rulers for magnitude, duration, slope and starting position;
Describe data qualitatively using above rulers and a direction indicator
qualitative descriptions
MATCHER:
Compare the qualitative labels of corresponding linear data pieces from the two data sets
match record
INTERPRETER
Automatic interpretation of the match record
list of differences between time variant data
Alert operator of differences

Figure 5.1: Summary of the process by which the comparative analysis tools can be used for on-line fault detection. DSIMP, QUAL and MATCHER are FORTRAN computer routines that were described in Chapter 3. The data base was described in Chapter 2. The terms in italics are the outputs from each step.

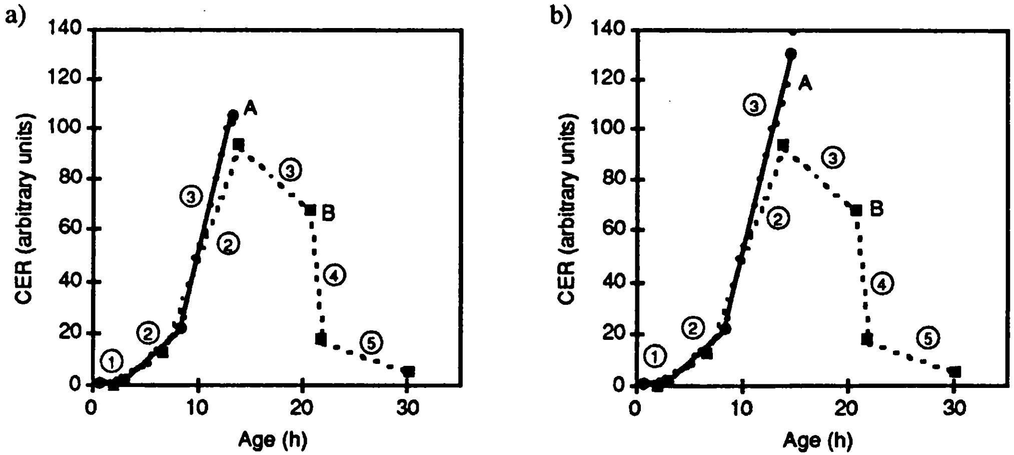
The time invariant data, ie the batch sheet information and any descriptive data, are recorded in the data base as described in Chapter 2. This process is similar to the use of batch sheets but is more structured. The information in the data base record is then compared with that in the standard data base record. Any differences in descriptive data values are immediately brought to the attention of the operator as these are indicative of variations in the performance of the process. Differences in the batch sheet information are useful in fault diagnosis which is discussed in the next section.

The time variant data are simplified by piecewise linearisation using DSIMP. All the data pretreatment discussed in Chapters 3 and 4, ie the removal of extraneous data points and the calculation of goodness of fit values (used in the fitting of linear data segments), must be carried out automatically. The simplified time profiles of the current fermentation and the standard fermentation are described qualitatively using the algorithm coded in QUAL and compared using MATCHER. The result of each comparison is a *match record* which indicates whether the lines in the profiles were similar or not. A computer routine must be written to automatically interpret the match record. When any differences between the current fermentation and the standard are detected in the match record the operator must be informed by an alarm.

An alternative to the above procedure for the comparison of time variant data is to compare the raw data of the current fermentation with the linearised standard profile. If the mean absolute deviation between the raw data and the linearised data is greater than the goodness of fit value a fault is indicated.

Process specific features of the data can also be included in the comparison procedure. In Chapter 4 correlations between event times in variables were highlighted by the linearisation of the time variant data, for example the point at which pH control ceased corresponded with glucose exhaustion. Computer routines should be developed to search for these correlations during operation. Violation of the standard correlations indicates a variation in the process and the operator must be informed.

This on-line fault detection process was simulated using data from the aFGF fermentations (Chapter 4). Batch C447 was used as the 'standard' data set and batch C442 was used to simulate an on-line process. The 'on-line' comparison of the carbon dioxide evolution rate profiles is shown in Figure 5.2. After each data point was obtained in C442, DSIMP was used to simplify the data up to and including the most recent point. The simplified profile was then compared with the complete simplified profile of C447 using QUAL and



Qualitative Descriptions

Line	Dir	Mag	Dur	Slope	Start
A1	0	1	2	1	1
A2	1	2	5	4	1
A3	1	7	4	7	2
B1	1	2	4	4	1
B2	1	7	6	6	2
B3	-1	3	6	4	8
B4	-1	5	1	8	6
B5	-1	2	7	2	2

Line	Dir	Mag	Dur	Slope	Start
A1	0	1	2	1	1
A2	1	2	4	4	1
A3	1	8	5	6	2
B1	1	1	4	3	1
B2	1	6	6	6	1
B3	-1	2	5	4	7
B4	-1	4	1	8	5
B5	-1	1	7	2	2

Match Record

Lines	Mag	Dur	Slope	Start
A1 → -	-	-	-	-
A2 → B1	1	1	1	1
A3 → B2	1	0	1	1
- → B3,B5	-	-	-	-

Lines	Mag	Dur	Slope	Start
A1 → -	-	-	-	-
A2 → B1	1	1	1	1
A3 → B2	0	1	1	1
- → B3,B5	-	-	-	-

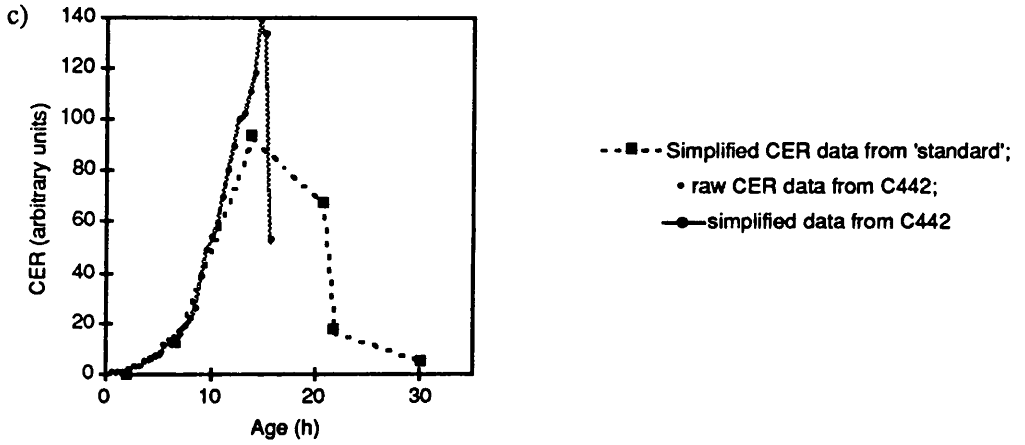


Figure 5.2: On-line detection of faulty carbon dioxide evolution rate data. a) the match record shows the fermentation was proceeding well at 13.13 h. The difference in duration between lines A3 and B2 does not indicate a problem because the duration of A3 was less than that of B2 and is expected to continue. b) the match record shows a deviation from the normal profile at 14.63 h, the magnitude of line A3 was larger than that of B2; the expected event time, ie a change to a negative slope, had been passed. c) manual observation detected the error two measurements later at 15.63 h.

MATCHER. A fault was detected in C442 at 14.63 h. Manual analysis of the CER data during actual operation of batch C442 had not detected the fault until 15.63 h, two samples later. It should be remembered that more than one variable must be consulted to corroborate the detection of faulty operation and the set point controlled variables must be included in this analysis. In batch C442 both the temperature and the dissolved oxygen profiles deviated significantly from their set points thus confirming the presence of a fault in the process.

The example demonstrates the ability of the computerised analysis techniques to detect faults during on-line operation. These techniques overcome the shortcomings of a manual process for the detection of faults: all data can be examined with relative ease, the presence of an expert is not required and the comparison process is consistent from batch to batch and from day to day. Another important advantage of the comparative analysis tools is that the line of reasoning followed in the fault detection process is easy to understand and is therefore more likely to be accepted by operating staff.

When the time variant data are considered alone, the computerised comparison techniques have two main advantages over the band profile method for fault detection. Firstly, the comparison of all variables is possible in a much shorter period of time. Secondly, the correlations between events in the variables, detected through the linearisation of the data, is a powerful tool in the detection of aberrant behaviour but cannot be exploited in the band profile method. These advantages alone should warrant the adoption of the comparative analysis tools. The data base forms a useful addition to the fault detection process as it allows the inclusion of time invariant data. If fault analysis is required, then the computerised techniques have considerable advantages over manual analysis, these are described in the next section.

The tools developed in this work provide the building blocks for an on-line fault detection process. A complete working system requires the development of computer routines to:

1. automatically remove extraneous data from the raw data files and, where necessary, calculate goodness of fit values for the linearisation routine;
2. interpret the match records, ie the results of the comparison of the time variant data;
3. identify correlations between the event times of variables in the standard fermentation and search for these correlations in in-progress fermentations.

An overall management routine is also required to schedule the operations, manage the data files, link the individual modules and provide the necessary output.

5.2 On-Line Fault Diagnosis

A second area in which the comparative analysis tools can be employed is the diagnosis of faulty operation. Once a fault has been detected in a process it is then necessary to find the cause of the fault, determine whether or not it is detrimental to the process, and decide if some action needs to be taken to rectify the problem. This process of fault diagnosis is generally based on previous experience of similar faults.

Much of the current work in fault diagnosis techniques is concentrating on the development of expert systems (Chen *et al.* 1989, Halme 1989, Karim and Halme 1989, Cooney *et al.* 1991, Morris *et al.* 1991) and neural networks (Cooney *et al.* 1991, Morris *et al.* 1991). These were discussed in Chapter 3. The disadvantages with expert systems include the extensive amount of work required to set up the system, they are totally unpredictable in situations for which rules are unavailable, and a different expert system must be created for each process. Neural networks suffer from being black box systems, ie the line of reasoning is not available to the operator or analyst, and, like the expert systems, a different neural network is required for each process.

Manual fault diagnosis is probably still the most widely used technique in the fermentation industry. The process requires experienced personnel and suffers from the usual problems of human bias and inconsistencies.

The comparative analysis tools, in combination with a new technique called Case Based Reasoning (CBR), are an alternative means of detecting and diagnosing faults in a processing environment.

Experts often reason about a current problem by determining similarities between it and an actual previous problem that has been encountered, the reasoning used in the previous problem is used as a precedent for the current problem (Koton 1988). This is the fundamental premise behind a branch of artificial intelligence known as Case Based Reasoning (CBR). Applications of CBR described in the literature include menu planning (Kolodner 1987), the diagnosis of medical complaints (Koton 1988) and the resolution of disputes (Kolodner and Simpson 1987). The common element is that all the problems are solved by reference to previous similar problems, or cases.

The concept of CBR could potentially be applied to the analysis of aberrant behaviour in a fermentation process. The 'cases' would be completed fermentations and the 'case base' would contain all the time variant and time invariant data from each completed fermentation. The computerised comparative analysis techniques would be instrumental in the implementation of Case Based Reasoning in fermentations.

During the developmental work and initial stages of a new fermentation process the case base would be developed. A standard fermentation would be identified using the data from all successful fermentations as described in the previous section. All non-standard fermentations, as determined by the comparative analysis tools, would be recorded in the case base as examples of faulty operation (cases). Some 'generic' faults may be identified during this process, for example if an increase in operating temperature generally resulted in a decrease in product yield this would be noted in the comments section of each case containing a temperature fault.

The fault diagnosis procedure involves:

1. detecting a fault;
2. searching the case base for a historical fermentation with a similar fault;
3. using the outcome of the historical fermentation to predict the outcome of the current fermentation;
4. deciding on the required course of action.

The detection of a fault proceeds as described in the previous section and summarised in Figure 5.1.

Once a fault has been detected the case base is searched to find a historical fermentation which exhibited a similar fault. The case base is likely to be large and thus some search parameters must be defined. The first step is to investigate the differences between the time invariant data, ie the data base information, of the current run and the standard as these may indicate the cause of the fault. These causes can then be used to index the case base, ie to find historical batches in which the same fault had occurred.

If the cause of the fault is not clear or the search is not successful, the faulty time variant data are used to index the case base. The fault detection procedure would have indicated which time variant data differed from the standard and these are then compared with the time variant data of the historical batches in the case base. Only those historical cases that had variations in the same variables as the current case are investigated.

Once a similar historical case has been found it can be used to predict what will happen in the current fermentation. If the behaviour of the historical fermentation, up to the current point, is similar to that of the in-progress one, a prediction of the behaviour of the current fermentation can be made by inferring the behaviour of the historical fermentation. The prediction required would normally be related to the productivity of the fermentation, for example the effect of the fault on the product yield. This information will be found in the time variant data or data base information of the historical batch.

The outcome of the prediction is then used to indicate whether or not action should be taken to rectify the situation or if it is advisable to terminate the batch. Expert comments provided for the historical batch may indicate what action is required.

Upon completion, the current batch is added to the case base as another example of faulty operation.

A significant amount of work is required in the development of a Case Based Reasoning tool for fermentations but once developed it can be used for all processes with each different process being sectioned into its own separate case base. The data base structure described in Chapter 2 would be the basic building block of the case base structure and the linearised profiles produced by DSIMP (Chapter 3) would be included within the structure. The case based reasoner must be linked to the fault detection system described in the previous section and must have access to the comparative analysis tools for the comparison of time variant data. The time taken to retrieve a similar case from memory is an important consideration for on-line work and could be prohibitory to the application of a case based reasoner.

The case based reasoning tool, in conjunction with the comparative analysis tools, offers significant advantages over the typical manual fault analysis: all data would be considered in the analyses, the absence of an expert would not be critical to the analysis of a fault, and the assessment of faults would not be influenced by human bias and inconsistencies. An important aspect of the Case Based Reasoning tools is that the operator can follow the line of reasoning in the diagnosis of any observed fault and can therefore make an informed judgement on the applicability of the result.

5.3 Implications for Downstream Processing

The comparative reasoning tools can be used to improve the information link between fermentations and their subsequent downstream processing operations.

In many research environments the fermentation recovery operations are developed separately from the fermentation itself: the fermentation technologists concentrate on finding the relationships between fermentation operating conditions and fermentation performance whilst the downstream engineers look at altering the conditions of the recovery operations so as to optimise recovery of the biomass or product as required. This is not an ideal situation. The performance of downstream processing operations is heavily dependent upon the quality of the output from the fermentation step as a sequence of product recovery operations cannot perform optimally if the feed varies from the design specification. The current lack of suitable on-line sensors precludes complete characterisation of a fermentation broth during operation and hence the ability to control a fermentation to achieve a consistent output is severely limited. Relationships between fermentation and recovery performance are poorly documented but, if available, could be used to improve the recovery of product from 'non-standard' fermentation broths.

In Chapter 4 the comparative reasoning tools were applied to a set of fermentation experiments to determine the effects of various operating condition changes on the progress of the fermentation. It is possible that these operating condition changes and other, perhaps unintentional, changes may have some effect on the performance of the downstream processing operations. The expansion of the comparative analysis techniques to include recovery data would allow the detection of cause-effect relationships, not only within the downstream operations, but also between fermentation and downstream processes.

During a production fermentation, knowledge accumulated from previous fermentations enables an operator to evaluate the current situation and adapt operating conditions so as to provide as consistent an output as possible (Section 5.2). The inclusion of downstream data within this 'case base' of historical fermentations would enable correlations between recovery performance and fermentation performance and could be used to indicate adaptations to downstream operations to cope with non-standard feed streams (ie fermentation output) based on previous examples.

Putting these ideas into practice would ensure better utilisation of the data resources available and would enable better communication between the different teams of workers. It is envisaged that this would enable a greater understanding of the processes studied and thereby improve the operation of the processes.

6 SUMMARY AND CONCLUSIONS

The aim of this work was to investigate the feasibility of automating the comparative reasoning techniques used by an expert in the analysis of fermentation data. Automation was considered desirable to overcome the problems inherent in manual analysis of the data:

1. not all data are routinely recorded;
2. it is difficult to keep track of all pieces of information when a large number of fermentations are being considered;
3. qualitative data are often ignored in the comparisons;
4. data are not readily available to other researchers;
5. the comparison of time variant data lacks consistency from one person to the next and even from day to day;
6. the consideration of all variables is time consuming.

Further requirements were specified prior to development of the automated comparative reasoning tools:

1. the tools must be applicable in both research and production environments;
2. no prior knowledge of the process should be required for use of the tools in a developmental environment;
3. the line of reasoning used must not be obscured from the user as he/she needs to interpret the results;
4. the system must not be process specific.

The tools developed for the comparative analysis of fermentation data were:

- a relational data base;
- DSIMP, a computer routine to simplify time variant data into piecewise continuous linear segments;
- QUAL, a computer routine to describe the linear data segments using the qualitative terminology of an expert;
- MATCHER, a computer routine to compare the qualitative description of one time profile with that of another time profile.

The analysis process begins with the preparation of the data in which the time invariant data are recorded in the data base and the time variant data sets are linearised by DSIMP. The simplified data from two fermentations are then described qualitatively using QUAL and compared using MATCHER. The results from this comparison are combined with the results from the comparison of the data base information of the two fermentations and interpreted manually.

The comparative reasoning tools were used in the analysis of a set of laboratory scale acidic fibroblast growth factor (aFGF) fermentations. Comparison of the data sets using these tools showed firstly that the aFGF fermentation was reproducible. A number of discrepancies in the data sets were explained by recourse to the comparison of the data base information. Most of these findings had not been identified in a previous manual analysis of the same data. The comparison tools were then used to show that a low inoculum concentration had very little effect on the process and, if glucose were sterilised separately from the bulk medium, the sterilisation conditions that would be experienced on scale up would not affect the performance of the fermentation. These conclusions concurred with those reached by manual analysis thus demonstrating the efficacy of the comparative analysis tools in a developmental environment.

The comparative analysis tools can also be applied to the on-line detection of faults and, with the aid of a Case Based Reasoning system, could feasibly be used to diagnose the cause of faults during production.

Data from downstream processing operations can be included in the comparative reasoning process enabling a greater understanding of the effects perturbations in the fermentation environment have on the subsequent recovery operations. This will result in a greater understanding of the downstream operations and should lead to improvements in the overall process.

The automated techniques have a number of advantages over a manual analysis process:

1. the data base guides the recording of the batch sheet data ensuring that all information is available for analysis;
2. the detection of differences between the data base information of two fermentations enables identification of possible causes of discrepancies in the data;
3. the identification of cause-effect relationships is facilitated by presenting all information to the analyst for consideration;

4. the comparisons of numerical data are more consistent because of the inclusion of uncertainty values which remain constant for each variable from one batch to another;
5. the comparison process is not influenced by prior expectations of what the data should look like;
6. automation of the comparison of time variant data removes the need to overlay all the various time profiles and the analyst is therefore able to spend more time determining why there are discrepancies in the data rather than establishing that differences are present;
7. the linearisation of the time variant data highlights correlations between events in variables of the same fermentation, violation of these correlations is useful in detecting and explaining aberrant behaviour.

The comparative reasoning tools can be applied to any fermentation process because they require no *a priori* knowledge of the process. This prevents loss of salience of the tools as new processes are developed or old ones improved and allows the tools to be applied to developmental processes.

The tools for the comparison of time variant data can be applied to any data sets that reflect the dynamics of the process. They also have no dependence on the magnitude of the data. However, extremely noisy data and data in which the sampling frequency is low relative to the dynamics of the process can not be simplified by the linearisation routine. These problems would be a barrier to any data analysis programme. Variables that are under single value set point control throughout a fermentation cannot be compared using MATCHER and are treated in the data base.

The line of reasoning used by the routines is easily followed: a listing of appropriate qualitative descriptions and match records enables the user to see the differences in the time variant data and the reporting facilities of the data base ensure that all information is readily available to the analyst as required. This is extremely important if the new analysis tool is to be accepted by the users.

Human intervention is required in the final interpretation of the comparisons. This is an area where the human expert is considerably more adept than the computer and no advantages would be attained by automation.

The tools developed in this work do improve the comparative analysis of fermentation data. The consistency of the comparisons, the extra information obtained as a result of applying these techniques in a developmental environment, and the potential benefits of applying them in a production environment, justify the work required in setting up a fully integrated comparative reasoning package.

APPENDIX 1: Cubic Spline Smoothing of Fermentation Data

A number of fermentation researchers use curve fitting techniques to aid data analysis. In the early stages of this work a cubic spline fitting technique was investigated as a tool for the comparative analysis of fermentation data. The methods were found to be inadequate for the desired purposes: the methods were complicated firstly by the need to know the uncertainties in the data and, secondly, by the specification of smoothing parameters; it was also found that simply smoothing the data did not facilitate the comparison process. The cubic splines techniques are described below for completeness. A definition of cubic splines is given in Section A1.1 along with a list of references. The tools used in this work are described in Section A1.2. Practical details of the cubic splines software used are presented in Section A1.3. The cubic splines routines were applied to fermentation data to determine whether or not they would aid comparative analysis of the data; the results are presented and discussed in Section A1.4 and the conclusions of the work are summarised in Section A1.5.

A1.1 Introduction to Cubic Splines

The mathematical description of an experimental curve is a form of function approximation. There are a number of methods available for the approximation of functions, eg polynomial approximations: least squares fit, Legendre polynomials, Chebyshev polynomials, spline functions; and non-polynomial approximations: rational functions, exponential functions, logarithmic functions. The most common approximations are the polynomial functions.

A problem that often arises when trying to approximate a function by means of a polynomial of high degree is that the polynomial starts to oscillate between data points which is undesirable if the polynomial is to be used for interpolation and it makes numerical differentiation meaningless. Spline approximations can be used to avoid this problem as splines are more stable than polynomials with less possibility of wild oscillations between data points.

Spline functions are essentially a chain of polynomial arcs of a specified degree (d). The arcs link together smoothly (with continuity of the first d-1 derivatives) at a set of chosen abscissa values known as knots.

Two classes of splines are distinguished: interpolating splines and smoothing splines. An interpolating spline is a polynomial between each pair of adjacent data points, ie the spline function interpolates between data values. The polynomials are connected together at the data points and the derivatives (up to a degree specified by the order of the spline function) of the functions on both sides of a data point are set to be equal at that point, thus producing a smooth, piecewise continuous curve. A smoothing spline is similar to an interpolating spline except that the errors in the measurements are taken into consideration and the spline is not forced through every data point. In general a smoothing spline is constructed using a trade-off between goodness of fit and smoothness of the curve. Interpolating splines are in fact a special case of a smoothing spline with the goodness of fit set to 100% or smoothing factor set to zero. Smoothing splines are more relevant for experimental data where the presence of uncertainties and noise precludes the use of interpolating splines.

Cubic splines are the most common splines used in curve fitting. A cubic spline is a continuous function which has continuous first and second derivatives and each interval between knots, the points at which the polynomials meet, is represented by a polynomial of degree not exceeding three. The theory of cubic splines approximation can be found in Greville (1969), de Boor (1978), Schumaker (1981), and Silverman (1985).

Spline functions have been used in a number of instances to smooth experimental data (Reinsch 1967 and 1971, Wold 1974, Dierckx 1975, Craven and Wahba 1979, McWhirter 1981, Wegman and Wright 1983, Silverman 1985, Hutchinson and de Hoog 1985, Oner *et al.* 1986, Buono *et al.* 1986). The general procedure involves the determination of the spline function in each interval such that a cost function, which controls the smoothness of the curve and the goodness of fit, is minimised. The cost functions are usually based on those described by Reinsch (1967) and Craven and Wahba (1979).

There are two known examples of cubic spline routines developed specifically for fermentation data analysis: Erickson and co-workers (Oner *et al.* 1986, Buono *et al.* 1986) and Thornhill and colleagues at University College London (unpublished). The first of these is described briefly below and the second is described in more detail in Section A1.2 as these routines were further developed and utilised in this work.

Oner *et al.* (1986) and Buono *et al.* (1986) used carbon and available electron balances to determine the best fitting cubic spline to biomass, substrate and product concentration data from some experimental fermentations. The determination of carbon and available electron balances requires calculation of derivative quantities which in turn requires that a smooth curve be drawn through each of the data sets. Cubic splines were used to provide this smooth curve and the derivatives thereof. Closure of the carbon and available electron balances was used to determine how good the spline fits were, the best fit was that which resulted in the smallest deviation from one in the balances. The procedure was as follows. A smoothing parameter must be provided for each data set to guide the spline fitting. A window of possible smoothing parameter values was defined for each data set based on the number of data points. Different smoothing parameters result in different curves and different combinations of smoothing parameters for the different variables result in different values for the parameters of the balance equations. A set of numerical experiments was defined using a central composite design: in each experiment a different combination of smoothing parameters was used, the spline function for each variable was determined using Craven and Wahba's cost function (Craven and Wahba 1979), the derivative functions were calculated and appropriate values substituted into the balance equations. The closure of the balance equations was used to set up a response surface, the result from each numerical experiment contributed to a point on the surface. The minimum point on the surface identified the best closure of the balances and thus indicated the best combination of smoothing parameters (and best spline fits) for the variables under investigation.

Although the techniques gave adequate results there were two inherent problems. Firstly, the use of balance equations assumes considerable knowledge of the process as the kinetics of growth, substrate consumption and product formation must be known and it is not always possible to assume Monod kinetics as these workers did. When complex media are used it is usually not possible to obtain sufficient information to satisfy balance equations. It is also necessary to include variables other than those considered here to complete the balances, for example the carbon dioxide in the exit gas must be considered in a carbon balance. Secondly, each data point must be weighted for the spline determination. The procedure used by Buono *et al.* (1986) and Oner *et al.* (1986) was to hand-draw a smooth curve through the data and use the standard deviation of the raw data from this curve as the weighting for each data point. This precludes the use of this technique for on-line work and introduces a significant amount of human bias into the procedure.

A1.2 Cubic Spline Theory (MSPLIN)

The software package MSPLIN was developed by staff and students at University College London for the fermentation researcher who requires a smooth and continuous representation of fermentation time variant data. The routines fit cubic splines to data using Dierckx's algorithm (Dierckx 1975) with some modifications. The routines are more general than those described by Oner *et al.* (1986) and Buono *et al.* (1986) and can be applied on-line and off-line to fermentation data.

A brief account of the theory used in MSPLIN will be presented here. The practical aspects of using MSPLIN will be described in Section A1.3 and the applicability of MSPLIN as an aid to the comparative analysis of fermentation data is discussed in Section A1.4 with reference to some examples from *Escherichia coli* fermentations.

The unknown 'smooth' curve, ie that underlying the data, is designated g(x). The measurements (y) are then:

$$y(x) = g(x) + e(x) \qquad \text{(A1.1)}$$

where e(x) are the errors on each measurement.

The spline approximation to g(x) is s(x) which is defined on a set of m knots, $t_1<t_2<...<t_m$, in which the first and last x-coordinates of the data set coincide with knots t_4 and t_{m-3} respectively. It takes the form:

$$s(x) = \sum_{i=1}^{m-4} c_i b_i(x) \qquad \text{(A1.2)}$$

The coefficients of the spline equation, c_i, are determined by minimising a cost function described below.

The basis functions, $b_i(x)$, are known as normalised b-splines (Cox 1972, de Boor 1978). Each of these basis functions spans five knots, $t_i,t_{i+1},...,t_{i+4}$, and takes the form of a single localised positive hump which is zero when $x \le t_i$ or $x \ge t_{i+4}$ and attains its peak at $x=t_{i+2}$. The area between each knot is called an interval. Each of the intervals covering the data set, ie those between t_4 and t_{m-3}, has four b-splines contributing to the function describing that interval.

The spline function in each interval is chosen such that a cost function is minimised. This cost function controls the smoothness of the curve and the goodness of the fit. The cost function used in MSPLIN is:

$$p \sum_{i=1}^{n} w_i(s(x_i)-y_i)^2 + \sum_{r=5}^{m-4} [d_r]^2 \qquad \text{(A1.3)}$$

where $$d_r = (s^{(3)}(t_r))^+ - (s^{(3)}(t_r))^- \qquad \text{(A1.4)}$$

The superscripts + and - indicate the value of the third derivative either side of the knot t_r. In Equation A1.3 n is the number of data points. The weighting function, w_i, is the inverse of the variance of the errors, or noise, in the measurements. The error term is made up of a fixed error and a relative error.

The first term in Equation A1.3 is the distance between the fitted curve and the raw data points and is a measure of the goodness of the fit. The second term describes the size of the discontinuities in the third derivatives of the function and is a measure of the smoothness of the curve. The smoothing parameter, p, controls the trade-off between the smoothness of the curve and the goodness of the fit.

The distribution of the goodness of fit term is chi-squared when s(x) is a good approximation to the underlying function g(x). The smoothing parameter p is chosen such that the discontinuities in the third derivatives at the knots are minimised subject to the constraint that the weighted goodness of fit term in Equation A1.3 is less than the chi-squared value corresponding to a user-specified percentage fit. If the smoothness criterion cannot be met the number of knots is increased until a suitable value can be found. The coefficients in the spline equation are calculated by rearranging Equation A1.3 (details not presented here) and the appropriate values of the spline function, s(x), calculated using Equation A1.2.

The Dierckx algorithm implemented in MSPLIN is only useful if the noise statistics of the measurements are known. In other instances the GCV (Generalised Cross Validation) method of Craven and Wahba (1979) (and later improved by Utreras 1980, Silverman 1984, Hutchinson and de Hoog 1985 and de Hoog and Hutchinson 1987) would be more applicable. When derived variables are required the GCV method is more accurate as shown by an analysis of the errors between the 'true' function and the spline function (N.F. Thornhill, UCL, private communication).

A1.3 Practical Aspects of MSPLIN

The software package MSPLIN will perform the following tasks:

1. smooth a set of raw data using cubic splines;
2. smooth a portion of a raw data set;
3. divide two sets of raw data producing a smoothed result;
4. calculate first and second derivatives of the smoothed curves;
5. determine the integral of a smoothed curve between specified limits;
6. augment a low frequency data set using a related high frequency data set so that a more accurate representation of the low frequency data set can be obtained;
7. plot the resulting smoothed curves to the screen or on hard copy.

A fermentation user menu has been provided which allows the user to select options such as 'calculate smoothed specific rate data' or 'calculate smoothed RQ (respiratory quotient) data'. These selections automatically trigger the appropriate routines: specific rate data requires finding the first derivative of the data and then dividing it by the biomass concentration (or other variable); calculation of RQ requires division of the CER data by the OUR data.

The routines are written in FORTRAN and are available for use on the MicroVax 2000 and on IBM and IBM compatible personal computers.

The raw fermentation data are read from files where they are stored in two column format (time,value). The user must supply the percentage error in the data, the resolution of the data, ie the fixed error term, and a goodness of fit for the smoothed profile. A goodness of fit of 100% results in an interpolating spline and a goodness of fit of 0% results in a single cubic polynomial fit.

The user can specify one of two output formats: the first lists the smoothed fermentation data at regular time intervals in two column format (time,value) and can be read to a file for storage; the second details the parameters of the spline equations which was useful during the development of the routines.

A1.4 Results and Discussion

MSPLIN was used to smooth some time variant data from the *Escherichia coli* fermentations described in Chapter 4 (the details of the experiments are not required to understand the following work). The results are presented and discussed here in relation to their applicability for use in the comparative analysis of fermentations.

The use of MSPLIN requires some knowledge of the errors in the experimental data. Unfortunately accurate error evaluations are not usually available and the user must guess at appropriate values. In this work the initial fitting for each variable utilised the results of the error analyses described in Chapter 4. If the resulting fit was deemed inadequate adjustments were made until a suitable fit was obtained.

The characteristic S-shaped curves followed by many variables in batch fermentations are easily described by splines in most cases. An example is the *E. coli* biomass concentration curve shown in Figure A1.1. The depletion of the substrate concentration also follows an S-shaped pattern as shown in Figure A1.2. In this instance the calculated uncertainties in the raw data did not allow a suitable approximation to the data (Figure A1.2(a)) and alterations were required (Figure A1.2(b)).

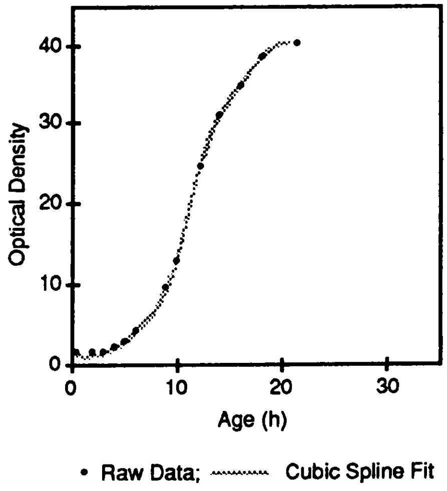


Figure A1.1: Cubic spline fits to optical density data from fermentation batch C447. The user inputs were: relative error 2%, resolution 0.1, and goodness of fit 80%.

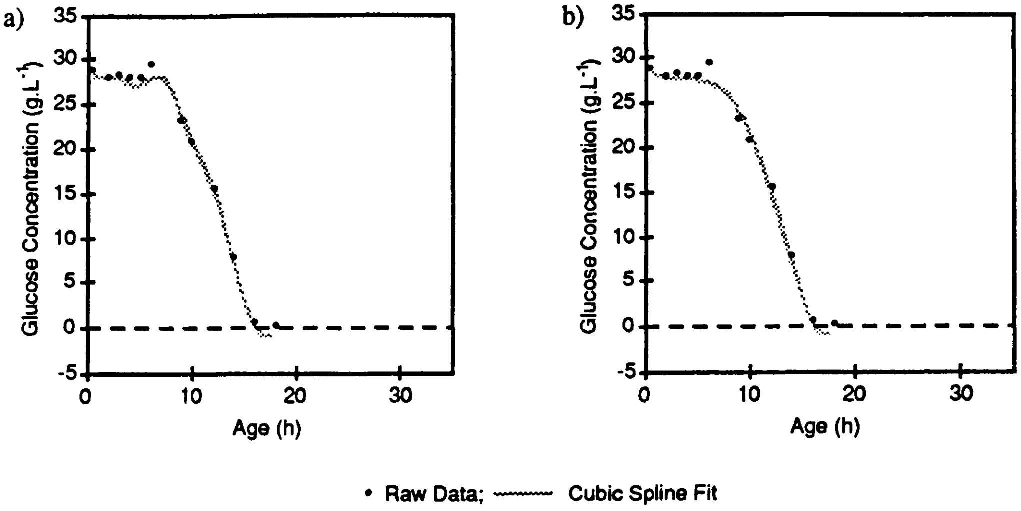


figure	(a)	(b)
relative error	2	5
resolution	0.1	0.1
goodness of fit	10	50

Figure A1.2: Cubic spline fits to glucose data from fermentation batch C447. The user inputs are summarised in the table and their effects seen in the figures. The fit in (a) is too tight. The fit in (b) is acceptable.

Carbon dioxide evolution rate (CER) profiles are somewhat more complex than the S-shaped curves and a number of problems were encountered when trying to approximate these profiles with cubic splines. The relative error in CER data is estimated to be 4% (Chapter 3). This figure was used in the MSPLIN routines. Trial and error was used to determine appropriate values for the resolution and goodness of fit. For one data set (C447) values of 0.5 for the resolution and 10% for the goodness of fit were found to give a reasonable fit (Figure A1.3(a)).

The effect of each of the input values on the resulting fit was evaluated by changing each value in turn by a small amount. The results are shown in Figure A1.3. A number of different combinations were found to give acceptable fits, although each of the fits was slightly different. This would make it difficult to determine which fit is best and the result would probably vary from user to user. Figure A1.3 also shows that small changes in any of the input values can have drastic effects on the resulting spline approximation. This makes finding a suitable fit by trial and error a difficult task.

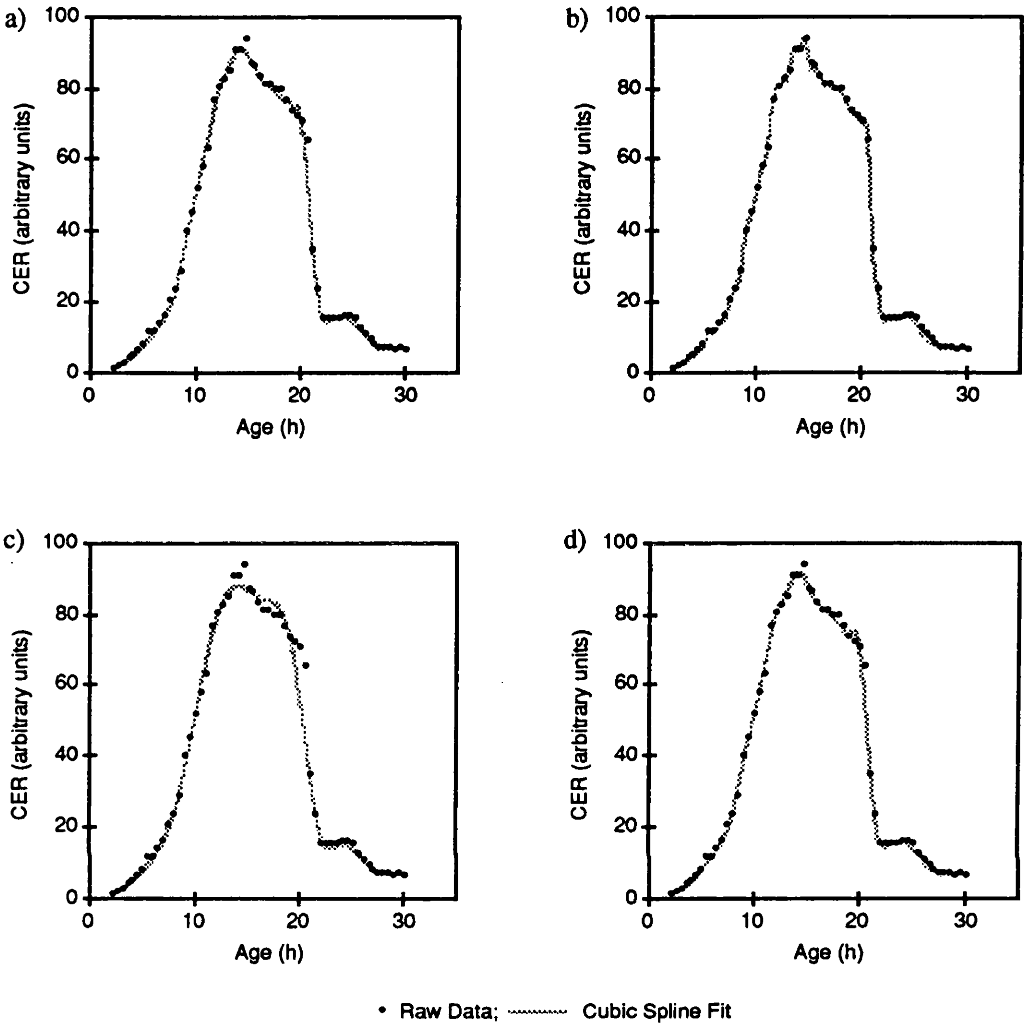


figure	(a)	(b)	(c)	(d)
relative error	4	4	7	4
resolution	0.5	0.1	0.5	0.5
goodness of fit	10	10	10	20

Figure A1.3: Cubic spline fits to carbon dioxide evolution rate data from fermentation batch C447. The user inputs are summarised in the table and their effects seen in the figures. (a) and (d) are considered to be acceptable fits. In (b) the data have been interpolated. The fit in (c) is too loose.

It was hoped that one set of input values could be used to smooth a particular variable for all fermentations. The input values used in Figure A1.3(a) were thus used to smooth other CER profiles. The results were poor as the spline fits followed the data too closely, ie the data were not smoothed (Figure A1.4). Suitable fits were obtained for two of these profiles (C443 and C446) by altering the input values (results not shown). No suitable fit could be obtained for the CER profile from batch C442 because of the rapid decrease in the CER value at about 15 h, the splines either interpolated the data or smoothed out the major peak.

The inability of splines to adequately smooth data in which rapid changes occurred was also observed when trying to smooth pH data from the *E. coli* fermentations. In the best fit that could be obtained the spline became oscillatory in an attempt to 'go around' the steep corner in the data (Figure A1.5). The fit is not a suitable approximation of the data.

Some of the problems alluded to here could be solved by implementing the GCV routines which assume no knowledge of the uncertainties in the data. The GCV routines would not enable the splines to adequately model data sets which change rapidly. The implementation of the GCV routines was not attempted as it was decided that cubic splines would not be suitable for the comparative analysis of fermentation data as described below.

A1.4.1 Applicability to the Comparative Analysis of Data

Cubic splines were investigated as a means of simplifying fermentation data to enable comparisons between two data sets. The most obvious means of comparison between two smoothed profiles is to perform some form of numerical differencing. For example, it would be possible to calculate the vertical difference between two profiles at specified time points, the sum of these differences could then be used as a measure of similarity. A number of problems were envisaged with such a technique as described below.

Firstly, a number of different smooth approximations can be obtained for any one profile (Section A1.4), the best of these is a matter of subjective opinion; a numerical comparison using any of these different approximations would obviously give different results for the similarity metric. The comparative procedure would therefore not be any better than the current visual comparison of data which is also a matter of subjective judgement.

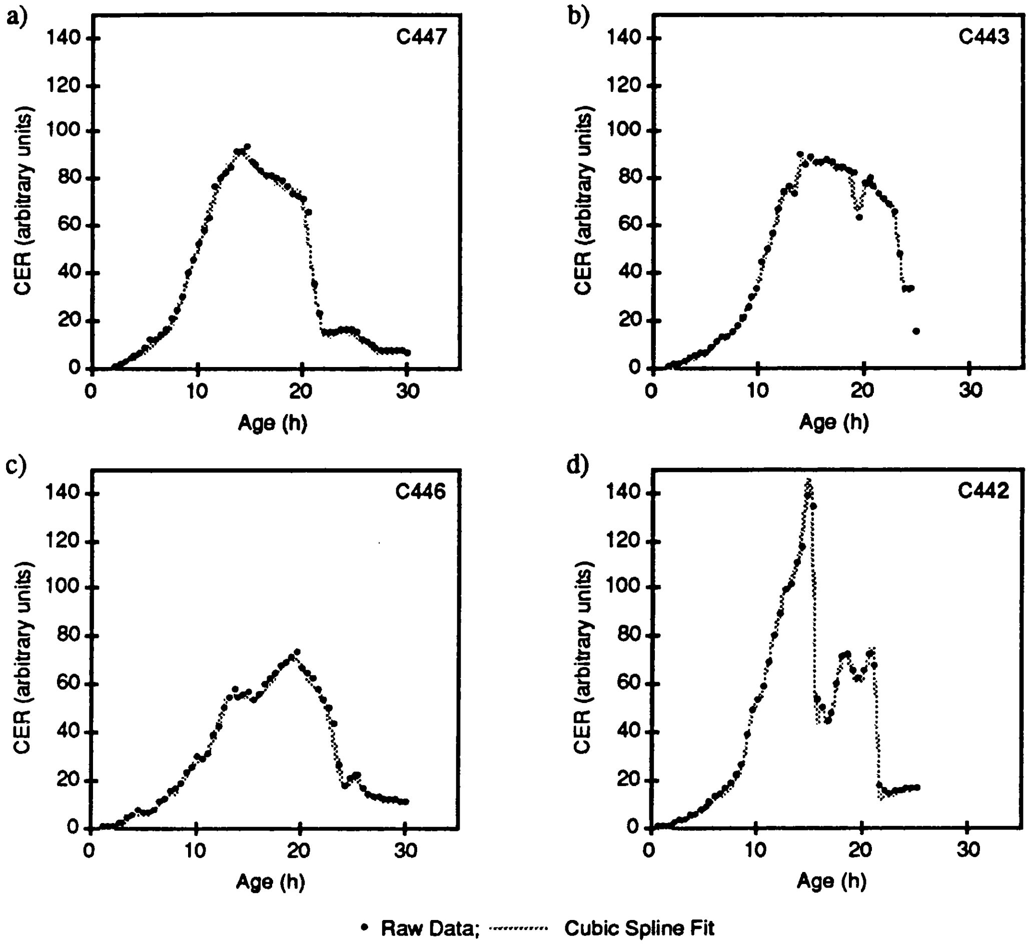


Figure A1.4: Cubic spline fits to carbon dioxide evolution rate data from different fermentation batches. The user inputs for each of the fits were: relative error 4%, resolution 0.5, and goodness of fit 10%. The fit to profile (a) was acceptable. The others followed the data too closely, ie the data were not smoothed, and thus the fits were not acceptable.

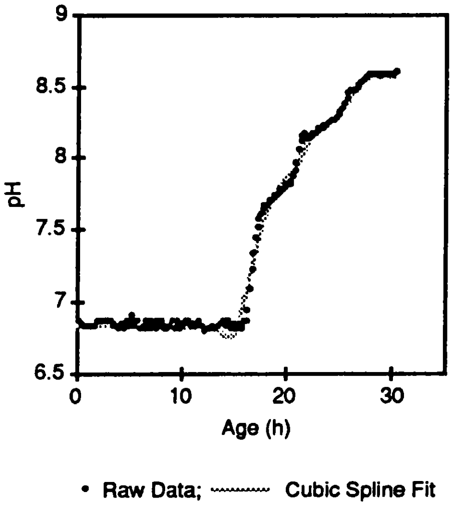


Figure A1.5: Cubic spline fits to pH data from fermentation batch C447. The user inputs were: relative error 0.01%, resolution 0.5, and goodness of fit 80%. The spline became oscillatory in an attempt to 'go around' the sharp corner. The fit was not acceptable.

Secondly, it is known that two fermentations may vary for only a short period of time; an overall similarity measure such as the one described here gives no information on where the processes differed and where they were similar making interpretation difficult.

Thirdly, the length of the lag phase of two fermentations often varies as a result of uncontrollable factors, this leads to slight differences in the time axis of the variables of these fermentations; numerical differences between the profiles would be strongly influenced by this shift and would not be a true representation of the similarity of the two profiles.

Another option for the comparative analysis of the splines data is to relax the description of the data using qualitative or semi-quantitative terminology as described in Section 3.1.2. However, this would require that the smoothed data be further simplified into data pieces utilising some feature of the data, eg linear or curved segments could be used as described in Section 3.1.1. The techniques that were developed in this work, and described in the body of the thesis, simplified the data using a linear segmentation and then compared a qualitative description of the simplified data. There appears to be no advantage in smoothing the data using cubic splines as the simplification was carried out in a single step.

A1.5 Conclusions

1. The MSPLIN package developed at UCL was able to smooth some fermentation data sets using cubic splines.

2. Data sets containing rapid changes were not adequately smoothed.

3. The results for all data sets were dependent on the user's choice of input error and smoothing parameters which often had to be determined by trial and error.

4. The 'best' fitting profile was chosen by the user, different users may choose different profiles. As a result of this subjective judgement, the comparison of data sets is not improved by smoothing them with splines.

5. A numerical comparison between two data sets smoothed by MSPLIN would show large differences if two profiles were shifted slightly in the time dimension as a result of differing lag periods.

6. A numerical comparison between two spline-smoothed data sets would provide a single measure of similarity, it would not indicate that the two profiles were similar over some portions of the profile and different over other portions.

7. The abstraction of the spline-smoothed data to a qualitative representation to facilitate comparison would require a further simplification step in which the profile is divided into smaller portions which could be described qualitatively. This simplification can be performed in a single step without prior smoothing of the raw data as described in the body of the text and thus it was concluded that no advantage is conferred on the comparative analysis procedure by the application of cubic splines.

APPENDIX 2: Outlier Detection for aFGF Off-Line Data

In any experimental work one needs to know the validity of the data. Bad data due to obvious blunders can be discarded immediately. Data that simply look bad cannot be thrown out unless something is obviously wrong. If bad points fall outside the range of normally expected random deviations they may be discarded on the basis of some consistent statistical data analysis.

Holman and Gajda (1978) suggest the use of Chauvenet's criterion to determine outliers. Using this method a reading is rejected if the probability of obtaining the particular deviation from the mean is less than 1/2n, where n is the number of measurements in a data set. The method is appropriate for small data sets, even down to n=2.

The steps involved in the outlier detection routine were:

1. calculate mean and standard deviation for each sample;
2. calculate deviation of each point from the mean and compare with the standard deviation;
3. eliminate points according to the above criterion;
4. recalculate the mean and standard deviation.

The algorithm was programmed using the SmartWare II programming language (Informix Software Inc., Menlo Park, CA) and the data were manipulated in the SmartWare II spreadsheet.

The data investigated were the off-line measurements of optical density, glucose concentration, dry cell weight and volumetric aFGF concentration. No outliers were found in these data.

APPENDIX 3: aFGF Calculations

A3.1 Conversion to a Specific Concentration

In the aFGF experiments the product concentrations were initially determined in volumetric terms, ie normalised units per litre of broth. When the product is intracellular, as was the case for the aFGF fermentations, it may be more meaningful to talk about the productivity of the cells. The aFGF values were therefore also calculated in specific terms, ie normalised units per gram of biomass.

The calculation of the specific aFGF for each sample was:

$$\text{aFGF}_s = (\text{mean aFGF}_v) / (\text{mean DCW}) \qquad \text{(A3.1)}$$

where

aFGF_s = specific aFGF concentration (normalised units.(g biomass)$^{-1}$)
aFGF_v = volumetric aFGF concentration (normalised units.L^{-1})
DCW = dry cell weight calculated from the correlation with optical density (g.L^{-1})

The mean aFGF_v and mean DCW values used in Equation A3.1 were the mean values calculated for each sample. The correlation between dry cell weight and optical density is discussed in Appendix 4.

A3.2 Uncertainties in the aFGF Measurements

Three or four aFGF_v values were obtained for each sample. The mean and standard deviation of each sample were calculated and an average standard deviation over the complete set of fermentations was obtained. The average standard deviation was thought to be more representative of the spread of the measurements than that calculated for each individual sample where only three or four measurements contributed to the calculations.

The aFGF$_s$ values were calculated as described in A3.1. Only one value was available at each sample point thus there was no means of calculating the standard deviation of the measurement. The uncertainties in the aFGF$_s$ values were obtained from the relative uncertainties in the dry cell weight and aFGF$_v$ values.

When the ratio of two values is obtained, the relative error in the result is calculated by adding the relative errors of the two component terms. Thus for aFGF$_s$ the relative errors in the mean aFGF$_v$ and mean DCW were required.

The method for obtaining the absolute error in the dry cell weight is presented in the next appendix (A4.2). This is converted to a relative error (δ_d) by dividing it by the mean dry cell weight.

The relative error in aFGF$_v$, for each sample, was calculated from:

$$\delta_v = (\Sigma \text{ abs(mean aFGF}_v \text{ - aFGF}_v)) / \Sigma \text{ aFGF}_v$$

The relative error in aFGF$_s$, for each sample, is therefore

$$\delta_s = \delta_v + \delta_d$$

and the absolute error in aFGF$_s$ is

$$\varepsilon_s = \delta_s * \text{aFGF}_s$$

These uncertainties were calculated for each sample using the SmartWare II spreadsheet (Informix Software Inc., Menlo Park, CA).

The aFGF time profiles are presented in Figures A3.1 and A3.2. The 'error bars' represent three average standard deviations about the mean for aFGF$_v$ and the mean plus or minus the absolute error for aFGF$_s$.

A3.3 aFGF Concentration at the Recommended Harvest Point

The aFGF fermentations are usually harvested about the time that the glucose concentration reaches 5 $g.L^{-1}$. In the experiments in Chapter 4 the fermentations were run past this point. The aFGF concentration was estimated at the recommended harvest point as a means of determining the productivity of the fermentation.

It was assumed that the aFGF concentrations were linear between each sample point and the aFGF harvest concentrations were determined by linear interpolation using the samples on either side of the recommended harvest time. The equations were obtained from Mendenhall and Sincich (1988, p.234).

For the volumetric aFGF values, where an average standard deviation was available, the following calculations yielded the required aFGF concentration and its associated uncertainty:

Let $a = (t_2-t_h) / (t_2-t_1)$

$b = (t_h-t_1) / (t_2-t_1)$

Then $aFGF_h = a * aFGF_1 + b * aFGF_2$

and $\sigma_h^2 = a^2 * \sigma_1^2 + b^2 * \sigma_2^2$

Where t_h = the recommended harvest time

t_1 = the time of the last sample before t_h

t_2 = the time of the next sample after t_h

$aFGF_h$ = the estimated aFGF concentration at t_h

$aFGF_1$ = the aFGF concentration at t_1

$aFGF_2$ = the aFGF concentration at t_2

σ_h^2 = the estimated variance in $aFGF_h$

σ_1^2 = σ_2^2 = the average variance in the measured aFGF values

For the specific aFGF values the uncertainties were given in terms of an absolute error. The calculation of the aFGF harvest value then simply involved determining the upper and lower limits of $aFGF_h$ by interpolating separately between the upper bounds and the lower bounds of the two samples either side of the harvest time.

The harvest values are indicated on Figures A3.1 and A3.2 and are listed in Table 4.7.

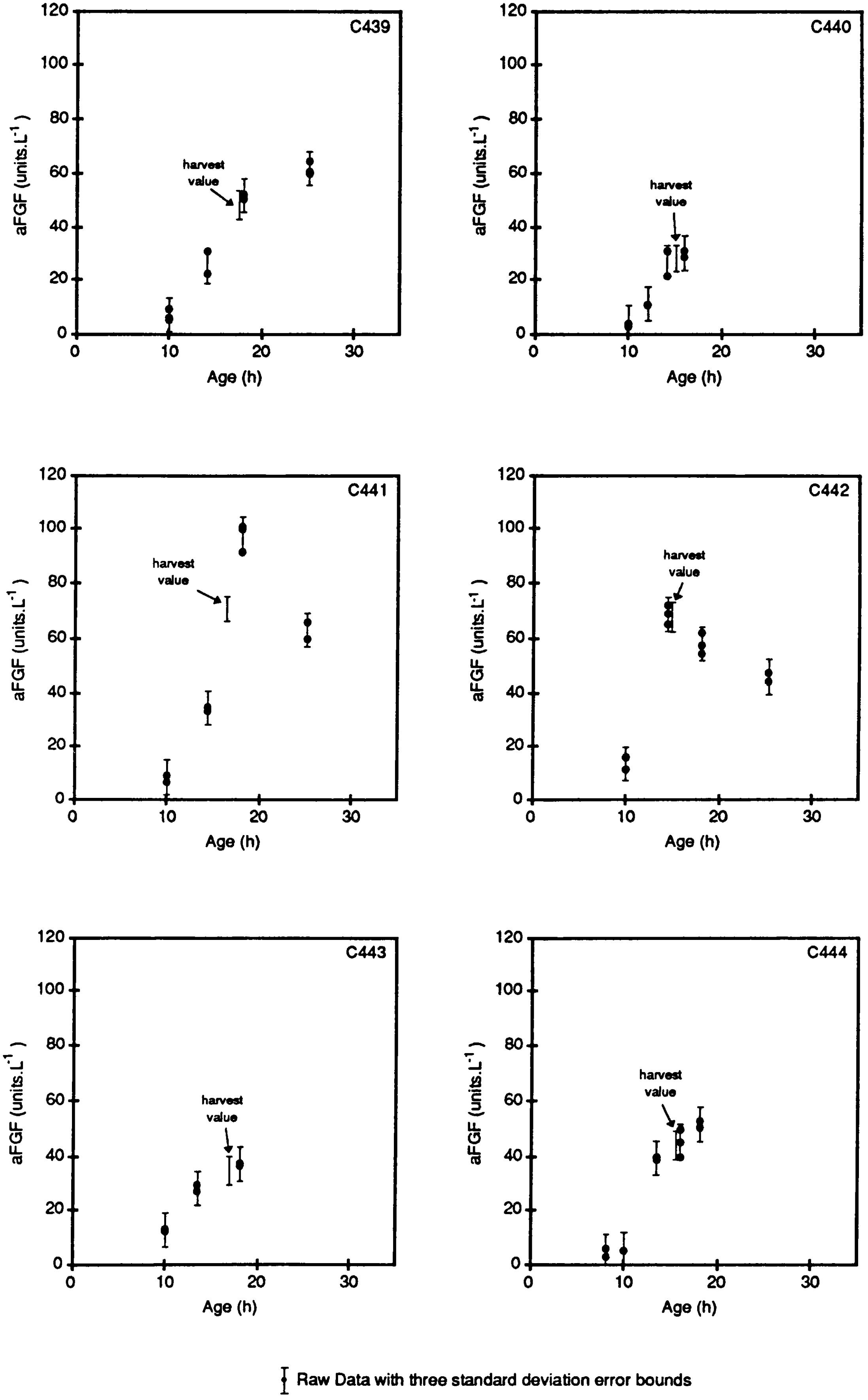


Figure A3.1(a): aFGF concentration profiles (normalised units.L^{-1}) for fermentations C439 to C444.

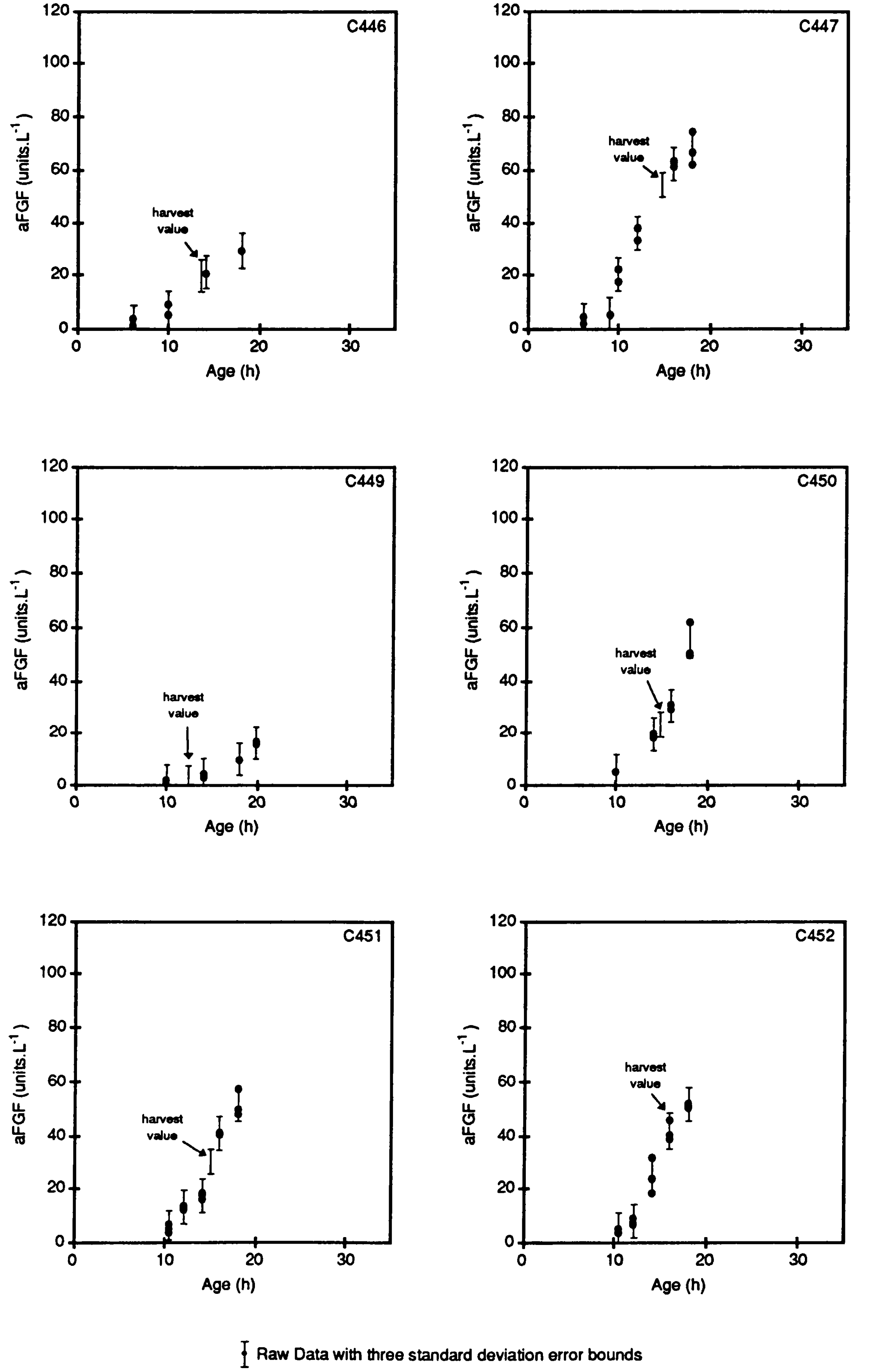


Figure A3.1(b): aFGF concentration profiles (normalised units.L^{-1}) for fermentations C446 to C452.

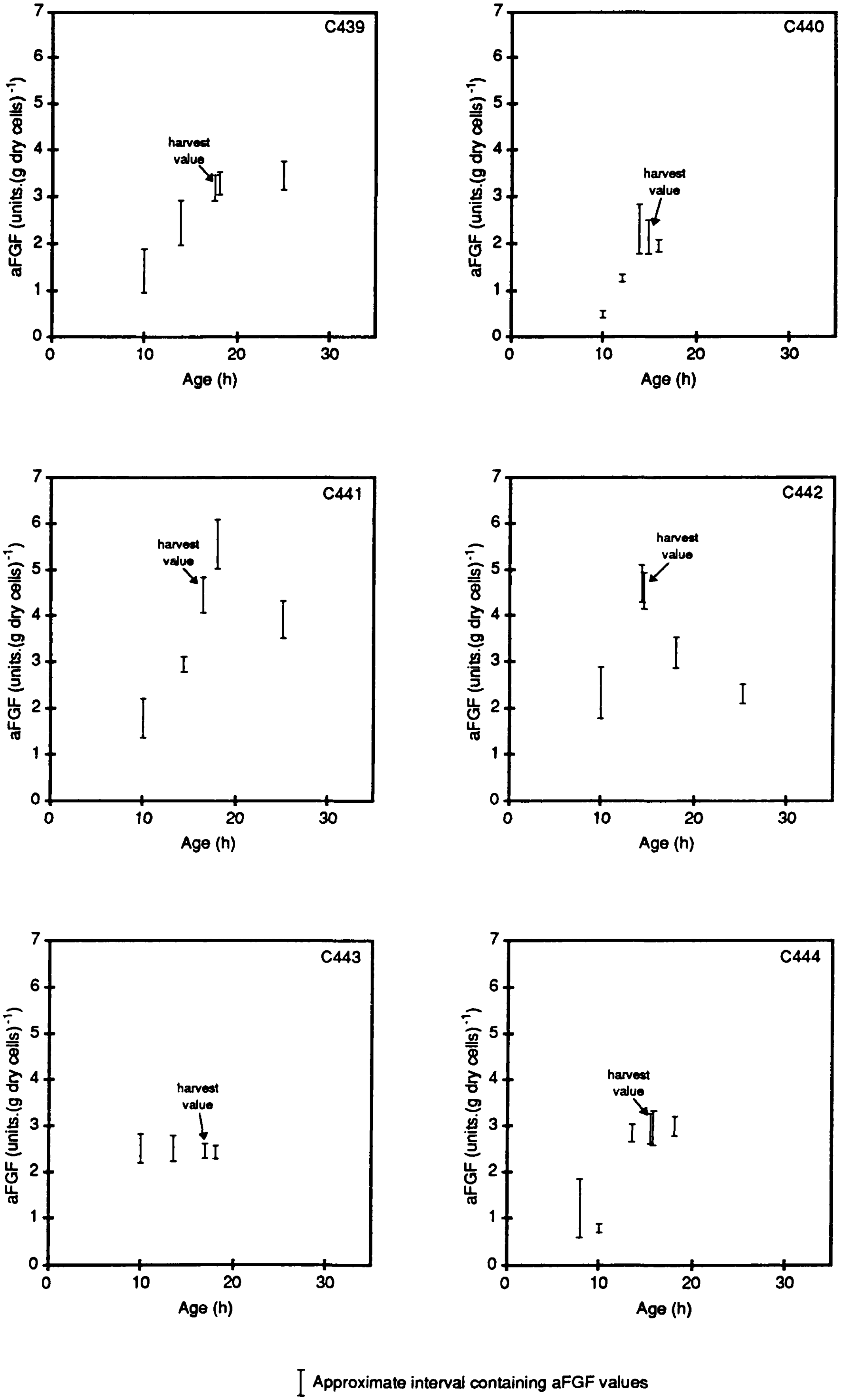


Figure A3.2(a): aFGF concentration profiles (normalised units per gram of dry cells) for fermentations C439 to C444.

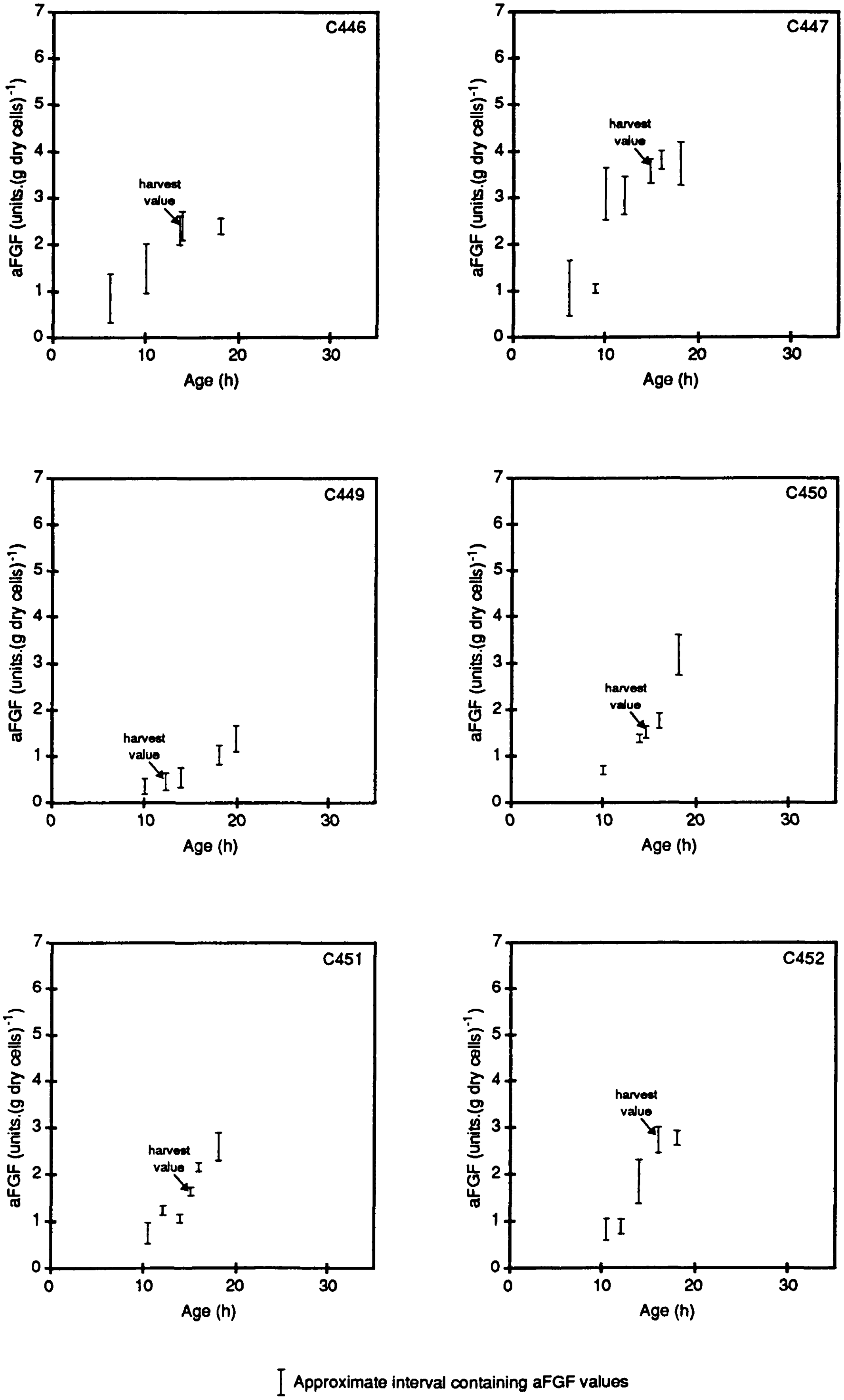


Figure A3.2(b): aFGF concentration profiles (normalised units per gram of dry cells) for fermentations C446 to C452.

APPENDIX 4: Correlation Between Optical Density and Dry Cell Weight in aFGF Fermentations

An experimentally determined linear relationship between optical density and dry cell weight is often used to infer biomass concentration in a broth from the relatively straight forward optical density measurement. This was attempted in the aFGF fermentations described in Chapter 4. A rough plot of the dry cell weight versus optical density for all the fermentations suggested the possibility of a difference in the correlation for the batch that had undergone a sixty minute sterilisation with glucose *in situ* (C449). This was examined in A4.1. The uncertainties in the dry cell weight values are determined in A4.2.

A4.1 A Statistical Analysis of the Correlations Between OD and Dry Cell Weight

The possibility of a difference in the correlations between optical density and dry cell weight in the aFGF fermentations was investigated using a statistical test described by Mendenhall and Sincich (1988).

Three variables were defined: the dependent variable, dry cell weight (y); and the independent variables, optical density (x_1) and batch number (x_2-x_{10}). The batch number is a logical variable and is interpreted as follows:

$$x_2=\begin{cases}1 \text{ if C439}\\ 0 \text{ if not}\end{cases} \qquad x_3=\begin{cases}1 \text{ if C440}\\ 0 \text{ if not}\end{cases} \qquad x_4=\begin{cases}1 \text{ if C441}\\ 0 \text{ if not}\end{cases}$$

$$x_5=\begin{cases}1 \text{ if C442}\\ 0 \text{ if not}\end{cases} \qquad x_6=\begin{cases}1 \text{ if C443}\\ 0 \text{ if not}\end{cases} \qquad x_7=\begin{cases}1 \text{ if C444}\\ 0 \text{ if not}\end{cases}$$

$$x_8=\begin{cases}1 \text{ if C446}\\ 0 \text{ if not}\end{cases} \qquad x_9=\begin{cases}1 \text{ if C447}\\ 0 \text{ if not}\end{cases} \qquad x_{10}=\begin{cases}1 \text{ if C450}\\ 0 \text{ if not}\end{cases}$$

Note that if all of x_2 to x_{10} are 0 then the batch number is C449. C451 and C452 were not included as no dry cell weight measurements were made on these fermentations.

If a single line were adequate to describe the optical density - dry cell weight correlation for all the fermentations, the equation of that line would be:

$$E(y) = \beta_0 + \beta_1 x_1 \qquad \text{(A4.1)}$$

where E(y) is the expected value of the dry cell weight and the βs are the parameters of the first-order model.

If the correlations were in fact different, the equation representing all the lines would be:

$$E(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_{10} x_{10} + \\ \beta_{11} x_1 x_2 + \beta_{12} x_1 x_3 \ldots + \beta_{19} x_1 x_{10} \qquad \text{(A4.2)}$$

The null hypothesis for the test was that a single line adequately describes the optical density - dry cell weight correlation for all the fermentations, whilst the alternative hypothesis was that at least one of the fermentations differs in its relationship between optical density and dry cell weight:

H_o: $\beta_2 = \beta_3 = \ldots = \beta_{18} = \beta_{19} = 0$
H_a: at least one of β_2 to β_{19} differs from 0

The null hypothesis was tested by fitting the complete model (Eq. A4.2) and the reduced model (Eq. A4.1) using the regression capabilities of RS/1 (Release 4, BBN Software Products Corporation, Cambridge, MA, USA) and then conducting an F test on the reduction in the residual sum of squares caused by the fitting of the complete model. The results of the fits are presented in Tables A4.1 to A4.4.

SOURCE	DF	SUM OF SQUARES	MEAN SQUARE	F VALUE	R-SQUARE
regression	19	31.719	1.669	83.521	0.978
residual	36	0.720	0.020		

Table A4.1: Portion of analysis of variance table for the 'complete model', ie separate lines fitted to all fermentations. (DF = degrees of freedom).

COEFFICIENT	VARIABLE NAME	FITTED COEFFICIENT
β_0	intercept	-1.76
β_1	OD	0.47
β_2	C439	2.43
β_3	C440	1.76
β_4	C441	2.85
β_5	C442	2.68
β_6	C443	1.89
β_7	C444	2.18
β_8	C446	0.88
β_9	C447	2.25
β_{10}	C450	2.76
β_{11}	OD*C439	0.01
β_{12}	OD*C440	0.00
β_{13}	OD*C441	-0.02
β_{14}	OD*C442	-0.05
β_{15}	OD*C443	0.10
β_{16}	OD*C444	-0.01
β_{17}	OD*C446	0.09
β_{18}	OD*C447	-0.01
β_{19}	OD*C450	-0.04

Table A4.2: Coefficients table for the fitting of separate lines to all the fermentations.

SOURCE	DF	SUM OF SQUARES	MEAN SQUARE	F VALUE	R-SQUARE
regression	1	30.959	30.959	1130.425	0.954
residual	54	1.479	0.027		

Table A4.3: Portion of analysis of variance table for the 'reduced model', ie a single line fitted to all fermentations. (DF = degrees of freedom).

COEFFICIENT	VARIABLE NAME	FITTED COEFFICIENT
β_0	intercept	0.33
β_1	OD	0.46

Table A4.4: Coefficients table for the fitting of one line to all fermentations.

The test statistic is:

$$F = ((SSE_1 - SSE_2) / (k - g)) / MSE_2 \qquad (A4.3)$$

The rejection region is $F > F_\alpha$ (choose α=0.05)

where: SSE_1 = sum of squared errors for the reduced model = 1.479

SSE_2 = sum of squared errors for the complete model = 0.720

MSE_2 = mean squared error for the complete model = 0.020

k-g = number of β parameters specified in H_o = 19-1 = 18

and for the comparison with the F Tables

v_1 = k-g = degrees of freedom for the numerator = 18

v_2 = n-(k+1) = degrees of freedom for the denominator = 56-20 = 36

n = sample size = 56

From Equation A4.3 F = 2.11 which is larger than the value in the F Tables, $1.84 < F_{.05} < 2.01$ (Mendenhall and Sincich 1988). H_o was therefore rejected and it was possible to conclude that at least one of the fermentations differs in its relationship between optical density and dry cell weight.

As it was observed earlier that the optical density - dry cell weight correlation for fermentation C449 appeared to differ from the other correlations, the above procedure was repeated with C449 being excluded from the model: factors with x_{10} were removed from Equation A4.2; when all x_2 to x_9 are zero then the batch is C450. The results of the fits are presented in Tables A4.5 to A4.8.

SOURCE	DF	SUM OF SQUARES	MEAN SQUARE	F VALUE	R-SQUARE
regression	17	34.190	2.011	83.322	0.978
residual	32	0.772	0.024		

Table A4.5: Portion of analysis of variance table for the 'complete model', ie separate lines fitted to all fermentations except C449. (DF = degrees of freedom).

COEFFICIENT	VARIABLE NAME	FITTED COEFFICIENT
β_0	intercept	1.00
β_1	OD	0.43
β_2	C439	-0.33
β_3	C440	-1.00
β_4	C441	0.09
β_5	C442	-0.08
β_6	C443	-0.87
β_7	C444	-0.58
β_8	C446	-1.88
β_9	C447	-0.51
β_{10}	OD*C439	0.06
β_{11}	OD*C440	0.05
β_{12}	OD*C441	0.02
β_{13}	OD*C442	-0.01
β_{14}	OD*C443	0.14
β_{15}	OD*C444	0.03
β_{16}	OD*C446	0.13
β_{17}	OD*C447	0.03

Table A4.6: Coefficients table for the fitting of separate lines to all the fermentations except C449.

SOURCE	DF	SUM OF SQUARES	MEAN SQUARE	F VALUE	R-SQUARE
regression	1	33.852	33.852	1463.217	0.968
residual	48	1.110	0.023		

Table A4.7: Portion of analysis of variance table for the 'reduced model', ie a single line fitted to all fermentations except C449. (DF = degrees of freedom).

COEFFICIENT	VARIABLE NAME	FITTED COEFFICIENT
β_0	intercept	0.55
β_1	OD	0.46

Table A4.8: Coefficients table for the fitting of one line to all fermentations except C449.

The following values were substituted into Equation A4.3

SSE_1 = sum of squared errors for the reduced model = 1.110

SSE_2 = sum of squared errors for the complete model = 0.772

MSE_2 = mean squared error for the complete model = 0.024

k-g = number of β parameters specified in H_o = 17-1 = 16

and for the comparison with the F Tables

v_1 = k-g = degrees of freedom for the numerator = 16

v_2 = n-(k+1) = degrees of freedom for the denominator = 50-18 = 32

n = sample size = 50

From Equation A4.3 F = 0.88 which is smaller than the value in the F Tables, $1.84 < F_{.05} < 2.01$ (Mendenhall and Sincich 1988). There is no evidence to suggest that the optical density - dry cell weight correlation differs for the remaining fermentations.

The equations describing the relationship between optical density and dry cell weight are therefore:

$$\hat{y} = -1.76 + 0.47x_1 \qquad \text{for C449} \qquad (A4.4)$$

$$\hat{y} = 0.55 + 0.46x_1 \qquad \text{for the other fermentations} \qquad (A4.5)$$

where $\hat{y}$ is the predicted value of the dry cell weight.

The correlation equation for C449 was obtained by performing linear regression on the data from that fermentation only. The correlation for the other fermentations was that obtained in Table A4.7. The correlations are shown in Figure 4.6.

The regression also gave values for the errors on the parameter estimates, these are summarised in Table A4.9.

BATCH	SLOPE	INTERCEPT
C449 others	0.47±0.04 0.46±0.01	-1.76±0.65 0.55±0.22

Table A4.9: Parameter values for the regression equations and their errors.

A4.2 Uncertainties in the Dry Cell Weight Measurements

Each optical density reading was converted to a dry cell weight value using the appropriate correlation from A4.1. For each sample the standard deviation was determined and the average standard deviation over the full set of experiments was calculated to be 0.2.

For the simplification routines there were two goodness of fit criteria: three times the average standard deviation and the range of values at each sample. The larger of these for each sample was the governing criterion for fitting the linear data pieces.

There was obviously some error associated with the regression equations (Table A4.9) thus the range of values for each sample was modified to account for this: the lower value had the uncertainty subtracted from it and the upper value had the uncertainty added to it thereby increasing the range. The uncertainties were calculated as follows.

For all fermentations except C449:

regression equation: $DCW = 0.55 + 0.46*OD$

where

DCW = dry cell weight
OD = optical density

When adding two values their absolute errors are added. When multiplying two values their relative errors are added.

Let ε_i = absolute error in intercept (0.55) = 0.22

ε_s = absolute error in slope (0.46) = 0.01

δ_s = relative error in slope = 0.01/0.46 = 0.02

δ_o = relative error in OD = (abs(mean OD - OD))/OD

Then $\varepsilon_d = \varepsilon_i + (\delta_s + \delta_o)*0.46*OD$

where ε_d is the absolute error in the dry cell weight value. Thus the range of values at each sample point is given by:

minimum DCW - ε_d to maximum DCW + ε_d

Similarly for batch C449:

regression equation: DCW = -1.76 + 0.47*OD

ε_i = absolute error in intercept (-1.76) = 0.65

ε_s = absolute error in slope (0.47) = 0.04

δ_s = relative error in slope = 0.04/0.47 = 0.09

δ_o = relative error in OD = (abs(mean OD - OD))/OD

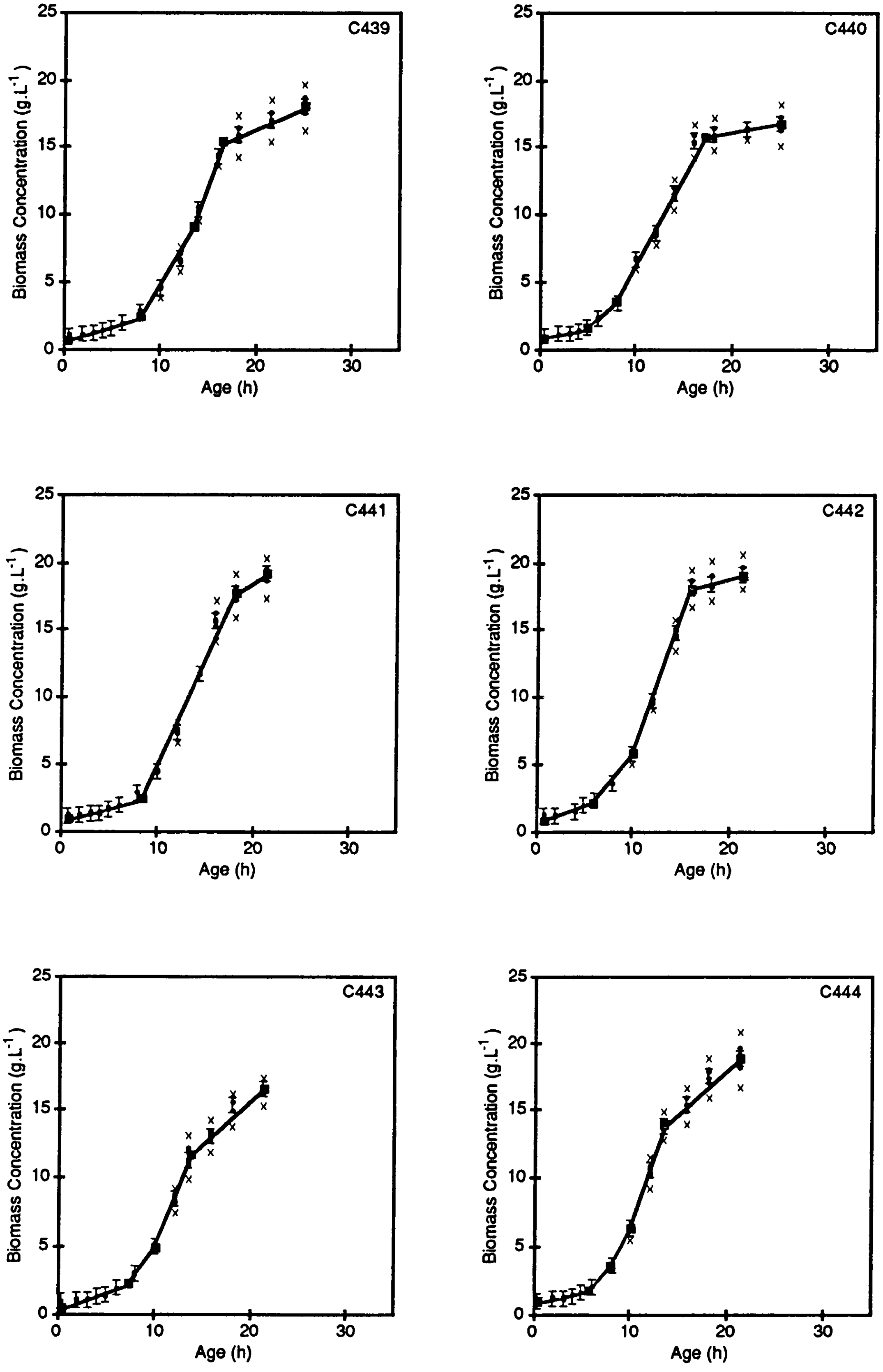
$$\varepsilon_d = \varepsilon_i + (\delta_s + \delta_o)*0.47*OD$$

These uncertainties were calculated for each sample using the SmartWare II spreadsheet (Informix Software Inc., Menlo Park, CA).

APPENDIX 5: Time Variant Data From aFGF Fermentations

The following graphs depict the time variant data from the aFGF fermentations described in Chapter 4. Where appropriate the data have been linearised using DSIMP.

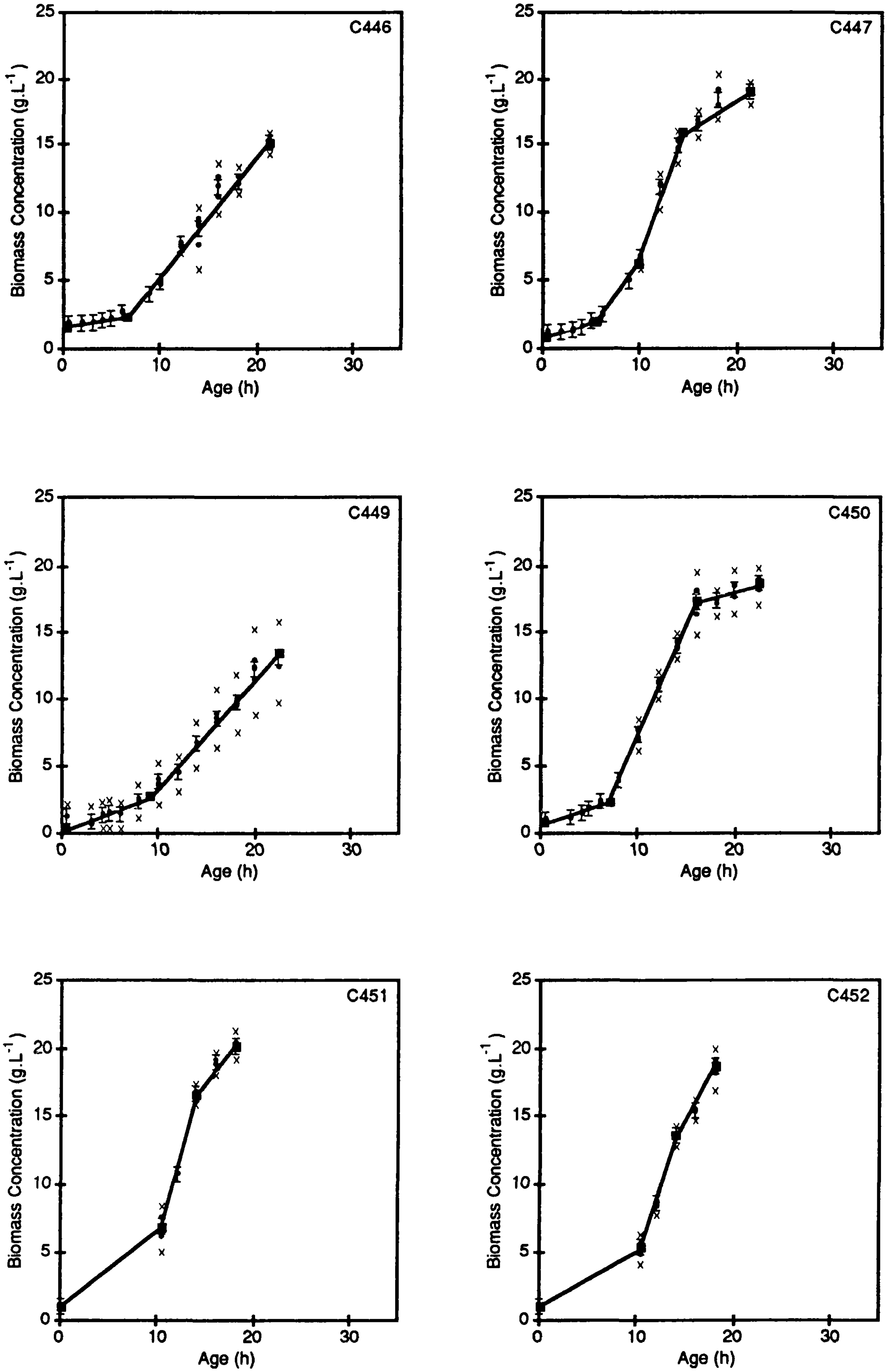
The aFGF yield data were presented in Appendix 3.



Raw Data with three standard deviation error bounds; Simplified Data

When the uncertainty bounds on an individual point fall outside the error bar it is marked with an x

Figure A5.1(a): Biomass concentration profiles for fermentations C439 to C444.



Raw Data with three standard deviation error bounds; —■— Simplified Data

When the uncertainty bounds on an individual point fall outside the error bar it is marked with an x

Figure A5.1(b): Biomass concentration profiles for fermentations C446 to C452.

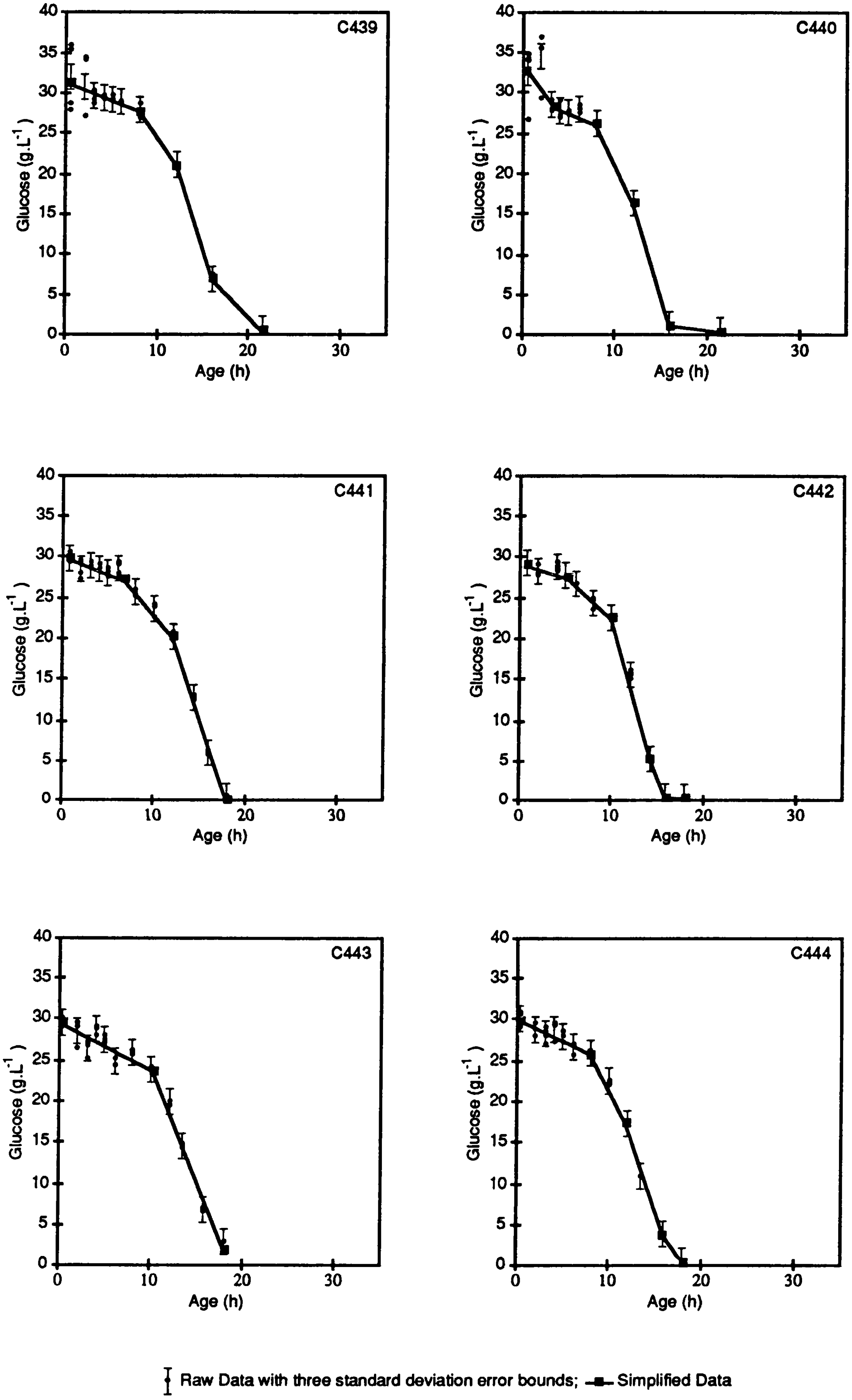


Figure A5.2(a): Glucose concentration profiles for fermentations C439 to C444.

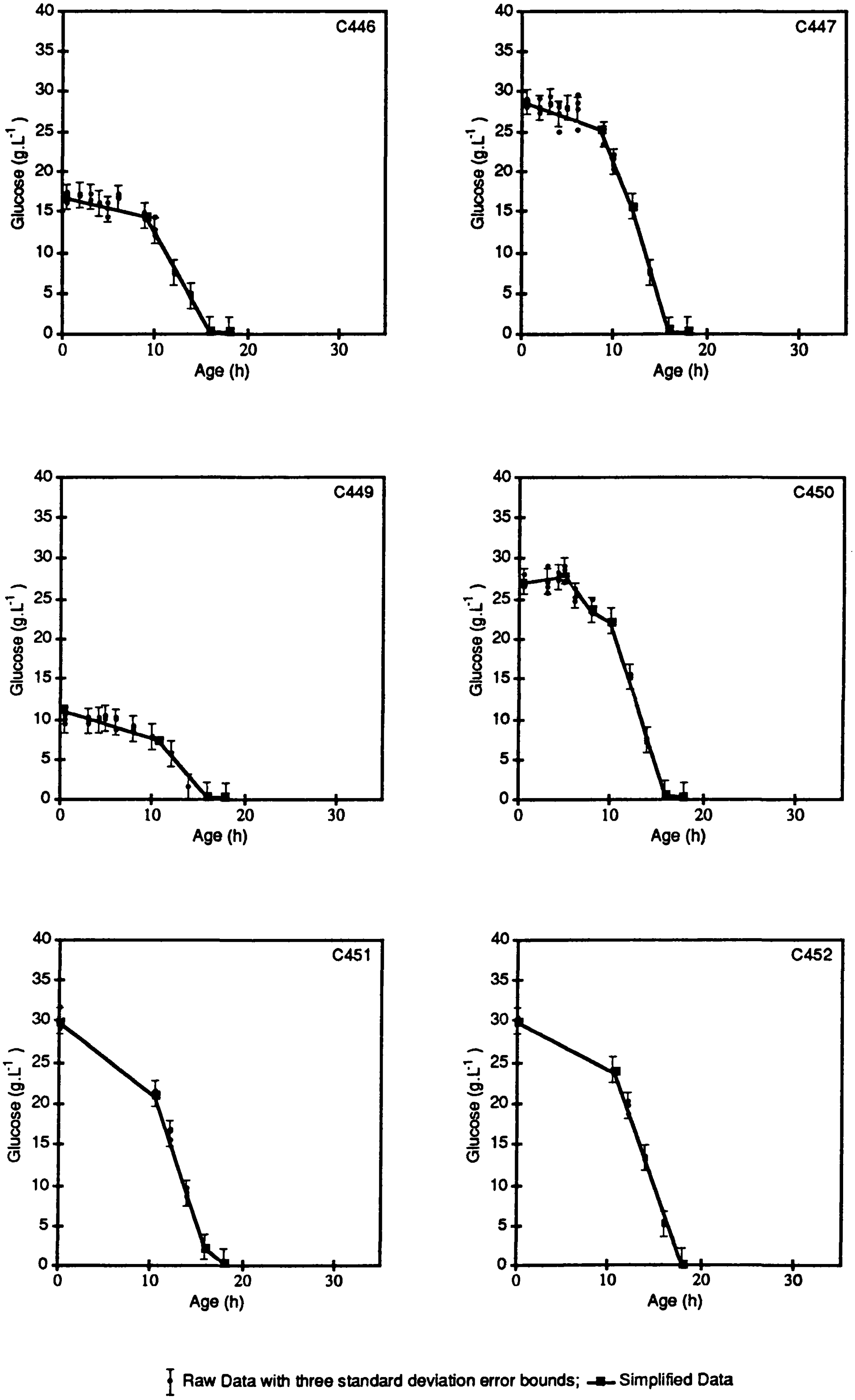


Figure A5.2(b): Glucose concentration profiles for fermentations C446 to C452.

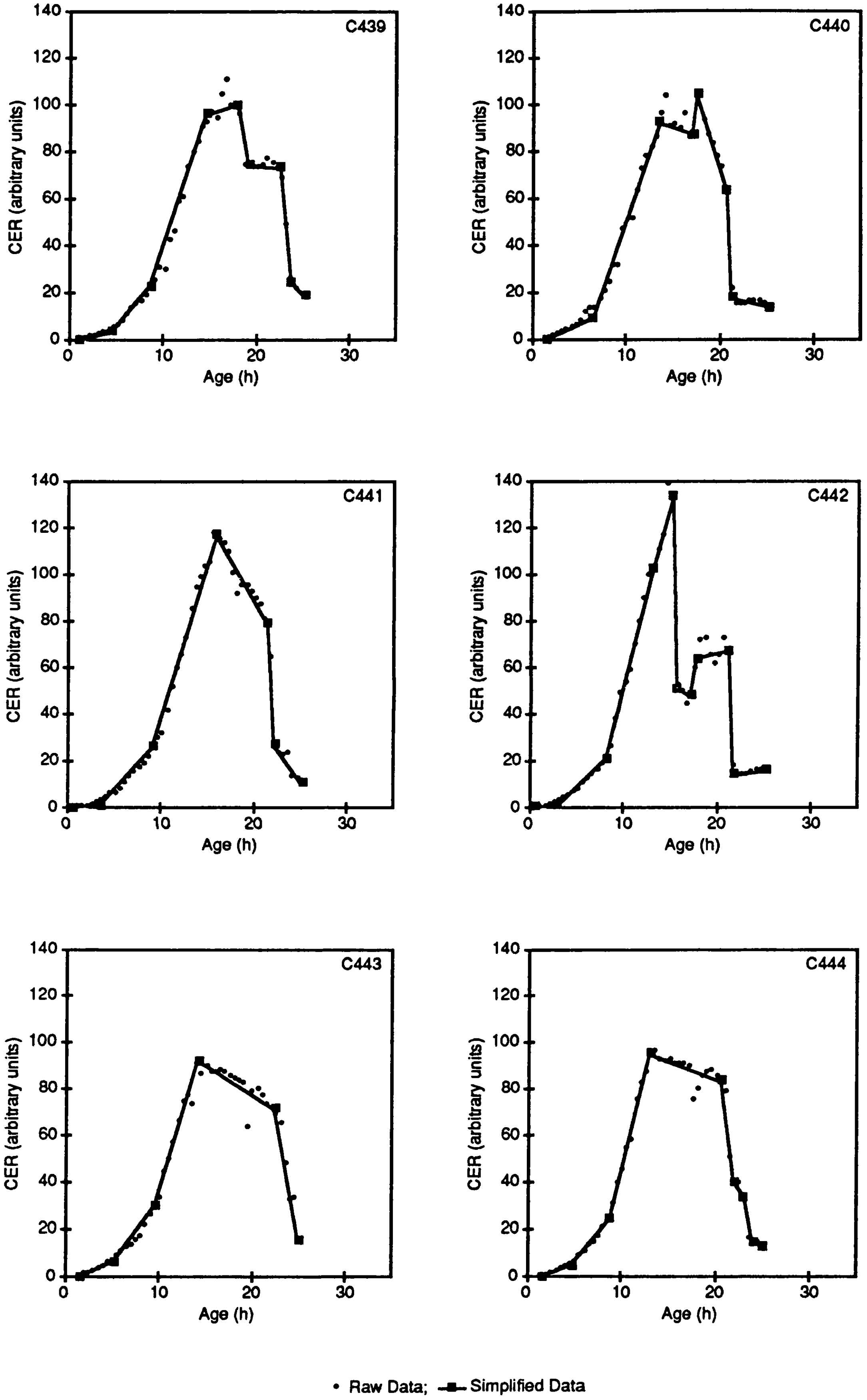


Figure A5.3(a): Carbon dioxide evolution rate profiles for fermentations C439 to C444.

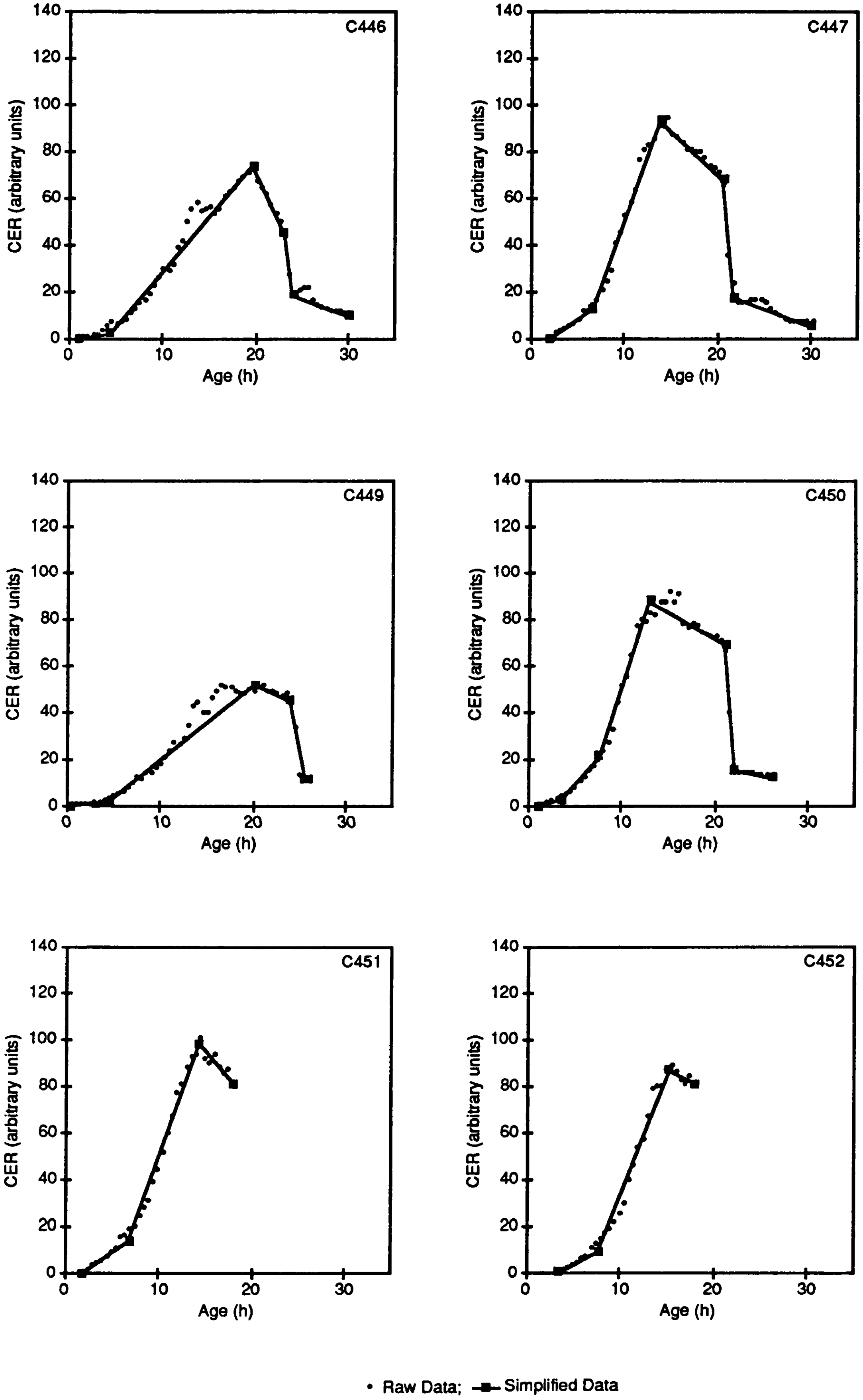


Figure A5.3(b): Carbon dioxide evolution rate profiles for fermentations C446 to C452.

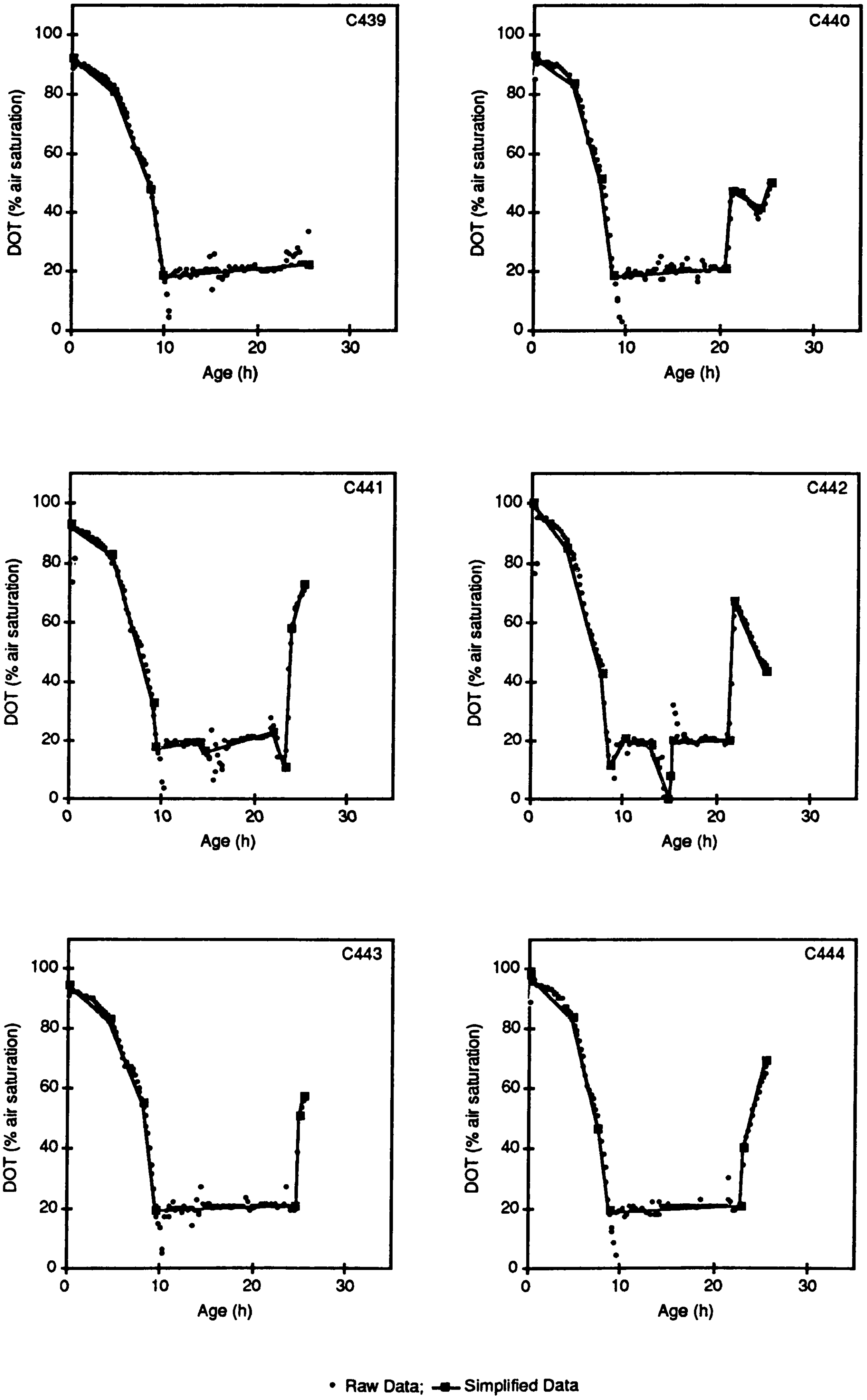


Figure A5.4(a): Dissolved oxygen tension (DOT) profiles for fermentations C439 to C444.

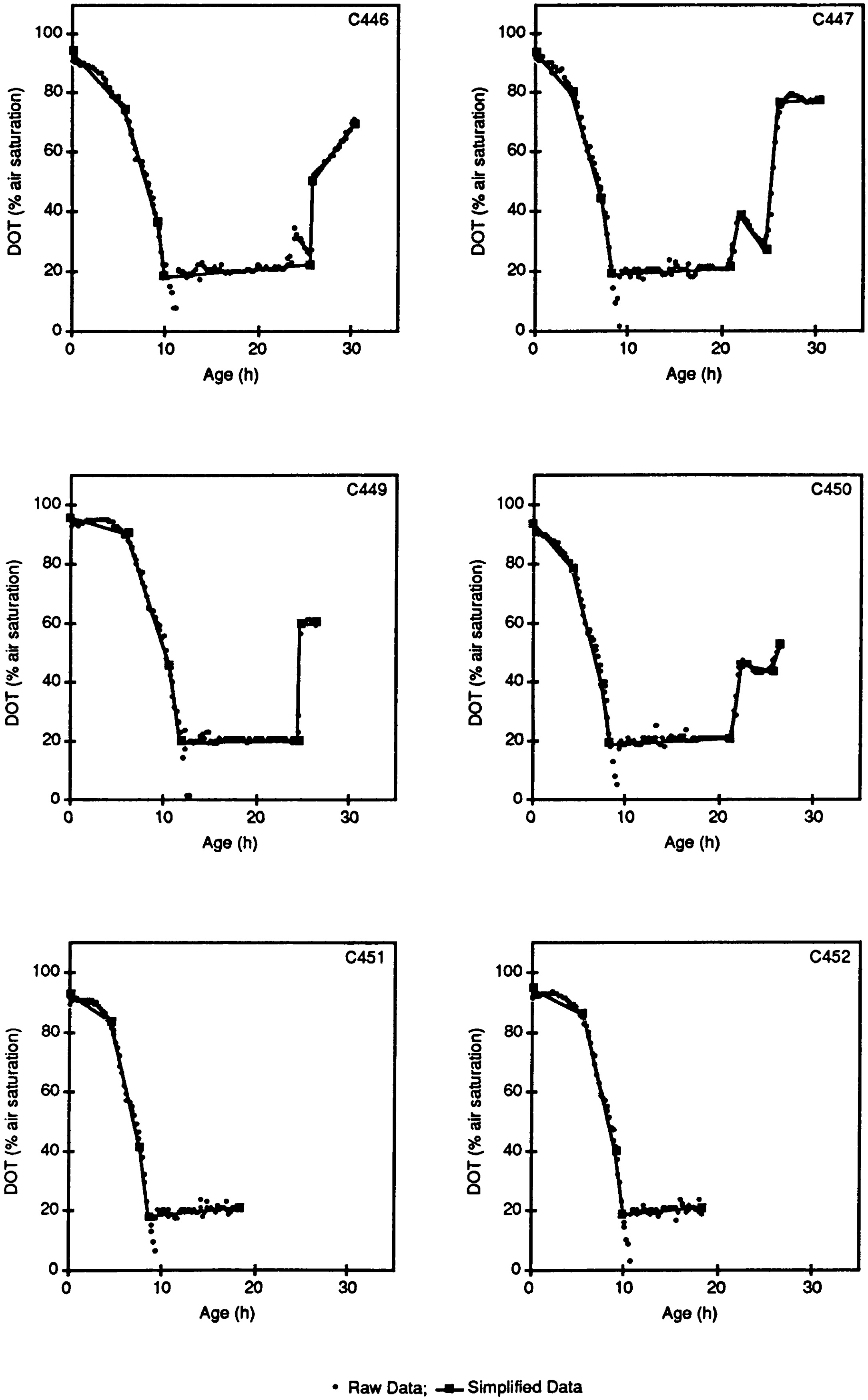


Figure A5.4(b): Dissolved oxygen tension (DOT) profiles for fermentations C446 to C452.

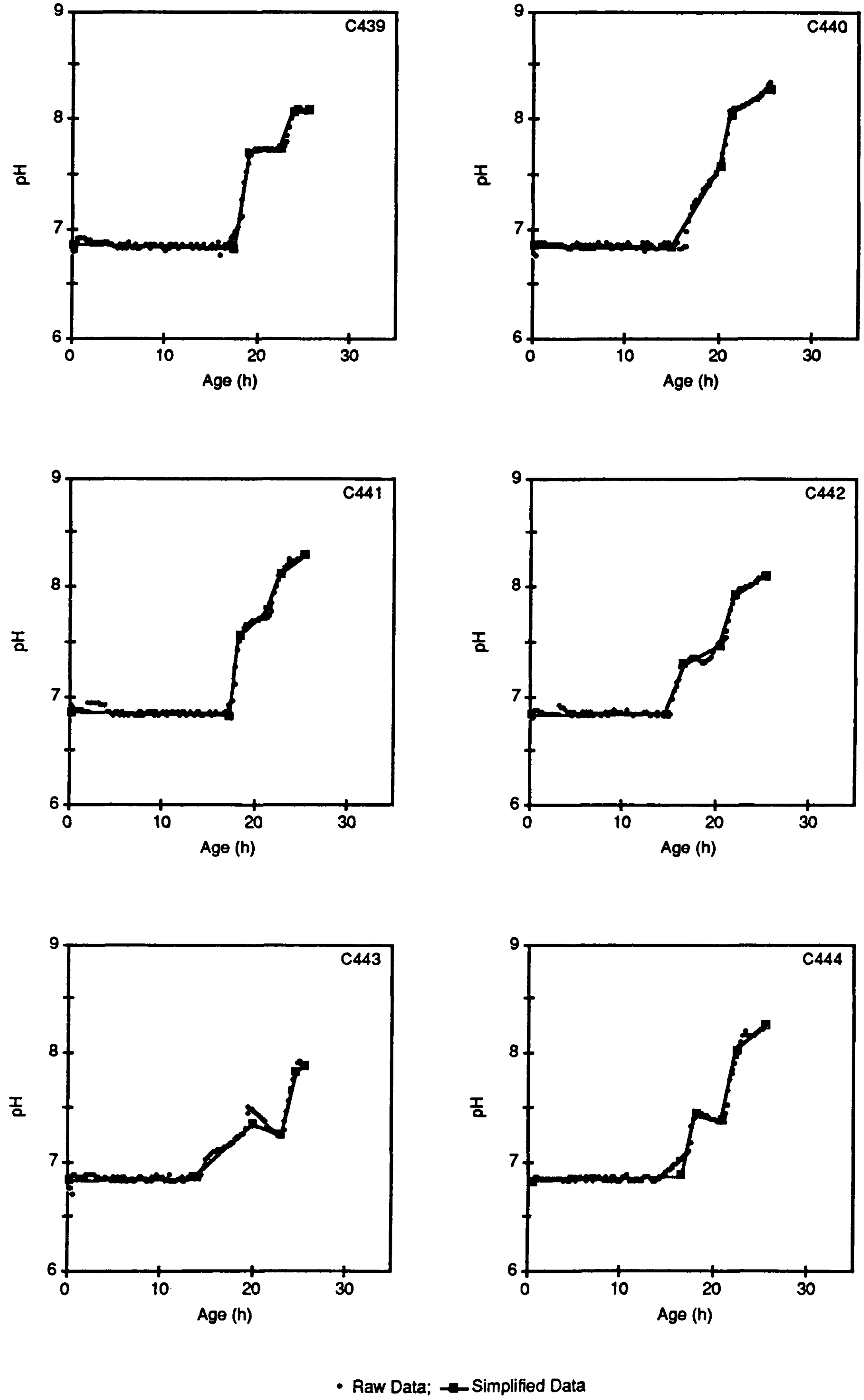


Figure A5.5(a): pH profiles for fermentations C439 to C444.

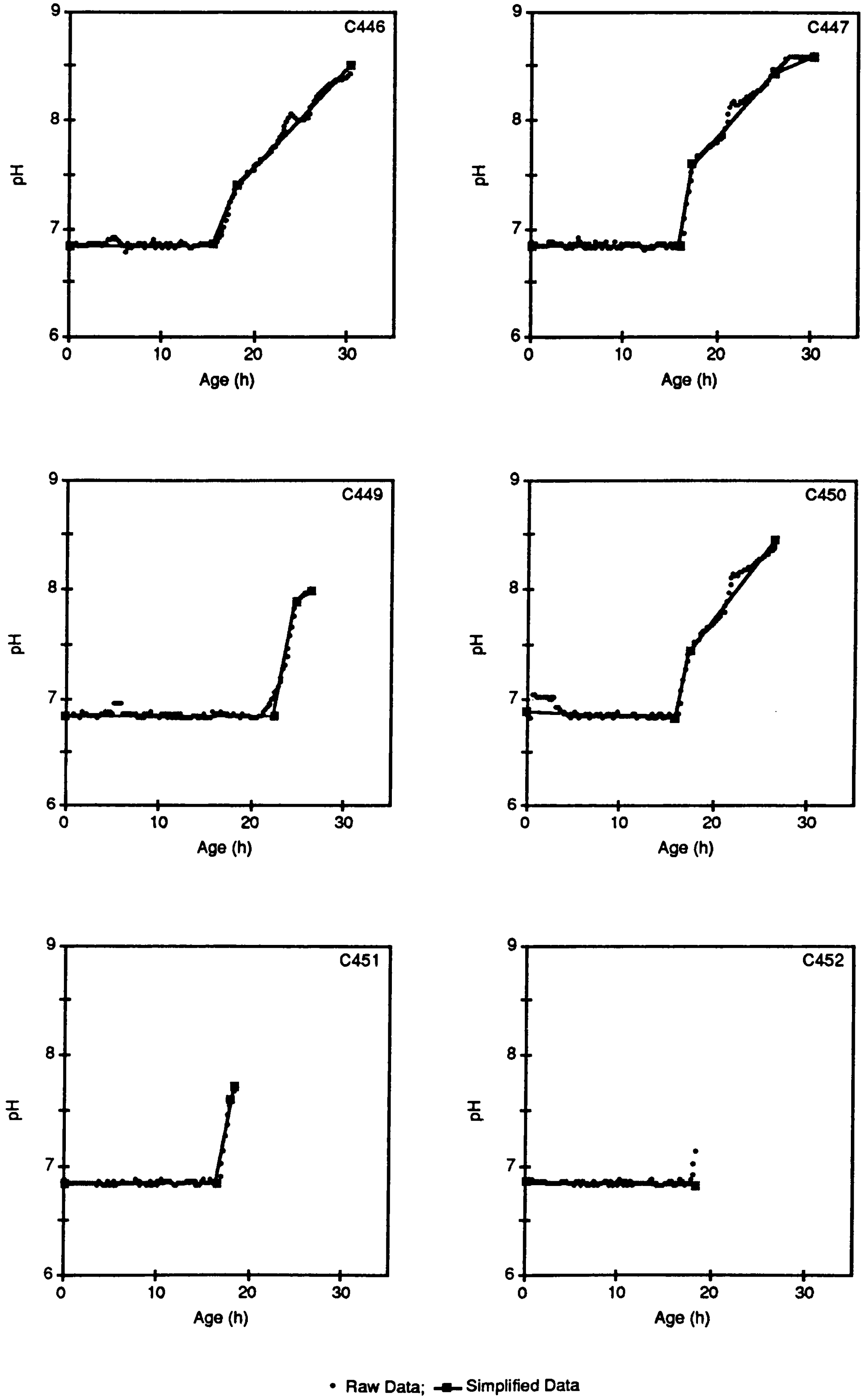


Figure A5.5(b) pH profiles for fermentations C446 to C452.

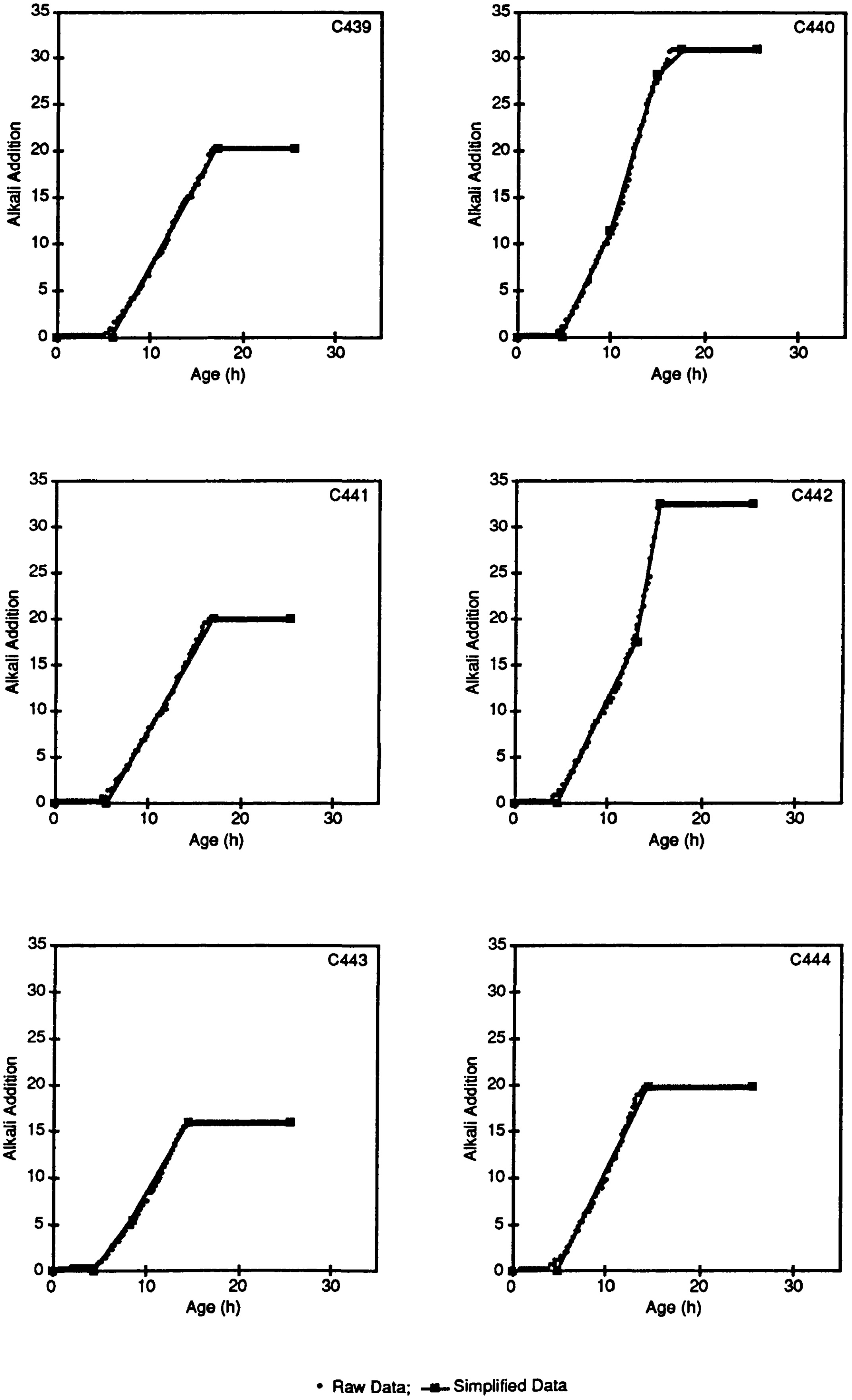


Figure A5.6(a): Alkali addition profiles for fermentations C439 to C444.

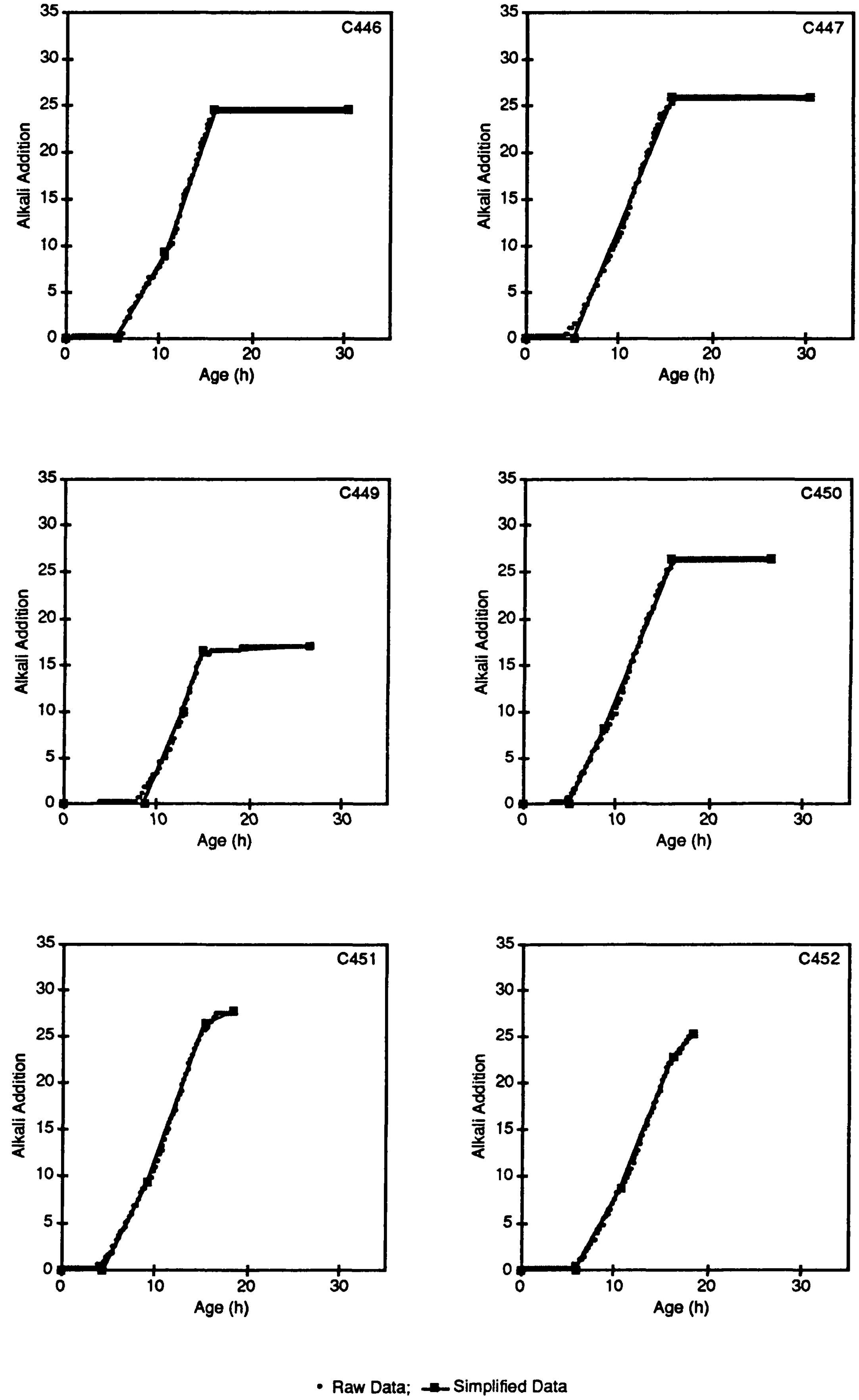


Figure A5.6(b): Alkali addition profiles for fermentations C446 to C452.

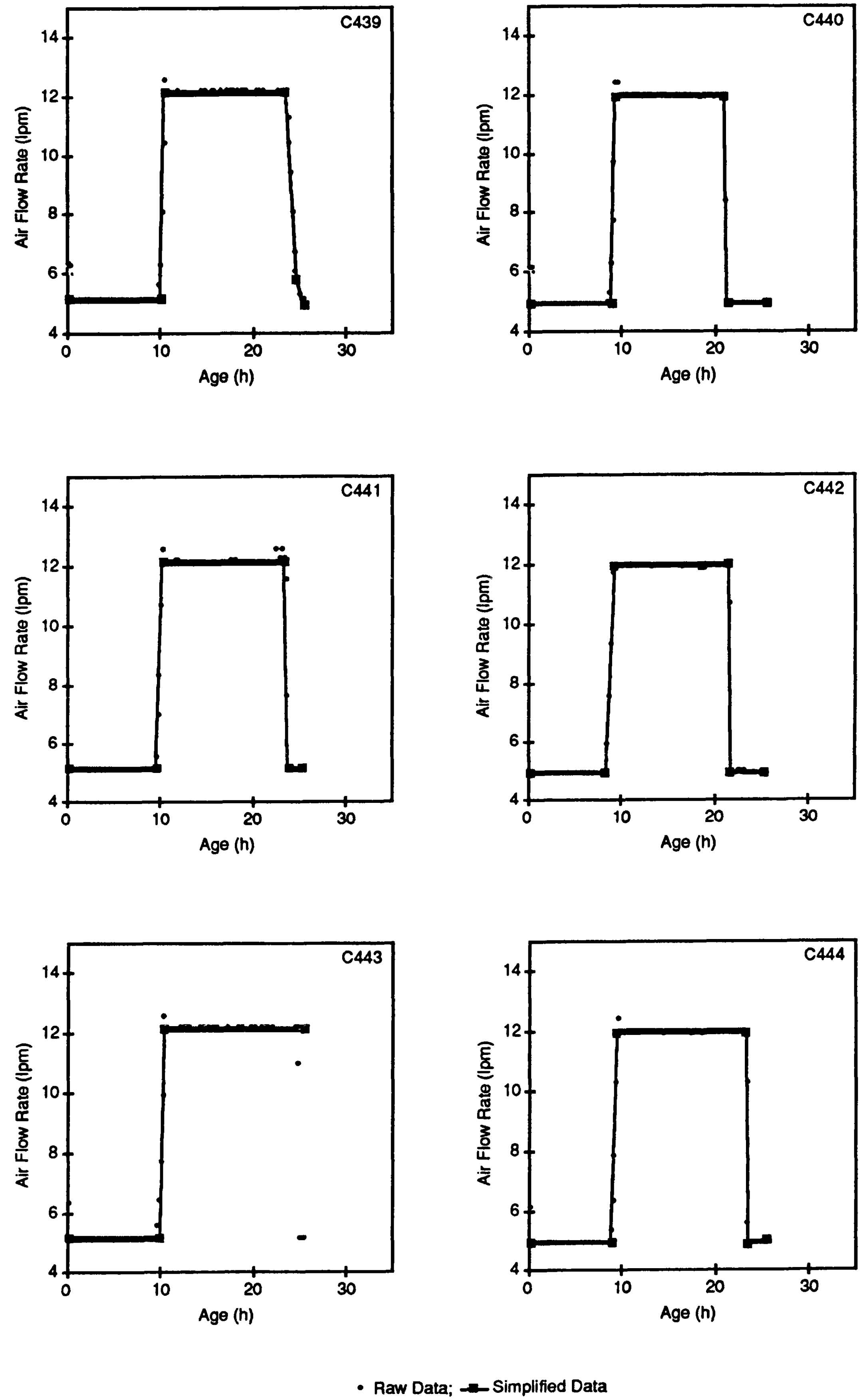


Figure A5.7(a): Air flow rate profiles for fermentations C439 to C444.

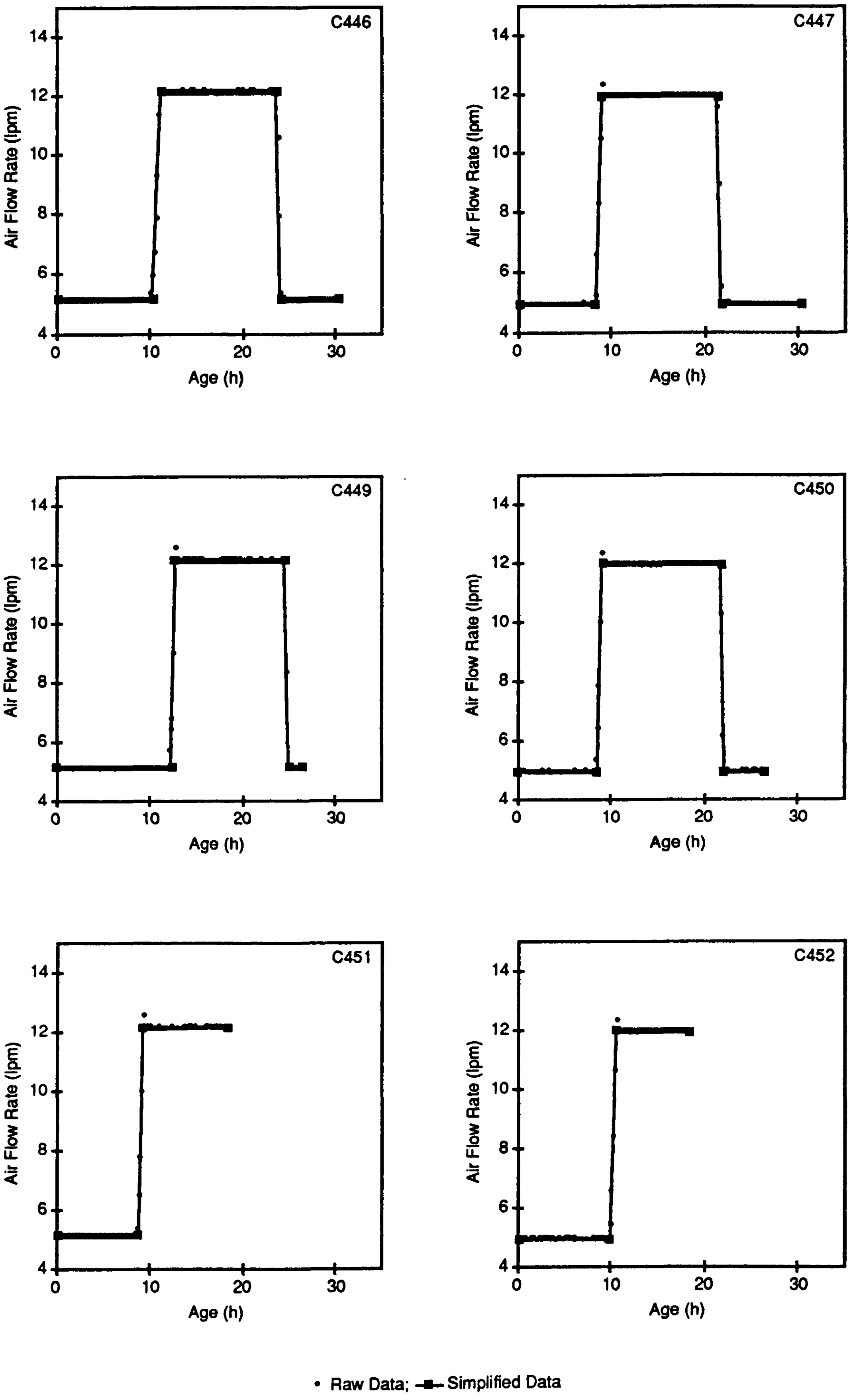


Figure A5.7(b): Air flow rate profiles for fermentations C446 to C452.

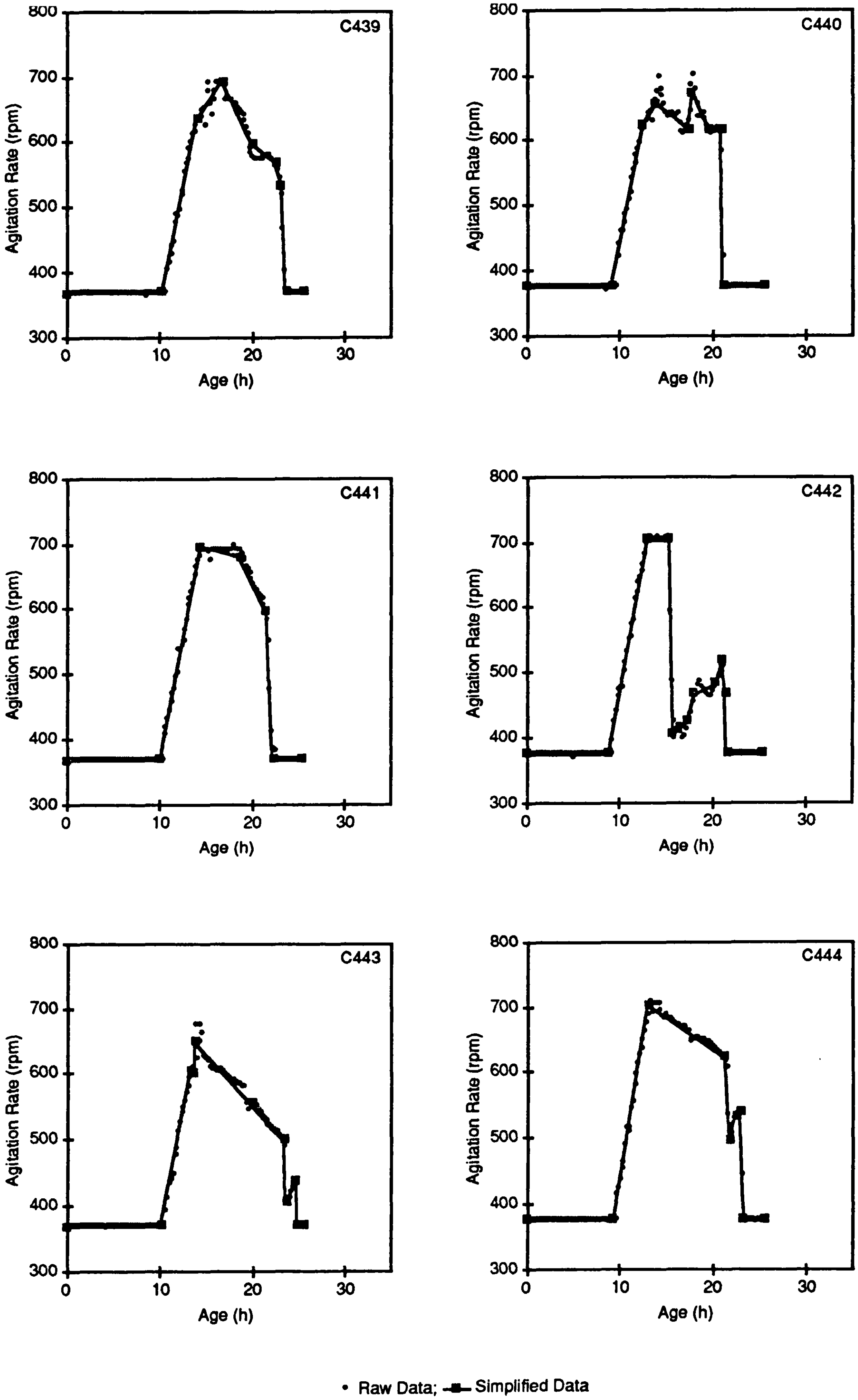


Figure A5.8(a): Agitation rate profiles for fermentations C439 to C444.

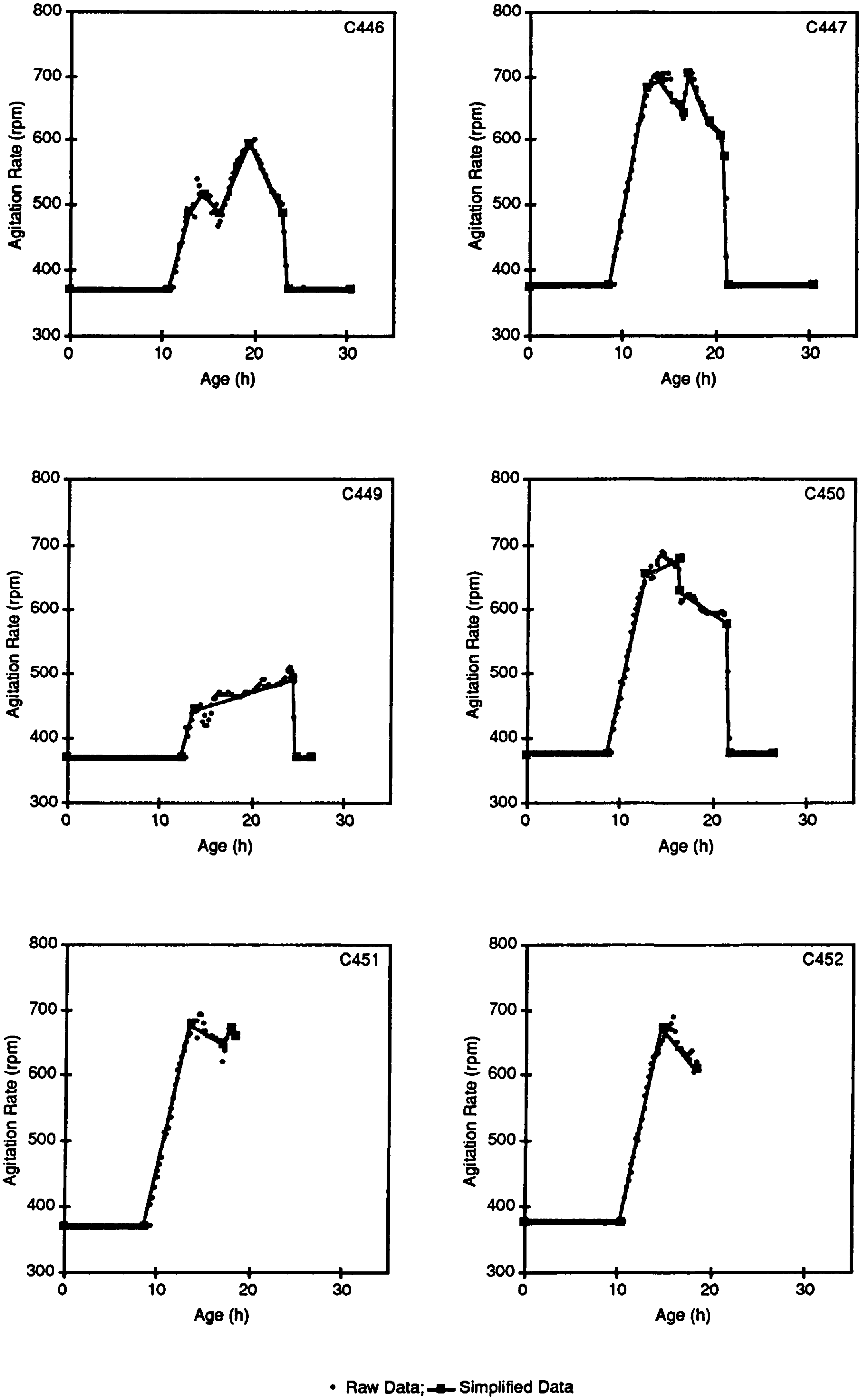


Figure A5.8(b): Agitation rate profiles for fermentations C446 to C452.

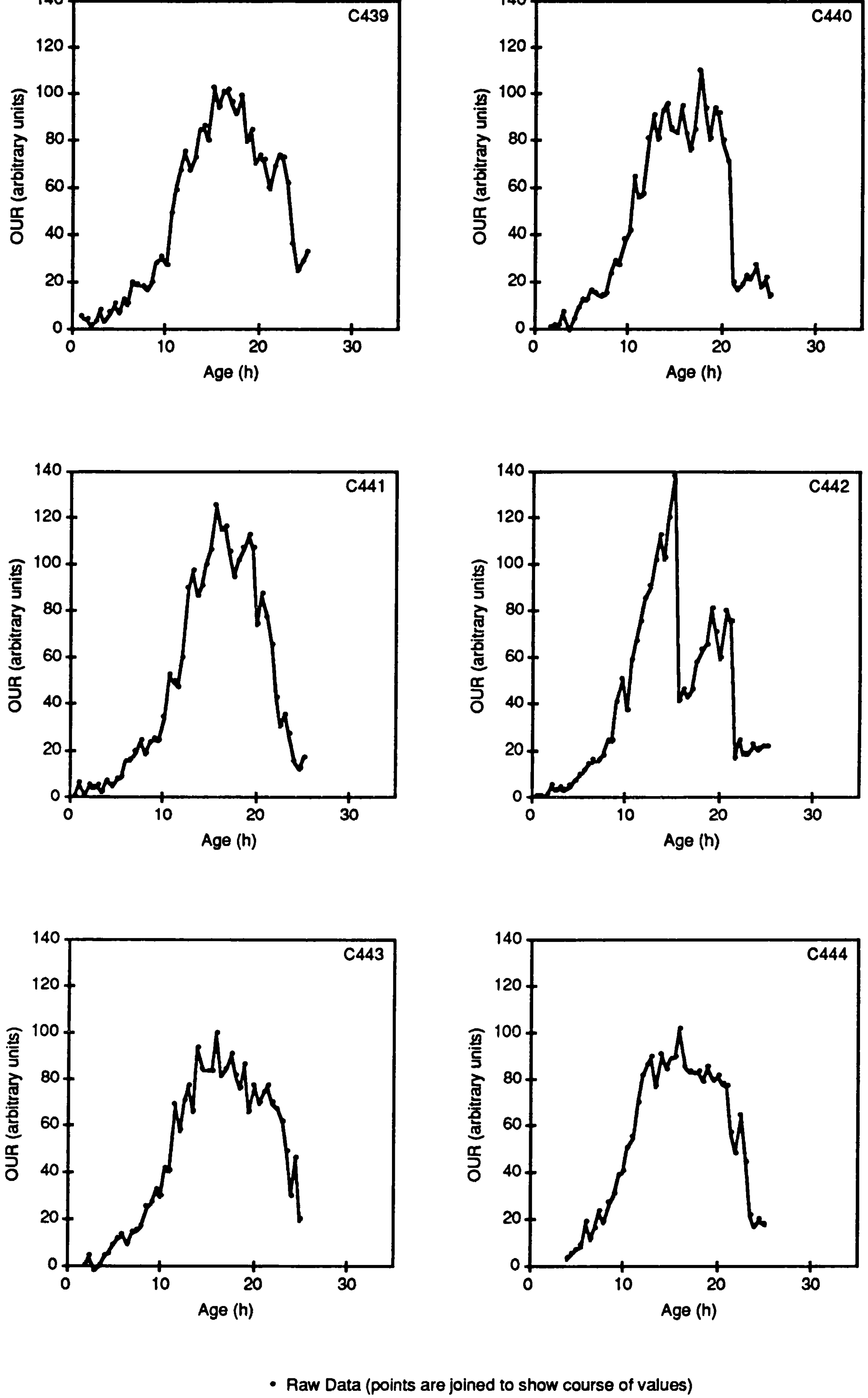


Figure A5.9(a): Oxygen uptake rate (OUR) profiles for fermentations C439 to C444.

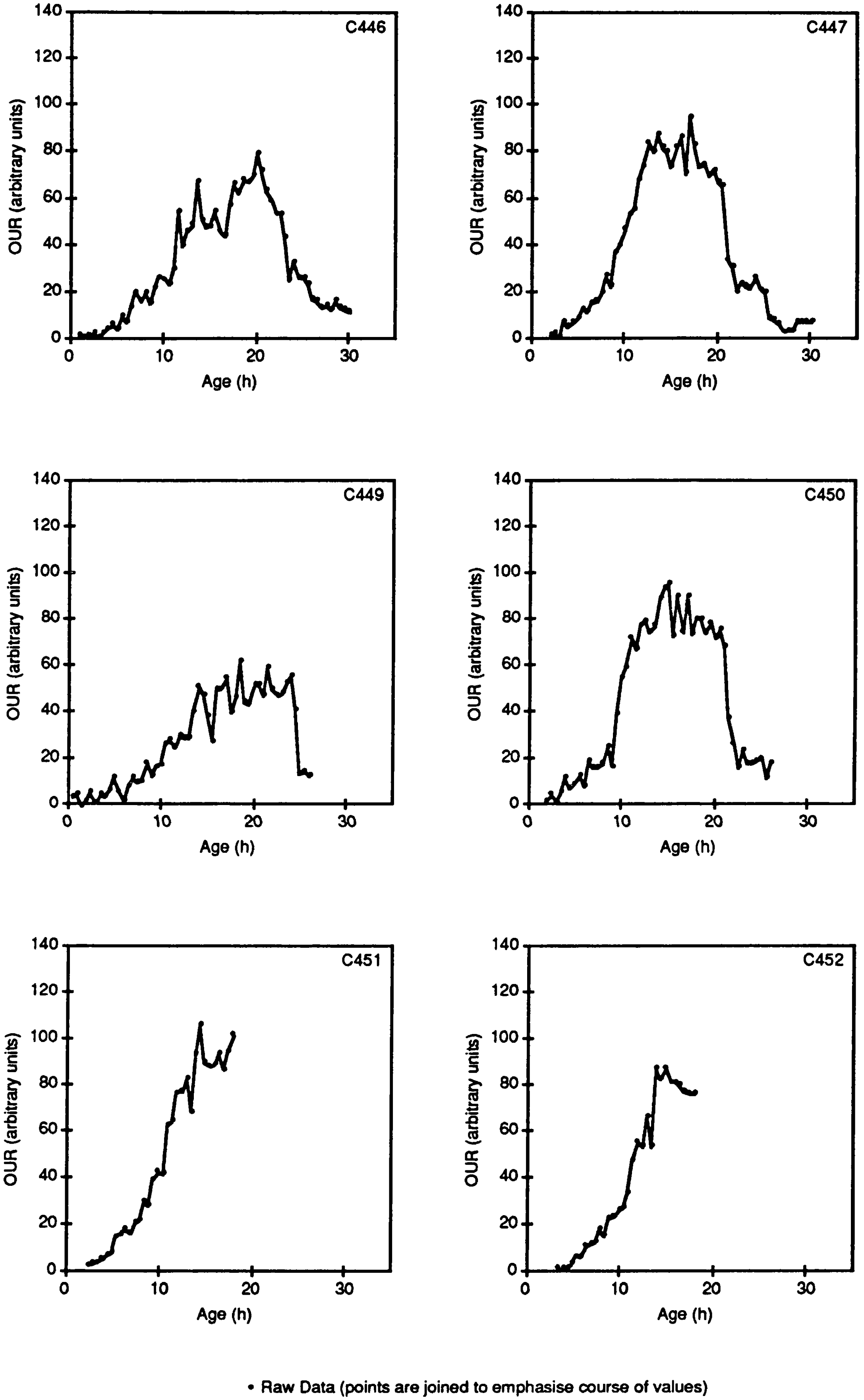


Figure A5.9(b): Oxygen uptake rate (OUR) profiles for fermentations C446 to C452.

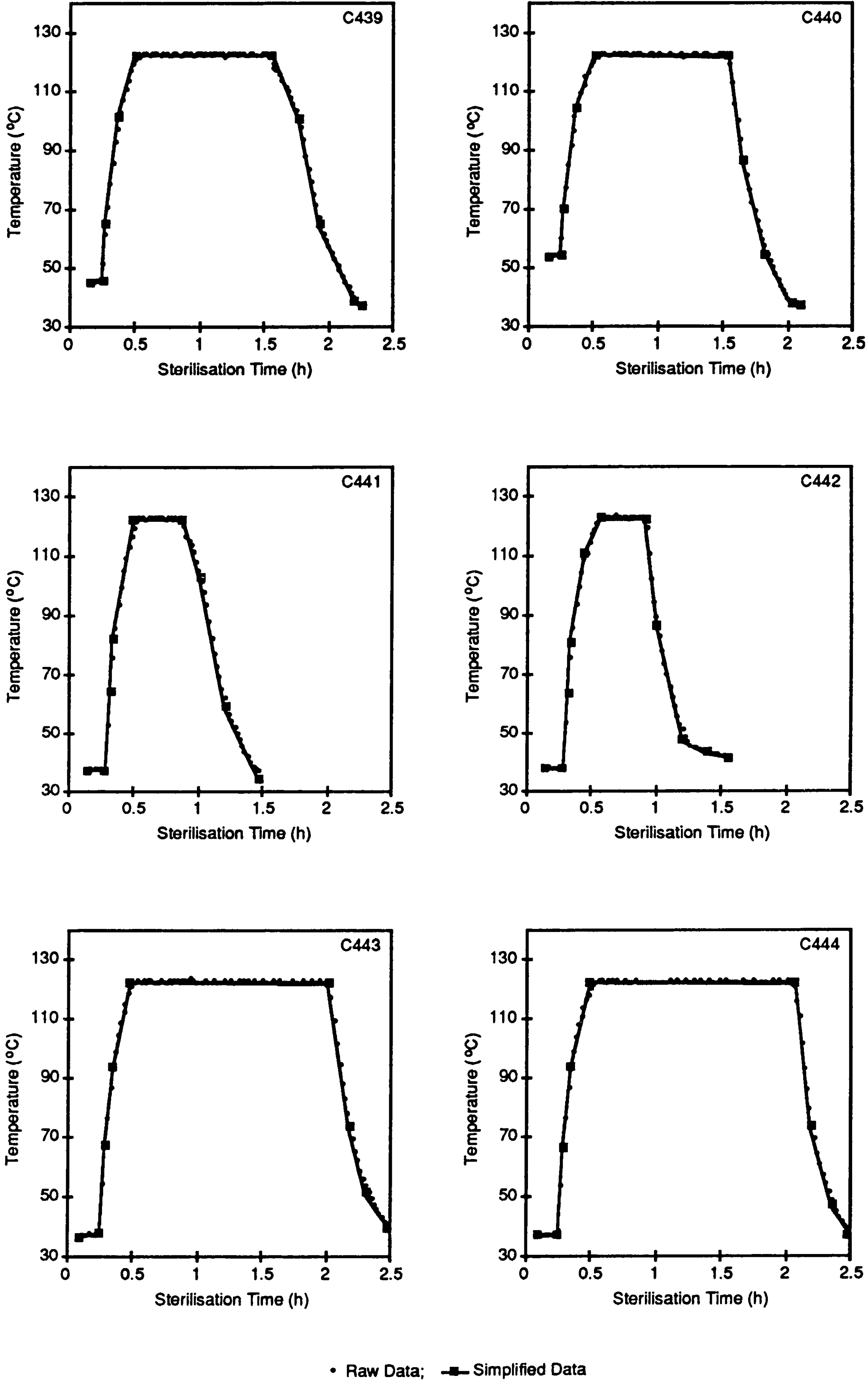


Figure A5.10(a): Sterilisation temperature profiles for fermentations C439 to C444.

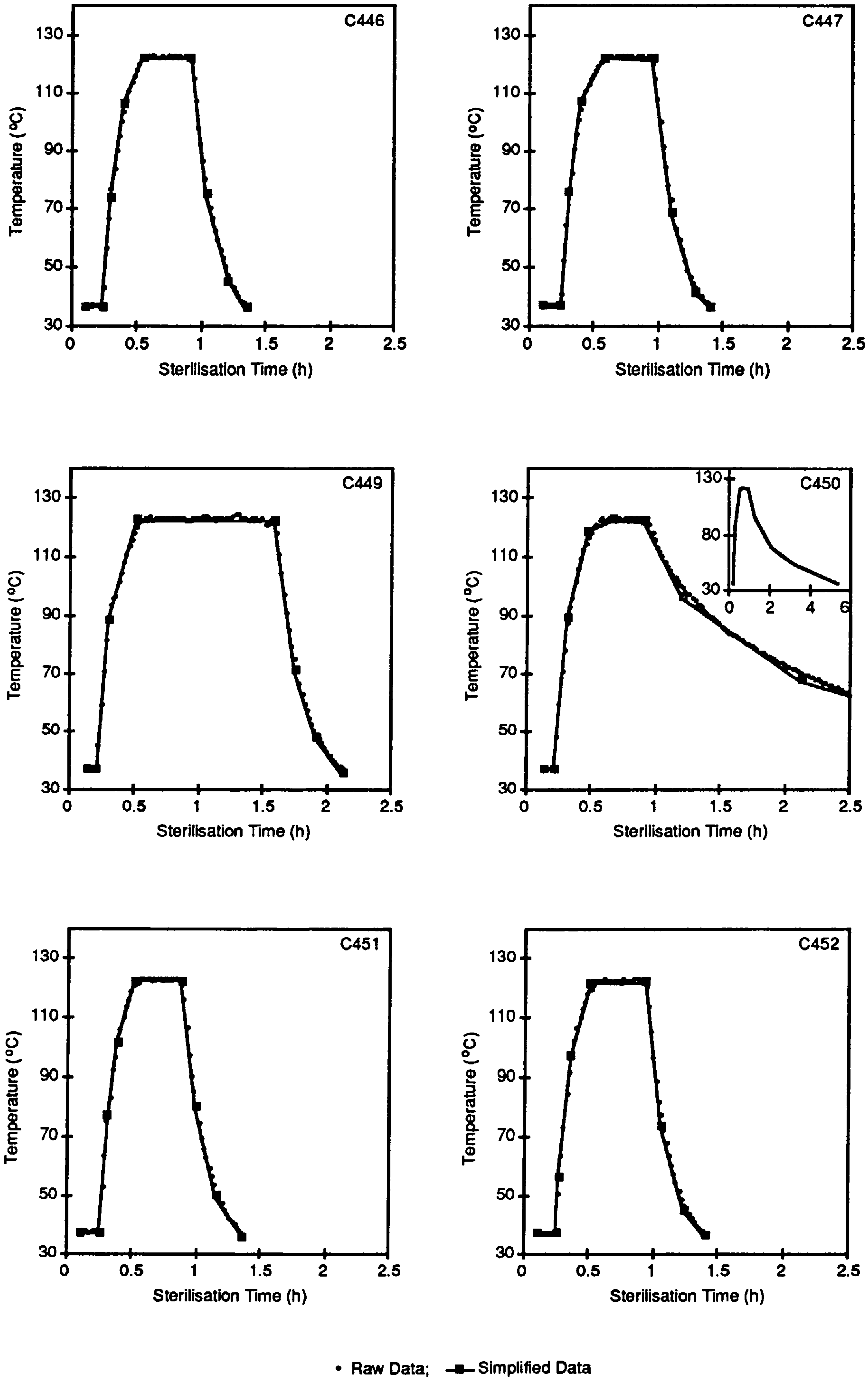


Figure A5.10(b): Sterilisation temperature profiles for fermentations C446 to C452.

REFERENCES

Bailey, J.E. and Ollis, D.F. (1986), 'Biochemical Engineering Fundamentals', Second Edition, McGraw-Hill.

Benterud, A. (1977), 'Vitamin Losses During Thermal Processing', in 'Physical, Chemical and Biological Changes in Food Caused by Thermal Processing', Hoyem, T. and Kvale, O. (eds), Applied Science Publishers, Chpt. 11, pp. 185-201.

Blanchard, S.M. and Barr, R.C. (1985), 'Comparison of Methods for Adaptive Sampling of Cardiac Electrograms and Electrocardiogram', Medical and Biological Engineering and Computing **23**: 377-386.

Bobrow, D.G. (1985), 'Qualitative Reasoning About Physical Systems: An Introduction', in 'Qualitative Reasoning About Physical Systems', Bobrow, D.G. (ed.), MIT Press, MA.

Buckland, B.C. (1984), 'The Translation of Scale in Fermentation Processes: The Impact of Computer Process Control', Bio/Technology **2**: 875-883.

Buckland, B.C. (1990), 'The Application of Computer Control to Improve Fermentation Processes', in 'Computer Control of Fermentation Processes', Omstead, D.R. (ed.), CRC Press, Florida, Chpt. 10, pp. 237-256.

Buono, M.A., Yang, S.S. and Erickson, L.E. (1986), 'Comparison of Two Methods of Selecting Smoothing Spline Functions for Estimation of Specific Rates in Fermentations', Chemical Engineering Communications **45**: 145-161.

Cantor, C.R. and Timasheff, S.N. (1982), 'Optical Spectroscopy of Proteins', in 'The Proteins', Neurath, H. and Hill, R.L. (eds), Academic Press, Third Edition, Vol. V, Chpt. 2, pp. 145-306.

Carleysmith, S.W. and Fox, R.I. (1984), 'Fermenter Instrumentation and Control', in 'Advances in Biotechnological Processes', Mizrahi, A. and van Wezel, A.L. (eds), Alan R. Liss N.Y. (publ.), Vol. 3, pp. 1-51.

Chen, Q., Wang, S. and Wang, J. (1989), 'Application of Expert System to the Operation and Control of Industrial Antibiotic Fermentation Process', in 'Computer Applications in Fermentation Technology', Fish, N.M., Fox, R.I. and Thornhill, N.F. (eds), Elsevier, pp. 253-261.

Cheung, J.T-Y. and Stephanopoulos, G. (1990a), 'Representation of Process Trends - Part I. A Formal Representation Framework', Computers and Chemical Engineering **14**: 495-510.

Cheung, J.T-Y. and Stephanopoulos, G. (1990b), 'Representation of Process Trends - Part II. The Problem of Scale and Qualitative Scaling', Computers and Chemical Engineering **14**: 511-539.

Clapp, K.P. and Ruel, G.J. (1991), 'Expert Systems in Bioprocessing', BioPharm, February 1991, pp. 28-35.

Codd, E.F. (1970), 'A Relational Model of Data for Large Shared Data Banks', Communications of the ACM **13**: 377-

Cooney, C.L., Raju, G.K. and O'Connor, G. (1991), 'Expert Systems and Neural Nets for Bioprocess Operation', International Symposium on Bioprocess Modelling and Control, Newcastle, U.K., Jan. 21-22.

Corbett, K. (1985), 'Design, Preparation and Sterilization of Fermentation Media', in 'Comprehensive Biotechnology', Moo-Young, M. (ed.), Pergamon Press, Vol. 1, pp. 127-139.

Cox, M.G. (1972), 'The Numerical Evaluation of B-Splines', Journal of the Institute of Mathematics and its Applications **10**: 134-149.

Craven, P. and Wahba, G. (1979), 'Smoothing Noisy Data with Spline Functions', Numerische Mathematik **31**: 377-403.

de Boor, C. (1978), 'A Practical Guide to Splines', Springer-Verlag.

de Hoog, F.R. and Hutchinson, M.F. (1987), 'An Efficient Method for Calculating Smoothing Splines Using Orthogonal Transformations', Numerische Mathematik **50**: 311-319.

Deindoerfer, F.H. and Humphrey, A.E. (1959) 'Analytical Method for Calculating Heat Sterilization Times', Applied Microbiology **7**: 256-264.

Dhurjati, P.S. and Leipold, R.J. (1990), 'Biological Modeling', in 'Computer Control of Fermentation Processes', Omstead, D.R. (ed.), CRC Press, Florida, Chpt. 8, pp. 207-220.

Dierckx, P. (1975), 'An Algorithm for Smoothing, Differentiation and Integration of Experimental Data Using Spline Functions', Journal of Computational and Applied Mathematics **1**: 165-184.

Dreyfus, H.L. and Dreyfus, S.E. (1986), 'Mind Over Machine: The Power of Human Intuition and Expertise in the Era of the Computer', Basil Blackwell Ltd., Oxford.

Esener, A.A., Roels, J.A. and Kossen, N.W.F. (1981), 'On the Statistical Analysis of Batch Data', Biotechnology and Bioengineering **23**: 2391-2396.

Fastert, H.J. (1990), 'Supervisory Computer Systems', in 'Computer Control of Fermentation Processes', Omstead, D.R. (ed.), CRC Press, Florida, Chpt. 6, pp. 147-163.

Finkelstein, R. (1990), 'Multiuser Databases: The SQL', Byte, May 1990, pp. 136-150.

Fish, N.M., Fox, R.I. and Thornhill, N.F. (eds) (1989), 'Computer Applications in Fermentation Technology', Elsevier.

Fordyce, A.P., Rawlings, J.B. and Edgar, T.F. (1990), 'Control Strategies for Fermentation Processes', in 'Computer Control of Fermentation Processes', Omstead, D.R. (ed.), CRC Press, Florida, Chpt. 7, pp. 165-206.

Fox, R.I. (1984), 'Computers and Microprocessors in Industrial Fermentations', in 'Topics in Enzyme and Fermentation Biotechnology', Wiseman, A. (ed.), Ellis Horwood London (publ.), Vol. 8, Chpt. 4, pp. 125-174.

Fu, C., Wang, S. and Wang, J. (1989), 'A Modelling Approach to Trouble Diagnosis by Multilevel Fuzzy Functions and Its Application', in 'Computer Applications in Fermentation Technology', Fish, N.M., Fox, R.I. and Thornhill, N.F. (eds), Elsevier, pp. 411-420.

Ghosh S.P. (1989), 'Statistical Databases for Automated Engineering', in 'Statistical Process Control in Automated Manufacturing', Keats, J.B. and Hubele, N.F. (eds), Marcel Dekker, N.Y. (publ.), Chpt. 10, pp. 195-217.

Gimenez-Gallego, G., Rodkey, J., Bennett, C., Rios-Candelore, M., DiSalvo, J. and Thomas K.A. (1985), 'Brain-Derived Acidic Fibroblast Growth Factor: Complete Amino Acid Sequence and Homologies', Science **230**: 1385-1388.

Gimenez-Gallego, G., Conn, G., Hatcher, V.B. and Thomas, K.A. (1986), 'The Complete Amino Acid Sequence of Human Brain-Derived Acidic Fibroblast Growth Factor', Biochemical and Biophysical Research Communications **138**: 611-617.

Gottschalk, A. (1972), 'Interaction Between Reducing Sugars and Amino Acids Under Neutral and Acidic Conditions', in 'Glycoproteins: Their Composition, Structure and Function', Gottschalk, A. (ed.), Second Edition, Elsevier, Chpt. 3-2, pp. 141-157.

Greville, T.N.E. (1969), 'Theory and Applications of Spline Functions', Academic Press, New York, pp. 1-36.

Hale, J.C. and Sellars, H.L. (1981), 'Historical Data Recording for Process Computers', Chemical Engineering Progress, Nov. 1981, pp. 38-43.

Halme, A. (1989), 'Expert System Approach to Recognize the State of Fermentation and to Diagnose Faults in Bioreactors', in 'Computer Applications in Fermentation Technology', Fish, N.M., Fox, R.I. and Thornhill, N.F. (eds), Elsevier, pp. 159-168.

Holman, J.P. and Gajda, W.J. (1978), 'Experimental Methods for Engineers', Third Edition, McGraw-Hill, p. 65-66.

Hutchinson, M.F. and de Hoog, F.R. (1985), 'Smoothing Noisy Data with Spline Functions', Numerische Mathematik **47**: 99-106.

Karim, M.N. and Halme, A. (1989), 'Reconciliation of Measurement Data in Fermentation Using On-Line Expert System', in 'Computer Applications in Fermentation Technology', Fish, N.M., Fox, R.I. and Thornhill, N.F. (eds), Elsevier, pp. 37-46.

Kaufmann, A.K. and Gupta, M.M. (1991), 'Introduction to Fuzzy Arithmetic: Theory and Applications', Van Nostrand Reinhold, N.Y.

Keats, J.B. and Hubele, N.F. (eds) (1989), 'Statistical Process Control in Automated Manufacturing', Marcel Dekker.

Kolodner, J.L. (1987), 'Extending Problem Solver Capabilities Through Case-Based Inference', Proceedings of the 4th Annual International Machine Learning Workshop.

Kolodner, J.L. and Simpson, R.L. (1987), 'Problem Solving and Dynamic Memory', in 'Experience, Memory and Reasoning', Kolodner, J.L. and Riesbeck, C. (eds), Lawrence Erlbaum (publ.), Chpt. 6, pp. 99-114.

Konstantinov, K. and Yoshida, T. (1989), 'Physiological State Control of Fermentation Processes', Biotechnology and Bioengineering **33**: 1145-1156.

Konstantinov, K. and Yoshida, T. (1990a), 'An Expert Approach for Control of Fermentation Processes as Variable Structure Plants', Journal of Fermentation and Bioengineering **70**: 48-57.

Konstantinov, K. and Yoshida, T. (1990b), 'On-Line Monitoring of Representative Structured Variables in Fed-Batch Cultivation of Recombinant *Escherichia coli* for Phenylalanine Production', Journal of Fermentation and Bioengineering **70**: 420-426.

Koton, P. (1988), 'Reasoning About Evidence in Causal Explanations', Proceedings of the AAAI-88.

Lewis, M. (1990), 'SQL Base and SQL Vision', Personal Computer World, Feb. 1990, pp. 174-178.

Linemeyer, D.L., Kelly, L.J., Menke, J.G., Gimenez-Gallego, G., DiSalvo, J. and Thomas, K.A. (1987), 'Expression in *Escherichia coli* of a Chemically Synthesized Gene for Biologically Active Bovine Acidic Fibroblast Growth Factor', Bio/Technology **5**: 960-965.

Locher, G., Sonnleitner, B. and Fiechter, A. (1990), 'Pattern Recognition: A Useful Tool in Technological Processes', Bioprocess Engineering **5**: 181-187.

Love, P.L. and Simaan, M. (1988), 'Automatic Recognition of Primitive Changes in Manufacturing Process Signals', Pattern Recognition **21**: 333-342.

McIlraith, S.A. (1989), 'Qualitative Data Modeling: Application of a Mechanism for Interpreting Graphical Data', Computational Intelligence **5**: 111-120.

McWhirter, J.G. (1981), 'A Well-conditioned Cubic b-spline Model for Processing Laser Anemometry Data', Optica Acta **28**: 1453-1475.

Mendenhall, W. and Sincich, T. (1988), 'Statistics for the Engineering and Computer Sciences', Second Edition, Collier Macmillan Publishers.

Montague, G.A., Morris, A.J. and Ward, A.C. (1989), 'Fermentation Monitoring and Control: A Perspective', Biotechnology and Genetic Engineering Reviews **7**: 147-188.

Montgomery, D.C. and Friedman, D.J. (1989), 'Statistical Process Control in a Computer-Integrated Manufacturing Environment', in 'Statistical Process Control in Automated Manufacturing', Keats, J.B. and Hubele, N.F. (eds), Marcel Dekker, Chpt. 5, pp. 67-87.

Morris, A.J., Montague, G.A., Tham, M.T., Aynsley, M., Di Massimo, C. and Lant, P. (1991), 'Towards Improved Process Supervision - Algorithms and Knowledge Based Systems', International Symposium on Bioprocess Modelling and Control, Newcastle, U.K., Jan. 21-22.

Oakland, J.S. (1986), 'Statistical Process Control: A Practical Guide', Heinemann.

Omstead, D.R., Phillips, J.A. and Humphrey, A.E. (1990), 'Indirect Parameter Estimation', in 'Computer Control of Fermentation Processes', Omstead, D.R. (ed.), CRC Press, Florida, Chpt. 4, pp. 107-127.

Oner, M.D., Erickson, L.E. and Yang, S.S. (1986), 'Utilization of Spline Functions for Smoothing Fermentation Data and for Estimation of Specific Rates', Biotechnology and Bioengineering **28**: 902-918.

Pascal, F. (1989), 'A Brave New World?', Byte, Sept. 1989, pp. 247-256.

Pietka, E. (1991), 'Feature Extraction in Computerized Approach to ECG Analysis', Pattern Recognition **24**: 139-146.

Postlethwaite, B.E. (1989), 'A Fuzzy State Estimator for Fed-Batch Fermentation', Transactions of the Institution of Chemical Engineers, Part A, Chemical Engineering Research and Design **67**: 267-272.

Press, W.H., Flannery, B.P., Teukolsky, S.A. and Vetterling, W.T. (1986), 'Numerical Recipes. The Art of Scientific Computing', Cambridge University Press.

Reinsch, C.H. (1967), 'Smoothing by Spline Functions', Numerische Mathematik **10**: 177-183.

Reinsch, C.H. (1971), 'Smoothing by Spline Functions II', Numerische Mathematik **16**: 451-454.

Rumelhart, D.E. and McClelland, J.L. (1986), 'Parallel Distributed Processing: Explorations in the Microstructure of Cognition', Vol. 1, MIT Press, Cambridge, MA.

Schmid, F.X. (1989), 'Spectral Methods of Characterizing Protein Conformation and Conformational Changes', in 'Protein Structure: A Practical Approach', Creighton, T.E. (ed.), IRL Press, Chpt. 11, pp. 251-286.

Schumaker, L.L. (1981), 'Spline Functions: Basic Theory', John Wiley and Sons.

Silverman, B.W. (1984), 'A Fast and Efficient Cross-Validation Method for Smoothing Parameter Choice in Spline Regression', Journal of the American Statistical Association **79**: 584-589.

Silverman, B.W. (1985), 'Some Aspects of the Spline Smoothing Approach to Non-parametric Regression Curve Fitting', Journal of the Royal Statistical Society B **47**: 1-52.

Singh, V., Hensler, W. and Fuchs, R. (1989), 'Optimization of Batch Fermentor Sterilization', Biotechnology and Bioengineering **33**: 584-591.

Stephanopoulos, G. and Tsiveriotis, C. (1989), 'Toward a Systematic Method for the Generalization of Fermentation Data', in 'Computer Applications in Fermentation Technology', Fish, N.M., Fox, R.I. and Thornhill, N.F. (eds), Elsevier, pp. 169-178.

Svenson, D. and McLean, M. (1991), 'Computer-Integrated Electronic Batch Records and Pharmaceutical Documents: A First Step Toward Paperless Factories', Pharmaceutical Technology International, March 1991, pp. 31-34.

Thomas, K.A. and Gimenez-Gallego, G. (1986), 'Fibroblast Growth Factors: Broad Spectrum Mitogens with Potent Angiogenic Activity', Trends in Biochemical Sciences **11**: 81-84.

Tou, J.T. and Gonzalez, R.C. (1974), 'Pattern Recognition Principles', Addison-Wesley, MA.

Tufte, E.R. (1983), 'The Visual Display of Quantitative Information', Graphics Press.

Udupa, J.K. and Murthy, I.S.N. (1980), 'Syntactic Approach to ECG Rhythm Analysis', IEEE Transactions on Biomedical Engineering BME **27**: 370-375.

Utreras, D.F. (1980), 'Sur le Choix du Parametre d'ajustement dans le Lissage par Fonctions Spline', Numerische Mathematik **34**: 15-28.

Wang, N.S. and Stephanopoulos, G. (1986), 'Computer Applications to Fermentation Processes', in 'CRC Critical Reviews in Biotechnology' Vol. 2 (Issue 1), pp. 1-103.

Wegman, E.J. and Wright, I.W. (1983), 'Splines in Statistics', Journal of the American Statistical Association **78**: 351-365.

Wold, S. (1974), 'Spline Functions in Data Analysis', Technometrics **16**: 1-11.

Zadeh, L.A. (1965), 'Fuzzy Sets', Information and Control **8**: 338-353.