# Framework for online modeling, optimization and monitoring of bioprocesses

**Dissertation**

zur Erlangung des akademischen Grades

Doktoringenieur (Dr.-Ing.)

**genehmigt durch die**

Matematisch-Naturwissenschaftlich-Technischen Fakultät

(Ingenieurwissenschaftlicher Bereich)

der Martin-Luther-Universität Halle-Wittenberg

von Herrn Ezequiel Franco Lara

geb. am 4. Oktober 1969 in México City, México

Gutachter:

1. Prof. Dr. Andreas Lübbert
2. Prof. Dr. Markus Pietzsch
3. Prof. Dr. em. Karl Schügerl

Halle (Saale), 15.10.2002

**urn:nbn:de:gbv:3-000004115**
[http://nbn-resolving.de/urn/resolver.pl?urn=nbn%3Ade%3Agbv%3A3-000004115]

## *Chapters*

## Appendixes

# Acknowledgments

I acknowledge the Deutsche Akademischer Austauschdienst (DAAD) for the grant given to me to accomplish my Ph. D. work in Germany. They supported me not only financially, but also provided me in a constant manner with most valuable guidelines to live and study in Germany.

I would like to thank Maren Tornow for her inestimable friendship, her never-ending disposition and most valuable help to find a place where to stay in Halle and to come in contact with the university life at the MLU. Gracias!

I acknowledge Rimvydas Simutis for the introduction and training in the HybNet environment system and specially Rui Oliveira for his patience while working with me and his inestimable friendship: muito obrigado!

I would like to acknowledge my workgroup colleges, specially Martina Anwand for her endless dedication and good disposition in the laboratory and in general, for her support during the whole time of my stay at the Institute: thanks for the patience by correcting translations! Thanks also for the big help given and good laughing moments in the laboratory to Sylke Gutknecht and Barbara Friedling: you are really a big team to work with! I thank Dr. Kerstin Junghans for the big help given to go through the tortuous ways of bureaucracy at the end of my stay and for the most "valuable" non-scientific discussion we had.

I'm very grateful with Istvan Godó and Vytautas Galvanauskas for the revision of the first manuscripts of this thesis and their valuable opinions, comments and critics.

I acknowledge Dr. Volk and Dr. Hertel for the outstanding scientific support they gave to me throughout my Ph. D. work. I want to thank also the personal friendship they offer me and, it is for sure, I will never forget some memorial moments we spend together that made me experienced the pleasant and charming "countryside" of Germany.

I acknowledge Prof. A. Lübbert for offer and giving me the chance to work on my Ph. D. project. Without his excellent support and guidance "in Rat und Tat" this work would never became reality.

I acknowledge specially my family for being a source of strength and encouragement during my whole stay in Germany. It is not easy to live four years abroad home without you!

# Summary

The establishment of a framework for the online modeling, monitoring and optimization of bioprocesses is presented. The modeling and monitoring schemas comprise a mathematical formulation of the microbial system through neural network-based hybrid models. Under the concept of hybrid model is understood a set of non linear differential equations and neural network or neuro-fuzzy models, which are incorporated to account exclusively for key kinetic parameters, as the specific growth or specific production rate. The set of differential equations describes the mass balance relationships of the biotechnical process. Three different microbial cultivations were carried out, two of them comprising recombinant strains of the bacteria *Escherichia coli* and another a recombinant strain of the yeast *Kluyveromyces lactis*. The first part of this work describes the hybrid modeling of specific kinetic rates in a multi-substrate batch cultivation of *Escherichia coli* B pUBS520 p12023. The method takes advantage of available *a priori* knowledge and theoretical considerations found in specialized literature. After experimental evidence gain, the reduction of the data demanded for training purposes is demonstrated. The use of neural network-based techniques is established as milestone for model development. In these applications in particular, the neural networks are used as "black box" model components. These black-box models associate certain known and measurable process input variables to other output variables of the process, whose values are usually not known or not measureable. A complex relationship between them is supposed to occur, but would be only described by the neural network after a proper training procedure. The second part describes the off-line modeling and optimization of the production of the viral capsid complex VP1-DHFR using a strain of *Escherichia coli* BL21 under fed-batch conditions. The optimization task is fulfilled using an evolutionary procedure: first actualization of a model based on available information and *a priori* knowledge; the actualized approach was used in a model-based optimization; a control experiment is performed to test the correctness of the model's assumptions, supplying additional data for model's improvement. It is demonstrated how this technique enables the estimation of key unmeasured variables for the process, like the specific growth rate. This formulation not only improves the description and understanding of the biological system, but also allows a model-based optimization of the operating conditions of the cultivation. The last two parts of this work are dedicated exclusively to the online implementation of all the aforementioned modeling, monitoring and optimization techniques. The first of these sections describes the on-line neuro-fuzzy modeling and optimization of the production of VP1-DHFR. It is proved here how this method can effectively deal with changing environments and that increased productivities were obtainable, as compared to optimizations employing conventional approaches. Although off-line training of neural network systems is usually a straightforward matter, online training of hybrid models confirmed its ability to circumvent slow convergence problems caused by the training phase. Moreover, validated at the optimized production of recombinant protein complex GAL80/HIS-TAG with *Kluyveromyces lactis*, the real-time estimation and monitoring of bioprocesses' profits through a soft-sensor is presented. The approach demonstrates the inherent plasticity of the neural network-based approach to infer complex kinetic rates, using exclusively control variables like temperature and process correlated measurements available online. Finally, being the reduction of the invested time for developing and improving a given process a fundamental demand in today's biotechnology, the developed online framework improved the gain of process knowledge at high learning rates. Such a methodology can be viewed either as an alternative or as an intermediary evolution step between pure empirical towards full formal mechanistic approaches.

# Zusammenfassung

Die Festlegung eines Bezugssystems und der Rahmenbedingungen für die online Modellierung, Beobachtung und Optimierung von Bioprozessen wird hier dargestellt. Die Modellierungs- und Beobachtungsschemata beinhalteten eine auf hybriden Modellen basierende mathematische Formulierung des mikrobiellen Systems. Unter dem Begriff „hybride Modelle“ versteht man die Zusammenstellung einer Reihe von nicht-linearen Differentialgleichungen und verschiedener „black-box“ Modelle. Während die Kinetik bzw. die spezifischen Wachstums- und Produktionsraten mit Hilfe künstlicher neuronaler Netze oder neuro-fuzzy Ansätzen dargestellt wurden, beschrieben die Differentialgleichungen die Massenbilanzen des Bioreaktors. Unter batch bzw. fed-batch Bedingungen wurden 3 verschiedene Mikroorganismenstämme kultiviert; 2 davon waren rekombinante *Escherichia coli*–Stämme. Zur Validierung der Ergebnisse wurde ein rekombinanter Stamm der Hefe *Kluyveromyces lactis* eingesetzt. Der erster Teil dieser Arbeit umfasst die hybride Modellierung einer multi-Substrat batch-Kultivierung von *Escherichia coli* B pUBS520 p12023 (MAK-33). Im Modell, wurde die spezifische Wachstumsrate durch einen neuronalen Netzansatz formuliert. Bei der Modellierungsprozedur wurde nachgewiesen, wie das verfügbare *a priori*-Prozesswissen systematisch genutzt werden kann und wie sich die Datenmenge für Trainingszwecke reduzieren läßt. Da im Voraus keine Daten über spezifische Substrataufnahmeraten zur Verfügung standen, wurden die Parameter des hybriden Modells indirekt mit Hilfe eines „Random Search“-Verfahrens identifiziert. In zweitem Teil der Arbeit werden die off-line Modellierung und Optimierung der Produktion des viralen Hüllproteinskomplex VP1-DHFR durch einen rekombinanten *Escherichia coli* BL21 Stamm präsentiert. Zur Optimierung der Prozessführung wurde eine evolutionäre Methodik genutzt: ein Modell wurde unter Verwendung des verfügbaren Wissens und vorhandener Datensätzen aktualisiert; dieses aktualisierte Modell wurde zur Optimierung des Prozesses eingesetzt; dann erfolgte ein Kontrollexperiment, um die Richtigkeit der Modellannahmen zu bestätigen bzw. um Daten zur weiteren Verbesserung des Modells zu gewinnen. Weil keine Daten über der Kinetik verfügbar waren, wurde die spezifische Wachstumsrate mit Hilfe eines künstlichen neuronalen Netzes beschrieben. Die letzten Teile dieser Arbeit sind der online-Anwendung der verschiedenen Techniken zur Modellierung, Beobachtung und Optimierung gewidmet. Bei der Produktion der Protein VP1-DHFR wurde eine neuro-fuzzy-Modellierung angewendet, die auf wechselnde Bedingungen reagieren kann, so das im Vergleich zu klassischen Optimierungsansätzen, höhere Produktivitäten erzielt werden konnten. Probleme wie die langsame Konvergenz des Identifizierungsverfahrens, die beim off-line-Training künstlicher neuronaler Netze auftreten, werden durch die online Training von hybride Modelle deutlich vermindert. Am Beispiel der Produktion des rekombinanten Proteinkomplexes GAL80/HIS-TAG durch die Hefe *Kluyveromyces lactis*, wurde ein soft-Sensor vorgestellt, der die Echtzeitschätzung und die Beobachtung von Profitfunktionen bzw. die Beurteilung der Leistung eines Prozesses erlaubte. Gezeigt wurde auch die inhärente Plastizität eines neuronalen Netzes, die komplexe Kinetik eines Prozesses nur auf der Basis online gemessener prozesskorrelierter Variablen zu charakterisieren. Mit Hilfe hybrider Modelle kann das vorgestellte Bezugssystem selbst komplexe Prozesse in einer geringen Anzahl von Optimierungsschritten verbessern. Die Methodik kann dazu verwendet werden, schnell und kostengünstig, d. h. mit geringem experimentellen Aufwand, die optimale Führung bei der Prozessentwicklung zu erzielen oder laufende Produktionsprozesse zu optimieren.

# Chapter 1

# Introduction

## 1.1 Recombinant protein production systems

The importance of applied biotechnology using recombinant microorganisms is increasing more and more, especially in the production of biologicals for human and veterinary pharmaceutical purposes. Based on recombinant technology, a multibillion dollar industry was born with the potential to affect our lives, being the production of recombinant proteins the most commercially important (Swartz, 1996). The most successful application of recombinant DNA (rDNA) technology has taken place on the *Escherichia coli* bacteria: considering the combination of a profound knowledge of the physiology and genetic, a well developed ability to alter the microorganism and rapidly determine the consequences of these alterations, the cultivation of recombinant *Escherichia coli* bacteria strains has presented advantageous features for enabling the production of recombinant protein products at low costs (Swartz, 1996). Additionally, since the bacteria have rapid growth rates and the ability to rapidly metabolize substrates as well as to produce heterologous proteins from relatively simple cultivation media, higher productivities can be achieved, as compared with other expression systems. Regarding all recombinant products available from this novel industry, human pharmaceuticals play the predominant role concerning their production and marketing.

Concerning the production process, where the most important goal is to attain an effective production of high levels of recombinant proteins, the organism may be subjected to stressful situations. Efficient production is usually obtained by growing the culture quickly to a high cell density, e. g., 30 to 50 g (dry weight)/liter culture or higher, and then inducing product formation (Yee and Blanch, 1992). Under these circumstances, the maximum oxygen assimilation rate of the microorganism can exceed or take to the limit the fermentor's oxygen delivery capacity. Therefore, in order to avoid oxygen depletion, it is common to limit the carbon and energy source while entering the protein production phase. Commonly, glucose is used simultaneously as carbon and energy source and its limitation results in a decrease of the specific growth rate.

The production phase is a demanding and arduous process for the cells, especially when the induction of the recombinant protein expression through a promoter occurs. With this action, a complex internal process is accomplished, initiating a metabolic demand for the protein production. The ideal promoter may be regulated for a minimal protein expression during the high cell density cultivation phase and be capable of rapid transcription after the induction takes place, without affecting other metabolic processes. This situation requires an accurate process optimization strategy to enhance process performance, which is usually the improvement of the protein production. Especially in the case of the recombinant protein production phase, the metabolic resources should be concentrated towards the protein expression. Figure 1.1 depicts an idealized picture of the features for high cell density cultures.

Process optimization strategies search basically for the cell's most adequate environmental conditions for product formation. Temperature and pH for high density bacterial growth may not be optimal for recombinant protein production (Kopetski *et al.*, 1989). For some products, it has been shown that lower temperatures often favors the production of valuable soluble proteins (native proteins). It is suspected that such low temperatures displace the equilibrium between the native product and its inactive counterpart, protein in inclusion bodies (Tsai *et al.*, 1995). Additionally, slow metabolic rates can lead also to less cytoplasmic aggregation (inclusion bodies formation) and, therefore, reduce the requirement of later protein folding. Finally, low temperature production phases have also been used as an alternative to diminished rDNA protein degradation, problem that, otherwise, can be solved by creating fusion complex proteins that may be more resistant to proteolysis.

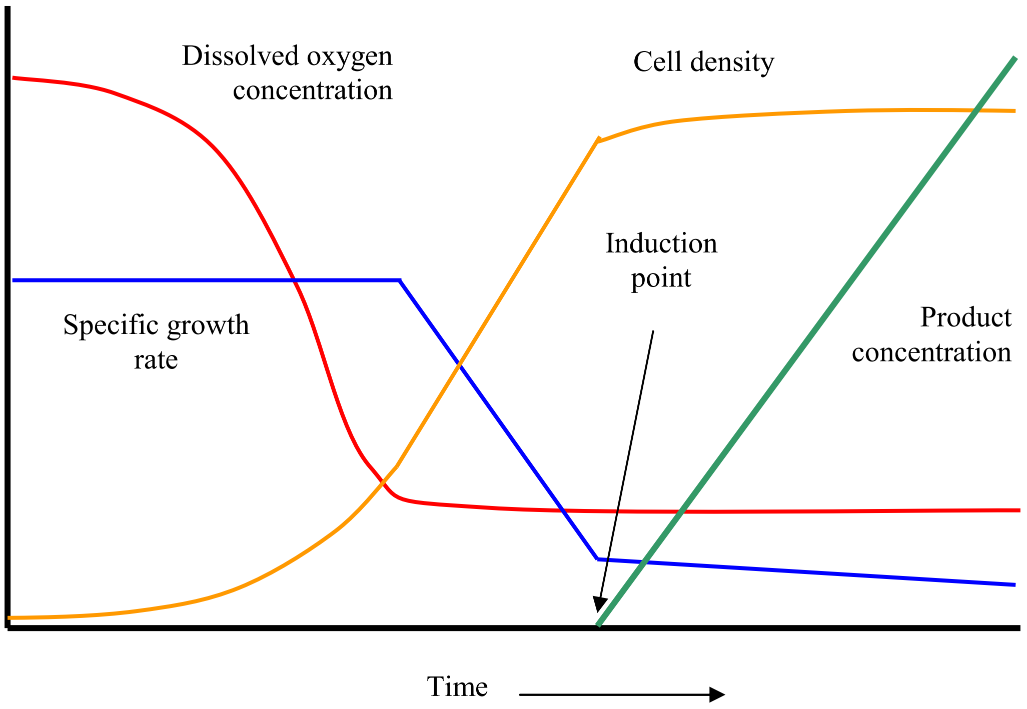

**Figure 1.1** Schematic representation of the idealized growth and production kinetics for high-cell density cultures (Swartz, 1996).

## 1.2 Model-based design of bioprocesses

Modeling is understood as a quantitative exploitable formulation of our current knowledge about the process under consideration. It takes a lot of effort and is thus justified only when it does help to solve open questions of significant importance (Lübbert, 2000). In the model-based design of bioprocesses, process improvement is the main goal to be achieved through accurate long term prediction of the state of the biological system. This technique is generally used to enhance, in a systematic and optimized way, the quantitative production, the product quality or to reduce the operating costs of a fermentation system through manipulating the fermentation process variables.

Considering the current developments in science and technology, challenging problems in the monitoring, modeling and optimization of bioprocess require the use of advanced tools to be solved. As stated by Kim and Lewis (1998), today, an important goal is to formalize human-like decision making, behavior and performance into a rigorous system theory. According to this concept, artificial neural network, linguistic fuzzy logic techniques and experts systems can be considered under such novel techniques.

Development of rigorous models for a given biological reaction mechanism on a physical and chemical basis is still a costly procedure for the industry. This is mainly due to the complex nonlinear dynamic behavior and, in some cases, the incomplete knowledge about the structure of the kinetics involved in such systems. Some innovations performed in recent years are notable, with respect to modeling and optimizing bioprocesses.

The first significant innovation is that the *a priori* knowledge needs not necessarily be represented by classical mathematical models based on the mass balances and kinetic expressions of the Monod-type. It was shown that those parts of these models that are not completely understood, e.g. models for the process' kinetics, may be better represented in a data-driven form, e.g., by artificial neural networks, while the well known mass balances are still formulated by differential equation systems (Psichogios and Ungar, 1992; Schubert *et al.*, 1994; Van Can *et al.*, 1997). Then, the synergism between the different complementing single approaches is used to tackle the complex problem of modeling.

Artificial neural networks by themselves have been successfully utilized for system modeling in biotechnology (Lübbert and Simutis, 1994; Montague and Morris, 1994). In these applications in particular, the neural networks are used as "black box" model components. These black-box models associate certain known and measurable process input variables to other output variables of the process, whose values are usually not known or not measurable. A complex relationship between them is supposed to occur, but would be only described by the neural network after a proper training procedure. Lübbert and Simutis (1994) as well as Montague and Morris (1994) reported that such stand-alone neural network models require extended data records for the training of such systems.

As stated before, it is possible to combine data-driven models (black-box) with first principles descriptions in so called hybrid models. A hybrid model for a given process contemplates a set of non linear differential equations combined with neural or neuro-fuzzy representations. While the differential equations set describes the mass balances relationships, the neural network and/or the neuro-fuzzy models are used as numerically exploitable representations of key kinetic variables like the specific growth or specific production rate. As with stand-alone systems, a proper learning process is required to train the hybrid model.

Finally, another noteworthy innovation to consider in the field of bioprocess engineering is the evolutionary process optimization procedure proposed by Galvanauskas *et al.* (1998). This technique iteratively improves the process description while approaching the optimal feeding profile, or more generally, control profiles in fed-batch cultivation processes. The methodology guarantees a quick approximation to optimal process control profiles.

## 1.3 General objectives

The establishment of a framework for the *online* modeling, optimization and monitoring of bioprocesses is the main objective of the present work. It focuses principally on the development of such methods for improving recombinant protein production systems.

Under this scope, the employment of neural network-based hybrid models to represent biotechnological processes is to be considered. A hybrid model consists in set of differential equations describing the mass balance for the bio-process complemented with artificial neural network or neuro-fuzzy components that describe the corresponding kinetics. As with all data-driven procedures, the need to validate the neural network-based representation will be considered. This can be done employing the experimental evidence obtained when an optimization of the process is carried out using the hybrid approach. However, a main goal of this work is to reduce, as far as possible, the amount of experimental data required for proper model identification. It may be shown, that even when the data available for model identification is essentially very small, compared to that used by stand-alone neural network implementations, the obtained results are quite satisfactory when using a hybrid model. The last scenario implies then to bring down the number of experiments to an essential minimum. This situation can be attractive for the industrial application: the use of these methods involves a decrease in the research and development costs through the reduction of the

experimental effort involved in time and expendables. This applies either for the improvement of some old process strategies or for the establishment of new processes.
The development of the framework is conceived as an evolutionary process. The first step is the establishment of a process model for a real microbial fermentation system. This is used then to run a model-based mathematical optimization (offline optimization), where the time profiles of some control variables is computed. Afterwards, a validation of the accuracy of the model is accomplished. The validation is carried out after analyzing and comparing the experimental evidence obtained in a real fermentation process against the model's prediction. The acceptance, refinement or total modification of the model is determined by its suitability to accurately describe the process under consideration. As can be inferred, the technique takes advantage of the accumulative knowledge gained through the systematic approach of optimal process regimes. However, it is presumed that this learning and development process can be further enhanced, if the modeling and optimization procedures were performed online.
Online modeling and optimization would be of high benefit in situations where unexpected changes in the normal course of a process come into play. Such an optimization tool may be able to confront alterations that may not have been considered (like temperature control or feed pump failures) and where it is necessary to react also optimally. Under these circumstances, a very important subject is to validate the online optimization system performance. This can be accomplished inducing in an artificial manner some changes in the process environment and then monitoring of the responses of the performance index and penalty functions. Main idea behind is to carry out the validation using an independent soft sensor and additionally to verify its proper functioning on another microbial system. The essential innovation on the development of such alternative soft-sensor systems is that they are *exclusively* based and trained on common online measured or correlated process variables. The most substantial advantage of using neural network components in hybrid models is the online incorporation of process knowledge at high learning rates. This methodology overcomes the common slow convergence problem during the initial training stages allowing a careful monitoring of optimized recombinant protein production processes.

## 1.4 Outline of the thesis

The present thesis is organized in 7 chapters.

In chapter 1 (actual chapter), the motivations and objectives of this work are presented together with some introductory theoretical background and concepts that will be used later in the rest of the chapters. Focused on the importance to produce high valuable pharmaceuticals, a brief description of the general characteristics and the application of recombinant protein production processes using *Escherichia coli* is depicted. Furthermore, the usefulness of the model-based design of bioprocesses is highlighted. This is a systematic mathematical tool that can be employed to improve the performance of such systems. The particular use of neural network-based techniques is established as milestone of the work.

In chapter 2 the *Materials and Methods* are presented. Three different microbial cultivations were performed, two of them comprising recombinant strains of the bacterium *Escherichia coli* and another, a recombinant strain of the yeast *Kluyveromyces lactis*. Together with the composition of the culture media, two mathematical algorithms, the quasi-Newton and the chemotaxis techniques are presented, too. Model parameter identification and model-based optimization was performed in this work using the HybNet software package (Oliveira *et al*., 1996) where the aforementioned algorithms are available.

Chapter 3 addresses the hybrid modeling of a multi-substrate batch cultivation process. The first phase of the procedure was the establishment of a hybrid model for the process. Data obtained from a deterministic model (taken from literature) was used to pre-train the hybrid model. This data was supposed to roughly describe the main features of the real process. It was considered that this procedure would simplify the identification procedure of the hybrid model when confronted to experimental data. To validate the hybrid model, the modeling technique was demonstrated at the batch cultivation of an *Escherichia coli* B pUBS520 p12023 strain growing on glucose, lactose and glycerol. The structure of the hybrid models was split into the mass balance, expressed as a system of differential equations and several neural network sub-units describing the specific consumption rates of the different substrates. The specific growth rate was calculated from the single specific consumption rates. The technique serves as general core for the hybrid modeling present in the following chapters.

Chapter 4 describes the *off-line* modeling and optimization of the production of VP1-DHFR with a strain of *Escherichia coli* BL21 utilized as host system. A fundamental demand in today's biotechnology is the reduction of the invested time for development and improvement of a given process. The present optimization method makes use of the modeling technique described in chapters 3 to achieve this goal. As the data records from the process under consideration were very scarce at the beginning of such a development, the initial stages considered classical Monod-type model representations supposed to roughly describe the main features of the real process. This model was then transformed into a hybrid approach, which was used to optimize the control variables of the process. The hybrid approach combined *a priori* knowledge and information from the available process data. The consequent optimization experiments were also used to validate the model's accuracy. This optimization strategy is in concordance with the evolutionary process optimization procedure described by Galvanauskas *et al.* (1998). The procedure improved the process description while approaching the optimal feeding profile, or more generally, control profiles in fed-batch cultivation processes. The technique guarantees a quick approach to optimal process control profiles.

Chapter 5 describes the online neuro-fuzzy modeling and optimization of the production of the VP1-DHFR with *Escherichia coli* BL21 as host system. Here the techniques used in chapters 3 and 4 were combined in the development of an online application. Concerning the specific growth rate, it was mathematically formulated via a feed forward neural network model. This neural network acts essentially as an adaptive system able to learn from the online measured process variables. The training of the neural network component of the hybrid model occurred exclusively online. That situation improved the gain of process knowledge at high learning rates, overcoming the slow convergence during the initial training stages. Moreover, already available heuristic knowledge about the protein development was incorporated by means of a simple neuro-fuzzy expert system. The corresponding kinetic was included with simple linguistic rules-of-thumb. The main advantage of this type of fuzzy representation is that the results of the training can be made more transparent to the process engineer. Using this approach, it can be shown that increased productivities are obtainable with the online-identified hybrid model, as compared to optimizations employing conventional off-line approaches.

Chapter 6 describes the monitoring of bioprocesses through their performance indexes using a neural network-based soft sensor. This is an online application and was tested in two different recombinant microbial systems: the bacteria *Escherichia coli* and the yeast *Kluyveromyces lactis*. Both cultivations are run under optimal feeding strategies. For the case of the

*Kluyveromyces lactis* cultivation a specific growth rate control strategy is applied with constant temperature conditions. However, the *Escherichia coli* cultivation runs under artificially induced temperature shifts. Monitoring the changes of process performance index and penalty functions related to these environmental alterations allows the development of high performance control and quality strategies for the pharmaceutical industry. The approach is essentially able to characterize the kinetics of the bioprocess making use of process correlated variables available online like the optical density of the broth, its oxygen consumption rate and the carbon dioxide evolution rate. The core of the monitoring lies in the estimation of the key variable of the process, the specific growth rate, which is formulated via a feed forward neural network.

Finally, chapter 7 presents the general conclusions for this work.

## 1.5 References

1. Galvanauskas, V., Simutis, R., Volk, N., Lübbert, A. Model based design of a biochemical cultivation process. Bioprocess Engineering, 18 (1998) 227-234
2. Kim, Y. H.; Lewis, F. L.: High-level feedback control with neural networks, World Scientific Publishing Co. Pte. Ltd. (1998)
3. Kopetski, E.; Schumacher, G.; Buckel, P.: Control of formation of active soluble or inactive insoluble baker's yeast alpha-glucosidase PI in *Escherichia coli* by induction and growth conditions. Mol. Gen. Genet. 216 (1989) 149-155
4. Lübbert, A.; Simutis, R.: Adequate use of measuring data in bioprocess modeling and control. In: Trends in Biotechnology 12 (1994) 304-311
5. Lübbert, A.: Bubble Column Bioreactors in Bioreaction Engineering (Modeling and Control), Schügerl, K. and Bellgardt, K.-H. (Eds.).. Springer-Verlag Berlin Heidelberg (2000) 247-273
6. Montague, G.; Morris, J.: Neural-network contributions in biotechnology. Trends in Biotechnology 12 (1994) 312-324
7. Oliveira, R.; Simutis, S.; Lübbert, A.: HYBNET, a new tool for advanced process modeling. Proceedings of the 1st European Symposium on Biochemical Engineering Science, Dublin, Ireland, pp. 182-183 (1996)
8. Psichogios, D. C.; Ungar, L. H.: A hybrid neural network – first principles approach to process modeling. AIChE J. 38 (1992) 1499-1511
9. Schubert, J.; Simutis, R.; Dors, M.; Havlik, I.; Lübbert, A.: Bioprocess optimization and control: Application of hybrid modeling. J. Biotechnology. 35 (1994) 51-68
10. Swartz, J. R.: *Escherichia coli* recombinant DNA technology. In: *Escherichia coli* and *Salmonella*. Cellular and Molecular Biology. 2nd Edition. ASM Press, Washington, D. C. (1996) 1693-1711
11. Tsai, A. M.; Betenbaugh, M. J.; Shiloach, J.: The kinetics of RCC1 inclusion body formation in *Escherichia coli*. Biotech. & Bioeng. 48 (1995) 715-718
12. Van Can, H. J. L.; Te Braake, H. A. B.; Hellinga, C.; Luyben, K. C. A. M.; Heijnen, J. J.: An efficient model development for bioprocesses based on neural networks in macroscopic balances. Biotech. & Bioeng. 54 (1997) 549-566
13. Yee, L. and Blanch, H. W.: Recombinant protein expression in high density fed-batch cultures of *Escherichia coli*. Bio/Technology 10 (1992) 1550-1556

# Chapter 2

# Materials and methods

Three different microbial cultivations were carried out during this work, two of them comprising recombinant strains of the bacteria *Escherichia coli* and another one with a recombinant strain of the yeast *Kluyveromyces lactis*.

## 2.1 Microorganisms and media culture composition

### 2.1.1 Multi substrate batch cultivation of *Escherichia coli.*

A multi substrate cultivation was carried out with a strain of *E. coli* B pUBS520 p12023 under batch conditions. All experiments were performed in 0.5 L fermentation units of a multi-fermenter system SIXFORS (INFORS GmbH). The pH was controlled at 7 during the entire course of the fermentation using a 1M solution of $H_3PO_4$ and a 25% solution of $NH_4OH$. Excessive foam formation was suppressed with a 1M silicon-based anti-foam emulsion. Temperature was also maintained constant at 37°C during the fermentations.
The medium for the pre-culture and the fermentation had the following composition: $Na_2SO_4$ 2.0 g $L^{-1}$, $NH_4Cl$ 0.5 g $L^{-1}$, $KH_2PO_4$ 14.6 g $L^{-1}$, $(NH_4)_2SO_4$ 2.468 g $L^{-1}$, $NaH_2PO_4 \cdot H_2O$ 3.6 g $L^{-1}$, $(NH_4)_2$-H-Citrat 1.0 g $L^{-1}$ and 2 mL $L^{-1}$ trace elements solution. This trace elements solution was constituted by $CoCl_2 \cdot 6H_2O$ 0.18 g $L^{-1}$, $MnSO_4 \cdot H_2O$ 0.1 g $L^{-1}$, $CuSO_4 \cdot 5H_2O$ 0.16 g $L^{-1}$, $Na_2$-EDTA 20.1 g $L^{-1}$, $ZnSO_4 \cdot 7H_2O$ 0.18 g $L^{-1}$, $FeCl_2 \cdot 6H_2O$ 16.7 g $L^{-1}$ $CaCl_2{\cdot}2H_2O$ 0.5 g $L^{-1}$. Glucose, lactose and glycerol were used as carbon sources. The inoculum was prepared in shaking flasks at a temperature of 37°C. It consisted of 2 mL of the main bacterial suspension (prior stored at -72°C) within 100 mL of culture medium with a glucose concentration of 5 g $L^{-1}$. 1 mg $mL^{-1}$ of ampicillin and 50 mg $mL^{-1}$ of canamycin were added. After an incubation time of 4 h, the flasks were centrifuged and the cells were re-suspended in sterile tap water.
Measurements of the concentrations of biomass, glucose, lactose and glycerol were performed off line. The dry biomass concentration was estimated via its correlation with the optical density of the culture at a wavelength of 600 nm. The analysis was carried out with a UV scanning photometer (UV-2102 PC, Shimadzu Corp.). Glucose concentrations were measured with an enzymatic glucose analyzer YSI Model 2700 (Yellow Springs Instrument Co. Inc., USA). Afterwards, the present lactose was hydrolyzed with β-galactosidase. The glucose concentration of the sample was measured again. The lactose concentration was estimated from the difference between the original and the resultant glucose concentration coming from the hydrolyzed lactose. Glycerol was measured using the enzymatic measurement kit for glycerol from Boehringer (UV-Test 148 270, Boehringer Mannheim GmbH, Mannheim, Germany), which is based on the UV-measurement of the quantity of NADH, equivalent to the quantity of glycerol present in the sample.

### 2.1.2 Production of the virus capsid protein construct VP1-DHFR with *Escherichia coli* BL21

The production of the protein complex VP1-DHFR was carried out with a strain of *Escherichia coli* BL21 under batch and fed-batch conditions. The product is a genetic construct from the fusion of the viral capsid protein of the murine polyoma virus (VP1) and the enzyme dihydrofolate reductase (DHFR, EC 1.5.1.3). The microorganism contains an ampicillin resistant plasmid pBR322, responsible for expressing the viral capsid protein under the control of the tac-promoter. The over-expression of the recombinant protein is induced employing 1.5 mmol of IPTG (Isopropyl β-D-Thiogalactopyranosidase). All experimental runs took place in a 10 L fermenter, Biostat C+ (B. Braun Biotech International). The pH was controlled during the whole course of the fermentation using a 1M solution of $H_3PO_4$ and a 25% solution of $NH_4OH$. Foam formation was suppressed with a 1M silicon anti-foam emulsion.

The medium for the pre-culture and the fermentation had the following composition: $KH_2PO_4$ 13,3 g $L^{-1}$, $(NH_4)_2HPO_4$ 4 g $L^{-1}$, citric acid 1,7 g $L^{-1}$, EDTA 8,4 mg $L^{-1}$, $CoCl_2 \cdot 6H_2O$ 2,5 mg $L^{-1}$, $MnCl_2 \cdot 4H_2O$ 15 mg $L^{-1}$, $CuCl_2 \cdot 2H_2O$ 1,5 mg $L^{-1}$, $H_3BO_3$ 3 mg $L^{-1}$, $Na_2MoO_4 \cdot 2H_2O$ 2,5 mg $L^{-1}$, $ZnSO_4 \cdot 7H_2O$ 2 mg $L^{-1}$, $FeSO_4 \cdot 7H_2O$ 200 mg $L^{-1}$ and 2 mL $L^{-1}$ trace elements solution. This trace elements solution was constituted by $CaCl_2{\cdot}2H_2O$ 0.5 g $L^{-1}$, $ZnSO_4{\cdot}7H_2O$ 0.18 g $L^{-1}$, $MnSO_4{\cdot}H_2O$ 0.1g $L^{-1}$, $Na_2$-EDTA 20.1g $L^{-1}$, $FeCl_3{\cdot}6H_2O$ 16.7g $L^{-1}$, $CuSO_4{\cdot}5H_2O$ 0.16 g $L^{-1}$, $CoCl_2{\cdot}6H_2O$ 0.18 g $L^{-1}$. Glucose was used as sole carbon source.

The inoculum was prepared in shaking flasks at a temperature of 37°C. It consisted of 2 mL of the main bacterial suspension (stored at -72°C) within 100 mL of culture medium with a glucose concentration of 5 g $L^{-1}$. 1 mmol of ampicillin was added. After an incubation time of 4 h, the flasks were centrifuged and the cells were re-suspended in sterile tap water.

Measurements of the concentrations of the most important state variables, biomass, glucose and the product activity were performed off line. The estimation of the biomass concentration was made based on the correlation between dry biomass weight and the optical density of its corresponding suspension at a wavelength of 600 nm. The analysis was carried out in a UV scanning photometer (UV-2102 PC, Shimadzu Corp.). In the case of glucose, its concentrations were estimated with the enzymatic glucose analyzer YSI Model 2700 (Yellow Springs Instrument Co., Inc. , USA).

The enzymatic activity of the DHFR was utilized to estimate the product concentration. To apply this method, any probe taken from the bio-system should contain the same amount of biomass. Therefore, the volume of the sample (in mL) varies from sample to sample and is calculated from the reciprocal of the sample's measured optical density at a wavelength of 600 nm and then multiplied with a factor of 10 ($V_{SAMPLE} = 10/OD_{600}$). The sediment obtained after centrifugation of the probe was frozen at –72°C. Later, the cells were disrupted with glass pearls (0.5 g, 0.4 mm diameter) in an oscillating mill (Retsch Type MM2; Hann, Germany). The enzymatic activity of DHFR of the raw enzymatic extract was determined with the method reported by Ginkel *et al.* (1997). The measurement was corrected considering also the NADPH-Oxidase activity. A unit of activity is equal to 1 µmol of transformed substrate per minute.

Besides temperature, head pressure and culture weight, on line measurements of $O_2$ and $CO_2$ in the vent line were performed. The $O_2$ concentration was measured with a paramagnetic analyzer (OXOR 610, Maihak), while for $CO_2$ an infrared absorption analyzer (UNOR 610, Maihak) was used.

### 2.1.3 Production of the recombinant protein complex GAL80/HIS-TAG with *Kluyveromyces lactis* RUL 1888 D80ZR-pEAHG80

The yeast *Kluyveromyces lactis* RUL 1888 D80ZR-pEAHG80 was grown aerobically on glucose as only carbon source. The product, the recombinant protein GAL80 (Zenke *et al.*, 1999) is merged with a HIS-TAG and was constitutively expressed by an ADG promoter. The fermentation was run under fed-batch in a 10 L fermenter, Biostat C+ (B. Braun Biotech International). The pH was controlled during the entire course of the fermentation using a 1M solution of $H_3PO_4$ and a 25% solution of $NH_4OH$. Foam formation was diminished with a 1M silicon anti-foam emulsion.

For the cultivation of *Kluyveromyces lactis*, two different culture media were prepared. The pre-culture (Culture medium I) was prepared with a yeast-based complex medium, being slightly changed for the feeding solution. Culture medium II corresponded to the cultivation medium used in the fed-batch production process.

### *Culture medium I*

To prepare 1 L of cultivation media for the pre-culture, 6.7 g Yeast Nitrogen Base (Bacto Yeast Nitrogen Base w/o Amino Acids, Fa. DIFCO) were diluted in 850 ml distillated water and later sterilized in an autoclave. 50 mL of an amino acid solution were then added. The amino acid solution consisted of 625 ml distillated water containing the following amino acids: Adenine 140 mg, Histidine 480 mg, Tryptophane 480 mg, Arginine 480 mg, Methionine 480 mg, D/L Leucine 720 mg, D/L Isoleucine 720 mg, Tyrosine 180 mg, Phenylalanine 600 mg, D/L Valine 720 mg, D/L Threonine 720 mg. After the preparation the yeast nitrogen base solution resulted in the following composition: $(NH_4)_2SO_4$ 5g $L^{-1}$, $H_2PO_4$ 1g $L^{-1}$, $MgSO_4 \cdot 7H_20$ 0.5g $L^{-1}$; $ZnSO_4 \cdot 7H_2O$ 0.0004 g $L^{-1}$; $CuS0_4 \cdot 5H_2O$ 0.0004g $L^{-1}$, $CaCl_2 \cdot 2H_2O$ 0.1g $L^{-1}$, $Na_2MoO_4 \cdot 2H_2O$ 0.0002 g $L^{-1}$; $H_3BO_3$ 0.0005g $L^{-1}$, KI 0.0001g $L^{-1}$; NaCl 0.1g $L^{-1}$; FeCl 0.0002g $L^{-1}$; $MnSO_4 \cdot H_2O$ 0.0004g $L^{-1}$. The media also contained the following vitamins: Biotin 0.002 mg $L^{-1}$; Calcium Pantothenate 0.4 mg $L^{-1}$; Thiamin HCl 0.4 mg $L^{-1}$; Prydoxine HCl 0.4 mg $L^{-1}$; Para-amino Benzoic Acid 0.2 mg $L^{-1}$; Folic Acid 0.002 mg $L^{-1}$; Inositol 2 mg $L^{-1}$; Niacin 0.4 mg $L^{-1}$; Riboflavin 0.2 mg $L^{-1}$. The substrate concentration was adjusted in a separate solution. This was also separately autoclaved and after cooling added to the yeast-based complex solution.

### *Culture medium II*

To prepare 1 L of this medium, 1.7 g of yeast nitrogen base (Bacto Yeast Nitrogen Base w/o Amino Acids and Ammonium Sulfate, Fa. DIFCO) and 5g of $(NH_4)_2SO_4$ were dissolved in 850 ml of distilled water and autoclaved. 50 ml of the amino acid solution used for the preparation of the culture media I, were also added. The substrate concentration was adjusted with a separate solution prepared with the required amount of glucose in enough distilled water to complete 100 g of solution. This was also separately autoclaved and after cooling added to the yeast nitrogen base solution.

### *Feeding solution*

A solution containing 20 g $L^{-1}$ of a yeast-based complex medium (Bacto Yeast Nitrogen Base w/o Amino Acids and Ammonium Sulfate, Fa. DIFCO) and 5 g $L^{-1}$ $(NH_4)_2SO_4$ was autoclaved. Later, 133 mL of the amino acid solution used for the preparation of the culture medium I, were also added. A separate solution was prepared with the required amount of glucose in enough distilled water to fix the concentration in 100 or 200 g $L^{-1}$ of glucose. This was also separately autoclaved and after cooling added to the yeast nitrogen base solution.

## 2.2 Identification and optimization algorithms

Model parameter identification and model-based optimization was performed in this work using the HybNet program (Oliveira *et al.*, 1998). Two algorithms contained in this application were used for model identification and for model-based optimization. In the particular case, the model parameters of the multi-substrate cultivation of *E. coli* were identified using the *quasi-Newton* algorithm. The *chemotaxis* algorithm was used in the case of the cultivation processes with *E. coli* for producing the VP1-DHFR recombinant protein and in the case of the production process of the GAL80/HIS-TAC recombinant protein using *Kluyveromyces lactis*. In both cases, the chemotaxis algorithm was used for the identification of the model parameters, the weights of the neural network component describing the specific growth rate and for the identification of the weights of the neuro-fuzzy component representing the specific protein production rate.

Moreover, chemotaxis was also used to estimate the optimal induction time and to train the weights of the feed-forward neural network used to calculate feeding profiles in the model-based optimization of the *E. coli* BL21 cultivation and the model parameters of the soft-sensor of the *Kluyveromyces lactis* fermentation.

### 2.2.1 Quasi-Newton algorithm

The quasi-Newton algorithm (Bronstein *et al.*, 1999) is a gradient-based iterative optimization technique. It considers the general minimization problem:

$$f\,(\,\underline{\boldsymbol{\theta}}\,) = \text{min!} \quad \text{for} \quad \underline{\boldsymbol{\theta}} \in \Re^{\text{n}} \tag{2.1}$$

where $f\,(\,\underline{\boldsymbol{\theta}}\,)$ is a given differentiable function with the parameter set $\underline{\boldsymbol{\theta}}$. The solution to the problem is to find an approximation parameter set $\underline{\boldsymbol{\theta}}$* that minimizes equation 2.1. Beginning with a first guess $\underline{\boldsymbol{\theta}}^1 \in \Re^{\text{n}}$, the parameter set is changed iteratively according to:

$$\underline{\boldsymbol{\theta}}^{k+1} = \underline{\boldsymbol{\theta}}^{k} + \alpha_k \underline{\text{d}}^{k} \qquad (\,k = 1, 2, \ldots\,) \tag{2.2}$$

where $\underline{\text{d}}^{k} \in \Re^{\text{n}}$ is the *directional vector* and $\alpha_k$ the so-called *gain parameter*. For the specific case of the quasi-Newton algorithm, the directional vector is given by:

$$\underline{\text{d}}^{k} = -\,\text{M}_k \nabla f\,(\,\underline{\boldsymbol{\theta}}^{k}\,) \qquad (\,k = 1, 2, \ldots\,) \tag{2.3}$$

$\text{M}_k$ is a symmetric, positive matrix that approximates, in an iterative form, the inverse Hessian-matrix. The first guess is usually done defining $\text{M}_1$ = I (the identity matrix) and applying the following matrix correction:

$$\text{M}_k = \text{M}_{k-1} + (\underline{\text{v}}^{k}\,\underline{\text{v}}^{k\,\text{T}})/(\underline{\text{v}}^{k\,\text{T}}\,\underline{\text{v}}^{k}) - [(\text{M}_{k-1}\,\underline{\text{w}}^{k})(\text{M}_{k-1}\,\underline{\text{w}}^{k})^{\text{T}}]/\underline{\text{w}}^{k\,\text{T}}\,\text{M}_k\,\underline{\text{w}}^{k} \tag{2.4}$$

where, $\underline{\text{v}}^{k} = \underline{\boldsymbol{\theta}}^{k} - \underline{\boldsymbol{\theta}}^{k-1}$ and $\underline{\text{w}}^{k} = \nabla f\,(\,\underline{\boldsymbol{\theta}}^{k}\,) - \nabla f\,(\,\underline{\boldsymbol{\theta}}^{k-1}\,)$ for ( $k$ = 1, 2, ... ). The gain factor is calculated from the minimization of:

$$f\,(\,\underline{\boldsymbol{\theta}}\, - \alpha\,\text{M}_k \nabla f\,(\,\underline{\boldsymbol{\theta}}^{k}\,)) = \text{min} \qquad \alpha \geq 0 \tag{2.5}$$

### 2.2.2 Chemotaxis algorithm

The chemotaxis algorithm is an iterative random-search optimization technique. The following description is referred to the calculation of the parameters (weights and biases) of feed-forward neural networks. It considers the general minimization problem presented in equation 2.1 with the profit function, $f(\underline{\boldsymbol{\theta}})$. The technique can be resumed as follows:

1) Begin with a first parameter guess $\underline{\boldsymbol{\theta}}^{1} \in \Re^{n}$. According to the chemotaxis algorithm, a mutation in the parameters must be done randomly choosing an increment in the parameters, $\Delta\underline{\boldsymbol{\theta}}$**.** The parameter set is changed iteratively according to:

$$\underline{\boldsymbol{\theta}}^{k+1} = \underline{\boldsymbol{\theta}}^{k} + \Delta\underline{\boldsymbol{\theta}} = \underline{\boldsymbol{\theta}}^{k} + \gamma\,\underline{\boldsymbol{\theta}}^{k} \qquad (k = 1, 2, ..., \mathrm{M}) \qquad (2.6)$$

   where $\gamma$ is the so-called *mutation parameter*.
2) Calculate a new profit function, $f(\underline{\boldsymbol{\theta}}^{k+1})$. Calculations must be done until $f(\underline{\boldsymbol{\theta}}^{k+1}) \neq f(\underline{\boldsymbol{\theta}}^{k})$
3) Calculate the function ε according to:

$$\varepsilon = f(\underline{\boldsymbol{\theta}}^{k+1}) - f(\underline{\boldsymbol{\theta}}^{k}) \qquad (k = 1, 2, ..., \mathrm{M}) \qquad (2.7)$$

4) To minimize the profit function, a decision must be made to continue calculations.
   a) If $\varepsilon = 0$ after a predefined maximal number of iterations M is exceeded, then change randomly the mutation parameter and continue from step 2 again.
   b) If $\varepsilon > 0$, then go to step 1, change the first parameter guess $\underline{\boldsymbol{\theta}}^{1}$, and initiate the procedure again.
   c) If $\varepsilon < 0$, then continue with the calculations using Equation 2.6 until a predefined minimization criteria for the profit function $f(\underline{\boldsymbol{\theta}})$ is reached or a number of predefined maximal iterations steps M is exceeded.

# Chapter 3

# Hybrid modeling of a multi-substrate fermentation

## ABSTRACT

A hybrid modeling approach to describe a multi-substrate fermentation is presented here. The model consists of a system of differential equations describing the mass balance of the process. Complementary to the mass balance, the kinetics variables of the biological system are represented by neural network sub-models. All these components are combined to form a network-structured global hybrid model. Main objective of this study is to provide a reliable model for a complex bio-technical process. The method is tested and validated on the batch multi-substrate cultivation of a recombinant strain of *Escherichia coli*. The role played by the structure of the neural network models and their the different activation functions in the modeling performance of the hybrid model is also investigated. A comparison between the different kinetic sub-models is presented and set against the off-line measurements to asses the modeling accuracy of the different sub-models. A screening process for selecting the most appropriate models is also considered. Typical non-linearities present in these kind of bioprocesses, like diauxic growth, are adequately represented through this kind of approach.

## 3.1 Introduction

Design, development, optimization and control of bioprocesses have been lately successfully accomplished utilizing model-based methods. These require an accurate mathematical model for the microbial system under consideration. The biological phenomena involved in such processes, specially the kinetics, are highly complex and nonlinear. As regards to this, much attention has been paid to neural network systems as an alternative to represent the kinetics of bioprocesses. The motivation for using neural networks is based on their ability to learn from any complex data set and to filter noisy signals. In the present chapter, several hybrid models of a multi-substrate fermentation were tested to determine their overall performance to describe the bioprocess. The structure of the hybrid models was split into the mass balance, expressed as a system of differential equations and several neural network sub-units describing the specific consumption rates of the different substrates. The specific growth rate is inferred from these variables, *i. e.* it is considered the result of the individual specific growth rates for each single substrate.

The first phase of the modeling procedure was the establishment of a hybrid model for the process. Data obtained from a deterministic model (taken from literature) was used to pre-train the hybrid model. This data was supposed to roughly describe the main features of the real process. It was considered that this “pre-training” would simplify the identification procedure of the hybrid model considering that only a few real data were available. To validate the hybrid model, the modeling technique was demonstrated and validated at the optimized batch cultivation of an *Escherichia coli* B pUBS520 p12023 strain growing on glucose, lactose and glycerol. The optimization goal of this cultivation was the maximization of the biomass concentration at the end of the batch fermentation. The different substrates were supposed to be consumed sequentially. The experimental data acquired was used for the refinement, retraining and validation of the hybrid models.

## 3.2 Hybrid modeling of the process

The proposed hybrid model consisted of a system of differential equations describing the mass balance of the system in batch operation modus. The mass balance considered the components biomass ($C_X$) and the substrates glucose ($S_1$), lactose ($S_2$) and glycerol ($S_3$) concentrations. Biomass was considered unsegregated. The mass balance was given by:

$$\mathrm{d}C_X/\mathrm{d}t = \mu\, C_X \tag{3.1}$$

$$\mathrm{d}S_1/\mathrm{d}t = -\, \rho_1\, C_X \tag{3.2}$$

$$\mathrm{d}S_2/\mathrm{d}t = -\, \rho_2\, C_X \tag{3.3}$$

$$\mathrm{d}S_3/\mathrm{d}t = -\, \rho_3\, C_X \tag{3.4}$$

The specific growth rate ($\mu$) was regarded as the result of the superposition of the single specific growth rates ($\mu_{\mathrm{Si}}$) for each substrate,

$$\mu = \mu_{S2} + \mu_{S2} + \mu_{S3} = Y_{X/S1}\, \rho_1 + Y_{X/S2}\, \rho_2 + Y_{X/S3}\, \rho_3 \tag{3.5}$$

Concerning the different specific substrate uptake rates ($\rho_i$), they were described by artificial neural networks (ANNs). The ANN is an adaptive kinetic process module for the estimation of the specific substrate uptake rates $\rho_i$ during the whole cultivation process. It was supposed that interactions like catabolite repression and limitation of growth by substrates are considered within the different ANN units. For that purpose, the different state variables of the system were considered as input to the neural network sub-models. The variable time was also included as input of the neural network. Figure 3.1 depicts a schema of the basic general hybrid model structure.

In the Figure 3.1, the kinetic block delineates a single or a combination of artificial neural networks accounting for the substrates consumption rates ($\rho_i$). Many sub-models can represent the kinetic of the process, but only 8 of them were selected to be tested.

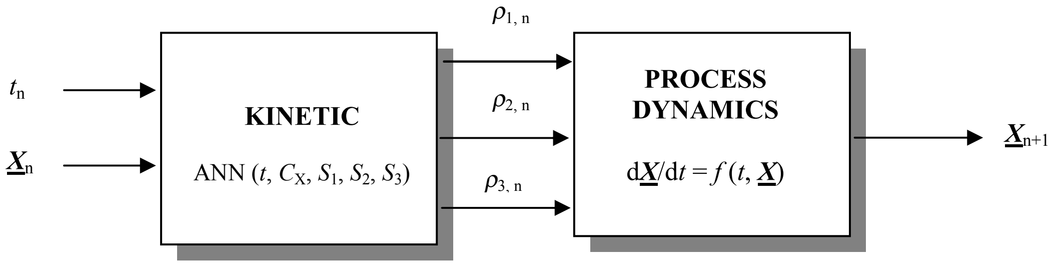

**Figure 3.1** General schematic representation of the hybrid model for the multi-substrate fermentation.

The selected 8 models proceeded from 4 basic model structures. To formulate a given neural network sub-model, some theoretical aspects or interactions between the biomass and the substrates could be considered. The structure of the model was explicitly defined once the possible interactions were defined.

As example of these interactions, the reported repression of the lactose consumption in the presence of glucose was considered for the first model structure (Bellgardt, 1991). Additionally, the possible repression of glycerol consumption in the presence of glucose was also taken into account. In the same manner, the model considered any possible interaction between lactose and glycerol too.

Once the possible interactions were established, the input for the neural network sub-models were precisely the variables associated in the aforementioned interactions. In the case of the first model, multiple possible interactions between all associated variables were considered. Therefore, time, biomass and all the substrate concentrations were taken as inputs of the neural network sub-model. The output of the model are the different specific consumption rates for each substrate. Table 3.1 resumes the basic model structures considered in this work. The associated function and the interaction considered in their formulation are also indicated.

| Basic model structure | Representation | Interactions considered |
|---|---|---|
| 1 | $\rho_{1,2,3} = f(t, C_X, S_1, S_2, S_3)$ | • The single specific consumption rates consider multiple interactions between all variables |
| 2 | $\rho_1 = f(t, C_X, S_1)$<br>$\rho_2 = f(t, C_X, S_2)$<br>$\rho_3 = f(t, C_X, S_3)$ | • The single specific consumption rates depend only on the biomass concentration and the substrate considered |
| 3 | $\rho_1 = f(t, C_X, S_1)$<br>$\rho_2 = f(t, C_X, S_1, S_2)$<br>$\rho_3 = f(t, C_X, S_1, S_3)$ | • The specific consumption rate for glucose is only dependent on the biomass and the glucose concentration<br>• The specific consumption rate for lactose is dependent on the biomass and the possible interactions between the glucose and lactose concentration<br>• The specific consumption rate for glycerol is dependent on the biomass and the possible interactions between the glucose and glycerol concentration |
| 4 | $\rho_1 = f(t, C_X, S_1)$<br>$\rho_{2,3} = f(t, C_X, S_1, S_2, S_3)$ | • The specific consumption rate for glucose is only dependent on the biomass and the glucose concentration<br>• The specific consumption rates for lactose and glycerol is dependent on the biomass and the possible interactions between the glucose, lactose and glycerol concentration |

**Table 3.1** Basic model structures and the interactions considered for their formulation.

Regarding the activation function of the neural network models, some continuous non-linear functions were used, specifically the sigmoid, the hyperbolic tangent and the radial basis functions (Kim and Lewis, 1998). Two types of neural networks were tested, the so-called standard neural network and the general neural network (Oliveira *et al.*, 1996). The first utilized exclusively sigmoid activation functions on all layers, while the later can support a combination of different activation functions per layer.

The 8 models that were tested in this work, were derived from the 4 basic model structures presented in Table 3.1. To each structure corresponded 2 different models. The input/output structure itself was maintained, but the activation function and the number of inner layers used was different for each model. The nomenclature and architecture of all kinetic neural network sub-models is detailed in Table 3.2. Their corresponding activation functions, layers and number of nodes per each layer are also presented.

The first phase of the modeling procedure was the establishment of a hybrid model for the process. Data obtained from a deterministic model (taken from literature) was used to pre-train the hybrid model. As information source for training the hybrid model, 15 sets of simulated data from the deterministic model (see Appendix A1) were built. 10 of these sets were used for model training and 5 for independent model validation.

Afterwards, data sets from 3 different experiment runs were used for the refinement of the hybrid model. Points from all experimental sets were randomly chosen for training and validation of the model. For data partitioning, a ratio of 75/25 for Training/Validation was fixed, *i. e.* ¾ of the total of the available data was used for training while the rest ¼ was used for validation.

The training of the hybrid models was carried out using the quasi-Newton algorithm (see *Materials and methods*, section 2.2.1) provided in the HybNet software (Oliveira *et al.*, 1996). The initial value guess for the weights of the neural network models was made randomly. The learning process minimized the overall least squared error between modeled and measured state variables. The model identification objective function ($J_{\text{IDEN}}$) considered was:

$$J_{\text{IDEN}}\,(y,\,Y) \equiv \sum_{j=1}^{M}\sum_{i=1}^{N}\left(\frac{y_i - Y_i}{\text{KN}_j}\right)_j^2 \longrightarrow \text{min!} \tag{3.6}$$

where $y$ are the measured state variables; $Y$ are the simulated state variables; $i$ is referred to the number of $N$ available data; $j$ is referred to the state variables under consideration and KN are pseudo-norms for the $M$ single state variables. The introduction of a pseudo-norm was done to constrict the influence of the single terms (weighting technique), because their real values lie in different numerical ranges. The values of these pseudo-norms were determined empirically.

| Structure | Model | Functionality representation and structure of the neural networks |
|---|---|---|
| 1 | M1s | $\rho_{1,2,3} = f(t, C_X, S_1, S_2, S_3)$ ⇒ ANN (IN = 5{Sig}, HN = 9{Sig}, ON = 3{Sig}) |
|  | M1g | $\rho_{1,2,3} = f(t, C_X, S_1, S_2, S_3)$ ⇒ ANN (IN = 5{Sig}, HN$_1$ = 4{HT}, HN$_2$ = 4{RB}, ON = 3{Sig}) |
| 2 | M2s | $\rho_1 = f(t, C_X, S_1)$ ⇒ ANN$_1$ (IN = 3{Sig}, HN = 3{Sig}, ON = 1{Sig})<br>$\rho_2 = f(t, C_X, S_2)$ ⇒ ANN$_2$ (IN = 3{Sig}, HN = 3{Sig}, ON = 1{Sig})<br>$\rho_3 = f(t, C_X, S_3)$ ⇒ ANN$_3$ (IN = 3{Sig}, HN = 3{Sig}, ON = 1{Sig}) |
|  | M2g | $\rho_1 = f(t, C_X, S_1)$ ⇒ ANN$_1$ (IN = 3{Sig}, HN$_1$ = 3{HT}, HN$_2$ = 3{RB}, ON = 1{Sig})<br>$\rho_2 = f(t, C_X, S_2)$ ⇒ ANN$_2$ (IN = 3{Sig}, HN$_1$ = 3{HT}, HN$_2$ = 3{RB}, ON = 1{Sig})<br>$\rho_3 = f(t, C_X, S_3)$ ⇒ ANN$_3$ (IN = 3{Sig}, HN$_1$ = 3{HT}, HN$_2$ = 3{RB}, ON = 1{Sig}) |
| 3 | M3s | $\rho_1 = f(t, C_X, S_1)$ ⇒ ANN$_1$ (IN = 3 {Sig}, HN = 3 {Sig}, ON = 1 {Sig})<br>$\rho_2 = f(t, C_X, S_1, S_2)$ ⇒ ANN$_2$ (IN = 4 {Sig}, HN = 4 {Sig}, ON = 1 {Sig})<br>$\rho_3 = f(t, C_X, S_1, S_3)$ ⇒ ANN$_3$ (IN = 4 {Sig}, HN = 4 {Sig}, ON = 1 {Sig}) |
|  | M3g | $\rho_1 = f(t, C_X, S_1)$ ⇒ ANN$_1$ (IN = 3{Sig}, HN$_1$ = 3{HT}, HN$_2$ = 3{RB}, ON = 1{Sig})<br>$\rho_2 = f(t, C_X, S_1, S_2)$ ⇒ ANN$_2$ (IN = 4{Sig}, HN$_1$ = 4{HT}, HN$_2$ = 2{RB}, ON = 1{Sig})<br>$\rho_3 = f(t, C_X, S_1, S_3)$ ⇒ ANN$_3$ (IN = 4{Sig}, HN$_1$ = 4{HT}, HN$_2$ = 2{RB}, ON = 1{Sig}) |
| 4 | M4s | $\rho_1 = f(t, C_X, S_1)$ ⇒ ANN$_1$ (IN = 3{Sig}, HN = 3{Sig}, ON = 1{Sig})<br>$\rho_{2,3} = f(t, C_X, S_1, S_2, S_3)$ ⇒ ANN$_2$ (IN = 5{Sig}, HN = 5{Sig}, ON = 2{Sig}) |
|  | M4g | $\rho_1 = f(t, C_X, S_1)$ ⇒ ANN$_1$ (IN = 3{Sig}, HN$_1$ = 3{HT}, HN$_2$ = 3{RB}, ON = 1{Sig})<br>$\rho_{2,3} = f(t, C_X, S_1, S_2, S_3)$ ⇒ ANN$_2$ (IN = 5{Sig}, HN$_1$ = 5{HT}, HN$_2$ = 2{RB}, ON = 2{Sig}) |

**Table 3.2** Nomenclature and architecture of the hybrid models. M1s model corresponds to the Model structure 1 Standard ANN, while M3g corresponds to the Model structure 3 General ANN. IN: Input Nodes; HN$_i$: Nodes in the ith Hidden Layer; ON: Output Nodes; Sig: Sigmoid activation function; HT: Hyper Tangent activation function; RB: Radial Basis activation function.

## 3.3 Results

### 3.3.1 Training and validation on artificially simulated process data

Table 3.3 shows the least squared errors from each model after 30 cycles of pre-training using the deterministic model-generated data-sets (see Appendix A1 for model). The training of hybrid models with general neural network units was more difficult than that of the standard neural networks. In general, the invested computational time was approximately 1.5 times greater for the general neural network models.

Additionally, standard neural networks seemed to perform better in the hybrid models than the more complex general approaches. Due to this, the general neural network models were omitted for further training. Moreover, it should be also noted that, until this step, none of the proposed models have properly modeled the theoretical diauxic growth phenomena described by the deterministic model (data not showed).

| Model | Training statistics | Validation statistics | Objective function |
|---|---|---|---|
| M1s | Biomass = 0.004018<br>Glucose = 0.002938<br>Lactose = 0.005181<br>Glycerol = 0.002760 | Biomass = 0.025436<br>Glucose = 0.002976<br>Lactose = 0.030871<br>Glycerol = 0.030844 | 1.4629 E-2 |
| M1g | Biomass = 0.007017<br>Glucose = 0.006784<br>Lactose = 0.008684<br>Glycerol = 0.004692 | Biomass = 0.023833<br>Glucose = 0.007436<br>Lactose = 0.028900<br>Glycerol = 0.028246 | 4.8941 E-2 |
| M2s | Biomass = 0.006888<br>Glucose = 0.004637<br>Lactose = 0.009795<br>Glycerol = 0.004656 | Biomass = 0.026288<br>Glucose = 0.005085<br>Lactose = 0.029964<br>Glycerol = 0.030546 | 4.4926 E-2 |
| M2g | Biomass = 0.007248<br>Glucose = 0.004725<br>Lactose = 0.008278<br>Glycerol = 0.004644 | Biomass = 0.028160<br>Glucose = 0.004813<br>Lactose = 0.034496<br>Glycerol = 0.031264 | 4.1085 E-2 |
| M3s | Biomass = 0.004547<br>Glucose = 0.002564<br>Lactose = 0.005883<br>Glycerol = 0.004196 | Biomass = 0.024919<br>Glucose = 0.002623<br>Lactose = 0.030023<br>Glycerol = 0.030281 | 1.9761 E-2 |
| M3g | Biomass = 0.004891<br>Glucose = 0.004109<br>Lactose = 0.005647<br>Glycerol = 0.003275 | Biomass = 0.026618<br>Glucose = 0.004494<br>Lactose = 0.031093<br>Glycerol = 0.031443 | 2.1446 E-2 |
| M4s | Biomass = 0.003511<br>Glucose = 0.002935<br>Lactose = 0.004667<br>Glycerol = 0.002328 | Biomass = 0.025071<br>Glucose = 0.002130<br>Lactose = 0.030288<br>Glycerol = 0.031189 | 1.2203 E-2 |
| M4g | Biomass = 0.007095<br>Glucose = 0.008518<br>Lactose = 0.009129<br>Glycerol = 0.004995 | Biomass = 0.025267<br>Glucose = 0.009800<br>Lactose = 0.028180<br>Glycerol = 0.028976 | 6.0142 E-2 |

**Table 3.3** Hybrid models performance after 30 cycles of training.

### 3.3.2 Training and validation on experimental process data

As planned, 3 experimental runs were carried out, twice each. The data was made available for further training of the hybrid models that exhibited better performance, in this case, only for the standard type. For validation and training, random points were sorted out from the experimental data-sets, who appeared to be noisy. A ratio of 75/25 for Training/Validation was fixed. Table 3.4 shows the least squared errors from each selected standard model. It seems that the accumulated knowledge obtained first by pre-training the models, helps considerably the new training procedure with the experimental data. A higher value for the minimum is reached (compared to the pre-trained), but in a very fast procedure (only 10 cycles). The diauxic growth after the retraining procedure was modeled quite accurate too. All alternative standard models revealed to be able to represent very good the experimental data (see Figures 3.2-3.5). Therefore, in the lack of experimental evidence, this seemed to be a good principle for training a hybrid model. Consequence of this approach is a faster training, because some historical data is incorporated to improve the description of the actual situation of the system. As shown in Table 3.3, after retraining the model that gave better results was M3s_exp.

As enumerated before, a good description of diauxic growth was successfully achieved by all of the four chosen hybrid models. Figure 3.2 depicts the modeled and measured biomass for one of the experimental assays. Two phases of growth can be clearly distinguished, the first of them from the begin of the cultivation and finishing at about 2.6 h. The second from 2.6 h until the end of the fermentation. As can be seen in Figure 3.3 the first substrate to be consumed is glucose followed by lactose (Figure 3.4). The consume of the late remains reppressed while glucose is present in the reactor, but was activated after 2.6 h, precisely the time when there was no more glucose remaining in the system. The diauxic growth phenomenon is described accurately through the expected sequential consumption of the two substrates, which is elsewhere detailed (Bellgardt, 1991). As can be stated, the four models give a good description of these states variables.

| Model | Training statistics | Validation statistics | Objective function |
|---|---|---|---|
| M1s_exp | Biomass = 0.028045<br>Glucose = 0.016581<br>Lactose = 0.047170<br>Glycerol = 0.061395 | Biomass = 0.040405<br>Glucose = 0.016385<br>Lactose = 0.040616<br>Glycerol = 0.054187 | 1.822087 E-1 |
| M2s_exp | Biomass = 0.028524<br>Glucose = 0.008667<br>Lactose = 0.048160<br>Glycerol = 0.060940 | Biomass = 0.039465<br>Glucose = 0.011287<br>Lactose = 0.043897<br>Glycerol = 0.048595 | 1.789387 E-1 |
| M3s_exp | Biomass = 0.029588<br>Glucose = 0.008539<br>Lactose = 0.045587<br>Glycerol = 0.060279 | Biomass = 0.039737<br>Glucose = 0.012050<br>Lactose = 0.035053<br>Glycerol = 0.047925 | 1.721528 E-1 |
| M4s_exp | Biomass = 0.027455<br>Glucose = 0.017637<br>Lactose = 0.046619<br>Glycerol = 0.062545 | Biomass = 0.066791<br>Glucose = 0.056241<br>Lactose = 0.088586<br>Glycerol = 0.109199 | 1.844107 E-1 |

**Table 3.4** Hybrid models performance after retraining with experimental data.

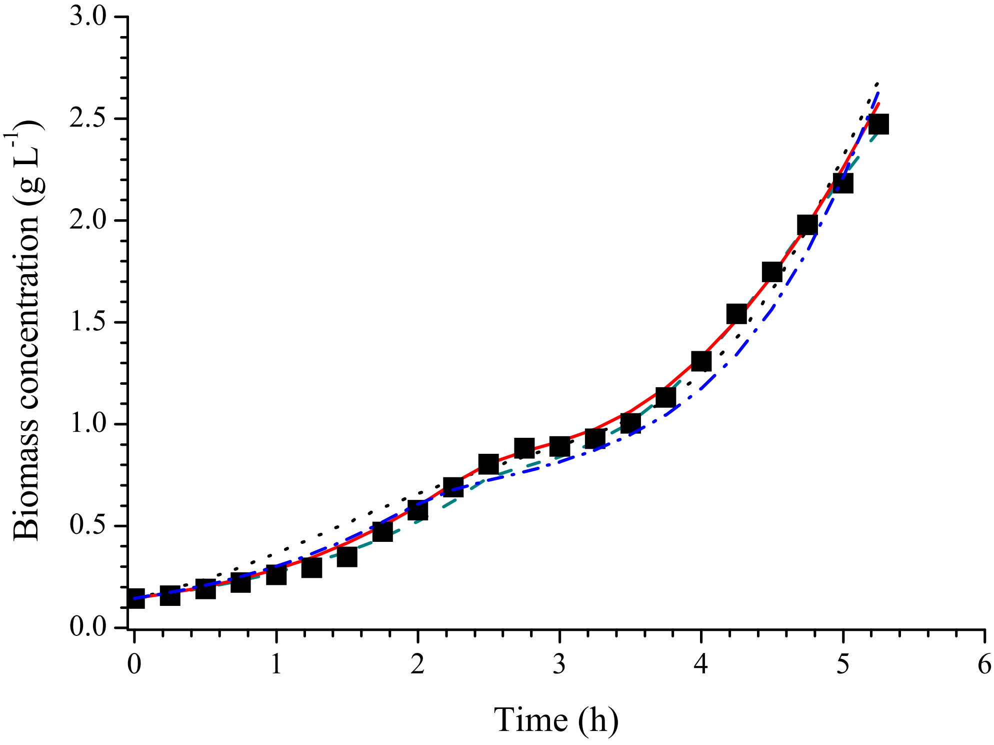

**Figure 3.2** Experimental (■) and simulated courses for biomass concentration for the different hybrid models: M1s_exp (···), M2s_exp (---), M3s_exp (—), M4s_exp (—·—).

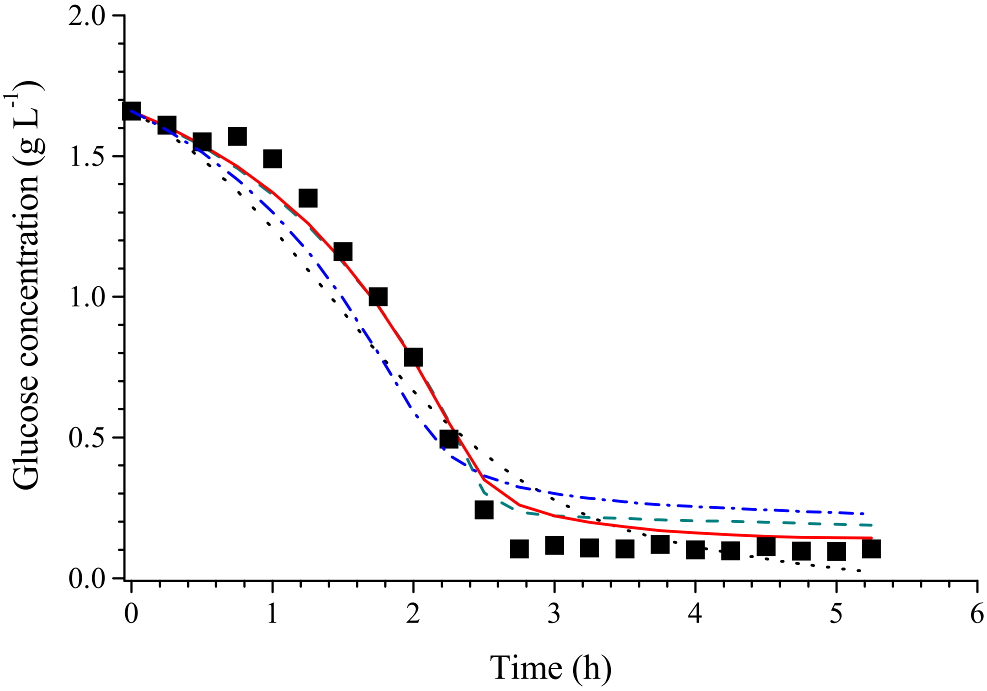

**Figure 3.3** Experimental (■) and simulated courses for glucose concentration for the different hybrid models: M1s_exp (···), M2s_exp (---), M3s_exp (—), M4s_exp (—·—).

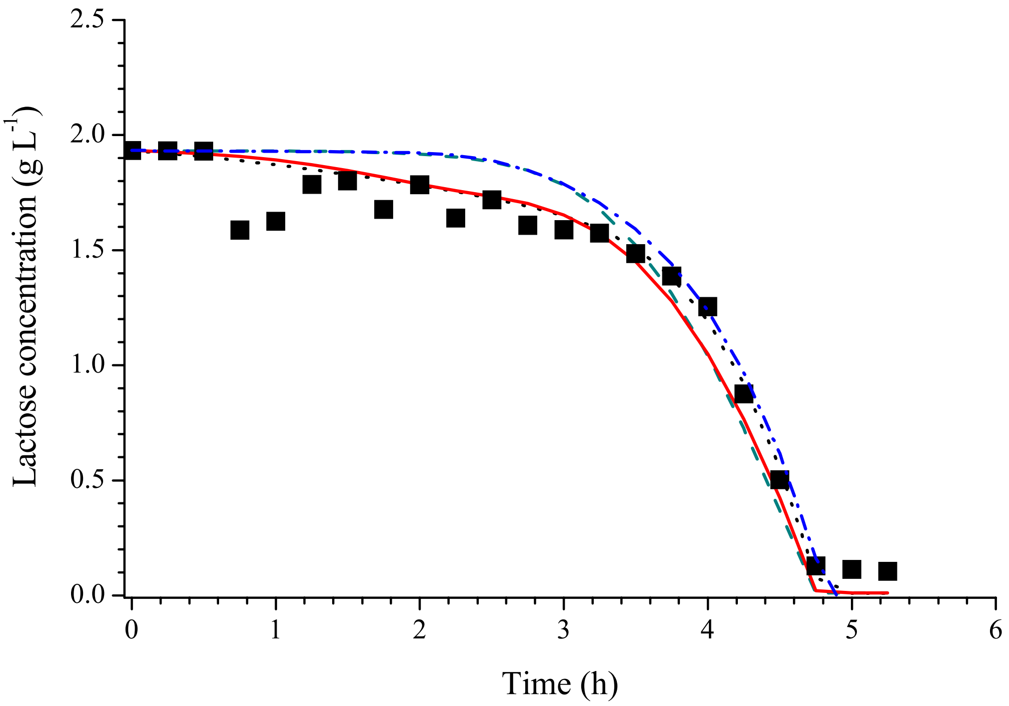

**Figure 3.4** Experimental (■) and simulated courses for lactose concentration for the different hybrid models: M1s_exp (···), M2s_exp (---), M3s_exp (—), M4s_exp (—·—).

However, in the case of glycerol, no evident and explicit interpretation of its results can be expressed, as compared to those for glucose and lactose. Even when the hybrid models seem to be able to describe the general tendency of the glycerol concentration, it is suspected that the glycerol concentration data was treated as a noisy signal and its influence removed. Simutis and Lübbert (1997) already described the role played for a non-relevant variable in the modeling performance a given hybrid model. The negative effect of this variable on the modeling efficiency is usually filtrated by the artificial neural network contained in the model. Nevertheless, in some most extreme cases this variable can even affect critically the overall modeling performance of the neural network.

Figure 3.5 presents the simulated and measured concentration of glycerol for the *E. coli* cultivation and for the different hybrid models. As can be seen, the glycerol was almost constant all over the fermentation duration and no clues indicating its consume could be determined from these data.

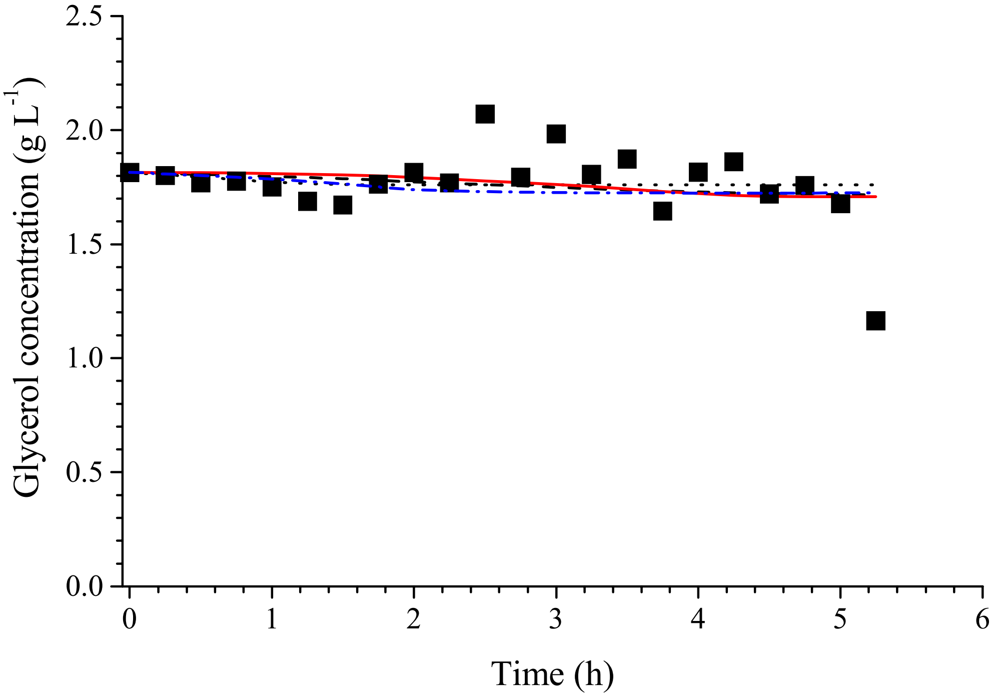

**Figure 3.5** Experimental (■) and simulated courses for glycerol concentration for the different hybrid models: M1s_exp (···), M2s_exp (---), M3s_exp (—), M4s_exp (—·—).

Core of the hybrid model, was the determination of the individual specific substrate uptake rates for glucose ($\rho_1$), lactose ($\rho_2$) and glycerol ($\rho_3$). As stated before, the training procedure was performed adjusting the neural network weights to produce a reliable kinetics estimator. Even when the training was a pure numerical fitting of the state variables, the obtained results appear to have a good agreement with some theoretical and experimental evidence, therefore a brief discussion of them will follow.

Figure 6 shows the modeled and measures of the glucose uptake rate ($\rho_1$). It is clear that for three of the four models (exception is model 1), a common typical behavior is achieved. The specific consumption rate for glucose is almost constantly high until the glucose concentration rapidly decreases to a level near zero at about 2.6 h (see Figure 3.3). Such a comportment is described in detail by Bellgardt (1991) for the case of an *E. coli* cultivation with glucose and lactose as substrates, which are consumed sequentially. Bellgardt also points out that on this phase the cells reach their maximum growth rate possible under the provided conditions for glucose.

The last statement is confirmed when considering that the main influence to the specific growth rate (Equation 3.5) in this stage is given precisely by the specific consumption rate for glucose, $\rho_1$. At the time when the glucose is exhausted, the contribution of the glucose uptake rate ($\rho_1$) to the specific growth rate is no longer important. So it can be stated here, that the first phase of the cultivation is accurately and exclusively described by the specific glucose consumption rate contribution, as was originally meant by the formulation of the majority of the models. Unfortunately, in the case of model 1, the situation can not be likely explained.

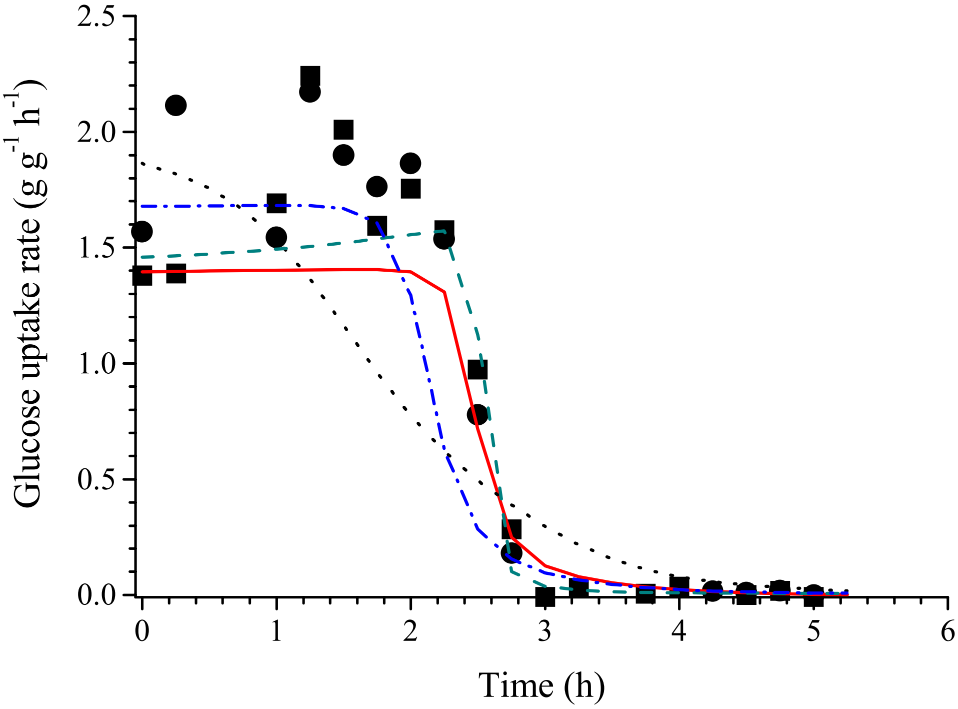

**Figure 3.6** Experimental (■, ●) and simulated courses for the glucose uptake rate ($\rho_1$) for the different hybrid models: M1s_exp (···), M2s_exp (---), M3s_exp (—), M4s_exp (—·—).

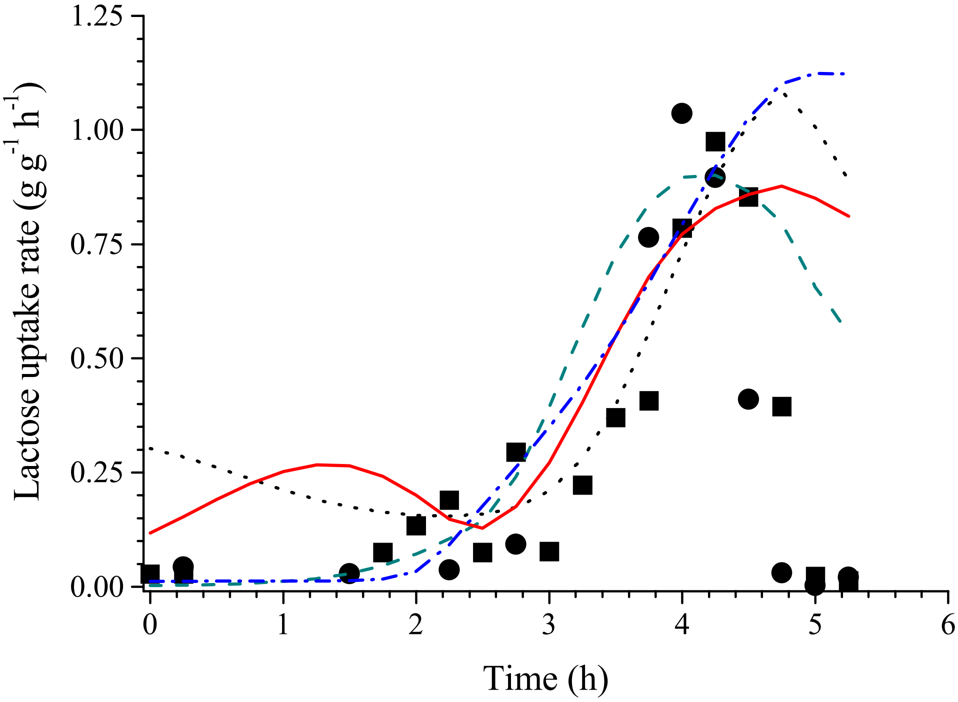

**Figure 3.7** Experimental (■, ●) and simulated courses for the lactose uptake rate ($\rho_2$) for the different hybrid models: M1s_exp (···), M2s_exp (---), M3s_exp (—), M4s_exp (—·—).

In reference to the lactose uptake rate ($\rho_2$), similar explanations can be raised as those for glucose. Figure 3.7 depicts the modeled and measured data for the lactose uptake rate. It seems that after a period of repression (due to the presence of glucose), the lactose consumption is activated after the cells adapted themselves to consume this new carbon source. This adaptation goes through a acceleration phase until an apparently second maximum growth rate is reached. No measurable contribution to the growth rate is given during the glucose-repressed period for lactose, even when some of the models predict such a situation. When lactose becomes limiting, the contribution of the lactose uptake rate ($\rho_2$) to the specific growth rate begins to decay. Again, as in the case of glucose, the main influence to the growth rate in this stage is given by the specific consumption rate for lactose, $\rho_2$.
No discussion about the specific consumption rate for glycerol is made here, because, as stated before, there was no experimental evidence of glycerol consumption.
Finally, Figure 3.8 presents the measured and modeled specific growth rate ($\mu$) for each of the models. Three of them evidence a similar behavior with the exception of model 1. The consistency of these results is based on the particular contributions coming from the different specific substrate uptake rates. The results also manifested the advantage of the inclusion of available theoretical knowledge. Between all the proposed and retrained models, model 2 and 3 can be highlighted as the best suited among all, not only because of their better performance (see Table 3.4), but also due to their well defined structure based in previous theoretical works.

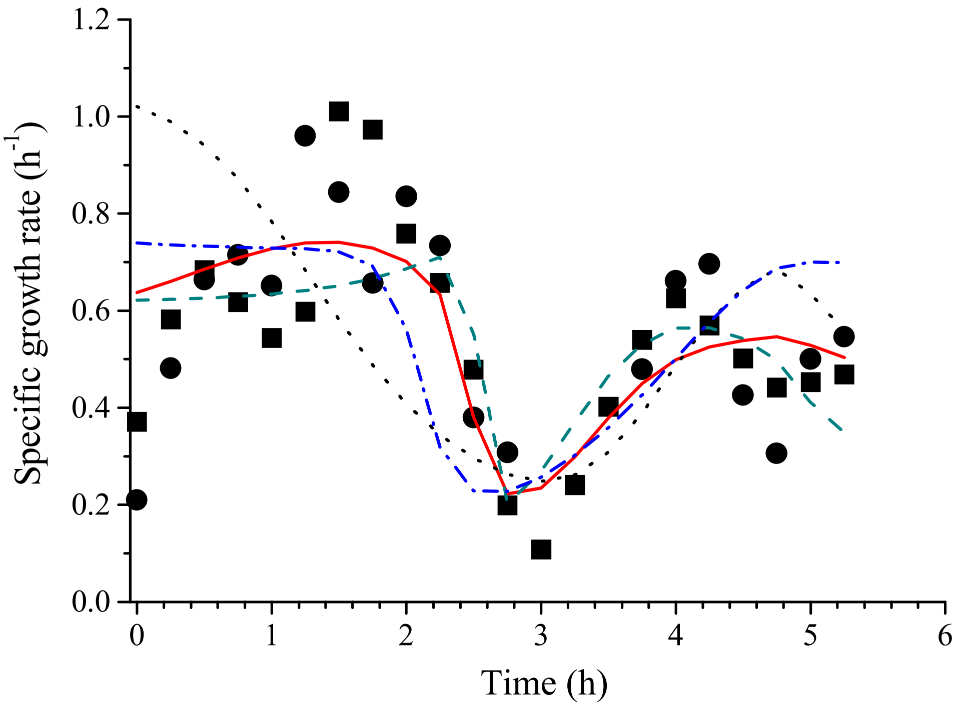

**Figure 3.8** Experimental (■, ●) and simulated data for the specific growth rate ($\mu$) for the different hybrid models: M1s_exp (···), M2s_exp (---), M3s_exp (—), M4s_exp (—·—).

## 3.4 Conclusions

Transfer of knowledge from a theoretical deterministic model into a hybrid model is of great advantage in the formulation of the latter. To describe a real process, the approximation of an artificial neural network as adaptive non-linear estimator of bioprocess kinetics has been used before (Montague and Morris, 1994). However, some clear advantages are acquired while coupling first principles balances and black box models, like the presented here. The first of these advantages is the easiness to set down the kinetics of the model by means of a relative simple neural network. A second advantage is the direct inclusion of the possible interactions between the states as part of the model's structure. The results also showed how flexible such models can be: in the case of noisy signals, they can even be filtered out at no evident cost in the model accuracy.

## 3.5 Nomenclature

| | | |
|---|---|---|
| $C_X$ | Biomass concentration | [g L$^{-1}$] |
| $S_1$ | Glucose concentration | [g L$^{-1}$] |
| $S_2$ | Lactose concentration | [g L$^{-1}$] |
| $S_3$ | Glycerol concentration | [g L$^{-1}$] |
| $\mu$ | Specific growth rate | [h$^{-1}$] |
| $\rho_1$ | Specific glucose uptake rate | [g g$^{-1}$ h$^{-1}$] |
| $\rho_2$ | Specific lactose uptake rate | [g g$^{-1}$ h$^{-1}$] |
| $\rho_3$ | Specific glycerol uptake rate | [g g$^{-1}$ h$^{-1}$] |
| $Y_{X/S1}$ | Glucose yield coefficient | [g g$^{-1}$] |
| $Y_{X/S2}$ | Lactose yield coefficient | [g g$^{-1}$] |
| $Y_{X/S3}$ | Glycerol yield coefficient | [g g$^{-1}$] |

## 3.6 References

1. Bellgardt, K.-H.: Growth of Microorganisms, in Biotechnology, 2nd. Ed. (Rehm, H.-J., Reed, G., Eds.), Vol. 1, pp 150-154, VCH (1991)

2. Bronstein, I. N.; Semendjajew, K. A.; Musiol, G.; Mühlig, H.: Taschenbuch der Mathematik. 4th Ed., Verlag Harri Deutsch (1999)

3. Kim, Y. H.; Lewis, F. L.: High-level feedback control with neural networks, World Scientific Publishing Co. Pte. Ltd. (1998)

4. Montague, G., Morris, J.: Neural-network contributions in biotechnology, TIBTECH, Vol. 12, pp. 312-324 (1994)

5. Oliveira, R.; Simutis, S.; Lübbert, A.: HYBNET, a new tool for advanced process modeling. Proceedings of the 1st European Symposium on Biochemical Engineering Science, Dublin, Ireland, pp. 182-183 (1996)

6. Simutis, R., Lübbert, A.: Exploratory analysis of bioprocess using artificial neural network-based methods, Biotechnol. Prog., pp. 479-487 (1997)

# Chapter 4

# Hybrid model-based optimization of the production of the viral capsid fusion protein VP1-DHFR

## ABSTRACT

A key requirement during the development of new production processes for recombinant proteins is to reduce the development time as far as possible. This obliges to bring the number of experiments to a minimum and to make most efficient use of a priori knowledge and the data from the remaining experiments. This chapter presents an approach that is based on hybrid process models for the fermentation part of these systems that optimally combine *a priori* knowledge and information from the available process data. Additionally, making used of the aforementioned model, an evolutionary model-based optimization technique is presented. The method is tested and validated on the production of the native fusion protein VP1-DHFR with *E. coli* as host microorganism. The optimization is essentially an information driven procedure: it uses all current available knowledge to formulate a mathematical representation of the bio-system which is used later to maximize the total amount of the viral protein complex at the end of the fermentation, optimizing the fed-batch working variables. The use and effectiveness of a hybrid model as a suitable convenient mathematical formulation for the bio-process is demonstrated too. Choosing a batch fermentation as reference and considering that both processes were run under the same optimization constrains. The amount of recombinant protein obtained by using this strategy, increased almost five-fold compared to the optimized conventional batch culture.

# 4. HYBRID MODEL-BASED OPTIMIZATION OF THE PRODUCTION OF THE VIRAL CAPSID FUSION PROTEIN VP1-DHFR

## 4.1 Introduction

The reduction of the invested time for developing and improving a given process is a fundamental demand in today's biotechnology being also a critical issue in biochemical industries, in particular in connection with processes in which new recombinant proteins are to be produced. Direct consequence of this challenge is the search for optimal solutions that allow, at the same time, a diminution in the number of expensive experiments too. Here it is considered the transfer of the product formation part of these processes from the biochemical or microbiological laboratory into a pilot scale fermenter. In the industry, this development is usually performed on a trial-and-error base, where a considerable number of experiences is indispensable to develop an efficient cultivation process. In order to reduce the development expenditure (time and cost) the number of such experiments must be kept as small as possible. This is only possible by systematic exploitation of all available knowledge about the process, careful design of the necessary experiments and exhaustive analysis of the data obtained therein.

The determination of optimal strategies has been carried out in different ways to improve fed-batch bio-processes. If a mathematical model of the system is available, all these processes can be optimized employing the method of Pontryagin (Pontryagin *et al.*, 1962). The application of the maximum principle is mainly restricted by the complexity of the model and the constrains of the process. Its application is specially complicated when considering the case of microbial systems with highly non-linear dynamics.

Two developments performed in recent years are notable respecting optimizing bio-processes. The first is that the *a priori* knowledge needs not necessarily be represented by classical mathematical models based on the mass balances and kinetic expressions of the Monod-type. It was shown that those parts of these models that are not completely understood, e.g. models for the process' kinetics, may be better represented in a data-driven form, e.g., by artificial neural networks, while the well known mass balances are still formulated by differential equation systems (Psichogios and Ungar, 1992; Schubert *et al.*, 1994; Van Can *et al.*, 1997). Such hybrid models, however, are believed to need extended data records for the training of their neural network components.

Artificial neural networks by themselves have been lately utilized with relatively high success for system modeling in biotechnology (Lübbert and Simutis, 1994; Montague and Morris, 1994). These black-box models associate certain known and measurable process input variables to other output variables of the process, whose values are usually not known or not measurable. A complex relationship between them is supposed to occur, but would be only described by the neural network after a proper training procedure.

The second development is the evolutionary process optimization procedure proposed by Galvanauskas *et al.* (1998). This procedure iteratively improves the process description while approaching the optimal feeding profile, or more generally, control profiles in fed-batch cultivation processes. As the data records from the process under consideration are very scarce at the beginning of such a development, this method was based on classical model approaches. This technique guarantees a quick approximation to optimal process control profiles.

In the present case, a model based optimization technique is applied to the production of a recombinant native protein. It consist in an iterative sequence of model identification and/or improvement followed by an optimization of the control variables using the identified model. This sequence is in agreement with the EOT (Estimation-Optimization-Task) methodology founded by Loeblein *et al.* (1999) and with the aforementioned experimental design of bioprocesses proposed by Galvanauskas *et al.* (1998). However, the main difference lays in the core of the approach, which uses a hybrid model to describe the bio-system.

As discussed before, the so called hybrid models are those based on a combination of a neural system and first principle formulations. The mass balance, in form of a differential equation system, can be complemented with a neural network that characterizes a badly known or, in some cases, very complex and high non-linear kinetic aspects of the bio-system. It is a common practice to use real data to validate the hybrid model ability to replicate the process under consideration. Validation is done presenting the hybrid model a dataset not used for identification purposes and later to evaluate its performance under this situation. Hybrid models were shown to be particularly useful in describing detailed interrelationships between variables that influence the profit of the production process. From this point of view, it would be of advantage to use hybrid models to extend the classical Monod-based approaches. The question is whether such hybrid models can be used from the beginning of such a development, i.e. even when only fairly scarce data are available. This problem is discussed in this work and the results are demonstrated on the example of a system producing the recombinant virus capsid fusion protein VP1.

Virus particles are considered interesting vectors for gene transfer. Most parts of the envelopes of virus particles are made from proteins, the capsid proteins. Such capsids can be produced artificially using appropriate recombinant protein expression systems. The example considered in this work is the fusion protein VP1. The protein is formed by the viral capsid protein 1 (VP1) and the dihydrofolate reductase (DHFR), which is inserted in the DE-loop of the protein. The fusion protein was expressed by *E. coli* bacteria. Such a fusion protein has the additional convenience that the concentration of the capsid protein can be measured via the activity of the attached enzyme. Measuring the enzymatic activity of the insert gives quantitative evidence of the expression of the native VP1 fusion protein, that serves to investigate the influence of the insert in the formation of pentamers (Braun *et al.*, 1999).

## 4.2 Model-based optimization and process description

The optimal estimation of control trajectories is a subject of crucial importance for the effective and economical operation of fed-batch fermentations. Such an estimation is specially desired for those cases where the operator can directly influence the process performance. The amount of product, its quality or even the reduction of fermentation duration are process variables subsceptible to be enhanced. Usually, the associated dynamical optimization problem can be solved with different methods coming from the optimal control (Wu *et al.*, 1982; Ponnuswamy *et al.*, 1987; Takamatsu *et al.*, 1988; Chang and Lai, 1992; Denbrigh, 1968; Edgar and Lapidus, 1972; San and Stephanopoulos, 1989; Lee and Ramirez, 1996). In general, it can be considered that the determination of optimal control is a constrained optimization problem. However, the solution of this problem is a very difficult task, mainly due to the complex kinetics and non-linear dynamics correlated with the growth of the living microorganisms. Additionally, in most of the cases, the profit functional defining the optimization task is also a complex function.

Moreover, the use of classical methods like dynamical programming or Pontryagin's maximum principle is commonly restricted to simple models (Pontryagin *et al.*, 1962; Park and Ramirez, 1988). Such a restriction arises mainly due to the mathematical expenditures required for its implementation. However, the description of biotechnological processes and their associated constraints problem has been done using highly non-linear complex models. This is the main reason why numerical optimization, based on a process model, is employed regularly to estimate the optimal trajectories of control variables. The accomplishment of the optimization task is done by transforming the problem into an algebraic system. After this transformation, the state and control variables are parameterized and therefore, suited to be solved numerically. Concerning the adjoined problem in the identification of parameters, random search techniques are gaining more and more attention, because they have proved to be fairly reliable and easy to implement, due to their uncomplicated evolution concept.

Simutis and Lübbert (1997) compared different random search methods applied to the optimization of the working conditions in technical bio-processes. Fundamental in these methods is that trajectories of the control variables can be formulated via a general mathematical function, like a polynomial representation. The free parameters of the representation can be varied in a random fashion until a predefined optimization criterion is fulfilled.

The described general model-based technique leads to an optimization scheme, which consists mainly of two parts. In the first step, a model for the process is established. This is done either through the identification of its parameters or through the modification of its structure after analyzing the available information. In the second part, the actualized model is considered adequate for numerical optimization of the control variable trajectories. As an example of this procedure, the sequence of identifying a model and optimizing the fed-batch conditions of a chemical process is reported by Loeblein *et al.* (1999). This strategy is known as Estimation-Optimization-Task (EOT) and its conceptual frame is presented in Figure 4.1.

The depicted EOT method, when applied and validated in an iterative manner, gives an approximation of the optimum state of the process. However, it should be taken into account that the actualized model is considered appropriate within a given working domain and it is extended for the actual optimization step. Resuming, the EOT method considers the accumulated knowledge from previous experiments and the information concerning the actual optimization domain. This technique is in accordance with the experimental design of bio-processes proposed by Galvanauskas *et al.* (1998), which utilizes a process design method, but oriented to the performance of the process itself.

Following essential aspects of the delineated procedures can be highlighted:

1. The choice of an experimental domain where the process can be enhanced. Its basis is a predefined profit function inside the optimization framework.
2. Existance of a model's accurancy criterium. The fulfillement of this criterium is basic to assure accurate long term prediction.
3. Cyclical calculation of optimal control trajectories. This strategy fulfills the moving horizon principle of the model predictive control (Garcia *et al.*, 1989) and can be applied for fed-batch processes with the additional reduction of the total process development time.

The EOT technique is also applicable in some situations that require the use of black-box models. However, it can be easily extended to cases where black-box representations are only single components of a more general model, like for example hybrid approaches. The numerical solution of the optimization problem produces a long term prediction for the process course. It supplies additionally an important criterium for the evaluation of the accurancy of the black-box or hybrid model.

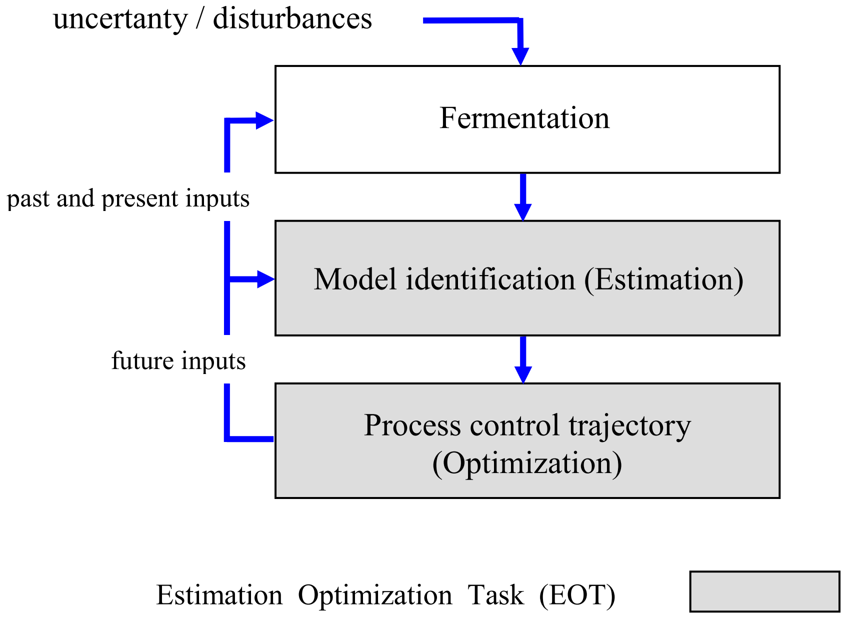
uncertanty / disturbances
Fermentation
past and present inputs
Model identification (Estimation)
future inputs
Process control trajectory (Optimization)
Estimation Optimization Task (EOT)

**Figure 4.1** Schematization of the Estimation-Optimization-Task (EOT) by Loeblein et al. (1999).

To investigate the application of the described optimization techniques in a fermentation, the production of polyomavirus-like particles was taken as example. This is an already well-characterized system (Salunke *et al*., 1986; Eckhart, 1991), established as a model system to investigate the gene transfer. Murine polyomavirus is a subspecies of the papovaviridae family which are non-enveloped, double-stranded DNA viruses with circular genomes of 5.3 kb in size. The crystal structures of VP1 in the virus shell as well as a proteolytically truncated form of pentameric VP1 have been reported recently (Stehle *et al*., 1994; Stehle and Harrison, 1997). *In vitro* studies demonstrated that purified VP1 can form virus-like particles consisting of 72 pentamers (Salunke *et al*., 1986). This feature makes the protein attractive for *in vitro* packaging of DNA (Slilaty *et al*., 1982) and gene transfer experiments (Forstova *et al*., 1995). The capsomer VP1 can be produced in recombinant form in *Escherichia coli* cells.
In the present case, *Escherichia coli* BL21 was utilized as host system. The microorganism contains an ampicillin resistant plasmid pBR322, responsible for the expression of a viral capsid protein using a tac-promotor. The over-expression of the recombinant protein was induced employing 1.5 mmol of IPTG (Isopropyl β-D-Thiogalactopyranosidase). After induction, the viral capsid protein of the murine polyomavirus (VP1) is fused with the enzyme dihydrofolate reductase (DHFR, EC 1.5.1.3). The native recombinant protein complex (VP1+DHFR) possesses a measurable enzymatic activity, which can be taken as a mass equivalent of the product. Therefore, any essay directed towards the determination of the DHFR activity should be interpreted as a quantitative description of the produced protein amount and that is why, the product amount is characterized hereon in units of activity (U).
Regarding the optimization, Galvanauskas *et al*. (1998) proposed to proceed iteratively. In each cycle one makes use of the available knowledge about the process under consideration. This knowledge is taken into account to propose an optimal set of control profiles and to predict the behavior of the process. The control profiles are immediately applied in a subsequent experiment, in order to validate the assumptions formulated in the model.

# 4. HYBRID MODEL-BASED OPTIMIZATION OF THE PRODUCTION OF THE VIRAL CAPSID FUSION PROTEIN VP1-DHFR

The first optimization step was based on a classical unstructured model for the batch production of VP1-DHFR proposed by Volk *et al.* (1998). The model consisted of a mass balance for the different state variables of the system (biomass, substrate and product activity) complemented by classical Monod-like growth kinetics with temperature dependence. The protein production rate term was described by a Moser-like kinetic term correlated with the specific growth rate and the specific protein activity. The protein production rate and the specific protein activity were also complex functions of the temperature (see Appendix A2). Volk *et al.* (1998) showed that the induction time ($t_I$) and the temperature after the induction with IPTG were the most important optimization variables to enhance the productivity of the system.

In the present case, the fermentations were carried out under fed-batch conditions trying to avoid effects of substrate inhibition and to bring an additional increment in the productivity by means of an optimal substrate feeding function. Optimization goal ($J_{OPT}$) was thus to produce as much of the native protein as possible within a predefined cultivation time, $t_f$:

$$J_{OPT} = P(t_f)\, W(t_f) \rightarrow \max \tag{4.1}$$

where $P(t_f)$ is the product activity per weight and $W(t_f)$ the culture weight. The key operational variables were the feeding rate $F(t)$ and the initial concentrations of substrate and biomass. Since the development of the particular product required the addition of the inducer IPTG, one extra variable to be optimized is the induction time, $t_I$.

Also, due to their capability to properly describe complex non-linear relationships, an artificial neural network was employed to compute the optimal feeding profile. In this case, the neural network generated a complex time function $F(t)$ that maximized the profit function described by Equation 4.1. The complete mathematical description of the bio-system working under fed-batch conditions is detailed in the Appendix A2.

To determine the optimal feeding profile an evolutionary technique was used: the chemotaxis algorithm (San and Stephanopoulus, 1989; Simutis and Lübbert, 1997). This was implemented for training the neural network that defines the feeding function. The training was carried out on the HybNet Software (Oliveira *et al.*, 1996; Oliveira *et al.*, 1998) and a description of the method can be found in the *Materials and Method* chapter, section 2.2.2.

The optimal induction time ($t_I$) was also determined with the chemotaxis algorithm (see Figure 4.2). The parameter initial guess was defined as the half of the total fermentation time ($0.5 \bullet t_f$). The parameter could be changed in the interval, $t_I \in [0,t_f]$.

A temperature profile consisting of two phases was chosen for the cultivation (see Figure 4.2). Before induction, the temperature was maintained constant at 37°C, i.e. at the temperature of the maximal specific growth rate ($\mu_{\max|37°C}$) for *E. coli*. After induction and until the end of the fermentation, the temperature descended linearly from 37°C to 25°C.

The optimization variables found with the described methodic are presented in Figure 4.2. An experimental fed-batch fermentation run was carried out using these profiles. The results are depicted in Figure 4.3, where modeled variables, biomass and glucose, together with their measured counterparts can be seen.

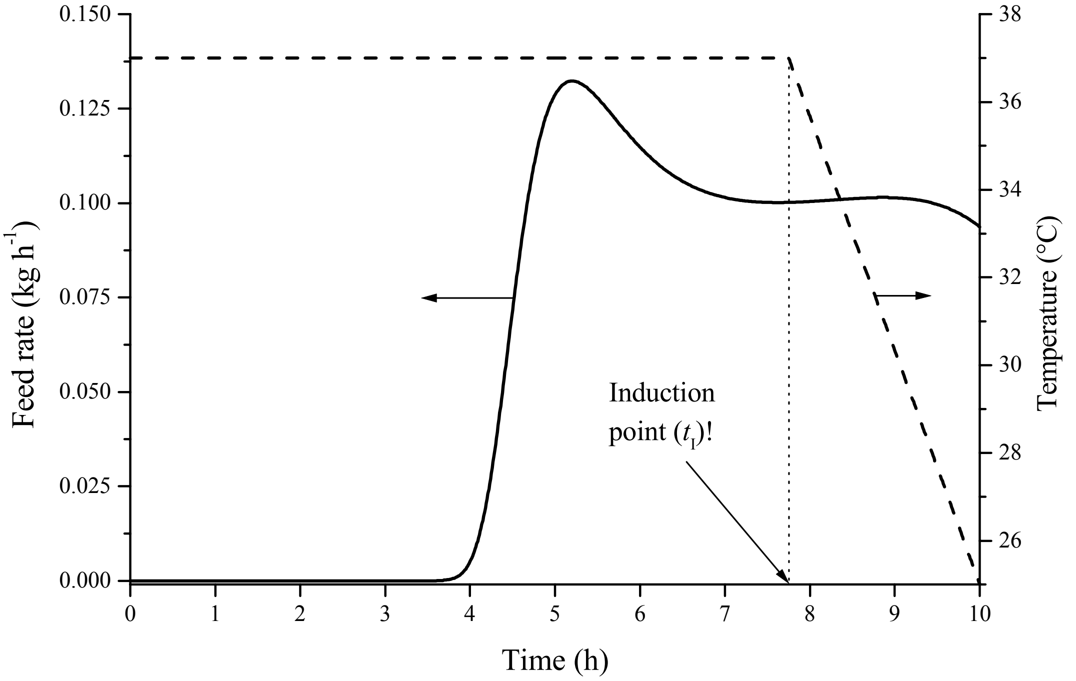

**Figure 4.2** Optimal induction point and temperature-, feeding rate trajectories for the fed-batch production of VP1-DHFR.

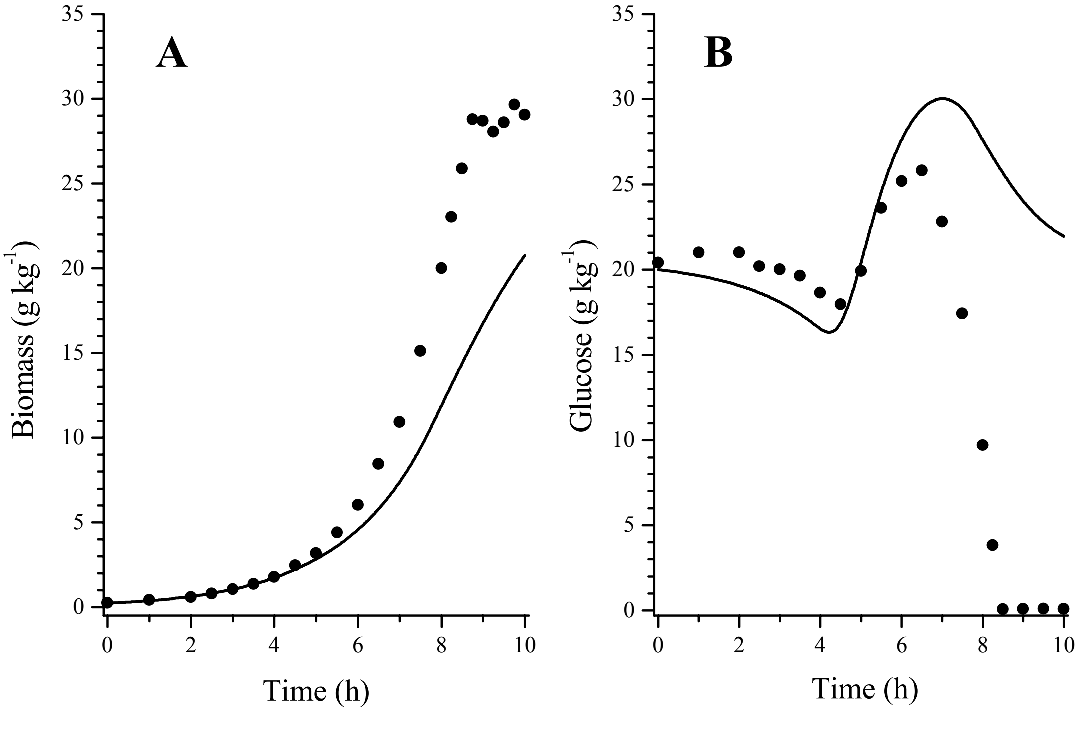

**Figure 4.3** Predicted (—) and measured (•) values for biomass (A) and glucose (B) concentration, after the first fed-batch optimization.

Concerning the model, an apparent deviation between predicted and experimental measurements was constated. The model was obviously not able to adequately describe the features of the fed-batch fermentation and should be improved. Especial attention was paid to the specific growth rate term, which was suspected to play the key role of this discrepancy.
However, in agreement with the results reported by Volk *et al.* (1998), the induction time was situated at the end part of the fermentation. Both optimization experiences exhibited long lasting pre-inductive phase with unrestricted growth of biomass on the substrate. After the induction with IPTG, the product was formed mainly as native protein. It is suspected that the decreasing temperature displaces the equilibrium between the native product and its inactive counterpart, protein in inclusion bodies (Tsai *et al.*, 1995). It seemed that the proposed optimization procedure favored the accumulation of the native fusion protein.
Moreover, the experimental run supplied some additional useful information coming from the analysis of the exhaust gases, oxygen and carbon dioxide. These process data has been traditionally used for the dynamic indirect measurement of the specific growth rate, making use of relevant correlated variables like the Oxygen Uptake Rate, OUR and the Carbon dioxide Production Rate, CPR (Chéruy and Flaus, 1994). Therefore, it was advisable that any modification of the model towards the improvement of its prediction capabilities, should also include the information delivered by the gas analysis of the outflow fermentation gases. One alternative method to do this, was the so called "hybrid modeling", discussed in the next section.

## 4.3 Hybrid modeling of the production of the fusion protein

As stated by some authors (Psichogios and Ungar, 1992; Schubert *et al.*, 1994; Van Can *et al.*, 1997), the main feature of the so called hybrid modeling approach is the combination of "white box" models (containing the available knowledge usually in form of mass balance equations) and "black box" models like ANN (describing the unknown or very complex aspects like bio-system kinetic). The resulting hybrid model can be structured either in a parallel or in a serial architecture, the last of them, to be used in this work, is represented in Figure 4.4.

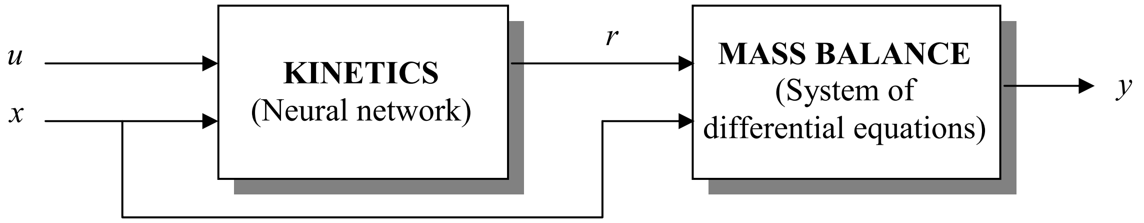

**Figure 4.4** Serial architecture for a hybrid model consisting on two components: a "black-box" model (neural network) and a "white-box" model (mass balance equations).

One of the advantages using the hybrid modeling technique was the easiness to incorporate all available information (state and on-line measured correlated variables) into a robust "black-box" component that can completely replace the formulation for the specific growth rate (see Appendix, Eq. A2.10). This component was a feed-forward neural network with a single hidden layer (6 nodes with sigmoid activation functions), that was thought to be able to improve the actual estimation of the specific growth rate, $\mu$. The approach makes use of all those variables which presented a significant influence in the estimation of specific growth rate in a direct or indirect manner. Following variables were tested as the most important input elements of the ANN (see *Nomenclature*): $t$, $C_X$, $S$, $P$, $pO_2$, $T(t)$ and $F(t)$.

OUR and CPR can be measured as well as modeled. The specific growth rate can be formulated also as an indirect function of the measured OUR, CPR and viceversa. Such kind of formulations are usually described as software sensors or shorter, soft-sensors. As described by Chéruy and Flaus (1994), the specific growth rate can be dynamically determined based on OUR and CPR measurements from the expressions:

$$\text{OUR (t)} = \beta_1\, \mu\, C_X + \beta_2\, C_X \qquad (4.2)$$

$$\text{CPR (t)} = \gamma_1\, \mu\, C_X + \gamma_2\, C_X \qquad (4.3)$$

The basic idea in use of these soft-sensors was to force convergence of in the estimation of the specific growth rate. This can be done including the influence of the model to measurement error of OUR and CPR into the identification procedure for the specific growth rate. This function is consistent insofar the modeling error of the states variables and those of the OUR and CPR tend to be minimal.

The identification procedure of the hybrid model consists on estimating the specific growth rate function that minimizes a given modeling error criterion. This criterion can be mathematically defined by:

$$J_{\text{IDEN}}\,(y, Y) \equiv \sum_{j=1}^{M}\sum_{i=1}^{N}\left(\frac{y_i - Y_i}{\text{KN}_j}\right)_j^2 \longrightarrow \text{min!} \qquad (4.4)$$

where $y$ are the measured OUR, CPR and state variables; $Y$ are the simulated OUR, CPR and state variables; $i$ is referred to the number of $N$ available data; $j$ is referred to the state variables under consideration and KN are pseudo-norms for the $M$ single state variables together with OUR and CPR. The introduction of a pseudo-norm was done to constrict the influence of the single terms (weighting technique), because their real values lie in different numerical ranges. The values of these pseudo-norms were determined empirically.

The chemotaxis algorithm was also used to identify the parameters of the neural network and those of the modeled OUR and CPR. The data used for calculating the identification criteria were the measured data coming from the fed-batch optimization. The whole identification procedure is depicted in Figure 4.5.

Estimated and measured variables obtained after the identification of the hybrid model are shown in Figure 4.6. As can be asserted, the results were satisfactory and much better than the case of the Monod-type approach of the specific growth rate. The figure exhibits the high accuracy achieved in the estimations of the specific growth rate and the correspondent state variables $C_X$, $S$ and $P$.

As pointed out before, the computations to be performed during the optimization did not only led to improved control profiles for the key parameters, they also provided an enhanced dynamic process model. This could not only be used to estimate the state variables during the off-line optimization. It also forms the base for on-line state estimation and software sensors for various other variables.

An example is provided in Figure 4.7. Here a software sensor for the current oxygen uptake rate and for the partial pressure of dissolved oxygen is displayed. This example was chosen since the oxygen uptake rate can also be calculated from the off-gas analysis, therefore it was possible to compare the accuracy that could be obtained with the software sensors. The same accuracy of an indirect measurement was also achieved for the variable CPR (not showed). Hence, it was possible to monitor these variables during the production process quite accurately.

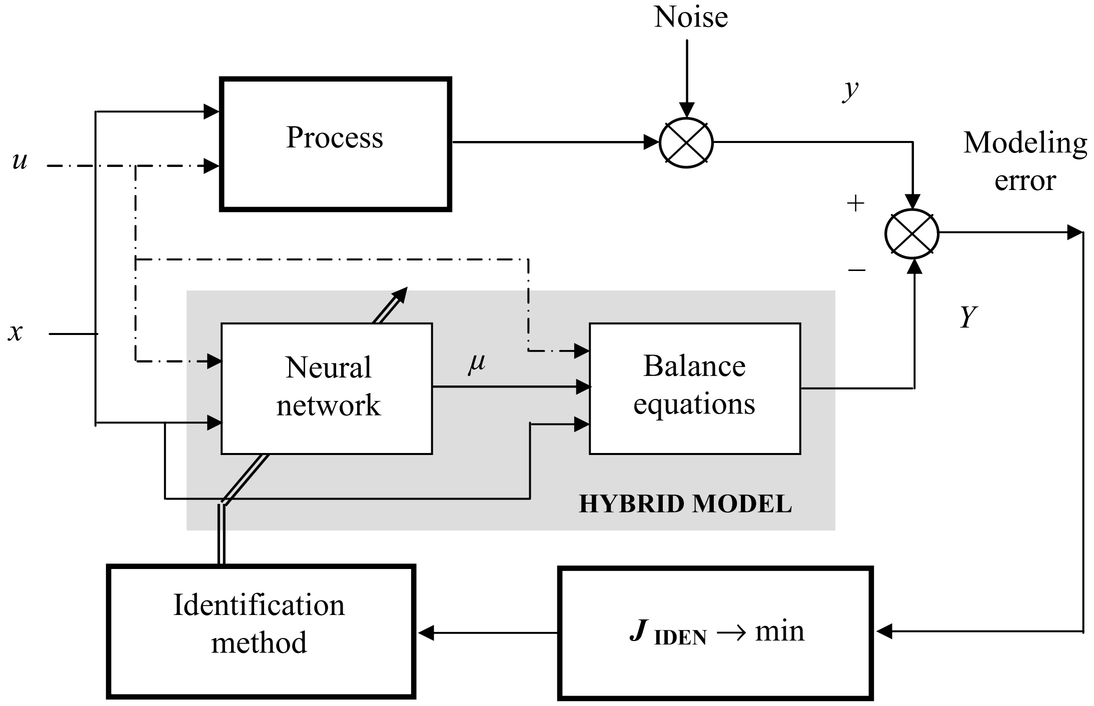

**Figure 4.5** Schematic representation of the identification procedure of the hybrid model.

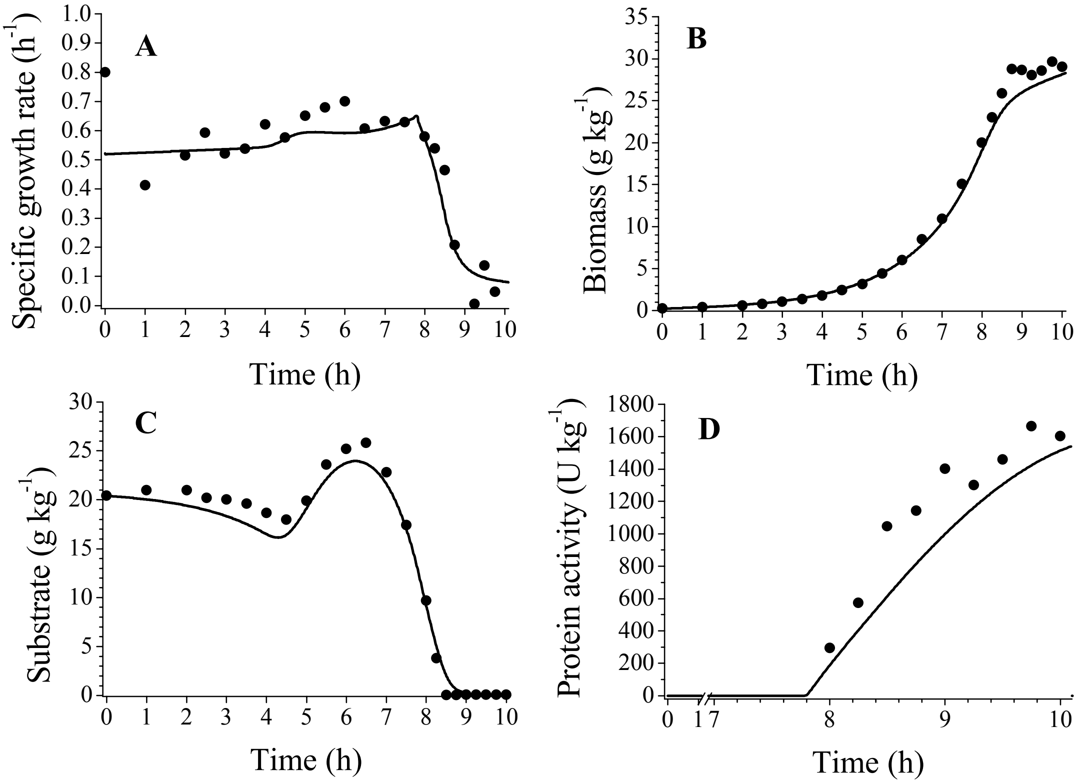

**Figure 4.6** Identified (—) and measured (•) values for the specific growth rate function (A) and its state correlated variables: biomass (B) and substrate (C) concentration, and protein (D) activity.

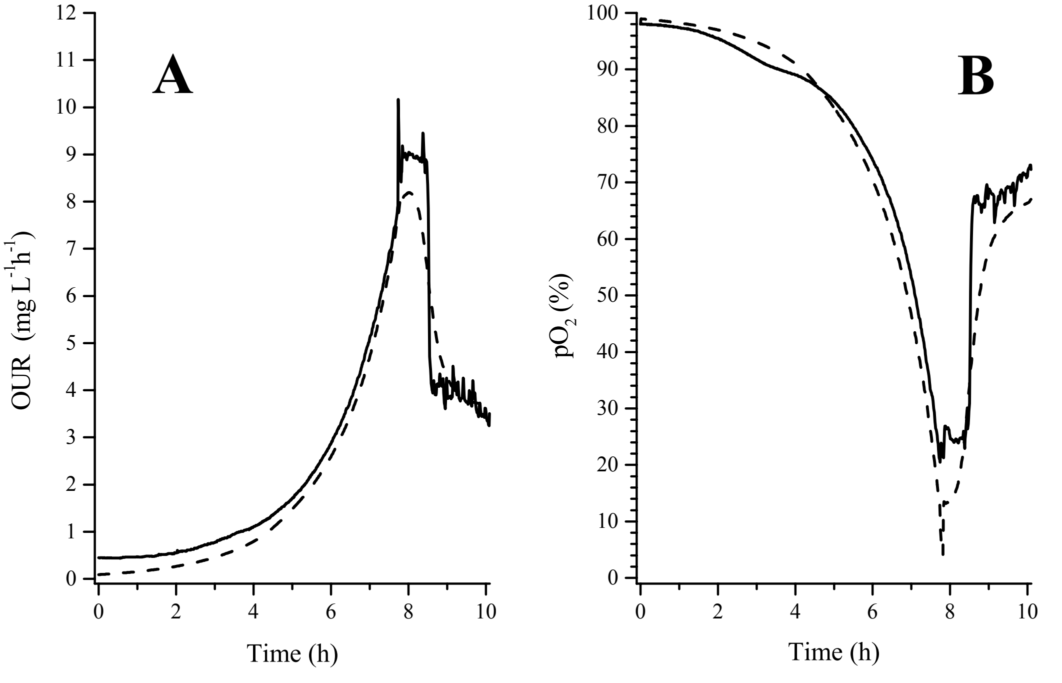

**Figure 4.7** Modeled (- - - ) and measured (—) oxygen uptake rate (A) and $pO_2$ (B) variables.

## 4.4 Optimization using a hybrid model

To justify the use of a hybrid model as an alternative to classical models, once the model was identified, it was considered to be suited to optimize the operating conditions of the bioprocess. Just like it was described before, the substrate feed rate and the induction point were optimized, but this time using the identified hybrid model. The temperature profile is a function of the optimized induction time. Figure 4.8 presents a schematic representation of this optimization cycle procedure.

After calculation, the new optimized variables were set up in a fermentation (see Fig. 4.9) and the experimental data obtained was used for independent validation of the hybrid approach. In this new optimization cycle, the induction time was displaced towards the end of the fermentation. As consequence, longer pre-inductive phases were obtained in comparison to its former two optimized fermentations. It seemed that with the use of the hybrid model the accumulation of biomass in the pre-inductive phase of the fermentation was forced .

Regarding the substrate concentration and especially its high concentration at the end of the fermentation, the phenomenon can be explained through the defined optimization task (Equation 4.1). As stated before, the optimization task was to maximize the total amount of product at a given predefined time, but not considering the presence of substrate at the end of the fermentation. A future optimization cycle may then consider the simultaneous optimization of the substrate concentration on the post-inductive phase.

Concerning the use of the aforementioned evolutionary optimization strategy, Figure 4.10 compares the different time trajectories obtained for the profit function from batch to fed-batch working conditions (using a classical model) and, under fed-batch conditions, from using a classical to a hybrid approach. The increase in the process performance from the referred optimized batch fermentation to the first optimized fed-batch was about a factor of 2.6, while an improvement of about 1.8 was obtained from one fed-batch experiment to the next.

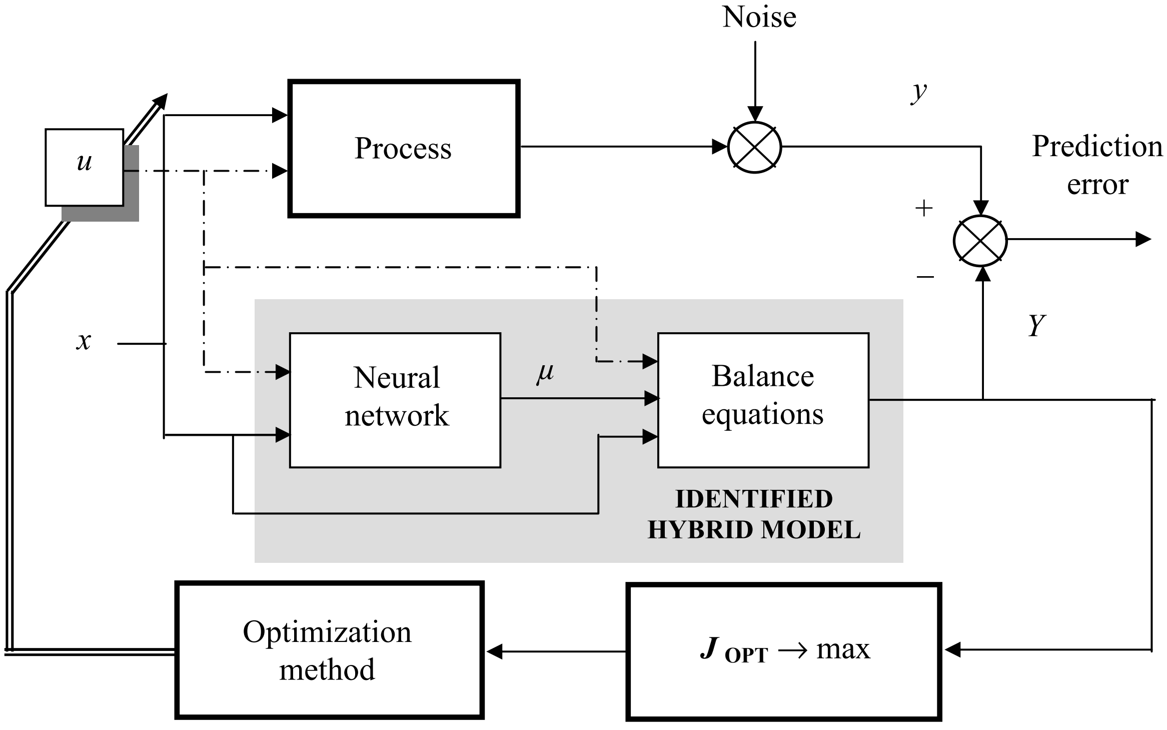

**Figure 4.8** Single optimization cycle making use of an identified hybrid model.

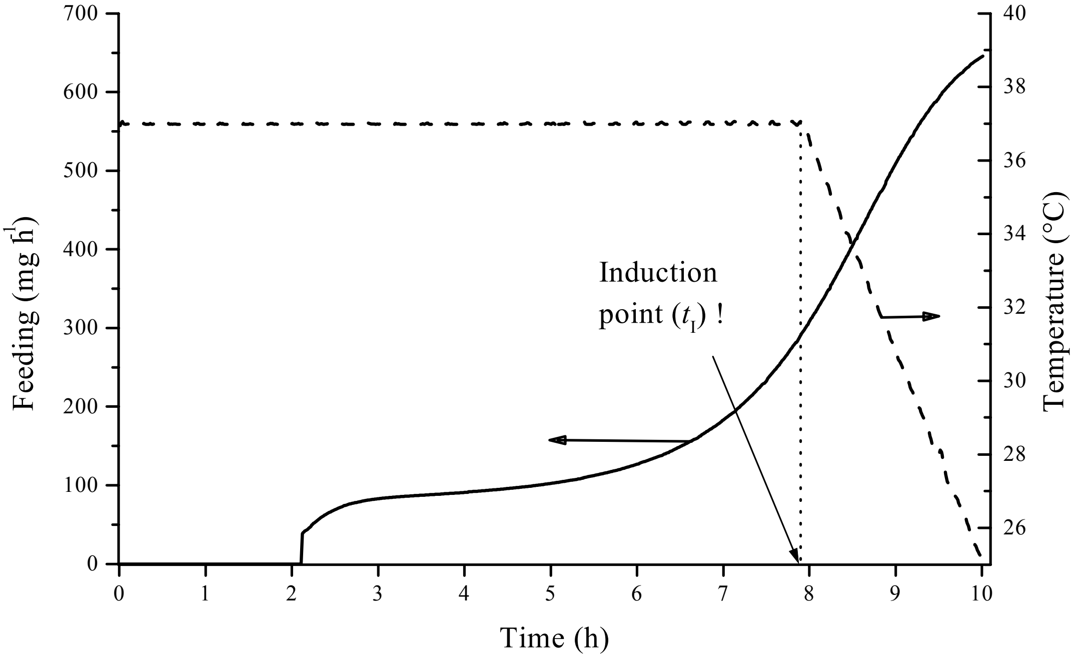

**Figure 4.9** Recalculated optimal induction point and temperature-, feeding rate trajectories for the fed-batch production of VP1-DHFR based on a hybrid model approach.

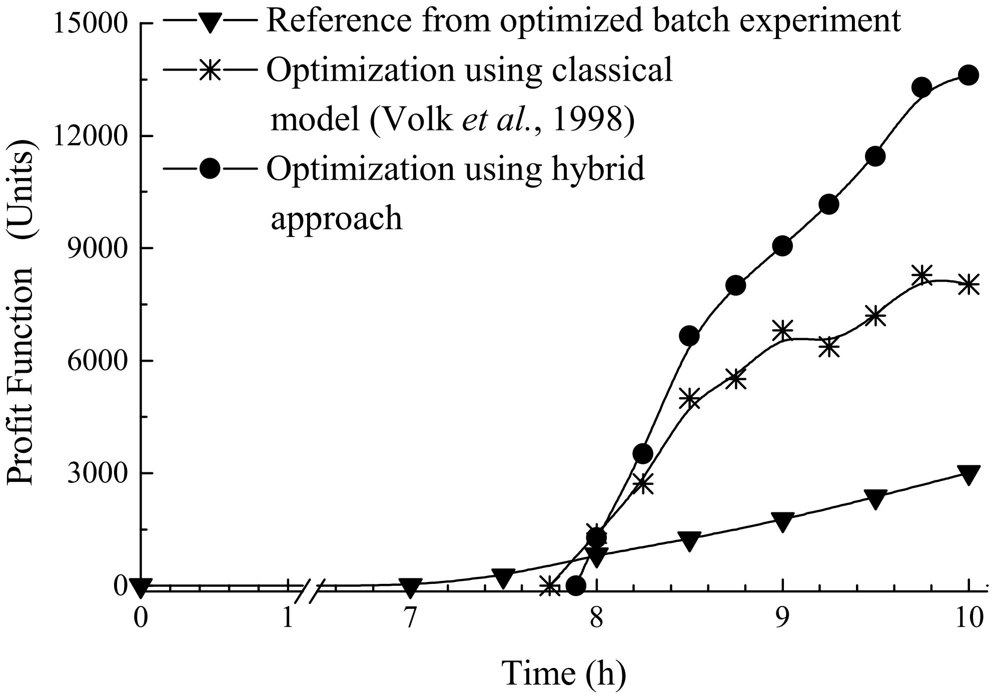

**Figure 4.10** Comparison of the time trajectories for the profit function with the sequential optimization steps from the evolutionary optimization strategy.

Just like in the first optimization step, a validation of the current model was necessary. This could be done employing the results coming from the optimization using the hybrid approach, which are depicted in Figure 4.11. The predicted values for the most important state variables of the process achieved good accuracy when compared to their experimental counterparts. This was especially clear in the case of substrate concentration trajectory. In the case of the rest of the product concentration there existed a process-model mismatch between the measurement and estimation. This situation was thought to be caused by a small error trend in biomass prediction at the end of the fermentation, precisely where the post-inductive phase took place.

The described information-driven optimization can be brought into play in an iterative fashion towards the establishment or improvement of any given bio-process. Each cycle consisted on the following steps:

1. Actualization of the model of the bio-system. This is done either modifying the structure of the model or adjusting its parameters using available *a priori* knowledge and concrete process data.
2. The actualized model may be set up for optimization purposes. In the present case, the optimization goal was to achieve a maximal amount of protein at the end of the fermentation.
3. Carry out a validation experiment. This would be used to independently validate the accuracy of the model approach. The collected data is also suited for improvement of the model in a new optimization cycle.

Usually only a few of these optimization cycles were needed to establish a proper and faithful model structure. If necessary, for the most part of the applications some later improvements can be reached only by re-identification of the model's parameters.

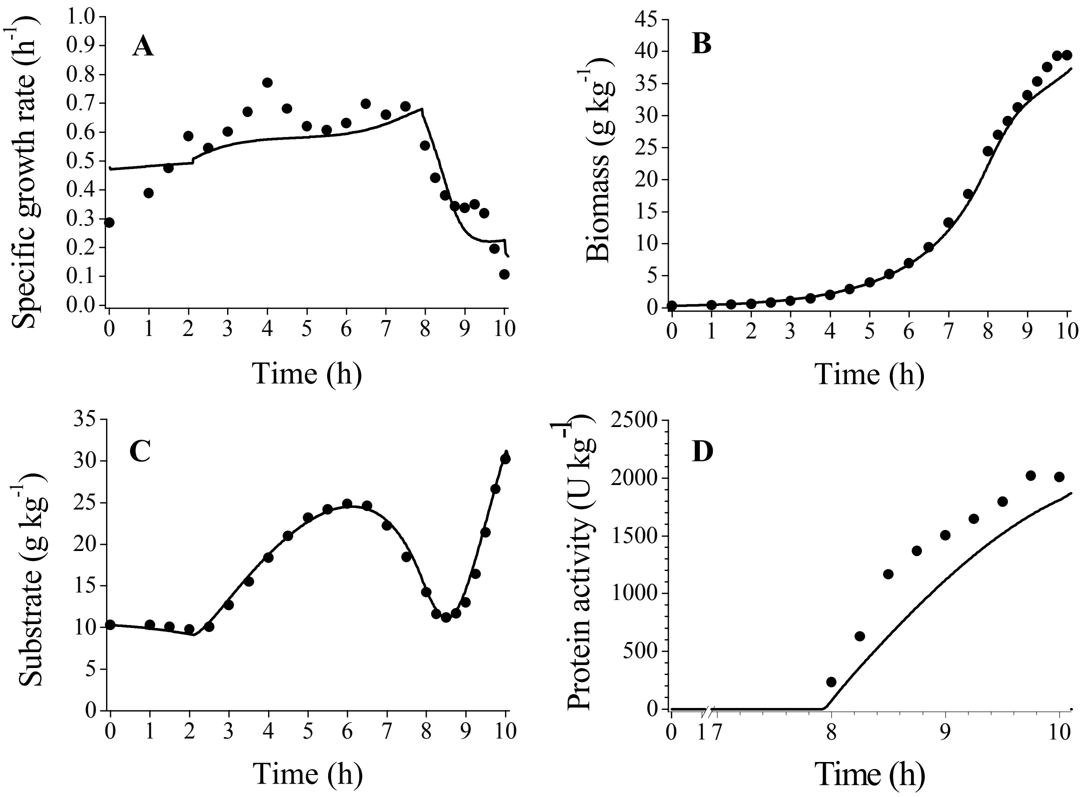

**Figure 4.11** Predicted (—) and measured (•) values for the specific growth rate (A) and its correlated state variables biomass (B) and substrate (C) concentration, and protein(D) activity for the fed-batch optimization using a hybrid model approach.

## 4.5 Conclusions

Exemplified on the production of the recombinant protein complex VP1-DHFR with *E. coli*, it was demonstrated that with the use of an information driven optimization, it can be possible to reduce the number of experiments and human effort required for optimizing a fermentation. This evolutionary optimization procedure bases its functioning principle in the recurrent incorporation of knowledge into a formulation that serves to model the bio-system under consideration. In the present case, the optimization task was to maximize the final amount of the native form of the VP1-DHFR protein complex. This was done manipulating the process working conditions through control variables that affect directly the protein expression. To integrate the available information, artificial neural networks were utilized to formulate the specific growth rate. The neural network unit (representing only the high non-linear kinetics) was part of a more general hybrid model of the fed-batch fermentation. Optimized were the substrate feed rate and the induction time for expressing the protein. A linear temperature profile as function of the induction time was also utilized to optimize the fed-batch process. The hybrid approach was demonstrated to be suited for the proposed model-based optimization. An additional advantage in the use of hybrid models is the reduction of the amount of information required for proper training, in comparison with stand-alone neural network models. Random search algorithms were used in the identification of the hybrid model and in the estimation of the control variables. No apparent numerical instabilities or restrictions in the number of these control variables or their complexity seemed to be present using these methods. The performance of the system was rapidly increased with just a few experiments following sequential optimization steps. Each new optimization cycle was used simultaneously as validation criterion for the accuracy of the current model approach. Resuming, during the proposed procedure the process model was improved in an evolutionary way. This means that:

1) The model was changed using currently available knowledge and validated on the collected measurement data.
2) The model able to describe the actual available data set, was selected to:
    a) Form the base of an accurate bio-process description.
    b) Propose the optimal control profiles for the next validation experiment and thus to improve the process performance.

## 4.6 Nomenclature

| | | |
|---|---|---|
| CPR | Carbon dioxide Production Rate | [mg kg$^{-1}$ h$^{-1}$] |
| $C_X$ | Biomass concentration | [g kg$^{-1}$] |
| $F$ | Feed rate | [kg h$^{-1}$] |
| $J_{IDEN}$ | Identification criteria | |
| $J_{MOD}$ | Optimization goal | [Units of activity] |
| OUR | Oxygen Uptake Rate | [mg kg$^{-1}$ h$^{-1}$] |
| $P$ | Product concentration | [Units kg$^{-1}$] |
| $pO_2$ | Partial pressure of dissolved oxygen | [%] |
| $r$ | Vector of kinetic variables | |
| $S$ | Substrate concentration | [g kg$^{-1}$] |
| $t$ | Time | [h] |
| $t_F$ | Time at the end of fermentation | [h] |
| $t_I$ | Induction time | [h] |
| $T$ | Temperature | [°C] |
| $u$ | Vector of input/control variables | |
| $W$ | Liquid reaction weight | [kg] |
| $x$ | Vector of state variables | |
| $y$ | Vector of modeled outputs variables | |
| $Y$ | Vector of measured outputs variables | |
| $Y_{X/S}$ | Yield coefficient Biomass/Substrate | [g g$^{-1}$] |

**Greek symbols**

| | | |
|---|---|---|
| $\beta_1$ | Model constant for CPR | [mg g$^{-1}$] |
| $\beta_2$ | Model constant for CPR | [mg g$^{-1}$ h$^{-1}$] |
| $\gamma_1$ | Model constant for OUR | [mg g$^{-1}$] |
| $\gamma_2$ | Model constant for OUR | [mg g$^{-1}$ h$^{-1}$] |
| $\mu$ | Specific growth rate | [h$^{-1}$] |

## 4.7 References

1. Braun, H.; Boller, K.; Löwer, J.; Bertling, W. M.; Zimmer, A.: Oligonucleotide and plasmid DNA packaging into polyoma VP1 virus-like particles expressed in *Escherichia coli.* Biotechnol. Appl. Biochem. 29 (1999) 31–43
2. Chang, J. S.; Lai, J.: Computation of Optimal Temperature Policy for Molecular Weight Control in Batch Polymerisation Reactor. Ind. Eng. Chem Res. 31 (1992) 861
3. Chéruy, A.; Flaus, J.M.: Des mesures indirectes à l'estimation en ligne. In: Capteurs et mesures en biotechnologie (Ed. Boudrant, J.; Corrieu, G.; Coulet, P.) pp. 444-484. France: Technique et Documentation – Lavoisier 1994 (In French)
4. Denbigh, K. G.: Optimum Temperature Sequences in Reactors. Chem. Eng. Sci. 8 (1968) 125.
5. Eckhart, W.: Polyomaviridae and their replication. In: Fundamental Virology, 2nd ed. B.N. Fields and D.M. Knipe eds. New York: Raven Press (1991) 727-741.
6. Edgar, T. F.; Lapidus, L.: The Computation of Optima Singular Bang-Bang Control II: Nonlinear Systems. AlChEJ 18 (1972) 780.
7. Forstova, J.; Krauzewicz, N.; Sandig, V.; Elitott, J.; Palkova. Z.; Strauss, M.; Griffin, B.E.: Polyomavirus pseudocapsids as efficient carriers of heterologous DNA into mammalian cells. Hum. Gene Ther. 3 (1995) 297-306.
8. Galvanauskas, V., Simutis, R., Volk, N., Lübbert, A. Model based design of a biochemical cultivation process. Bioprocess Engineering, 18 (1998) 227-234
9. Garcia, C.E; Prett, D.M; Morari, M.: Automatica, 25 (1989) 335.
10. Ginkel S.Z., Dooley T.P., Suling W.J., Barrow W.W.: Identification and cloning of *Mycobacterium avium folA* gene, required for dihydrofolate reductase activity. FEMS Microbiology Letters 156 (1997) 69 – 78
11. Lee, J.; Ramirez, W. F.: On-line optimal control of induced foreign protein production by recombinant bacteria in fed batch reactors. Chem Eng. Res. 51 (1996) 521
12. Loeblein, C.; Perkins, J. D. ;Srinivasanb, B.; Bonvin, D.: Economic performance analysis in the design of on-line batch optimization systems. Journal of Process Control, 9 (1999) 61-78
13. Lübbert, A.; Simutis, R.: Adequate use of measuring data in bioprocess modeling and control. In: Trends in Biotechnology 12 (1994) 304-311
14. Montague, G.; Morris, J.: Neural-network contributions in biotechnology. Trends in Biotechnology 12 (1994) 312-324
15. Oliveira, R.; Simutis, R.; Azevedo, F.; Lübbert, A.: HYBNET, a new tool for advanced process modeling. In: B. Glennon et al. (Ed.): Proceedings of the 1st European Symposium on Bioch. Eng. Sci., Dublin, Ireland. (1996) 182-183

16. Oliveira, R.; Simutis, R.; Feyo de Azevedo, S.; Lübbert, A.: Hybnet, an advanced tool for process optimization and control. In: Proc.Int.Conf.Computer Applications in Biotechnology, T. Yoshida, ed., IFAC Publ., Elsevier Science Ltd. (1998) 315-320

17. Park, S.; Ramirez, W. F.: Optimal production of secreted protein in fed-batch bioreactors. AIChEJ 34 (1988) 1550-1558

18. Patkar, A.; Seo, J. H.; Lim, H. C.: Modeling and Optimization of Cloned Invertase Expression in *Saccharomyces cerevisiae*. Biotech. & Bioeng. 41 (1993) 1066-1074

19. Ponnuswamy, S. R.; Shan, S. L.; Kparissides, C. A.: Computer Optimal Control of Batch Polymerization Reactor. Ind. Eng. Chem. Res. 26 (1987) 2229

20. Pontryagin, L. S.; Boltyanskii, Y. G.; Gramkrelidze, R. V. Mishchenko, E. F.: The mathematical theory of optimal processes. New York: Wiley Interscience 1962

21. Psichogios, D. C.; Ungar, L. H.: A hybrid neural network – first principles approach to process modeling. AIChE J. 38 (1992) 1499-1511.

22. Salunke, D.; Caspar, D.L.D.; and Garcea; R.L.: Self-assembly of purified polyomavirus capsid protein VP1. Celi 46 (1986) 895-904.

23. San, K. Y.; Stephanopoulos, G.: Optimization of a Fed-Batch Penicillin Fermentation Processes. Biotechnol. Bioeng. 34 (1989) 72.

24. Schubert, J.; Simutis, R.; Dors, M.; Havlik, I.; Lübbert, A.: Bioprocess optimization and control: Application of hybrid modeling. J. Biotechnology. 35 (1994) 51-68

25. Simutis, R.; Lübbert, A.: A comparative study on random search algorithms for biotechnical process optimization. J. Biotech. 52 (1997) 245-256

26. Slilaty, S.N.; Berns, K.I.; and Aposhian, H.V.: Polyoma-like particles: characterization of the DNA encapsidated *in vitro* by polyoma empty capsids. J. Biol. Chem. 257 (1982) 6571 -6575.

27. Stehle, T.; Van, Y.; Benjamin, T. L.; Harrison, S.C.: Structure of murine polyomavirus complexed with an oligosaccharide receptor fragment. Nature 369 (1994) 160-163.

28. Stehle, T.; Harrison, S.C.: High-resolution structure of a polyomavirus VP1-oligosaccharide complex: implications for assembly and receptor binding. Emboj. 76 (1997) 5139-5148.

29. Takamatsu,T,; Shioya, S.; Okada Y.: Molecular Weight Distribution - Control in a Batch Polymerization Reactor. Ind. Eng Chem. Res. 27 (1988) 93

30. Tsai, A. M.; Betenbaugh, M. J.; Shiloach, J.: The kinetics of RCC1 inclusion body formation in *Escherichia coli*. Biotech. & Bioeng. 48 (1995) 715-718

31. Van Can, H. J. L.; Te Braake, H. A. B.; Hellinga, C.; Luyben, K. C. A. M.; Heijnen, J. J.: An efficient model development for bioprocesses based on neural networks in macroscopic balances. Biotech. & Bioeng. 54 (1997) 549-566

32. Volk, N.; Hertel, T.; Lübbert, A.; Modellgestütze Optimierung der Produktion des Virus-Hüllproteins VP1-DHFR mit *E. coli* BL21. Chem. Technik 4 (1998) 192-197

33. Wu, G. Z. A.: Denton, L. A.; Laurence, R. L.: Batch Polymerization of Styrene-Optimal Temperature Histories. Polym. Eng. Sci. 22 (1982) 1

# Chapter 5

# Online modeling and optimization of the production of the viral capsid protein VP1-DHFR using a neuro-fuzzy approach

## ABSTRACT

A hybrid model to describe the production of the viral capsid protein VP1-DHFR is presented. This consists in set of differential equations describing the mass balance for the bio-process complemented with artificial neural network or neuro-fuzzy components that describe the corresponding kinetics. Considering the specific growth rate as the key variable for the process, this was mathematically formulated with a feed forward neural network. This model acts essentially as a dynamical adaptive system, able to learn from the online measured process variables. This kind of online learning method improves the gain of knowledge at high learning rates, overcoming slow convergence during the initial training stages. Heuristic knowledge about the complex protein production was formulated through a neuro-fuzzy expert system using a set of simple rules-of-thumb. Using this approach it can be shown that increased productivities are obtainable with the online-identified hybrid model, as compared to optimizations employing conventional approaches. These results exhibit the potential for application in the biochemical industry. Furthermore, the generality of this method can be also extended to a variety of processes and products.

## 5.1 Introduction

The use of recombinant protein systems for producing new drugs in the pharmaceutical industry has gained much attention in the last time. Particularly, evolutionary model supported process design has been applied recently in an increasing number of such production processes. In the early stages of the development, simple heuristic kinetic approaches, like Monod formulations, meet the initial accuracy requirements to roughly describe the kinetics of the process. However, when further development proceeds, it is usual that simple extensions to inhibitions or repression terms do not suffice. Under this consideration, instead of going into a long term theoretical investigation of the metabolism of the bioprocess, the alternative of a reliable data driven procedure is chosen. This procedure is advantageous in terms of reducing the development time and also by increasing the benefit/cost ratio.

The kinetic expressions for the biochemical conversion rates within the balance equations system are thus modeled by an artificial neural network. In order to incorporate already available heuristic knowledge about the protein development, the corresponding kinetic component is complemented with heuristic rules-of-thumb formulated by means of a simple neuro-fuzzy expert system. The parameters of the neuro-fuzzy kinetic model were identified within the environment of the entire balance equation system by means of a random search procedure, the chemotaxis algorithm. The main advantage of this type of network representation is that the results of the training can be made more transparent to the process engineer. The modeling procedure is illustrated at the example of the viral capsid protein construct (VP1-DHFR), which is expressed in recombinant *E. coli*. This construct is formed by the viral capsid protein 1 (VP1) and the insert dihydrofolate reductase (DHFR). The production process consists of two phases: a pre-inductive stage, where the biomass is maximized and a post-inductive phase, where the over-expression of the native recombinant protein is induced employing IPTG (Isopropyl β-D-thiogalactopyranosidase). The process is run in fed-batch operation modus and under optimized induction time, substrate feeding rate and temperature conditions.

## 5.2 Online state estimation of the bioprocess

The identification procedure of the system was based on the estimation of one of the key variables for the bioprocess, i.e. the specific growth rate, $\mu$. For that purpose, the specific growth rate was mathematically formulated via a feed forward neural network with 4 inputs (the modeled state variables biomass, substrate, protein and temperature) and a single hidden layer. The neural network model was complemented with a lag phase term. The neural model acts essentially as an adaptive system able to learn from the online measured process variables mentioned before. Concerning the specific protein production rate, available heuristic knowledge was incorporated in form of five rules-of-thumb implemented through a fuzzy artificial neural network system. The complete hybrid model of the system was formed by coupling the neural network and the neuro-fuzzy sub-models into the mass balance equation system of the bioreactor. Figure 5.1 depicts the structure of the hybrid model.

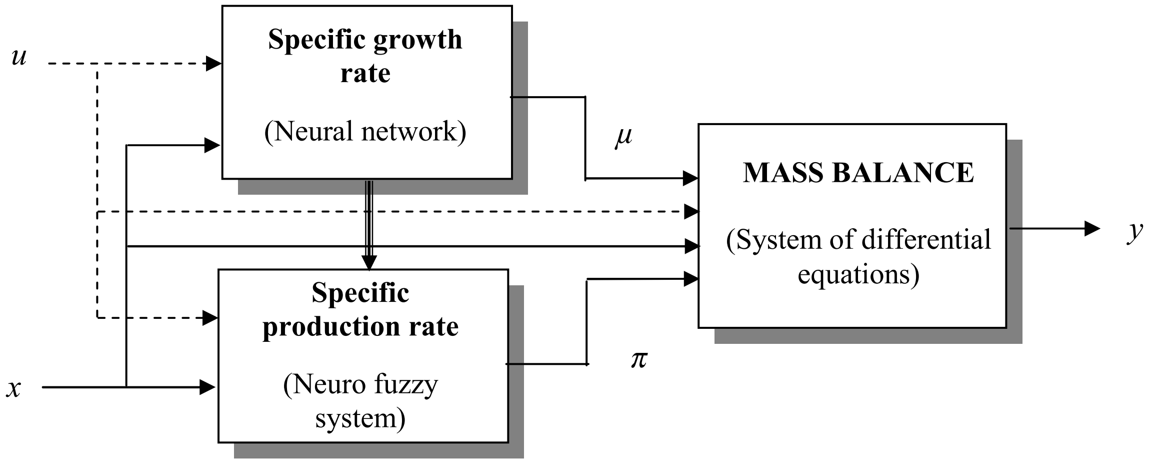

**Figure 5.1** Schematic representation of the hybrid model for the production of the viral capsid protein. The vector *u* represents the on-line measured or control variables (temperature, $CO_2$, etc.), the vector *x* represents the modeled state variables and the vector *y*, the output variables of the system after integration.

The training of the neural network component of the hybrid model occurred exclusively online. This was possible because data from the sampling/measuring procedure of three state variables (biomass, substrate and weight) were available at regular intervals during the course of the fermentation. This online training improved the gain of process knowledge at high learning rates, overcoming the slow convergence during the initial training stages (Kim and Lewis, 1998). Therefore, the methodology presented the further benefit of reducing the process development time and increasing the benefit/cost ratio.

A random search technique, the chemotaxis algorithm (Simutis and Lübbert, 1997), was used to identify the weights of the neural network model representing the specific growth rate. The fitting routine tunes the parameters each time an identification cycle is activated. This process minimizes the overall least squared error between modeled and measured state variables, i. e. the kinetics was inferred through the appropriate fit of the modeled state variables. The objective function ($J_{\mathrm{IDEN}}$) considered was:

$$J_{\mathrm{IDEN}}\,(y,\,Y) \equiv \sum_{j=1}^{M}\sum_{i=1}^{N}\left(\frac{y_i - Y_i}{\mathrm{KN}_j}\right)_j^2 \longrightarrow \text{min!} \qquad (5.1)$$

where the *y* are the online measured data for biomass and substrate concentration, and broth weight; *Y* are the simulated data for the aforementioned variables; *i* is referred to the number of *N* available data; *j* is referred to the state variables under consideration and KN are pseudo-norms for the *M* single state variables. The introduction of a pseudo-norm was done to constrict the influence of the single terms (weighting technique), because their real values lie in different numerical ranges. The values of these pseudo-norms are determined empirically.

Moreover, except for the case of the cultivation broth weight, the experimental measurements of the state variables were available at regular intervals but not continuously. That is, the frequency at which the biomass and glucose concentration measurements were made, was 50 times smaller than that of the true on-line measurements like the pH, temperature and the weight measurement itself. To overcome the problem of non synchronous data, an interpolation of the state variables with the same frequency of the online measurements was necessary. The interpolation procedure was very simple: the values of the off-line measured variables was kept constant until a new measurement was done. The new data was again kept constant until another measurement was available and so on. The result of this interpolation technique is a kind of step-like profile for the rendered pseudo-on-line biomass and glucose concentration variables (see Fig. 5.2).

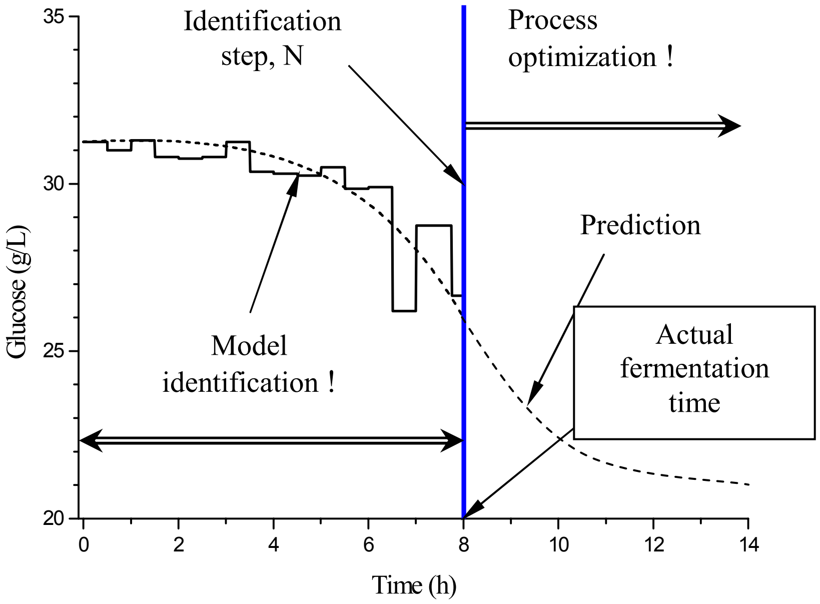

**Figure 5.2** Schematic representation of the on-line identification and optimization procedure.

Moreover, consistency between measurements and calculations was checked during the process. This was done through the on-line monitoring of the mass balance for carbon. The identification procedure is stopped after a predefined error criterion between measured and modeled state variables was achieved. Data of the measure of certainty ($r^2$) and the variance ($\sigma^2$) between modeled and measured variables for biomass and glucose concentration were used as dynamical fitting criterion and dynamical modeling performance. Based on these variables the training convergence of the model was supervised.

Finally, the specific protein production rate was described by a fuzzy artificial neural network. The specific protein production rate was set equal to zero before induction took place and represented by the neuro-fuzzy system after induction occurred. Off-line training was employed for the case of the fuzzy component, mainly due to the lack of on-line measurements for the protein activity. The specific protein production rate was correlated with the specific growth rate, the biomass and the specific protein activity by means of a set of 5 fuzzy rules presented in Table 5.1. The fuzzy expert system basically correlates the interest variables in form of conditional "*if ... then*" linguistic rules. The domains for these variables is expressed with "fuzzy" linguistic terms as "low", "medium" and "big". Both, rules and domains are determined empirically. This representation was chosen to integrate the available heuristic knowledge gained in previous experiments (Franco-Lara *et al.*, 2000) and simultaneously to improve the description of the kinetic data (equation A2.13). The resulting model allows the user to incorporate semi-quantitative information into a flexible mathematical representation. The complete model can be found in the Appendix A2.

1. IF ($\mu$ = LOW, $P/C_X$ = BIG, $C_X$ = BIG) THEN $\pi$ = LOW
2. IF ($\mu$ = LOW, $P/C_X$ = BIG, $C_X$ = MEDIUM) THEN $\pi$ = LOW
3. IF ($\mu$ = MEDIUM, $P/C_X$ = MEDIUM, $C_X$ = MEDIUM) THEN $\pi$ = MEDIUM
4. IF ($\mu$ = MEDIUM, $P/C_X$ = MEDIUM, $C_X$ =LOW) THEN $\pi$ = BIG
5. IF ($\mu$ = BIG, $P/C_X$ = LOW, $C_X$ = LOW) THEN $\pi$ = BIG

**Table 5.1** Set of fuzzy rules for the modeling the specific production rate, $\pi = f(\mu, C_X, P/C_X)$.

## 5.3 Online optimization of the bioprocess

Volk *et al.* (1998) showed that the induction time ($t_I$) and the temperature after the induction were the most important optimization variables to enhance the productivity of the VP1-DHFR system under batch operation modus. Furthermore, Franco-Lara *et al.* (2000) applied, under fed-batch operation, an evolutionary off-line optimization strategy using first a Monod-like. Later, the model was extended to a hybrid approach containing a feed forward neural network representation for the specific growth rate and a Pirt-like model for the specific protein production rate. The feeding rate profile and the induction point were also the optimized variables that maximized the total amount of native protein.

However, in the present case, during the fermentation process each model adjustment (identification) was followed by an optimization step. The performance index considered two objectives for the process optimization. The first aim was to maximize of the total amount of native protein at the end of the fermentation. The second objective was closed control of the substrate concentration in the post-inductive phase. This was included in the form of a hard constrain in the performance index (Patkar *et al.*, 1993). In the post-inductive phase the glucose concentration should not be greater than 1 g kg$^{-1}$. The optimization performance index ($J_{\mathrm{OPT}}$) to be maximized was given by:

$$J_{\mathrm{OPT}} = P(\mathrm{t_F}) \cdot W(\mathrm{t_F}) - f_{\mathrm{C}}(t) \tag{5.2}$$

where,

$$f_{\mathrm{C}}(t) = \mathrm{K_C} \cdot \sum \left(S(t) - 1.0\right)^2 \quad \text{if} \quad t > t_{\mathrm{I}} \text{ and } S(t) > 1.0 \tag{5.3}$$

$P(\mathrm{t_F})$ and $W(\mathrm{t_F})$ are the protein concentration and cultivation broth weight at the end of the fermentation time ($\mathrm{t_F}$). $f_{\mathrm{C}}(t)$ is a constrain function that penalizes glucose concentrations bigger than 1 g L$^{-1}$ after induction. $\mathrm{K_C}$ is a constant that can be set to control the influence of the constrain function on the optimization profit function.

Explorative experiments performed in the run-up to the process design in order to determine pH and temperature influences on the production of the desired protein showed that the temperature which is optimal for biomass growth is not the best for protein production after induction. Product development is preferred at lower temperatures. Hence, it is of advantage to decrease the temperature after induction. A temperature profile consisting of two phases was chosen for the cultivation. Before induction, the temperature was maintained constant at 37°C, i.e. at the temperature of the maximal specific growth rate ($\mu_{\max|37°C}$) for *E. coli*. After induction and until the end of the fermentation, the temperature descended linearly from 37°C to 25°C. This procedure, when applied in an iterative manner, led to the approximation of the optimal state of the system. Figure 5.2 exemplifies the general concept of the sequential online identification and optimization procedure.

As stated before, in the pre-inductive phase the model adjustment routine was let to run until a predefined model error criterion is fulfilled. From that point on, the induction time and the feeding profile were optimized. The induction time, $t_I$, was included as parameter in the mass balance unit of the hybrid model and marked the start of the production phase. The value for the induction time defined the moment at which the protein expression was optimally promoted. Because induction occurs only once in the whole process, its optimization could only be set up once too. Therefore, the induction time was calculated several times and monitored until no evident change in its value was reached and only then, set in function. The optimal induction time ($t_I$) was also determined with the chemotaxis algorithm. The parameter initial guess was defined as the half of the total fermentation time ($0.5 \bullet t_f$). The parameter could be changed in the interval, $t_I \in [0,t_f]$. Concerning the form of the temperature profile after induction, this is also determined once the optimal induction is estimated.

In contrast to the induction time, the optimization and usage of the feeding rate covered the whole time range of the fermentation. The feeding profile was represented by a simple feed-forward neural network. The activation function used in all layers was the exponential sigmoid function (Hilera and Martinez, 1995; Simutis *et al*., 1995). The structure of the neural network was fixed and consisted on three layers: the input layer with one input node, the hidden layer with 5 nodes and the output layer with just one node. The initial values for the neural network weights were chosen randomly and the chemotaxis algorithm was used again to identify them.

## 5.4 Results and discussions

In all experimental runs, the typical behavior of the specific growth rate, three different stages could be distinguished: an adaptation period at the start of the cultivation (lag phase), then the presence of a plateau corresponding to the maximum specific growth rate followed finally by an abrupt descend after induction took place. As can be seen in Figure 5.3, the hybrid approach modeled quite accurately the pattern of the specific growth rate during the course of the recombinant *E. coli* cultivation. The lag-phase lasted about 3 hours, followed by a 5 hours period of maximal growth rate. The post-induction period, with a duration of about 6 hours, was mainly characterized by low growth rates, caused by the descending temperature and the glucose substrate limitation.

Figure 5.4 presents the time trajectories for biomass and glucose concentrations. A proper description of the state variables was achieved by the hybrid model. Just like with specific growth rate, some phases were also distinguishable for the biomass and glucose concentrations. The first of these phases was the growth of the bacteria under batch operation in a period of about 4 hours. The temperature during this phase was maintained at 37°C, where the growth of *E. coli* achieved its highest value. At this point, the on-line feeding optimization strategy was activated. To avoid possible inhibition through high substrate concentrations, the feed maintained the glucose level under 15 g kg$^{-1}$. Besides, the feeding strategy supplied enough fresh substrate that prolonged unlimited growth until the induction point. This occurred 7.9 hours after the begin of the cultivation.

From the induction point on, glucose concentration was kept low to limit the bacterial growth. The substrate limitation was reflected in the biomass concentration behavior. After a 7.9 hours period of unlimited growth on glucose, biomass stopped to grow exponentially. The gradual deceleration of the specific growth rate in the post-inductive phase confirmed also this observation (see Figure 5.3). These phenomena are the combined results of both, the glucose limitation in the system and the temperature fall that occurred after induction with IPTG (see Figure 5.5). Even under these circumstances, high biomass concentrations in the order of 50 to 60 g kg$^{-1}$ were obtained.

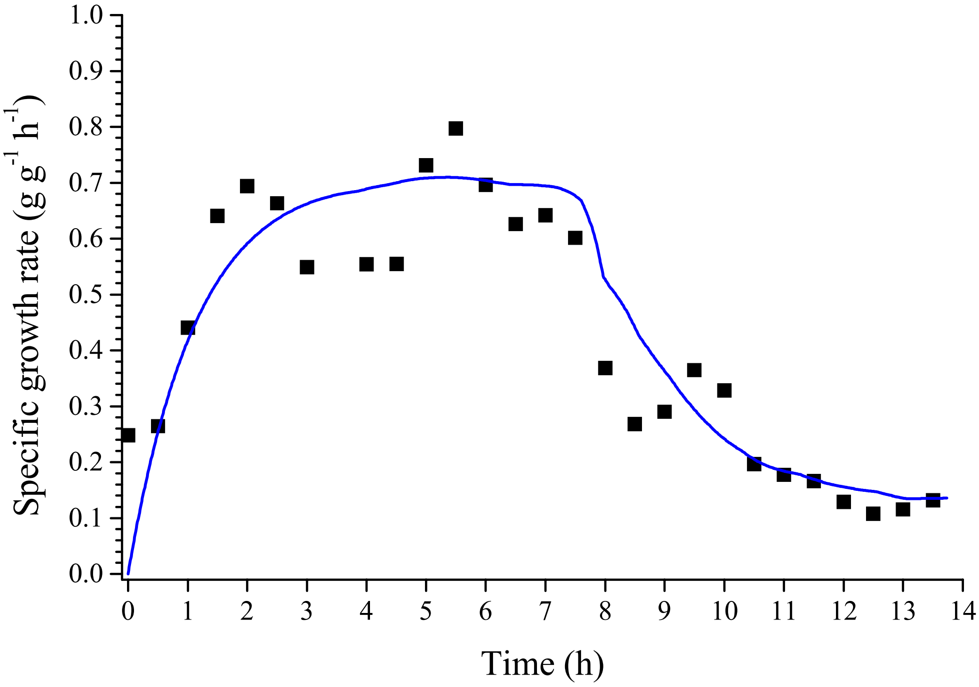

**Figure 5.3** Measured (■) and estimated (—) trajectory of the specific growth rate.

Figure 5.5 presents the on-line optimized trajectories of the variables feed rate and temperature. The time at which the induction took place is also indicated. The stepwise form of the feed rate function is a consequence of the iterative optimization process, that actualizes the feeding profile cyclically. After being activated, the feeding profile supplied enough substrate to maintain the aforementioned unlimited growth during the pre-inductive phase, but preventing the possible substrate inhibition by maintaining the glucose concentration under the level of 15 g kg$^{-1}$.

From the induction point on, the pattern of the feed profile changed drastically to eliminate the negative influence of the constraint function (Equation 5.3) on the optimization performance index (Equation 5.2). During the post-inductive phase, this could be only achieved by keeping the glucose concentration under 1.0 g kg$^{-1}$. The on-line optimization procedure fulfilled this task properly, as can be seen in Figure 5.4. The feed rate pattern was kept up around 600 g h$^{-1}$, slightly increasing only at the end of the fermentation.

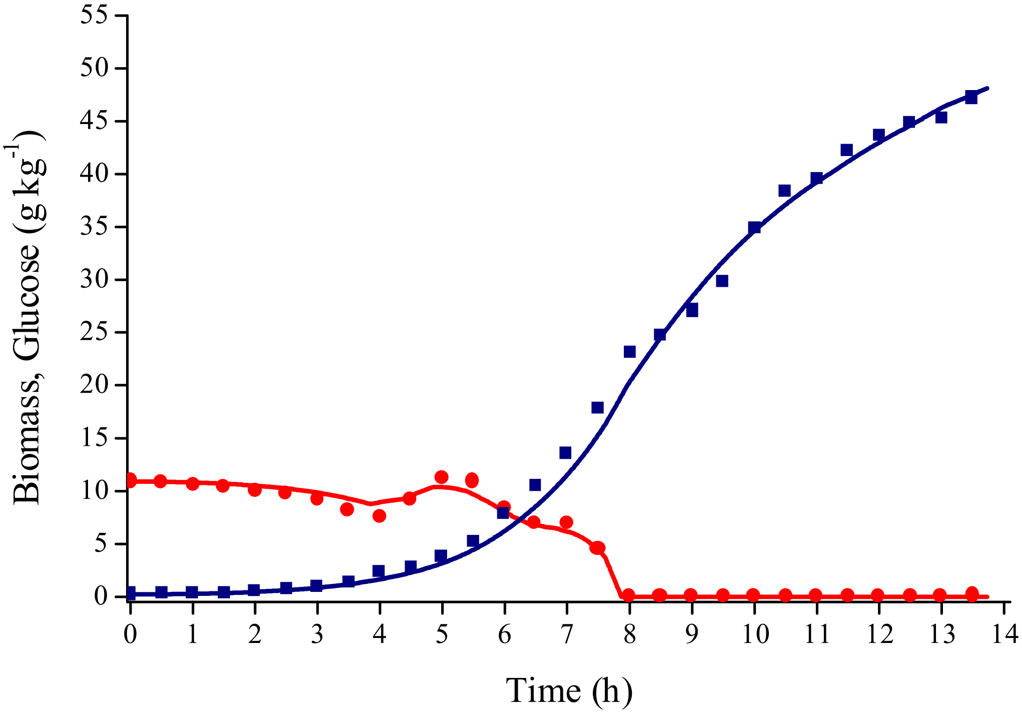

**Figure 5.4** On-line optimized trajectories for biomass and glucose concentrations. Symbols represent the measurements while lines account for the modeled variables.

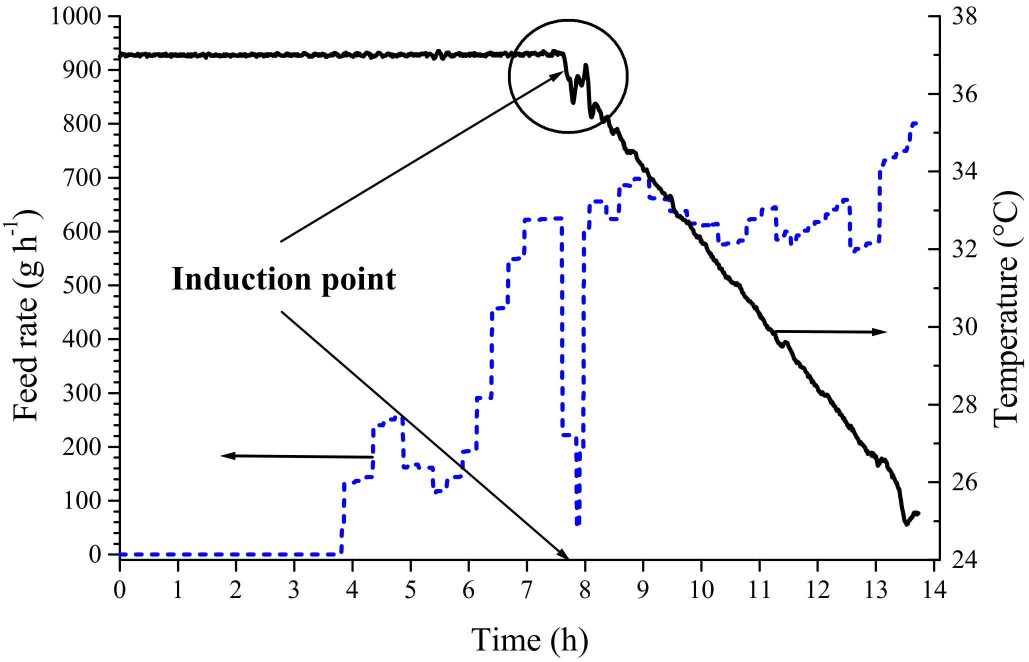

**Figure 5.5** On-line optimized trajectories for temperature (—) and feed rate (—). At the optimized induction point, temperature descends linearly until the end of the fermentation.

Consistency between measurements and calculations was also checked during the whole cultivation through the on-line monitoring of the mass balance for carbon. As inlet to the system was considered all carbon introduced to the system. It considered the initial substrate and biomass concentration in the bioreactor as well as the carbon from the feed. The balance was completed taking into account the outlet in form of evolved $CO_2$, samples and carbon converted to glucose and biomass in the bioreactor. Figure 5.6 presents the trajectories of the aforementioned components together with the *index of recovery*, i.e. the ratio between introduced and converted carbon and its transient behavior. The index of recovery was given by:

$$\text{Recovery (\%)} = (\text{Carbon}_{\text{IN}}/\text{Carbon}_{\text{OUT}}) * 100 \qquad (5.4)$$

The index of recovery is an indicator for the mismatch in carbon balance. Values lying over 95 %, considering the whole cultivation, indicate an acceptable mass balance of the system.
Furthermore, the reliability of the proposed hybrid model was one of the most important subjects to be tested during the identification procedure. As described before, real-time estimation of the specific growth rate was based on a pseudo-on-line measurement of biomass and glucose concentrations. For this reason, accuracy and performance of the hybrid model to describe these state variables was also continuously monitored. The model's reliability was tested through the monitoring of following statistical variables:

1. The measure of certainty ($r^2$) of measured to modeled state variables
2. The empirical standard deviation ($\sigma_i$) between model and measurements
3. The dynamical convergence of $r^2$ or $\sigma_i$

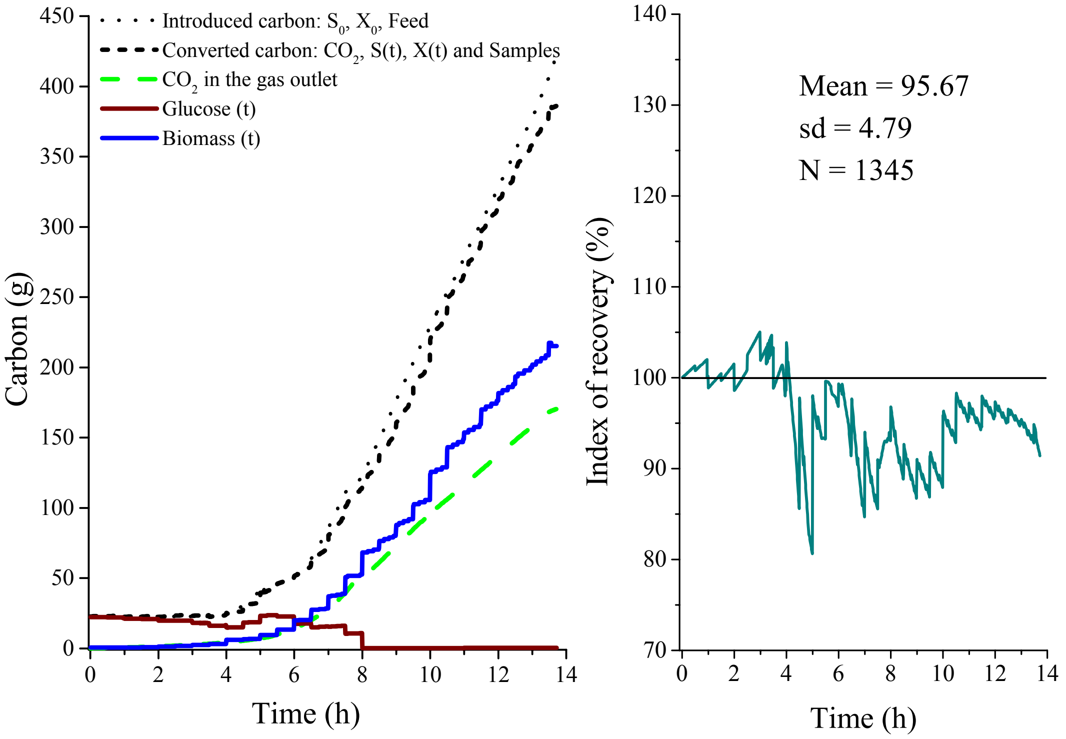

**Figure 5.6** Mass balance trajectories for carbon and its corresponding index of recovery.

In the case of the measure of certainty ($r^2$), it was used as a criterion of the linear dependence between the modeled and measured state variables. If these variables were consistent, *i. e.* if the model reflects the real behavior of the cultivation, then they could be described with a linear model between each other. The model should have also a normal standard deviation (Wolf, 1994; Weihs and Jessenberger, 1999). A more detailed description of the measure of certainty can be found in the Appendix A3.

The measure of certainty and its time derivative, the *dynamical fitting coefficient*, can also be used for statistical control of the modeling quality and for supervision of the process identification and optimization.

The idea behind is to implement some features of these variables as optimal switches for the identification process. For example, if a predefined fitting criteria is fulfilled during long periods, then there is no need of frequent training. The identification process can be enhanced by reducing the number of identification cycles or by reducing the required identification time. Concerning the reduction of the identification time, in some cases it is advisable to change from a high reliable, but time-consuming, to a more simple and faster identification procedure. For that purpose, some of the techniques coming from the statistical control of process can be set up.

Figure 5.7 depicts two examples of this technique. The graphics present the measurement of certainty and the dynamical fitting coefficient as function of the training time. Furthermore, some help indicators are included: the warning and control lines. They are used for the supervision of the process' variability, that is, to keep under surveillance the natural oscillations of the process (Weihs and Jessenberger, 1999).

For the process identification monitoring, the measurement of certainty was a more sensitive criterion than the dynamical fitting coefficient. In Figure 5.7, the measurement of certainty and the dynamical fitting coefficient are presented with a single warning and a single control limit. Due to their asymmetry, only the worsening of the process is controlled and supervised. That is, only changes diminishing the value of the measurement of certainty or changes towards negative values in the dynamical fitting coefficient produce a warning or a control call. The warning criteria to be attended are the following: if the warning limit of the measurement of certainty for the process is exceeded, then attention must be paid to the identification procedure; if the control limit is exceeded, the identification must be reinforced. This can be done either by augmenting the number of training cycles or, in some cases, varying the mutation factor of the method (e.g. chemotaxis). On the other hand, if the warning limit of the dynamical fitting coefficient for the process is exceeded, again, the identification must be reinforced, augmenting the number of training or varying the mutation factor of the method; but if the control limit is exceeded, the identification method itself must be changed for another alternate method. The main advantage of this switching is the acquirement of a fully automatic system for monitoring and controlling the bioprocess model's accuracy.

As criterion of model performance, values of the measure of certainty higher than 0.95 were considered as "very good". Values higher than 0.9, but smaller than 0.95 were only "good". Table A3.1 (see Appendix A3) presents a more detailed description of this goodness criterion.

In the present example, the measure of certainty went beyond the warning limit twice and the control limit only once.

Concerning the dynamical fitting coefficient, it went only once beyond its corresponding warning limit, exactly at the moment when the control limit for the measurement of certainty was exceeded. The correcting action taken was to increase the number of identification cycles of the identification procedure. This action enhanced the process description performance evidently and no further similar operations were necessary.

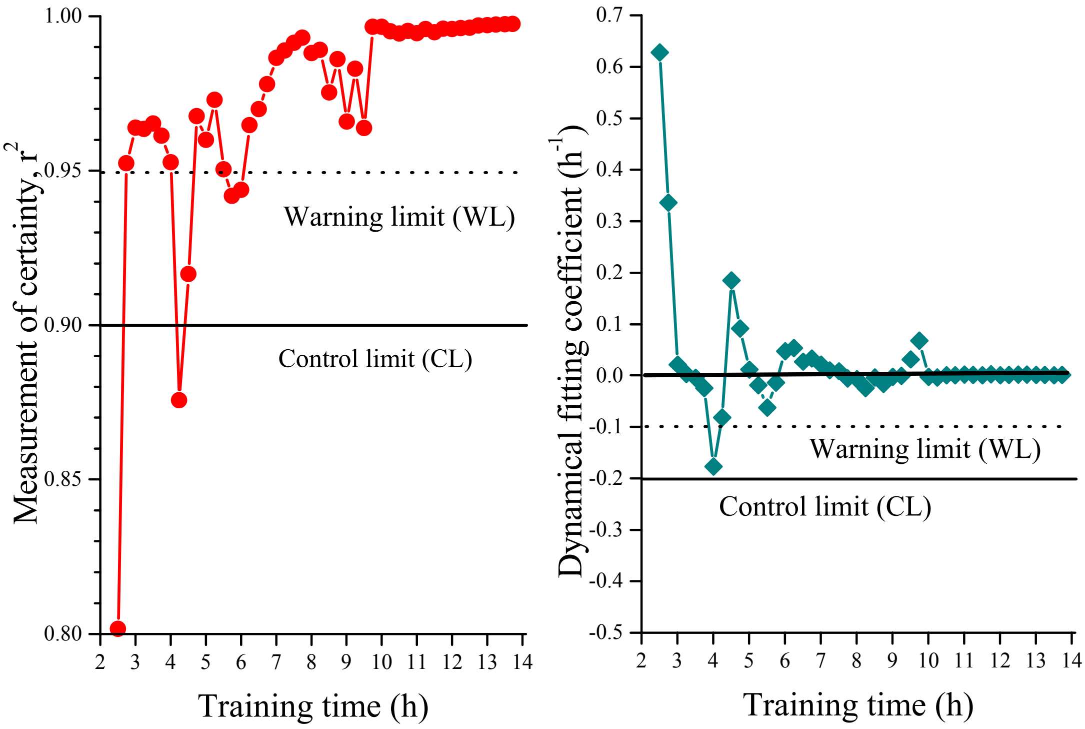

**Figure 5.7** Time trajectories for the measurement of certainty and its time derivative, the dynamical fitting coefficient and their corresponding warning and control limits.

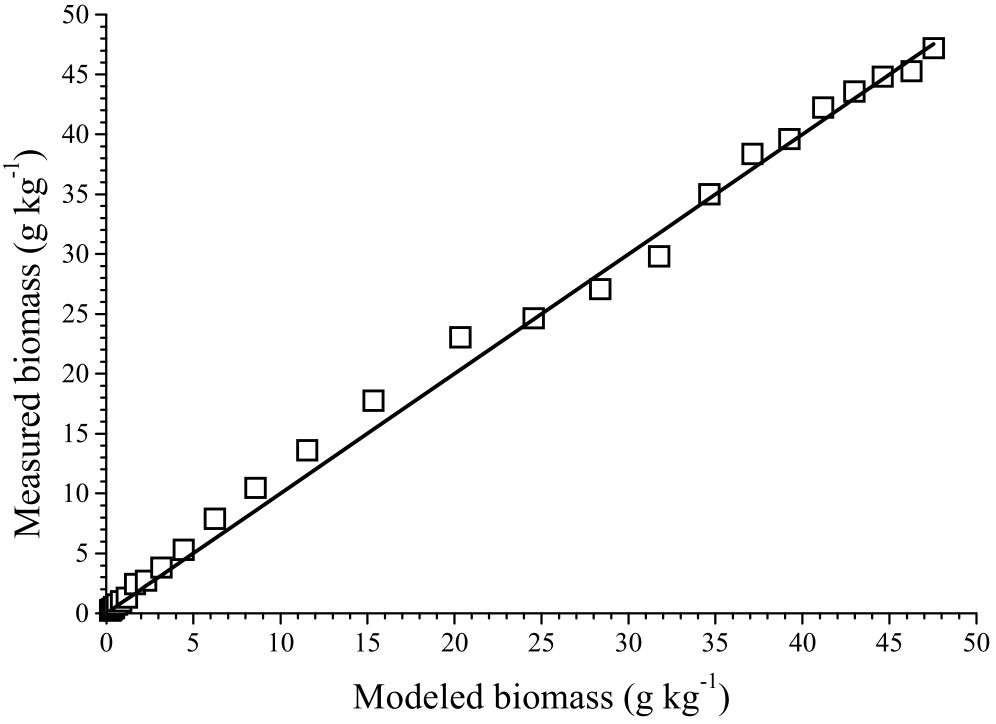

**Figure 5.8** Fitness between modeled and measured biomass concentration .

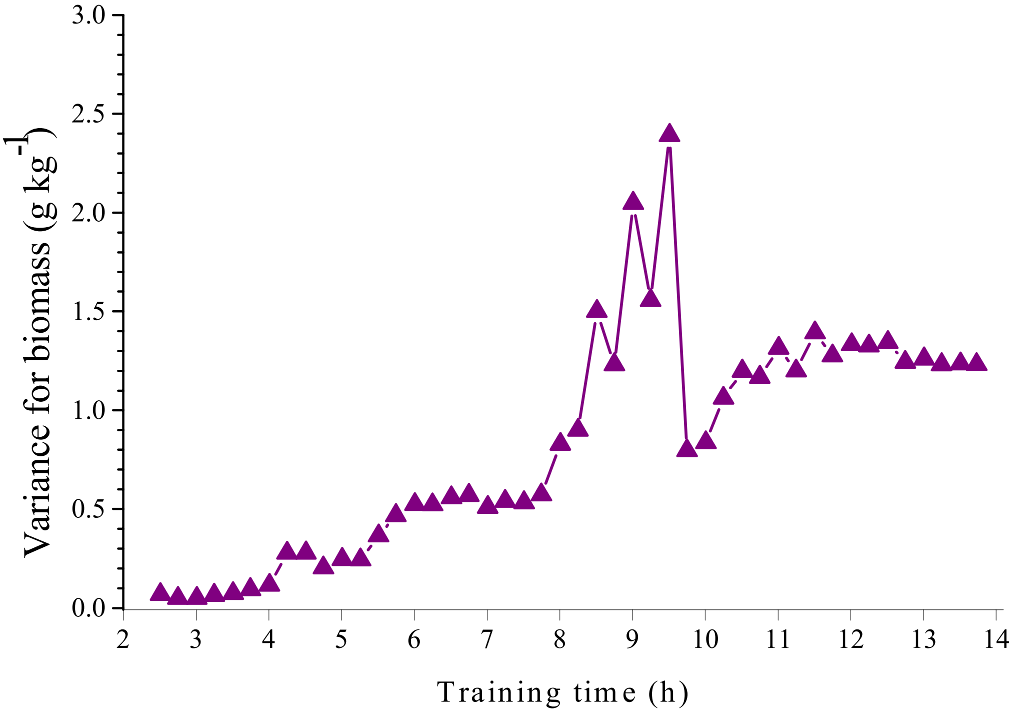

**Figure 5.9** Time dependent trajectory of the model to measurement variance for the biomass concentration as function of the training time.

Figure 5.8 shows the agreement between the modeled and measured biomass concentration. Its time dependent model to measurement variance is depicted in Figure 5.9. It can be stated that the hybrid model describes with high accuracy the process variable.

Concerning the variance for the biomass concentration, except for the period between 8 and 10 h (after induction), it increases in a constant manner, stabilizing at the end of the cultivation. The sudden increase was explained by the change in the specific growth rate estimation after induction: the hybrid model had to adapt itself to this new situation at expenses of its accuracy. The phenomenon could also be confirmed in the change on the pattern of the measure of certainty for the same period of time (see Fig. 5.7).

On the other hand, the specific protein production rate was formulated by a fuzzy artificial neural network with a set of five heuristic rules-of-thumb. After training the neuro-fuzzy system with raw data of the specific production rate, a non-normal distribution of the model to measurement residuals was obtained (data not shown). This indicated the presence of outliers or systematic error trends and therefore, an inter-quartile range (IQR) analysis was applied to the experimental data of the specific product production rate.

The IQR analysis is used to describe the distribution characteristics of a given population and to identify extreme features belonging to this population. An outlier in a population may be the result of a data entry error, a poor measurement or a change in the system that generated the data.

Formally, an outlier is any sample that exceeds more than 1.5 times the populations' inter-quartile range away from the top or the bottom of a notched Box-plot (Weihs and Jessenberger, 1999; The MathWorks, 1999). The notches in a Box-plot are graphic confidence intervals about the median of a given population. A more detailed description of the IQR and the notched Box-plot analysis can be found in Appendix A3.

After the statistical analysis of the quartile, the residuals' outliers of the specific protein production rate were eliminated and the neuro-fuzzy model was re-trained again, in order to confront it with the treated data. Figure 5.10 compares the measured and modeled values for the specific protein production rate from different experiments. The model delivered apparently a more compact distribution than that of their corresponding measurements.

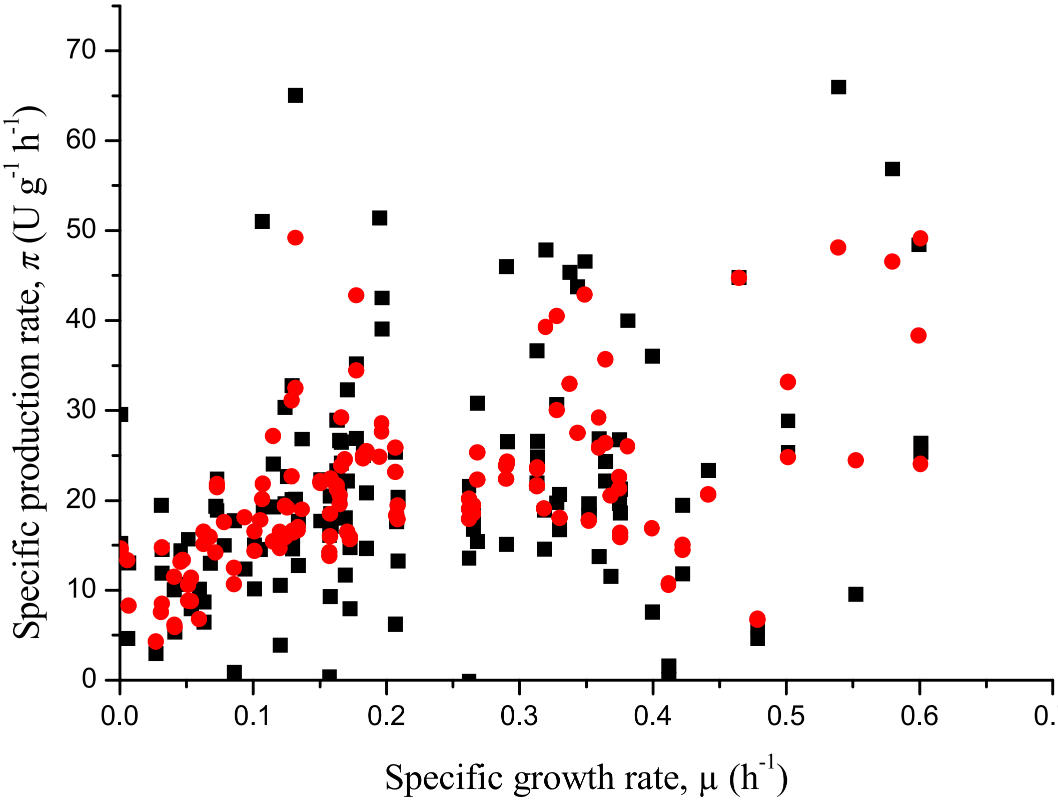

**Figure 5.10** Measured (■) and modeled (●) specific protein production rates.

The resulting re-trained model presented, as expected after the IRQ analysis, a normal distribution in the residuals of the estimations to measurements, as can be seen in Figure 5.11.A. Furthermore, Figure 5.11.B displays statistical data and a Gaussian fit for the resulting residuals distribution.

Another facet that can be highlighted in the neuro-fuzzy approach was its apparent ability to filter noisy data. As stated before, the model estimations of the specific protein production rate presented a more compact distribution than that of their corresponding measurements.

Figure 5.12 delineates actually the results obtained in the corresponding validation experiment. As performed with the described off-line optimization procedure, an independent experimental validation is also required for the online optimization. This was done using the re-trained hybrid model obtained after the IRQ analysis, which was considered to be suited for optimization purposes.

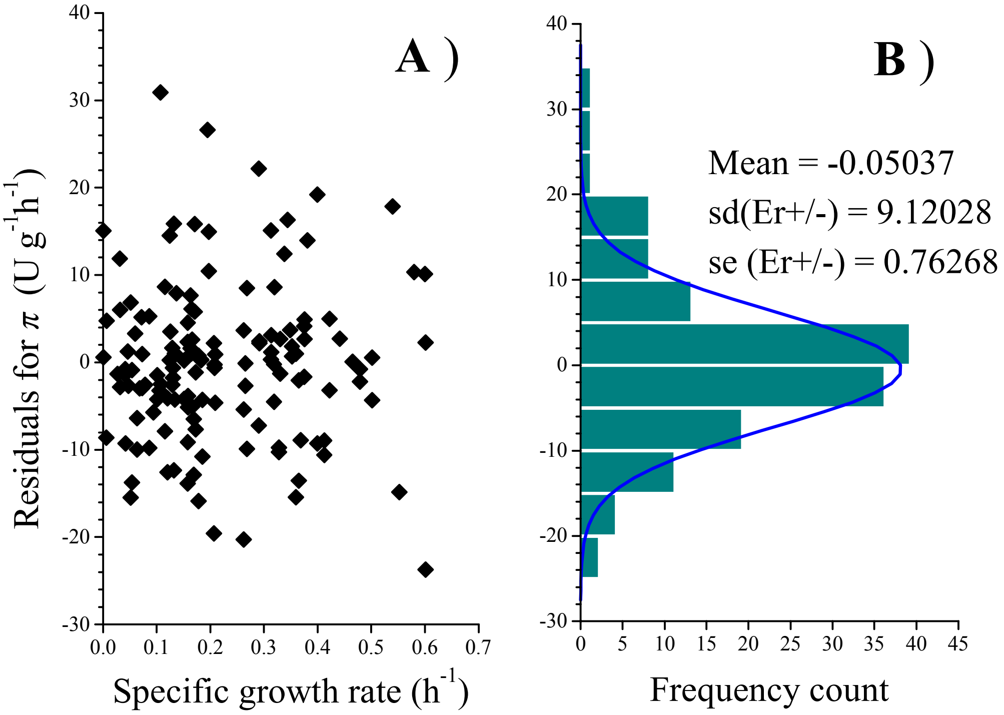

**Figure 5.11** A) Residuals of model to measurements for the specific production rate. B) Distribution of residuals model to measurements, statistical data and gaussian fit for the residuals.

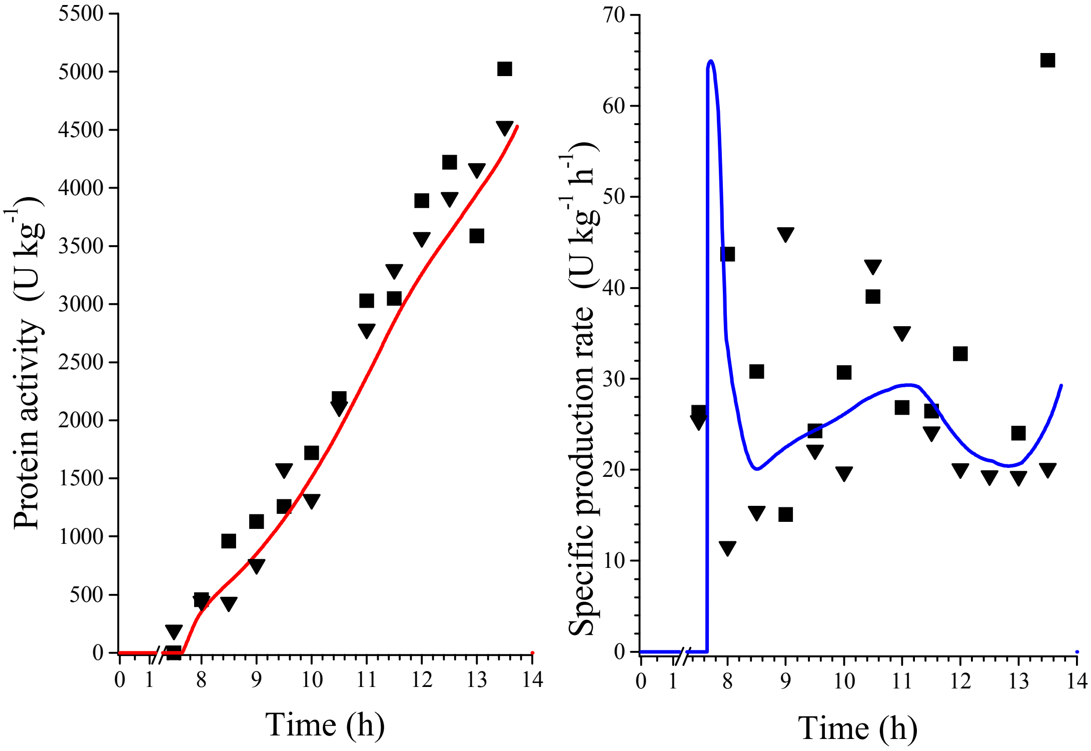

**Figure 5.12** Time trajectories for protein activity and its corresponding specific protein production rate. Symbols represent repeated measurements, while lines are the model estimates.

Figure 5.12 depicts the trajectories for the product activity and its respective specific protein production rate. The main features and tendencies in the estimation of the product activity were well described by the hybrid model. Moreover, a good agreement seemed to prevail in the description of the trajectory of the specific protein production rate with the neuro-fuzzy approach.

Only a few on-line optimization runs were needed to improve the results obtained by the off-line procedure (see Chapter 4). It is worth noting, that these last two off-line optimizations made use of a Pirt-like formulation for the specific protein production rate, in contrast to the actual neuro-fuzzy formulation. Figure 5.13 shows the profit function development obtained with two experiments using the on-line technique compared to the referenced off-line method (Franco-Lara *et al.*, 2000; Franco-Lara *et al.*, 2001). An exploratory off-line optimized batch run following the technique of Volk *et al.* (1998) is compared too. This batch fermentation was chosen as reference, considering that both processes were run under the physical optimization constrains (same maximum volume fixed for the batch process, $X(t = 0)$, etc.) except for the feeding rate.

The increase in the performance index obtained with the on-line procedure is about 2 to 3.5 times that of the best off-line optimized experiment. This can be explained not only based on the more reliable specific protein production rate estimation, performed by the neuro-fuzzy kinetic model, but also based on the implicit closer control of the specific growth rate after induction. Compared to the off-line optimization, the on-line model adjustment is capable to predict the state variables in a more reliable form. The availability of an actualized and trustworthy model also allowed an enhancement in the performance index of the bioprocess using the described on-line optimization procedure.

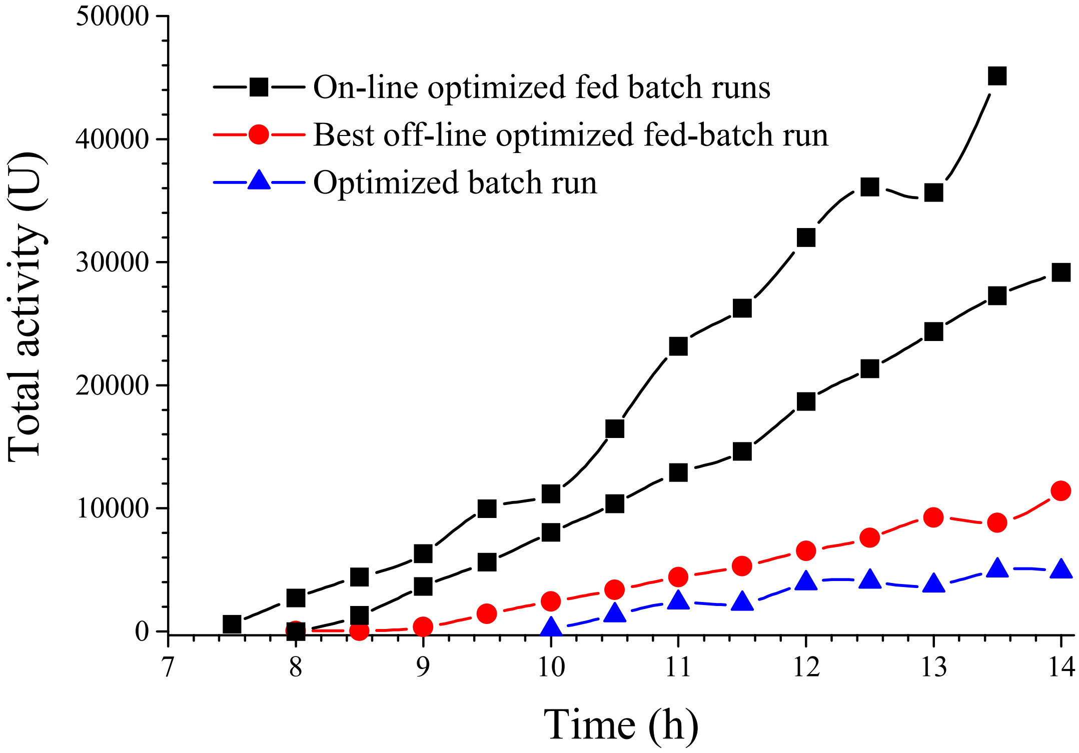

**Figure 5.13** Time trajectories for protein activity and its corresponding specific protein production rate. Symbols represent the measurements, while lines are the model estimates.

## 5.5 Conclusions

A hybrid model for the production of the construct VP1-DHFR has been developed. This consists in set of differential equations describing the mass balance for the bio-process complemented with artificial neural network and neuro-fuzzy components that describe the corresponding kinetics. A feed forward neural network described the specific growth rate, while a neuro-fuzzy component characterized the specific protein production rate formulated via a set of five fuzzy rules. The main advantage of this type of neuro-fuzzy representation is that the results can be made more transparent to the process engineer: it can integrate the available heuristic knowledge gained in previous experiments and simultaneously improved the description of the kinetic data. This bioprocess representation was developed as an on-line application. The approach was able to cope with tasks such as process modeling, supervision and optimization. Main characteristics of this method is the robustness of the system to successfully filter noisy signals due to poor measurements. Another remarkable aspect of this approach was its high speed learning feature, which improves the gain of process knowledge through the on-line training of the neural network. The neural component represented the key variable of the process: the specific growth rate. The on-line training resulted in further benefits as the reduction of the process design and development time. Mainly because this methodology is a data-driven type, the investment in long term theoretical investigations of the metabolism of the bioprocess can be partially avoided.

## 5.6 Nomenclature

| | | |
|---|---|---|
| $C_X$ | Biomass concentration | [g kg$^{-1}$] |
| $f_C$ | Constraint function | [Units of activity] |
| $J_{IDEN}$ | Identification criteria | |
| $J_{OPT}$ | Optimization goal | [Units of activity] |
| $K_{C1,2}$ | Constants for the constrain function | [U h kg g$^{-1}$] |
| $P$ | Product activity | [U kg$^{-1}$] |
| $r$ | Vector of kinetic variables | |
| $r^2$ | Measure of certainty | |
| $S$ | Substrate concentration | [g kg$^{-1}$] |
| $S_F$ | Substrate concentration in fresh feeding | [g kg$^{-1}$] |
| $t$ | Time | [h] |
| $t_F$ | Time at the end of fermentation | [h] |
| $t_I$ | Induction time | [h] |
| $T$ | Temperature | [°C] |
| $u$ | Vector of input/control variables | |
| $W$ | Liquid reaction weight | [kg] |
| $x$ | Vector of state variables | |
| $y$ | Vector of modeled outputs variables | |
| $Y$ | Vector of measured outputs variables | |

**Greek symbols**

| | | |
|---|---|---|
| $\mu$ | Specific growth rate | [$h^{-1}$] |
| $\pi$ | Specific production rate | [U $g^{-1}$ $h^{-1}$] |
| $\sigma_i$ | Empirical standard deviation | |

## 5.7 References

1. Franco-Lara, E.; Volk, N.; Lübbert, A.: Optimierung der Produktion rekombinanter Proteine mit hybriden Modellen (In German). Chemie Ingenieur Technik 72, 1+2: 110-114 (2000).
2. Franco-Lara, E.; Galvanauskas, V.; Volk, N.; Lübbert, A.: Model-based Optimization of the Cultivation Process for Recombinant Virus Capsid Proteins in *E. coli*. (Accepted for Publication) Proceedings of the 8th International Conference on Computer Applications in Biotechnology, Québec, Canada (2001).
3. Hilera, J. R.; Martínez, V. J.: Redes neuronales artificiales. Fundamentos, modelos y aplicaciones (In Spanish). Addison-Wesley Iberoamericana (1995).
4. Kim, Y. H.; Lewis, F. L.: High-level feedback control with neural networks, World Scientific Publishing Co. Pte. Ltd. (1998).
5. Lübbert, A.; Simutis, R.: Adequate use of measuring data in bioprocess modelling and control. Trends in Biotechnology 12: 304-311 (1994).
6. Nielsen, J.; Villadsen, J.: Bioreaction Engineering Principles. Plenum Press, New York (1994).
7. Oliveira, R.; Simutis, R.; Lübbert, A.: HYBNET, a new tool for advanced process modelling. Proceedings of the 1st European Symposium on Biochemical Engineering Science., Dublin, Ireland (1996).
8. Patkar, A.; Seo, J. H.; Lim, H. C.: Modeling and Optimization of Cloned Invertase Expression in *Saccharomyces cerevisiae*. Biotech. & Bioeng. 41, 1066-1074 (1993).
9. Simutis, R.; Lübbert, A.: A comparative study on random search algorithms for biotechnical process optimization.” Journal of Biotech. 52: 245-256 (1997).
10. Simutis, R.; Havlik, I.; Schneider, F.; Dors, M.; Lübbert, A.: Artificial Neural Networks of Improved Reliability for Industrial Process Supervision. IFAC Computer Applications in Biotechnology, 59-65 (1995).
11. The MathWorks, Inc.: Statistics Toolbox User’s Guide. The MathWorks, Inc. (1999).
12. Volk, N., T. Hertel, A. Lübbert: Modellgestütze Optimierung der Produktion des Virus-Hüllproteins VP1-DHFR mit *E. coli* BL21 (In German). *Chem. Technik* 4, 192-197 (1998).
13. Weihs, C.; Jessenberg, J.: Statistische Methoden zur Qualitätssicherung und –optimierung in der Industrie (In German). Wiley-VCH Verlag GmbH (1999).
14. Wolf, K.-H.: Aufgaben zur Bioreaktionstechnik (In German). Springer-Verlag Berlin Heidelberg (1994).

# Chapter 6

# Online monitoring of performance indexes employing a neural network-based soft-sensor

## ABSTRACT

A soft-sensor approach for monitoring of recombinant protein production systems is presented. This technique is based on a hybrid model that consist of a system of differential equations describing the mass balances of the system and a feed forward neural network component. The neural network accounts for the specific growth rate, which is exclusively inferred from online measured and estimated variables. The hybrid model is used to monitor and estimate the state variables and the performance indexes of the process. The approach is tested on two different recombinant microbial systems: the bacteria *Escherichia coli* and the yeast *Kluyveromyces lactis*. Both fermentations are run under optimal feeding strategies. Additionally, the *Escherichia coli* cultivation was carried out under artificially induced temperature shifts. Monitoring the changes of process performance index and penalty functions related to these environmental alterations allows the development of high performance control and quality strategies for the pharmaceutical industry.

## 6.1 Introduction

The requirement of high performance control constitutes a great challenge in the production of recombinant products in today's pharmaceutical industry. This is done with strict documented norms and quality strategies. However, the availability of only few sensors capable of providing reliable, direct monitoring of bioprocess variables (biomass, substrate and product concentration, etc.) has always be another challenge problem. The present work addresses this problematic through the design of a neural network based soft sensor, capable of substituting the lack of instrumental sensors just mentioned before. This approach is essentially able to characterize the kinetics of the bioprocess making use of process measurements available online like the optical density of the broth and the temperature. Variables like the biomass specific oxygen consumption rate and the biomass specific carbon dioxide evolution rate are defined here and utilized in the soft sensor. The methodology is applied in two fermentations with recombinant microorganisms: the optimized production of the construct VP1-DHFR in *Escherichia coli* and the optimized production of the complex formed by the protein GAL80 in *Kluyveromyces lactis*. Both microorganisms grew aerobically on glucose in fed-batch operation modus.

## 6.2 Recombinant protein production systems

### 6.2.1 *Escherichia coli* cultivation

The polyomavirus-like particles are a well-characterized system, established as a model for gene transfer studies. *In vitro* studies demonstrated that purified VP1 can form virus-like particles consisting of 72 pentamers (Salunke *et al*. 1986). This feature makes the protein attractive for *in vitro* packaging of DNA and gene transfer experiments. The polyomavirus-like particles VP1 can be produced in recombinant *Escherichia coli* bacteria.

In the present case, *Escherichia coli* BL21 was utilized as host system to produce a genetic construct. The construct was made of the viral capsid protein of the murine polyomavirus fused with the enzyme dihydrofolate reductase (DHFR, EC 1.5.1.3). In addition, this construct contained an ampicillin resistant plasmid pBR322, necessary for the expression of the viral capsid protein under control of a tac-promotor. The over-expression of the recombinant protein was induced by using 1.5 mmol of IPTG (isopropyl β-D-thiogalactopyranosidase). The native recombinant protein complex (VP1+DHFR) possesses a measurable enzymatic activity, which was proportional to the product concentration.

The production process of the recombinant viral capsid protein VP1 with *E. coli* consisted of two phases: a pre-inductive stage, where the main objective was to maximize the biomass production and a post-inductive phase, where the over-expression of the native recombinant protein was induced by adding IPTG.

Under fed batch conditions the phase before induction is characterized by the growth of the biomass at 37°C. Fresh substrate is added to the system following an optimized profile calculated for the feed rate. At an optimized induction point, IPTG is introduced to the system. This chemical induces the expression of the recombinant protein. Simultaneously to the induction, the temperature begins to decrease along a linear profile, from 37°C to 25°C. The recombinant product is formed as a native intracellular protein complex. The concrete performance index for the process is amount of native protein at a given predefined time ($t_F$).

Optimization goal ($J_{\mathrm{OPT}}$) was thus to produce as much of the native protein as possible within a predefined cultivation time, $\mathrm{t_f}$:

$$J_{\mathrm{OPT}} = P(\mathrm{t_f})\ W(\mathrm{t_f}) \rightarrow \max \tag{6.1}$$

where $P(\mathrm{t_f})$ is the product activity per weight and $W(\mathrm{t_f})$ the liquid reaction weight. A more detailed description of the kinetics of the process can be found in the Chapter 4, Section 4.3 and in the Chapter 5, Section 5.2.

However, to examine the performance of the employed online optimization under disturbances, the present process was run under artificially introduced temperature perturbation in the pre-inductive phase. The influence of these disturbances on the estimated process performance index was to be evaluated and monitored during the whole course of the fermentation.

## 6.2.2 *Kluyveromyces lactis* cultivation

The yeast *Kluyveromyces lactis* RUL 1888 D80ZR-pEAHG80 was grown aerobically on glucose as only carbon source. The product, the recombinant protein GAL80 (Zenke *et al.*, 1999) is merged with a HIS-TAG and was constitutively expressed by an ADG promoter. The fermentation was run under fed-batch modus where an optimized feeding profile was proposed for controlling the specific growth rate. A mathematical model for the system was adjusted cyclically and, based on it, an optimization was performed (Volk *et al.*, 2001). With this model, the best results were achieved by controlling the specific growth rate in order to decrease the intracellular metabolic overflow during the fermentation. The fermentation process was described by a bottleneck kinetics, similar to the yeast model proposed by Sonnleitner *et al.* (1986).

In the case of the product concentration $P(t)$, a linear correlation was established with the biomass concentration (Volk *et al.*, 2001):

$$P\ (t) = \alpha\ C_{\mathrm{X}}\ (t) \tag{6.2}$$

where $\alpha$ was a parameter to be identified.

The optimization task was to maximize the amount of biomass ($C_{\mathrm{X}}$) per fermentation time ($\mathrm{t_F}$) under fed-batch culture conditions. The optimization profit function $J'_{\mathrm{OPT}}$, can be described by:

$$J'_{\mathrm{OPT}} = C_{\mathrm{X}}/\mathrm{t_F} - f'_{\mathrm{C}}\ (t) = P/(\alpha\ \mathrm{t_F}) - f'_{\mathrm{C}}\ (t) \qquad \rightarrow\ \max \tag{6.3}$$

Where $f'_{\mathrm{C}}$ is the general penalization function:

$$f'_{\mathrm{C}}\ (t) = f_{\mathrm{C1}}\ (t) + f_{\mathrm{C2}}\ (t) + f_{\mathrm{C3}}\ (t) \tag{6.4}$$

$f_{\mathrm{C1}}\ (t)$ is the constraint function that penalizes deviations from the specific growth set point, $\mu_{\mathrm{SET}}\ (t)$ at which the fermentation was controlled. $f_{\mathrm{C2}}\ (t)$ and $f_{\mathrm{C3}}\ (t)$ are functions that penalize, respectively, any situation in which the physical limits for the feeding function or for the OUR may be exceeded. The single constraint functions $f_{\mathrm{C}1,2,3}$ are given by,

$$f_{\mathrm{C1}}\ (t) = \mathrm{K'_{C1}} \cdot \sum \left(\mu_{\mathrm{SET}}\ (t) - 0.13\right)^2 \quad \text{if} \quad \mu_{\mathrm{SET}} \neq 0.13\ \mathrm{h^{-1}} \tag{6.5}$$

$$f_{\mathrm{C2}}\,(t) = \mathrm{K}'_{\mathrm{C2}} \cdot \sum \left(F\,(t) - 0.8\right)^2 \qquad \text{if} \qquad F\,(t) > 0.8 \text{ kg h}^{-1} \qquad (6.6)$$

$$f_{\mathrm{C3}}\,(t) = \mathrm{K}'_{\mathrm{C3}} \cdot \sum \left(OUR\,(t) - 12.0\right)^2 \qquad \text{if} \qquad OUR\,(t) > 12.0 \text{ g kg}^{-1}\text{ h}^{-1} \qquad (6.7)$$

The $\mathrm{K}'_{\mathrm{C1,2,3}}$ are constants that were determined empirically to manipulate the influence of the single penalty function on the optimization profit function.

In this particular case, the feeding function was calculated from:

$$F\,(t) = \frac{\mu_{\mathrm{SET}}}{\mathrm{Y}_{\mathrm{X/S}}} \cdot C_{\mathrm{X}} \cdot \frac{W}{(\mathrm{S}_{\mathrm{F}} - S)} \qquad (6.8)$$

where the state variables biomass ($C_{\mathrm{X}}$) and glucose ($S$) concentration, and culture broth weight ($W$) were calculated online using the method proposed by Claes and Van Impe (1999). $\mathrm{Y}_{\mathrm{X/S}}$ and $\mathrm{S}_{\mathrm{F}}$ represent the yield coefficient and the glucose concentration in the fresh medium stream, respectively.

## 6.3 Online monitoring of the bioprocess performance indexes

The core of the monitoring procedure was based on the on-line estimation of the key variable for the bioprocess, the specific growth rate. For that purpose, the specific growth rate is formulated via a feed forward neural network with a single hidden layer and a single hidden node. The neural network model for the specific growth rate was complemented with a lag time term.

For both microbial systems, the Oxygen Uptake Rate (OUR) and the Carbon dioxide Production Rate (CPR) were estimated online. When these variables are expressed as a function of the total biomass present in the bioreactor, two new useful correlations can be defined: the biomass specific oxygen consumption rate, $q_XO_2$ and the biomass specific carbon dioxide evolution rate, $q_XCO_2$.

The biomass specific oxygen uptake rate, $q_XO_2$ is given by:

$$q_XO_2 = \frac{OUR}{C_{\mathrm{X}}(t)\;W(t)} \qquad (6.9)$$

where $OUR$ is the oxygen uptake rate, $C_{\mathrm{X}}(t)$ is the online estimated biomass concentration and $W(t)$ is the measured broth weight.

For the case of the biomass specific carbon dioxide production rate, $q_XCO_2$ it is estimated from:

$$q_XCO_2 = \frac{CPR}{C_{\mathrm{X}}(t)\;W(t)} \qquad (6.10)$$

where $CPR$ is the carbon dioxide production rate.

Four on-line available variables were considered as input of the neural network model for the specific growth rate. The optical density (OD), the biomass specific oxygen consumption rate ($q_XO_2$) and the biomass specific carbon dioxide evolution rate ($q_XCO_2$) were common for the estimation of the specific growth rate of both microorganisms. However, the fourth variable was different for each of the microbial system considered. In the case of the *Escherichia coli* cultivation the fourth variable was the temperature. For the *Kluyveromyces lactis* cultivation the forth variable was the $pO_2$.

### 6.3.1 Online training of the soft sensor

As with all online learning systems, the motivation and benefits for applying this procedure was the reduction in the process developing time and costs. In fact, as pointed out by Kim and Lewis (1998), the training of any neural network system should occur ideally online. This online training would result in high speed learning rates. It can use any initial set of parameters and assures good generalization properties. As modeling fitting criterion, the minimization of the overall least squared error between modeled and measured state variables defined in Equation 5.1 was taken into account.

However, the off-line measured data of the biomass and substrate concentration were available at regular intervals but not continuously. That is, the frequency at which the biomass and glucose concentration measurements were made, was 50 times smaller than that of the true on-line measurements like the pH, temperature and weight. To overcome the problem of non synchronous data, an interpolation of the state variables with the same frequency of the online measurements was necessary. This was done in the following manner (see Figure 6.1):

1. With the exception of the last available measurement point, all off-line data were linearly interpolated between each other (segment interpolation).
2. The last available data was kept constant until a new data came and then, also linearly interpolated with its previous counterpart.

The embodied data were used to identify in an online fashion the parameters of the neural network that minimize the identification criterion defined by Equation 5.1.

The chemotaxis algorithm was used to estimate the parameters of the neural network and those of the system of differential equations that describe the dynamics of the model of the system (see Appendixes A2 and A4).

Consistency between measurements and calculations was also checked during the whole cultivation through the on-line monitoring of statistical variables, like the measurement of certainty (for more details, see Section 5.4). This monitoring was extended to the performance indexes and, in the case of the *Kluyveromyces lactis* cultivation, to the penalty terms (Equations 6.5, 6.6 and 6.7). They were monitored during the fermentation as functions of the aforementioned training cycle time.

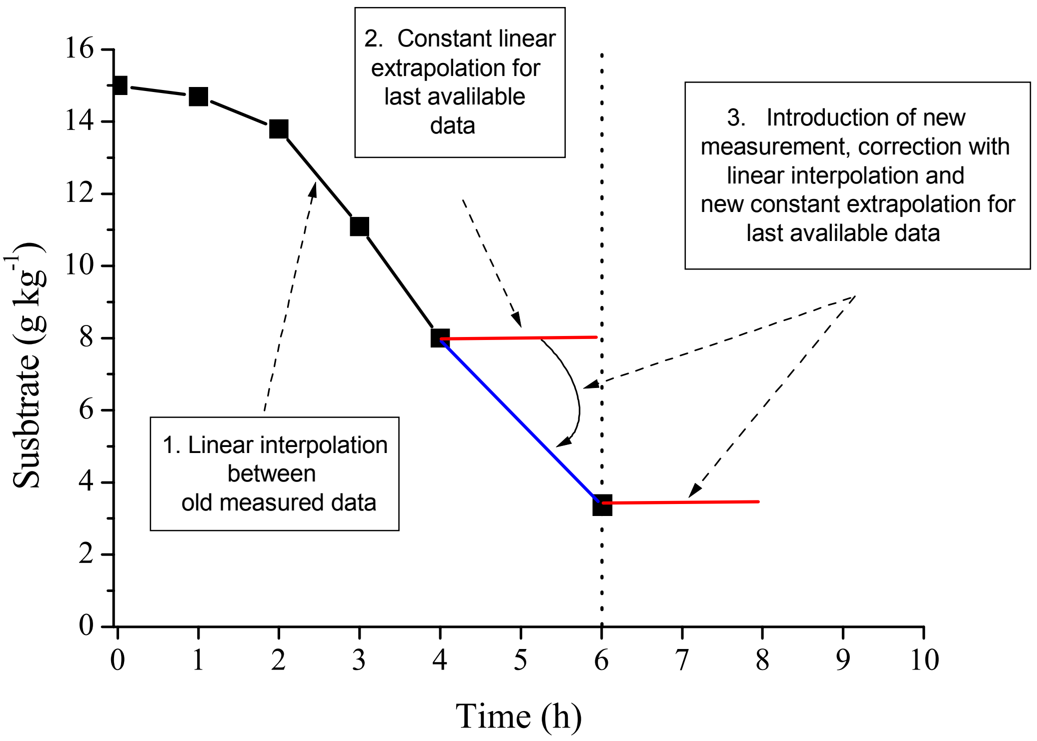

**Figure 6.1** Procedure for incorporating off-line measurements into the online database.

## 6.4 Results and discussion

A sensitivity analysis was performed for the soft sensor (Simutis and Lübbert, 1997). This examination showed that the variables temperature/p$O_2$, and OD were basic entities necessary for the proper functioning of the soft sensor.

Furthermore, the biomass specific oxygen consumption rate ($q_XO_2$) and biomass specific carbon dioxide evolution rate ($q_XCO_2$) complemented the inputs of the soft sensor. These variables may be used always in conjunction and not separately. The removal of one of these variables from the soft sensor improves the accuracy for estimating some of the state variables, but at expenses of worsening the estimation of others.

For both microbial systems, it was found that the biomass specific oxygen consumption rate was closely related with the accurate estimation of the biomass concentration. However, when used as single input for the soft sensor it provided a poor description of the substrate concentration. In contrast, the biomass specific carbon dioxide evolution rate accounted for high precision in the estimation of the substrate concentration. Nevertheless, it delivered a quite unsatisfactory estimation of the biomass concentration when it was used as single input for the soft sensor.

Compared with the classical soft sensors based on off-gas analysis (Chéruy and Flaus, 1994), the present approach exhibited the additional advantage of the easiness to incorporate temperature or p$O_2$ dependence in its formulation.

### 6.4.1 *Escherichia coli* cultivation

Figure 6.2 shows the measured temperature trajectory and the estimation of the expected final total protein activity (Equation 6.1). This estimation was done employing the soft sensor and is presented as function of the time. Every experimental point corresponds to a training cycle of the soft sensor. The data depicts the prediction of the expected final state for the total protein activity. The additional upper circle represents the expected protein activity without disturbances (optimal expected performance index). The lower circle represents the measured protein activity at the end of the fermentation, but after the disturbances took place.

As stated before, two different temperature drops were artificially introduced to test the performance of the online optimization. As can be seen in Figure 6.2, the temperature disturbances caused an alteration in the estimation of the expected protein activity. This estimation experienced abrupt changes in its value, which were additionally delayed in relation to the moments where the temperature drops took place.

After induction, the prediction of the final total protein activity presented a shift in its behavior, incrementing from around 9000 to 17000 Units of activity. It remained around this value with relative constant behavior. The upper circle in Figure 6.2 depicts the expected protein activity without temperature disturbances in the process (31500 Units). This is a prediction obtained using the hybrid model reported in Chapter 5, without considering any disturbance. The lower circle represents the total protein activity measured at the end of the fermentation (15800 Units) after the temperature disturbances occurred. In comparison, the soft sensor delivered a final estimation for the protein activity of about 16300 Units.

Even when an alternative optimization course for a disturbed process could be performed, it may be stated that the faithfulness in the estimation of the re-optimized performance index is strongly correlated with the accuracy of the model used, as will be discussed later.

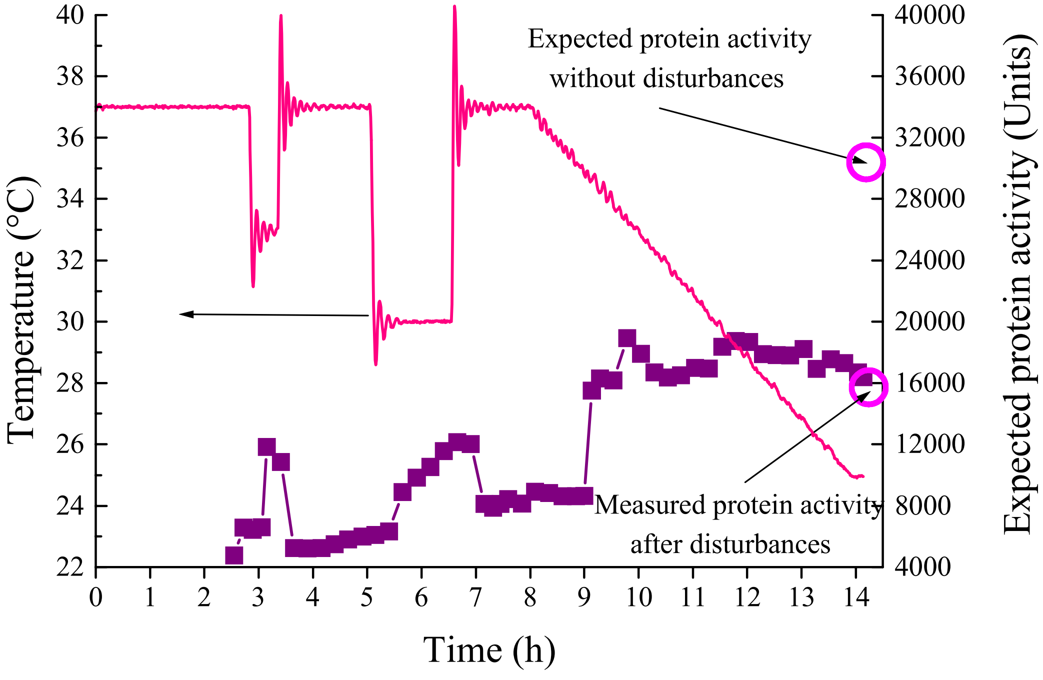

**Figure 6.2** Artificially disturbed temperature trajectory and online estimation of the final expected protein activity determined with the soft-sensor.

Figure 6.3 shows the results obtained for the measure of certainty in the estimation of biomass and substrate concentrations (see Chapter 5 for more details). As can be seen, the swift in the expected protein activity prediction seemed to be correlated not only to the temperature changes, but also to the accurate estimation of the substrate concentration variable. Inaccurate estimations of the performance index were prevalent when the measure of certainty for the substrate trespassed its warning limit.

The estimation of the biomass concentration was very reliable compared with that of the substrate concentration. In a period of about 4 h (from the 4th to the 8th fermentation hour), the control limit for good fitting (see Appendix 3, Table A3.1 for more details) was trespassed for the substrate concentration. From the 9th hour and on, both, the control and the warning limit for fitting were not violated anymore. This coincided with the aforementioned swift and stabilization of the profit function prediction. Then, in the particular case, the long term prediction capability of the soft-sensor was strongly affected by the accuracy of the substrate concentration estimation.

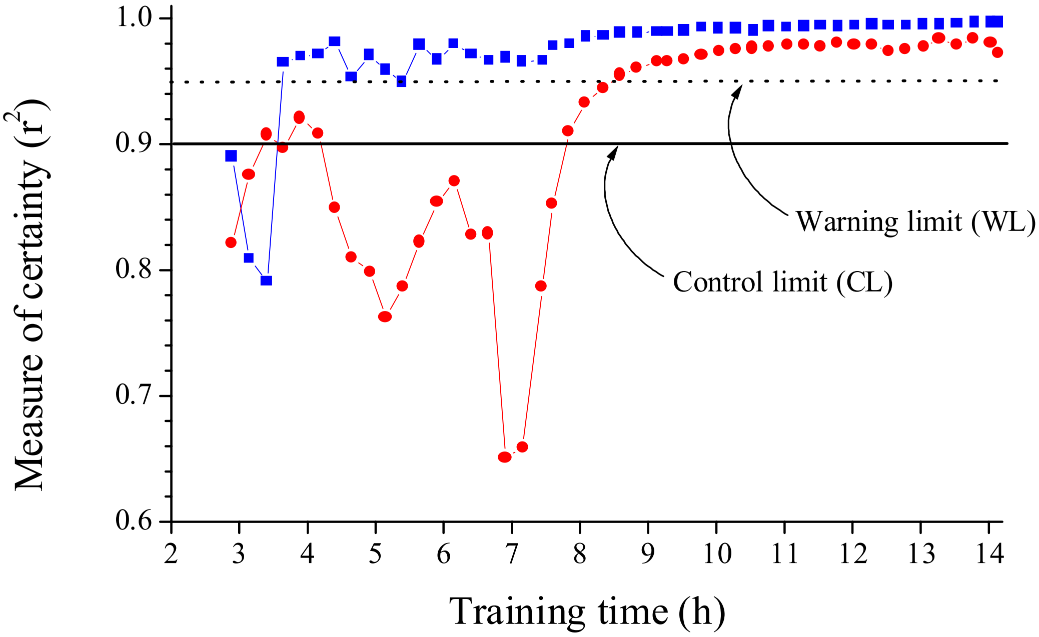

**Figure 6.3** Measure of certainty trajectories for the biomass (-■-) and substrate (-●-) concentration variables.

Finally, Figure 6.4 shows the estimated and measured biomass and substrate concentrations. The corresponding specific growth rate is also depicted. Symbols represent the measurements, while continuos lines account for the simulated values. The current estimation is actually the last calculation obtained on-line. As can be seen, the modeling of the specific growth rate by the neural network soft sensor provided very accurate final estimations of the state variables biomass and glucose concentrations.

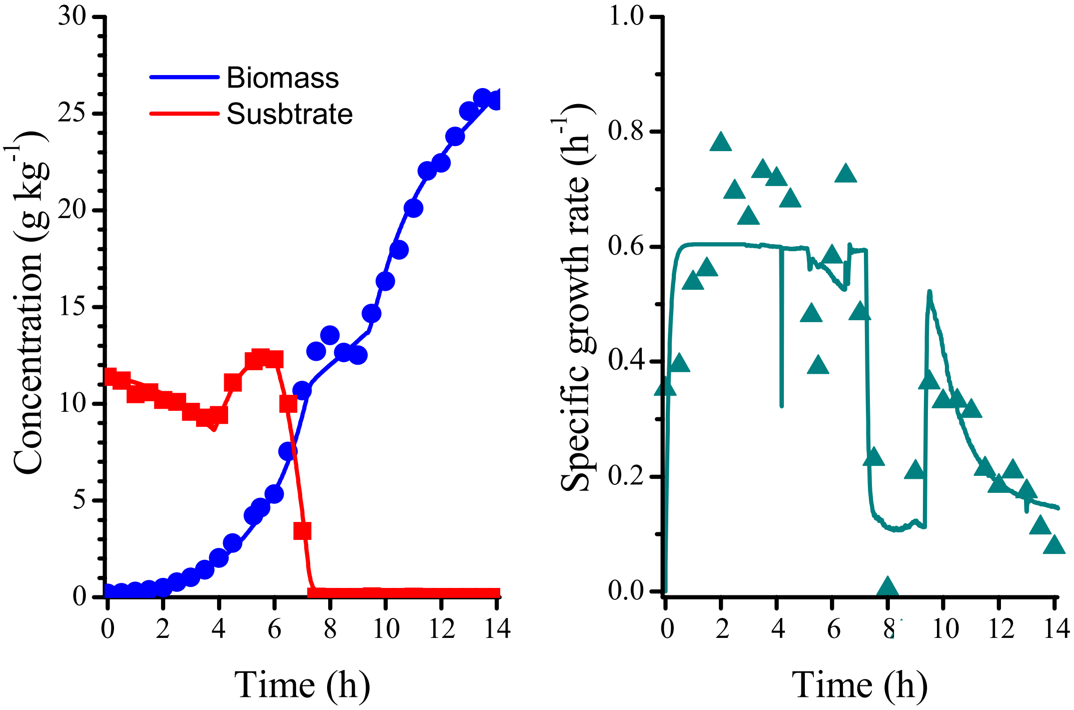

**Figure 6.4** Time trajectories of the concentration of biomass, substrate and the specific growth rate. Lines account for simulated values while symbols represent the measurements.

### 6.4.2 *Kluyveromyces lactis* cultivation

The fed-batch cultivation of *Kluyveromyces lactis* was run under specific growth rate-controlled conditions. The optimization goal was to maximize the biomass production per total fermentation time. The control of the specific growth rate was done mainly to avoid problems with the intracellular metabolic overflow during the fermentation. Figure 6.5 depicts the modeled and measured state variables, biomass, glucose and recombinant protein concentration, as well as the corresponding specific growth rate. As can be seen, the neural network-based soft sensor was able to properly describe the variables with high accuracy, even in the absence of measured data for a period of 9 h (from the 8th to the 17th fermentation hour).

In the cultivation, as described by Volk *et al*. (2001), the state variables of the process were adjusted cyclically. After every adjustment cycle an on-line re-optimization was performed by adjusting the feeding profile. The specific growth rate obtained is presented in Figure 6.5. As can be seen, the optimal set-point for the specific growth rate, $\mu=0.13$ $h^{-1}$ was only partially fulfilled. However the modeling of the specific growth rate can be considered satisfactory.

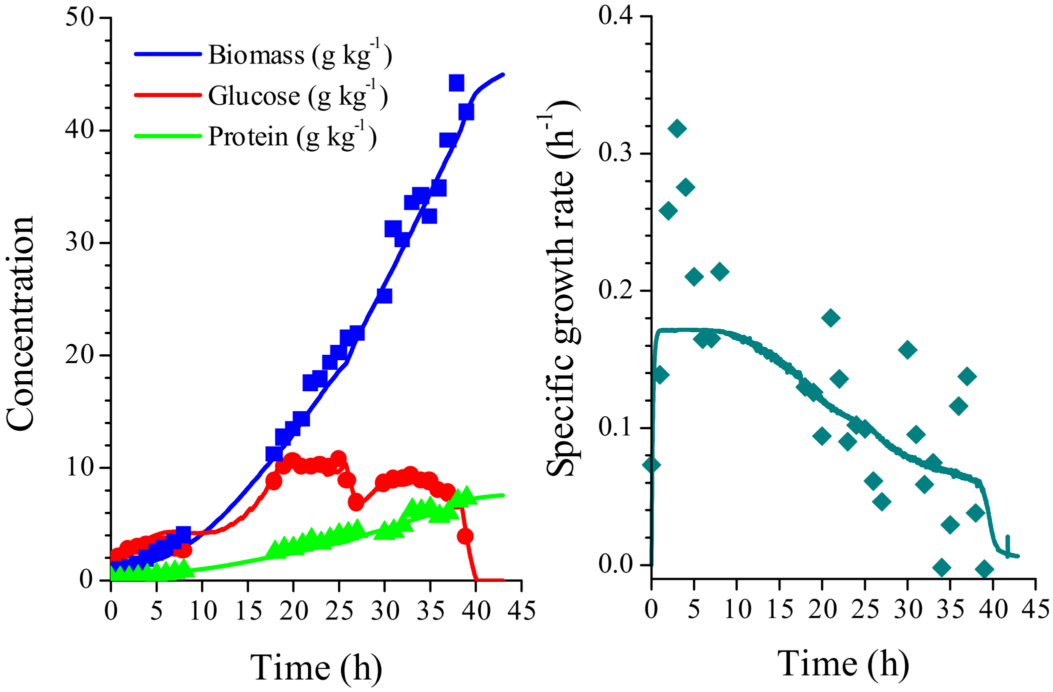

**Figure 6.5** Time trajectories of the concentration of biomass, glucose, recombinant protein and their correlated variable, the specific growth rate. Lines account for simulated values while symbols represent the measurements.

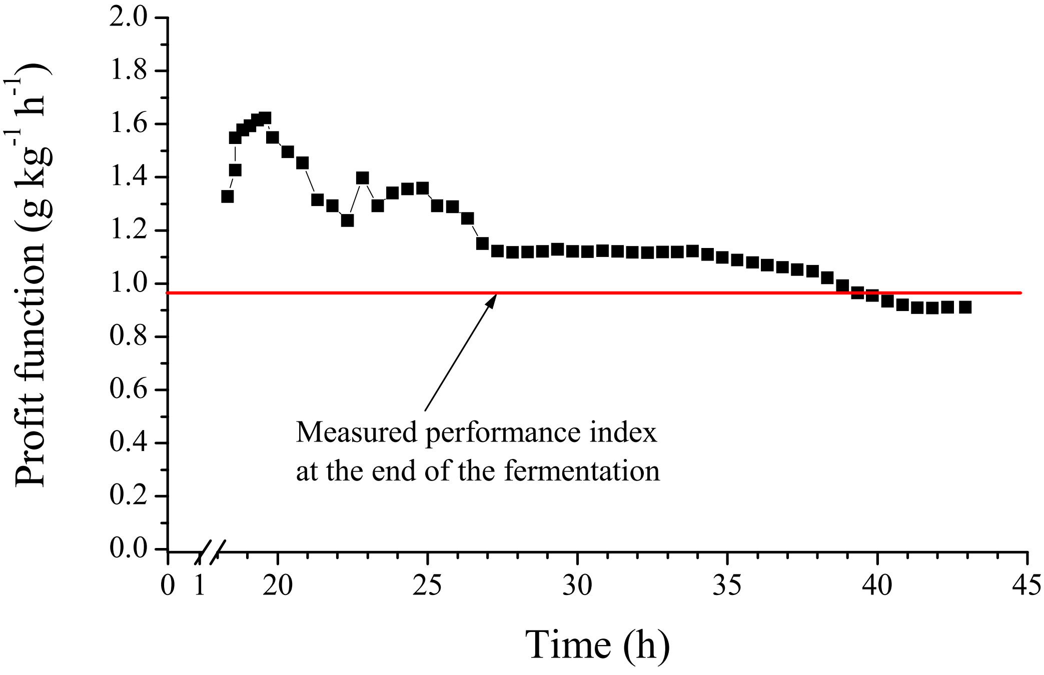

**Figure 6.6** Time trajectory of the expected performance index. The line represents the measured performance index at the end of the fermentation.

Figure 6.6 compares the expected values of the profit function using the neural network-based soft-sensor with the final measured total protein content per fermentation time. After a period of adaptation lasting about 20 h, the trajectory of the simulation for the expected performance index showed an almost constant trend in its prediction. The prediction also showed good agreement with the final measurement of the performance index obtained for the cultivation.

It was discussed that the optimization was constrained by controlling the specific growth rate at the set point, μ=0.13 $h^{-1}$ (see Equation 6.4). Figure 6.7 presents the different time trajectories for the penalties of the performance index, described by Equations 6.4, 6.5 and 6.6. As can be seen, the penalties for maximal OUR and feeding rate, went to zero at the end of the cultivation, because the performance index is taken into its optimal limits concerning the OUR and feeding rate variables.

Contrasting to them, the penalty of the controlled specific growth rate did not vanished and actually increased in the final part of the fermentation. It is worth to remark, that an apparent relationship between the estimated performance index (see Figure 6.6) and the vanishing of the feeding constraint may be established.

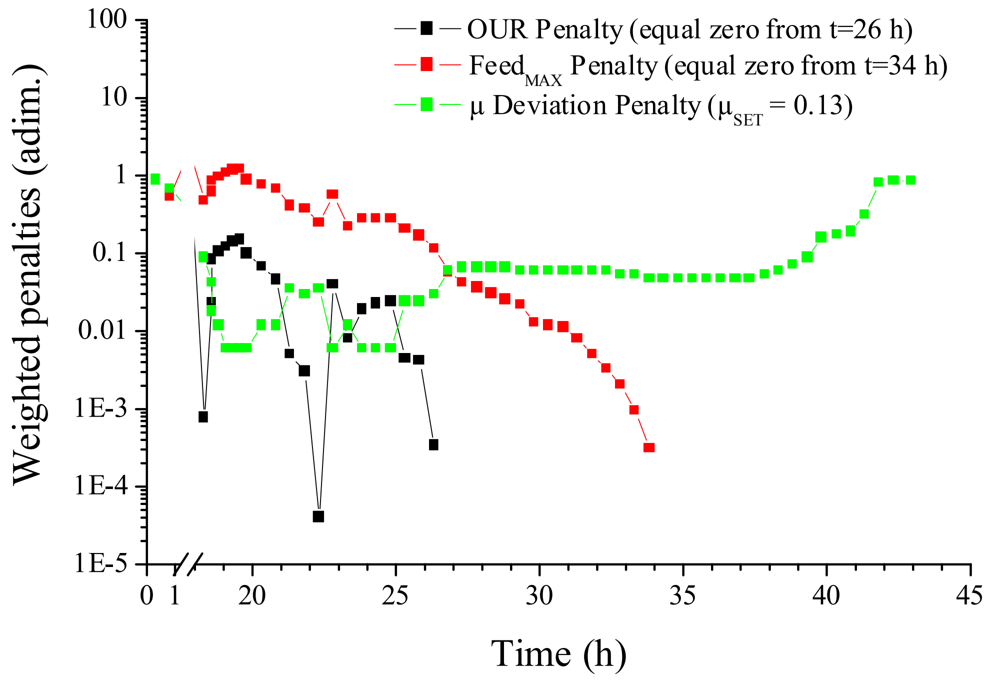

**Figure 6.7** Time trajectories for the optimization goal penalties. The penalty accounting for the maximal OUR disappeared from the 26th hour of cultivation time, while that of maximal feeding rate from the 34th hour.

## 6.5 Conclusions

A neural network-based soft-sensor was presented and tested for monitoring the performance index of two different microbial systems. The approach presented different operating characteristics, depending on the particularities of the process and microbial system under consideration. In the case of the *E. coli* cultivation, the soft-sensor presented a rather satisfactory modeling performance. To examine the efficiency of the proposed online optimization subjected to disturbances, the process was run under artificially introduced temperature perturbations. The prediction performance was highly dependent on the accuracy of the state variables estimation. In the present case, the substrate concentration estimation was very poor during the initial phase of the fermentation. Nevertheless, the modeling procedure was able to rectify in an online fashion, erroneous trends in the identification process of the state variables, improving its prediction ability. In the case of the *Kluyveromyces lactis* cultivation, the hybrid approach showed an excellent prediction and modeling capacity. In contrast with the *E. coli* cultivation, the yeast fermentation was growth at low rates and with no temperature changes. This situation facilitated the estimation of the specific growth rate for the process. Under such circumstances, a reliable extrapolation feature of the soft sensor was recognize, because missing data did not seem to affected the modeling performance of the approach. Moreover, the tracking of the performance index during the fermentation enhanced the comprehension of the dynamical relationships between the state and process variables.

## 6.6 Nomenclature

| | | |
|---|---|---|
| $CPR$ | Carbon dioxide production rate | [g kg$^{-1}$h$^{-1}$] |
| $C_X$ | Biomass concentration | [g kg$^{-1}$] |
| $f`_C$ | Constraint function | [Units of activity] |
| $F$ | Feeding rate of fresh media | [kg h$^{-1}$] |
| $J'_{IDEN}$ | Identification criteria | |
| $J'_{OPT}$ | Optimization goal | [Units of activity] |
| $K'_{C1,2,3}$ | Constants for the constrain function | [Variable] |
| $OUR$ | Oxygen uptake rate | [g kg$^{-1}$h$^{-1}$] |
| OD | Optical density | |
| $P$ | Product concentration | [g kg$^{-1}$] |
| $pO_2$ | Partial  pressure of dissolved oxygen | [%] |
| $q_XO_2$ | Biomass specific oxygen uptake rate | [mg g$^{-1}$ kg$^{-1}$h$^{-1}$] |
| $q_XCO_2$ | Biomass specific carbon dioxide production rate | [mg g$^{-1}$kg$^{-1}$h$^{-1}$] |
| $r^2$ | Measure of certainty | |
| $S$ | Substrate concentration | [g kg$^{-1}$] |
| $S_F$ | Substrate concentration in fresh feeding | [g kg$^{-1}$] |
| $t$ | Time | [h] |
| $t_F$ | Time at the end of fermentation | [h] |
| $T$ | Temperature | [°C] |
| $W$ | Liquid reaction weight | [kg] |
| $y$ | Vector of modeled outputs variables | |
| $Y$ | Vector of measured outputs variables | |
| $Y_{X/S}$ | Yield coefficient Biomass/Substrate | [g g$^{-1}$] |

**Greek symbols**

| | | |
|---|---|---|
| $\alpha$ | Yield coefficient Product/Biomass | [g g$^{-1}$] |
| $\mu$ | Specific growth rate | [h$^{-1}$] |
| $\mu_{SET}$ | Specific growth rate set-point | [h$^{-1}$] |

## 6.7 References

1. Bastian, G. and Dochain, D. On-line Estimation and Adaptive Control of Bioreactors. Elsevier Science Publishing B. V. (1990).
2. Chéruy, A. and Flaus, J. M. Des mesures indirectes à l'estimation en ligne. Published in Capteurs et mesures en biotechnologie by Boudrant et al. (In French). Technique et Documentation – Lavoisier (1994).
3. Claes, J.E. and Van Impe, J.F.: On-line estimation of the specific growth rate based on viable biomass measurements: experimental validation; Bioprocess Engineering 21, 389-395 (1999).
4. Franco-Lara, E.; Volk, N.; Lübbert, A. Optimierung der Produktion rekombinanter Proteine mit hybriden Modellen (In German) Chemie Ingenieur Technik 72 1+2: 110-114 (2000).
5. Simutis, R. and Lübbert, A. Exploratory Analysis of Bioprocesses Using Artificial Neural Network-Based Methods. Biotech. Prog. 13: 479-487 (1997).
6. Oliveira, R.; Simutis, R.; Lübbert, A. HYBNET, a new tool for advanced process modelling. Proceedings of the 1st European Symposium on Biochemical Engineering Science., Dublin, Ireland (1996).
7. Sonnleitner, B. and Käppeli, O. Growth of Saccharomyces cerevisiae is controlled by its limited respiration capacity; formulation and verification of a hypothesis. Biotechnol Bioeng 28: 927-937 (1986).
8. Volk, N.; Franco-Lara, E.; Galvanauskas, V.; Lübbert, A.: Model Supported Optimization of Fed-Batch Fermentations for Recombinant Protein Production. In: Recombinant Protein Production with Prokaryotic and Eukaryotic Cells. A Comparative View on Host Physiology. O.-W. Merten, D. Mattanovich, C. Lang, G. Larsson, P. Neubauer, D. Porro, P. Postma, J. Teixeira de Mattos, J. Cole (eds.). (Accepted) Kluwer Academic Publishers (2001)
9. Zenke, F.T., Kapp, L. and Breunig, K.D. Regulated phosphorylation of the Gal4p inhibitor Gal80p of Kluyveromyces lactis revealed by mutational analysis. Biol Chem 380: 419-430 (1999)

# Chapter 7

# General conclusions

## 7.1 General Conclusions

The establishment of a framework for the online modeling, monitoring and optimization of bioprocesses has been investigated. The utilization of neural network-based hybrid models was considered the backbone for process description. Under the concept of hybrid model is contemplated a set of non linear differential equations and neural network or neuro-fuzzy models. The differential equations set were used to described the mass balances relationships. The neural network models were incorporated to describe exclusively key kinetic aspects like the specific growth or specific product production rate.

In these applications in particular, the neural networks are used as "black box" model components. These black-box models associate certain known and measurable process input variables to other output variables of the process, whose values are usually not known or not measurable. A complex relationship between them is supposed to occur, but would be only described by the neural network after a proper training procedure. This kind of approaches offer some inherent advantages presented in the stand-alone neural network application, like the modeling of specific kinetic rates without using some *a priori* function relationships and/or model structure. There are several advantages in the employment of hybrid models that could be highlighted throughout the development of this investigation effort:

1. They permit the biochemical engineer to use extended data records and/or incorporate relevant heuristic knowledge of process dynamics in a single mathematical representation.
2. In contrast with the use of stand-alone neural network approaches, the data demand for training purposes can be reduced.
3. They can deal effectively with changing environments because of their flexibility in the description of nonlinear relationships (like microbial kinetics) and their model-free design.
4. Although off-line training of neural network systems is usually a straightforward matter, online training of hybrid models improves the process description circumventing the long time consuming and slow convergence of the training algorithm.
5. They permit the simulation of key unmeasured variables: the formulation not only improves the description and understanding of the biological system, but also allows a model-based optimization of the operating conditions of the system. Thus, in this way, partially avoiding the costly alternative of long term investigations of the biological and transport processes of the microbial systems.

The use of neural network-based techniques, established as milestone of this work, was tested at three different recombinant microbial systems either for modeling, monitoring and/or optimization purposes. Real-time estimation and monitoring of the specific growth and specific production rate, based on an pseudo-online measurements of biomass and substrate concentration was validated on several fed-batch fermentations. Being the reduction of the invested time for developing and improving a given process a fundamental demand in today’s biotechnology, the presented online modeling method improved the gain of process knowledge at high learning rates. Such a methodology can be viewed as an intermediary evolution step between pure empirical towards full formal mechanistic approaches. Concerning the online optimization, it could be shown that increased productivities were obtainable, as compared to optimizations employing conventional approaches.

At last, monitoring of bioprocesses and their performance indexes through a soft-sensor showed the inherent plasticity of the neural network-based approach to infer complex kinetic rates. They used exclusively control variables like temperature and process correlated measurements available online, like the optical density of the broth, its biomass specific oxygen consumption rate and the biomass specific carbon dioxide evolution rate.

# Appendix 1

## Model for the multi substrate cultivation of *Escherichia coli* (MAK-33)

The model consist of a system of differential equations describing the mass balance in batch operation modus. The mass balance considers the components biomass ($C_X$) and the substrates glucose ($S_1$), lactose ($S_2$) and glycerin ($S_3$). The uptake kinetics are described through a diauxic mechanism relating the three substrates under consideration. Biomass is considered unsegregated:

$$\frac{dC_X}{dt} = \mu \, C_X \qquad \text{(A1.1)}$$

The model for the substrates consumption is described by the equations (A1.2) to (A1.4):

$$\frac{dS_1}{dt} = -\rho_1 \, C_X \qquad \text{(A1.2)}$$

$$\frac{dS_2}{dt} = -\rho_2 \, C_X \qquad \text{(A1.3)}$$

$$\frac{dS_3}{dt} = -\rho_3 \, C_X \qquad \text{(A1.4)}$$

The kinetic relationships are derived in a heuristic way, taking into consideration basic correlations in a mechanistic form. Starting point is the consideration that the specific growth rate ($\mu$) is a result of the superposition of the additive growth rates ($\mu_{Si}$) of each single substrate,

$$\mu = \mu_{S1} + \mu_{S2} + \mu_{S3} = Y_{X/S1}\rho_{S1} + Y_{X/S2}\rho_{S2} + Y_{X/S3}\rho_{S3} \qquad \text{(A1.5)}$$

The substrate limitations were described by a Monod relationship, while that of the catabolite repression by a Haldane expression. Theoretically, the following interactions can take place:

- Catabolite repression of glucose consumption though glycerin
- Catabolite repression of lactose consumption though glucose
- Catabolite repression of lactose consumption though glycerin
- Catabolite repression of glycerol consumption though glucose
- Catabolite repression of glycerol consumption though lactose

Glucose uptake kinetics

The uptake of glucose represses the uptake of lactose. However, catabolite repression through glycerin must be taken into account and was expressed in the model with a Haldane term, while the glucose consumption is represented with a Monod term.

$$\rho_{S1} = \frac{\rho_{\max 1} * S_1}{S_1 + K_{S1} + \frac{S_3^2}{K_{I13}}} \tag{A1.6}$$

Lactose uptake kinetics

The catabolite repression of lactose consumption through glucose and glycerin is considered here. When present, glucose represses the activation of enzymes responsible for lactose consumption. After glucose is consumed, a lag time of 1.5 h in lactose consumption has been observed. The kinetics expresses the lactose limitation with a Monod term. The lactose repression, on the other hand, was represented with a lag-phase term coupled to a Haldane-type expression considering the blocking of the responsible enzyme for lactose consumption, due to the presence of glucose and glycerin. The lag-phase term is given by:

$$Philag = \left\{ \begin{array}{l} \arctan\left[ kp \frac{\frac{t-\infty}{t_{lag}-1}}{\pi + 0.5} \right]; S_1 > S_{Grenz} \\ \arctan\left[ kp \frac{\frac{t-t_{akt}}{t_{lag}-1}}{\pi + 0.5} \right]; S_1 = S_{Grenz} \\ \arctan\left[ kp \frac{\frac{t-t_{akt(S=0)}}{t_{lag}-1}}{\pi + 0.5} \right]; S_1 < S_{Grenz} \end{array} \right\} \tag{A1.7}$$

The repression of the enzyme responsible for lactose consumption (Haldane-type for glucose and glycerin) is multiplied with the lag-phase term to give the total lactose uptake rate:

$$\rho_{S2} = Philag * \frac{\rho_{\max 2} * S_2}{S_2 + K_{S3} + \frac{S_1^2}{K_{I21}} + \frac{S_3^2}{K_{i23}}} \tag{A1.8}$$

Glycerin uptake kinetics

Glycerol consumption can be modeled with a Monod approach, augmented with Haldane-type repression terms for glucose and lactose:

$$\rho_{S3} = \frac{\rho_{\max 3} * S_3}{S_3 + K_{S3} + \frac{S_1^2}{K_{I31}} + \frac{S_2^2}{K_{I32}}} \qquad \text{(A1.9)}$$

All parameters used in the partial models can be found in the Table A1.1

Table A1.1 Parameters of the heuristic model for the multi-substrate cultivation of *E. coli*

| Parameter | Symbol | Value | Units |
|---|---|---|---|
| Max. spec. uptake rate (Glucose) | $\rho_{max1}$ | 1.82 | $g\ g^{-1}\ h^{-1}$ |
| Max. spec. uptake rate (Lactose) | $\rho_{max2}$ | 0.6 | $g\ g^{-1}\ h^{-1}$ |
| Max. spec. uptake rate (Glycerin) | $\rho_{max3}$ | 0.62 | $g\ g^{-1}\ h^{-1}$ |
| Yield coefficient (Glucose) | $Y_{X/S1}$ | 0.4 | $g\ g^{-1}$ |
| Yield coefficient (Lactose) | $Y_{X/S2}$ | 0.62 | $g\ g^{-1}$ |
| Yield coefficient (Glycerin) | $Y_{X/S3}$ | 0.4 | $g\ g^{-1}$ |
| Saturation coefficient (Glucose) | $K_{S1}$ | 0.01 | $g\ L^{-1}$ |
| Saturation coefficient (Lactose) | $K_{S2}$ | 0.01 | $g\ L^{-1}$ |
| Saturation coefficient (Glycerin) | $K_{S3}$ | 0.01 | $g\ L^{-1}$ |
| Inhibition coefficient (Gluc←Glyc) | $K_{i13}$ | 100 | $g\ L^{-1}$ |
| Inhibition coefficient (Lac←Gluc) | $K_{i21}$ | 100 | $g\ L^{-1}$ |
| Inhibition coefficient (Lac←Glyc) | $K_{i23}$ | 100 | $g\ L^{-1}$ |
| Inhibition coefficient (Glyc←Gluc) | $K_{i31}$ | 100 | $g\ L^{-1}$ |
| Inhibition coefficient (Glyc←Lac) | $K_{i32}$ | 100 | $g\ L^{-1}$ |
| Constant | $k_p$ | 300 | |
| Lag-time | $t_{Lag}$ | 1.5 | h |
| Limit concentration | $S_{Grenz}$ | 0.1 | $g\ L^{-1}$ |

# Appendix 2

## Model for the fed-batch production of the native protein complex VP1-DHFR using *Escherichia coli* BL21.

# APPENDIX 2

The model is based on the formulation of Volk *et al*. (1998) for a batch process.

$$\frac{dC_{\mathrm{X}}}{dt} = \mu\, C_{\mathrm{X}} - \frac{F}{W} C_{\mathrm{X}} \tag{A2.1}$$

$$\frac{dS}{dt} = -\sigma\, C_{\mathrm{X}} + \frac{F}{W} (\mathrm{S}_{\mathrm{F}} - S) \tag{A2.2}$$

$$\frac{dP}{dt} = \pi\, C_{\mathrm{X}} - \frac{F}{W} P \tag{A2.3}$$

$$\frac{dO_2}{dt} = -\,\mathrm{OUR} + k_L a\, (O_2{*} - O_2) \tag{A2.4}$$

$$\frac{dW}{dt} = F\ \text{-}\ sample \tag{A2.5}$$

where,

$$F = F_{\text{Fresh Medium}} + F_{\text{Acid}} + F_{\text{Base}} + F_{\text{Antifoam}}\ \text{-}\ evap \tag{A2.6}$$

$$\mathrm{OUR} = \frac{1000}{22.4} \cdot \frac{273}{273 + \mathrm{T}} \cdot \frac{\mathrm{A}}{\mathrm{V}} \cdot \frac{32}{100} \left( \mathrm{O}_{2\ \text{Gas}}^{\ \ \text{Input}} - \mathrm{K} \cdot \mathrm{O}_{2\ \text{Gas}}^{\ \ \text{Output}} \right) \tag{A2.7}$$

with:

$$\mathrm{K} = \frac{(100 - \mathrm{O}_{2\ \text{Gas}}^{\ \ \text{Input}} - \mathrm{CO}_{2\ \text{Gas}}^{\ \ \text{Input}})}{(100 - \mathrm{O}_{2\ \text{Gas}}^{\ \ \text{Output}} - \mathrm{CO}_{2\ \text{Gas}}^{\ \ \text{Output}})} \tag{A2.8}$$

The $k_L a$ coefficient is a function of the airflow (A) and the stirrer speed (N):

$$k_L a = \mathrm{K}_1 \cdot \frac{\mathrm{A}^{\theta\,1}}{\mathrm{N}^{\theta\,2}} \tag{A2.9}$$

being $\theta_{1,2}$ two adjustable parameters to be identified.

The kinetics are given by:

$$\mu = \frac{(\mu_{\max|37^{\circ}C} * \alpha_1 \exp(\alpha_2 \mathrm{T})) \cdot S}{\mathrm{K}_{\mathrm{S}} + S} \cdot \left( 1 - \mathrm{e}^{-\,t/\tau} \right) \tag{A2.10}$$

where the last term describes the lag phase.

For the specific substrate uptake rate:

$$\sigma = \frac{\mu}{\mathrm{Y}_{\mathrm{X/S}}} \tag{A2.11}$$

For the specific production rate, if $t < t_{\mathrm{I}}$, then $\pi = 0$, else:

$$\pi = \frac{\mathrm{k_P}}{1+(\frac{\mathrm{K_{p\mu}}}{\mu})^5} * \left(1 - \frac{(P/C_{\mathrm{X}})}{P_{X\max}}\right) * \left(\frac{P/C_{\mathrm{X}}}{(P/C_{\mathrm{X}} + K_{SPX})}\right) \tag{A2.12}$$

In the case of a neuro-fuzzy representation of the specific production rate, if $t < t_{\mathrm{I}}$, then $\pi = 0$, else:

$$\pi_{\mathrm{FUZZY}} = f(\mu, C_{\mathrm{X}}, P/C_{\mathrm{X}}) \tag{A2.13}$$

where $\pi_{\mathrm{FUZZY}}$ is a function described by a fuzzy artificial neural network and a set of 5 fuzzy rules.

Parameters taken from Volk *et al.* (1998):

$\mu_{\max|37°\mathrm{C}} = 0.532\ \mathrm{h^{-1}}$<br>
$\alpha_1 = 0.54\ \mathrm{h^{-1}}$<br>
$\alpha_2 = 0.0825\ °\mathrm{C^{-1}}$<br>
$\mathrm{K_S} = 0.01\ \mathrm{g\ kg^{-1}}$<br>
$\mathrm{Y_{XS}} = 0.43\ \mathrm{g\ g^{-1}}$<br>
$\mathrm{k_P} = 92.86\ \mathrm{U\ g^{-1}\ h^{-1}}$<br>
$\mathrm{K_{p\mu}} = 0.602\ \mathrm{h^{-1}}$<br>
$P_{X\,max} = 92.6\ (-0.0096344\ \mathrm{T}^2 + 0.5386\ \mathrm{T} - 6.1123)$ Units $\mathrm{g^{-1}}$<br>
$K_{SPX} = 0.62\ (-0.0484\ \mathrm{T} + 2.26)$ Units $\mathrm{g^{-1}}$

## NOMENCLATURE

| | | |
|---|---|---|
| A | Airflow into the system | [kg $\mathrm{h^{-1}}$] |
| $C_{\mathrm{X}}$ | Biomass concentration | [g $\mathrm{kg^{-1}}$] |
| *evap* | Evaporation rate | [kg $\mathrm{h^{-1}}$] |
| $F$ | Total feed rate | [kg $\mathrm{h^{-1}}$] |
| $F_{\text{Fresh media}}$ | Feed rate of fresh media | [kg $\mathrm{h^{-1}}$] |
| $F_{\text{Acid}}$ | Feed rate of acid | [kg $\mathrm{h^{-1}}$] |
| $F_{\text{Base}}$ | Feed rate of base | [kg $\mathrm{h^{-1}}$] |
| $F_{\text{Antifoam}}$ | Feed rate of antifoam | [kg $\mathrm{h^{-1}}$] |
| $k_La$ | Volume specific mass transfer coefficient | [$\mathrm{h^{-1}}$] |
| $\mathrm{K_1}$ | Proportionality constant for $k_La$ | [$\mathrm{h^{-1}}$] |
| $\mathrm{K_S}$ | Monod constant | [g $\mathrm{kg^{-1}}$] |
| $\mathrm{K}_{SPX}$ | Saturation constant for product | [Units $\mathrm{g^{-1}}$] |
| $\mathrm{k_P}$ | Maximal specific production rate | [Units $\mathrm{g^{-1}\ h^{-1}}$] |

| | | |
|---|---|---|
| $K_{P\mu}$ | Saturation constant for growth rate | [h$^{-1}$] |
| N | Stirrer speed | [rev min$^{-1}$] |
| OUR | Oxygen Uptake Rate | [mg kg$^{-1}$ h$^{-1}$] |
| $O_2$ | Concentration of dissolved oxygen | [mg kg$^{-1}$] |
| $O_2*$ | Maximal concentration of dissolved oxygen | [mg kg$^{-1}$] |
| $\mathrm{O}_{2\ \mathrm{Gas}}^{\ \mathrm{Input}}$ | % of oxygen at the gas inlet | [%] |
| $\mathrm{O}_{2\ \mathrm{Gas}}^{\ \mathrm{Output}}$ | % of oxygen at the gas outlet | [%] |
| $\mathrm{CO}_{2\ \mathrm{Gas}}^{\ \mathrm{Input}}$ | % of carbon dioxide at the gas inlet | [%] |
| $\mathrm{CO}_{2\ \mathrm{Gas}}^{\ \mathrm{Output}}$ | % of carbon dioxide at the gas outlet | [%] |
| $P$ | Product concentration | [Units kg$^{-1}$] |
| $\mathrm{P}_{X\,max}$ | Maximal specific product concentration | [Units g$^{-1}$] |
| $S$ | Substrate concentration | [g kg$^{-1}$] |
| $\mathrm{S}_\mathrm{F}$ | Substrate concentration in fresh feeding | [g kg$^{-1}$] |
| *sample* | Sampling rate | [kg h$^{-1}$] |
| $t$ | Time | [h] |
| $t_\mathrm{F}$ | Time at the end of fermentation | [h] |
| $t_\mathrm{I}$ | Induction time | [h] |
| T | Temperature | [°C] |
| $V$ | Liquid reaction volume | [L] |
| $W$ | Liquid reaction weight | [kg] |
| $\mathrm{Y}_{\mathrm{X/S}}$ | Yield coefficient Biomass/Substrate | [g g$^{-1}$] |

### Greek symbols

| | | |
|---|---|---|
| $\alpha_1$ | Termical constant | |
| $\alpha_2$ | Termical constant | [°C$^{-1}$] |
| $\beta_1$ | Model constant for CPR | [mg g$^{-1}$] |
| $\beta_2$ | Model constant for CPR | [mg g$^{-1}$ h$^{-1}$] |
| $\gamma_1$ | Model constant for OUR | [mg g$^{-1}$] |
| $\gamma_2$ | Model constant for OUR | [mg g$^{-1}$ h$^{-1}$] |
| $\mu$ | Growth rate | [h$^{-1}$] |
| $\mu_{\mathrm{max}\vert 37°\mathrm{C}}$ | Maximal specific growth rate at 37°C | [h$^{-1}$] |
| $\theta_{1,2}$ | Exponential parameters for estimating $k_La$ | |
| $\sigma$ | Substrate consumption rate | [h$^{-1}$] |
| $\tau$ | Lag phase time | [h] |
| $\pi$ | Protein production rate | [h$^{-1}$] |

# Appendix 3

## Statistical estimation of fitness for linear models and interquartile range analysis (IQR)

## Statistical estimation of fitness for linear model

The measure of certainty is given by:

$$r^2 = 1 - \frac{\sum_{i=1}^{N} \left(Y_i - y_i\right)^2}{\sum_{i=1}^{N} \left(Y_i - \underline{y}\right)^2} \qquad \text{(A3.1)}$$

Where N is the number of measures, $Y_\mathrm{i}$ are the measured state variables, $y_\mathrm{i}$ the modeled state variables and $\underline{y}$ the mean of all measured state variables, given by:

$$\underline{y} = \frac{1}{N} \sum_{i=1}^{N} Y_\mathrm{i} \qquad \text{(A3.2)}$$

The empirical variance, $\sigma_Y^2$ is given by:

$$\sigma_Y^2 = \frac{1}{N-1} \sum_{i=1}^{N} \left(Y_i - \underline{y}\right)^2 \qquad \text{(A3.3)}$$

The empirical standard deviation, $\sigma_Y$ by:

$$\sigma_Y = \sqrt{\frac{1}{N-1} \sum_{i=1}^{N} \left(Y_i - \underline{y}\right)^2} \qquad \text{(A3.4)}$$

Table A4.1 present a numerical magnitude reference for the goodness of the measurement of certainty, $r^2$ (Wolf, 1994).

Table A4.1 Goodness for the measurement of certainty

| $r^2$ | Goodness |
|---|---|
| 0 | No functional relationship |
| ≥ 0.8 | Satisfactory |
| ≥ 0.9 | Good |
| ≥ 0.95 | Very good |
| 1.0 | Strong functional relationship |

## Box-Plot and interquartile range analysis (IQR)

The interquartile range analysis is used to describe the distribution characteristics of a given population and to identify extreme features belonging to this population. An outliner in a population may be the result of a data entry error, a poor measurement or a change in the system that generated the data.

> An *α-Quartile* $q_\alpha$, where $\alpha \in (0, 1)$, is a number that builds the $100 \cdot \alpha$ % of the population's feature smaller or equal to it, being also $100 \cdot (1-\alpha)$ % of the population bigger or equal to it. The 0.5-Quartile, $q_{0.5}$ is called *median*; the 0.25-Quartile, $q_{0.25}$ is called the *lower quartile* and the 0.75-Quartile, $q_{0.75}$ is called the *upper quartile*. The difference between lower and upper quartiles, $q_d := q_{0.75} - q_{0.25}$, is called the *quartile difference*.

The Box-Plot is a graphical representation of three quartiles: $q_{0.25}$, median and $q_{0.75}$. Its analysis proportionate information about the form, the situation and dispersion of population's distribution. Additionally, it helps to identify outliners in a given population.

> A *box*- (and *whisker*-) *Plot* is formed by a box having as vertical limits the lower quartile $q_{0.25}$ and upper quartile $q_{0.75}$ and an inner line representing the median. It present also additional lines called *whiskers* coming from the quartiles and including outer values that are:
>
> - Not bigger than $q_{0.75} + 1.5\,(q_{0.75} - q_{0.25})$ , and
> - Not smaller than $q_{0.25} - 1.5\,(q_{0.75} - q_{0.25})$ .
>
> All individuals from the population that are situated out of these ranges are designated as *outliners*. In the case of a population with normal distribution, the outliners build up to 0.7% of the total population.

# Appendix 4

## Model for the fed-batch production of the native protein complex GAL80/HIS-TAC, using the yeast *Kluyveromyces lactis* RUL 1888 D80ZR-pEAHG80.

The model considers the mass balances for biomass ($C_X$), substrate ($S$) and broth weight ($W$) with an additional algebraic description of the product ($P$):

$$\frac{dC_X}{dt} = \mu\, C_X - \frac{F}{W} C_X \qquad \text{(A4.1)}$$

$$\frac{dS}{dt} = -\sigma\, C_X + \frac{F}{W}(\mathrm{S}_\mathrm{F} - S) \qquad \text{(A4.2)}$$

$$P = \alpha\, C_X \qquad \text{(A4.3)}$$

$$\frac{dW}{dt} = F \text{ - } sample \qquad \text{(A4.4)}$$

where,

$$F = F_{\text{Fresh Medium}} + F_{\text{Acid}} + F_{\text{Base}} + F_{\text{Antifoam}} \text{ - } evap \qquad \text{(A4.5)}$$

The kinetics are given by:

$$\mu = \mu_{\text{ANN}} \cdot \left(1 - \mathrm{e}^{-t/\tau}\right) \qquad \text{(A4.6)}$$

where the last term describes the lag phase and being,

$$\mu_{\text{ANN}} = f(\mathrm{q_XO_2}, \mathrm{q_XCO_2}, \mathrm{OD}, \mathrm{pO_2}) \qquad \text{(A4.7)}$$

a function described by a single feed forward artificial neural network.

$$\sigma = \frac{\mu}{\mathrm{Y_{X/S}}} \qquad \text{(A4.8)}$$

## NOMENCLATURE

| | | |
|---|---|---|
| $C_X$ | Biomass concentration | [g kg$^{-1}$] |
| *evap* | Evaporation rate | [kg h$^{-1}$] |
| $F$ | Total feed rate | [kg h$^{-1}$] |
| $F_{\text{Fresh media}}$ | Feed rate of fresh media | [kg h$^{-1}$] |
| $F_{\text{Acid}}$ | Feed rate of acid | [kg h$^{-1}$] |
| $F_{\text{Base}}$ | Feed rate of base | [kg h$^{-1}$] |
| $F_{\text{Antifoam}}$ | Feed rate of antifoam | [kg h$^{-1}$] |
| OD | Optical density | |
| $P$ | Product concentration | [Units kg$^{-1}$] |
| $\mathrm{pO_2}$ | Partial  pressure of dissolved oxygen | [%] |

| | | |
|---|---|---|
| $q_XCO_2$ | Biomass specific carbon dioxide production rate | [mg g$^{-1}$ kg$^{-1}$ h$^{-1}$] |
| $q_XO_2$ | Biomass specific oxygen uptake rate | [mg g$^{-1}$ kg$^{-1}$ h$^{-1}$] |
| $S$ | Substrate concentration | [g kg$^{-1}$] |
| $S_F$ | Substrate concentration in fresh feeding | [g kg$^{-1}$] |
| *sample* | Sampling rate | [kg h$^{-1}$] |
| $t$ | Time | [h] |
| $W$ | Liquid reaction weight | [kg] |
| $Y_{X/S}$ | Yield coefficient Biomass/Substrate | [g g$^{-1}$] |

**Greek symbols**

| | | |
|---|---|---|
| $\alpha$ | Proportionality constant | [g g$^{-1}$] |
| $\mu$ | Specific growth rate | [h$^{-1}$] |
| $\mu_{ANN}$ | Specific growth rate described by an ANN | [h$^{-1}$] |
| $\sigma$ | Substrate consumption rate | [h$^{-1}$] |
| $\tau$ | Lag phase time | [h] |

# Curriculum Vitæ

*General*

Name: Ezequiel Franco Lara
Date of birth: October 4, 1969
Place of birth: México City, México
Citizenship: Mexican
Civil status: Divorced

*Present Position*

Ph. D. Student at the Martin-Luther-Universität Halle-Wittenberg, Institut für Bioengineering

*School Education*

**1975-81** Basic School in Guadalajara, México
**1981-84** Secondary School in Guadalajara, México
**1984-87** High School in Guadalajara, México

Diploma: 12th school year diploma 1987 with final mark 89,5/100

*Graduation*

**1987-93** Graduation in Chemical Engineering at the University of Guadalajara
Specialization: Plastics' technology
Theme of the thesis work: "Characterization of growth and distribution of immobilized yeast in alcoholic fermentation processes. Application of SEM and digital image analysis"
Diploma: Title of "Chemical Engineer" with final mark 93,4/100

**1994-97** Postgraduate studies in Chemical Engineering at the University of Guadalajara
Theme of the thesis work: "Validation of a kinetics model and monitoring of the morfological evolution of yeast in different bioreactors: Application of digital image analysis and artificial neural networks"
Diploma: Title of "Master in Science in Chemical Engineering" with final mark 95,2/100

*Grants*

**1992-93** Grant for young researcher from the University of Guadalajara
**1994-96** Grant from the National Council for Science and Technology (CONACyT 69683)
**1997-2001** Ph. D. grant from Deutscher Akademischer Austauschdienst (DAAD A/97/00417)

*Publications*

10 Publications

# Publications List

[1] Franco-Lara, E.; Volk, N.; Lübbert, A.: *Optimierung von Fermentationsprozessen mittels hybrider Modelle*. Proceedings from „DECHEMA-Jahrestagungen ´99". Band I, Dechema e.V. (1999) 236-237.

[2] Franco-Lara, E.; Volk, N.; Lübbert, A.: *Optimierung der Produktion rekombinanter Proteine mit hybriden Modellen*. Chemie Ingenieur Technik (72) 1+2 I, Wiley-VCH Verlag GmbH (2000) 110-114

[3] Franco-Lara, E.; Volk, N.; Hertel, T.; Lübbert, A.: *On line Model Adjustment and Optimization of the Production of a Viral Capsid Protein*. Proceedings from „Biotechnology 2000". Band I, Dechema e.V. (2000) 236-237

[4] Volk, N.; Franco-Lara, E.; Galvanauskas, V.; Lübbert, A.: *Model supported Optimization of Fed Batch Fermentations for the Recombinant Protein Production*. Proceedings from „Recombinant Protein Production with prokaryotic and eukaryotic cells", EFB event 103. European Federation of Biotechnology, Semmering, Austria (2000) 67

[5] Franco-Lara, E.; Volk, N.; Hertel, T.; Galvanauskas, V.; Lübbert, A.: *On line monitoring of profit functions employing a soft-sensor*. Proceedings from „DECHEMA-Jahrestagungen 2001", Dechema e.V. (2001) 452-453.

[6] Franco-Lara, E.; Volk, N.; Hertel, T.; Galvanauskas, V.; Lübbert, A.: *Neuro-fuzzy hybrid modeling of the production of a viral capsid protein*. Proceedings from „DECHEMA-Jahrestagungen 2001", Dechema e.V. (2001) 505-506

[7] Franco-Lara, E.; Volk, N.; Galvanauskas, V.; Lübbert, A.: *Model-Supported Optimization of Recombinant Protein Production Using Hybrid Models*. Proceedings from the 3rd European Congress of Chemical Engineering (Electronic version), Nuremberg (2001)

[8] Franco-Lara, E.; Galvanauskas, V.; Volk, N.; Lübbert, A.: *Model-Based optimization of the Cultivation Process for Recombinant Virus Capsid Proteins in E. Coli*. Proceedings from „8th International Conference on Computer Applications in Biotechnology ", Edited by Denis Dochain and Michel Perrier. Québec City (2001)

[9] Franco-Lara, E.; N. Volk; T. Hertel; V. Galvanauskas; A. Lübbert: *Model-Supported Optimization of Recombinant Protein Production Using Hybrid Models*. Chemie Ingenieur Technik (73), Wiley-VCH Verlag GmbH. (2000) 654-655

[10] Volk, N.; Franco-Lara, E.; Galvanauskas, V.; Lübbert, A.: *Model supported Optimization of Fed Batch Fermentations for the Recombinant Protein Production*. In: Recombinant Protein Production with Prokaryotic and Eukaryotic Cells. A Comparative View on Host Physiology. O.-W. Merten, D. Mattanovich, C. Lang, G. Larsson, P. Neubauer, D. Porro, P. Postma, J. Teixeira de Mattos, J. Cole (Eds.) (In Press). Kluwer Academic Publishers (Approximate date of publication: November 2001)

Ezequiel Franco Lara
Privada las lomas 765
Colonia Álamo Oriente
45560 Tlaquepaque, Jalisco
México

## Schriftliche Versicherung

Ich versichere hiermit, daß ich die vorgelegte Dissertation mit dem Titel

**Framework for online modeling, optimization and monitoring of bioprocesses**

selbständig verfaßt, sowie ohne unzulässige Hilfe Dritter und ohne Benutzung anderer als der angegebenen Hilfsmittel angefertigt habe. Die aus fremden Quellen direkt oder indirekt übergenommenen Gedanken sind also solche kenntlich gemacht.

Halle (Saale), 15. Mai 2002

Ezequiel Franco Lara
Privada las lomas 765
Colonia Álamo Oriente
45560 Tlaquepaque, Jalisco
México

## Schriftliche Erklärung.

Ich erkläre hiermit, daß die vorgelegte Dissertation mit dem Titel

**Framework for online modeling, optimization and monitoring of bioprocesses**

bisher bei keiner anderen Fakultät als Dissertation eingereicht worden ist. Darüber hinaus habe ich mich bisher auch mit keiner anderen Arbeit bei irgendeiner anderen Fakultät um eine Promotion beworben.

Halle (Saale), 15. Mai 2002