Nonlinear models speed up development of an enantioselective enzymatic process for a pharmaceutical intermediate

Abstract

Good process development can result in better production economics than can be achieved by operating in countries with low labour costs. To be able to produce the product most cost efficiently while fulfilling the requirements on product properties and other constraints, it is very advantageous to have mathematical models which relate the important variables in the process.

Process development of biotechnological processes is often carried out by performing a large number of experiments, for the simple reason that development of mathematical models of most biotechnological processes is too complicated. Physical modeling is not very effective, neither is empirical modeling with conventional linear statistical techniques. New techniques of nonlinear modeling have altered this situation entirely, and offer a tremendous advantage in quantitatively describing complicated biotechnological processes. Nonlinear models have successfully been used for a large number of processes in various industrial sectors. In the case described in this article, they turn out to be an order of magnitude better than linear models.

How these models in combination with appropriate mathematical tools help in efficient process development with much less experimentation is demonstrated with an example. Besides a little bit of theory, the article also explains the strengths and limitations of these new techniques.

Introduction

Pharmaceutical intermediate production technology has undergone a lot of progress over the last couple of decades. Biotechnological processes are better routes for several kinds of products including some kinds of pharmaceutical intermediates. A particularly interesting category is that of enantiomers. There are enzymes which catalyse the conversion of primarily one enantiomer, which is also an effective means of separating enantiomers.

Different biotechnological processes have different characteristics - different objectives, different variables, different kinds of constraints, different micro-organisms or enzymes. The process may be a batch process, a continuous process or a fed-batch one. However, some things are common to process development of various kinds of processes. A product needs to be produced with a given process, from specified raw materials such that the resulting product properties satisfy some conditions, typically upper and lower limits.

a_1 < property 1 < b_1
a_2 < property 2 < b_2
a_3 < property 3 < b_3
...
a_n < property n < b_n

The same methodology can be applied also to separation processes. For process development purposes of several kinds of biotechnological processes, the product concentration, concentrations of undesired side-products, viscosity, etc. can be considered to be the product properties, where the product is the liquid resulting at the end of the process. Besides these constraints, there are limits on other variables. Process variables like temperature, pH, stirring, etc. and feed characteristics like substrate concentration, microbial or enzyme concentration, aeration, etc. also have upper and lower limits. The objective of process development is to determine the best values of the feed characteristics and process variables such that the product properties are within desired limits, and preferably, a production economic variable (e.g. production rate, raw material consumption, energy efficiency, purity, number of defects, emissions or whatever) is maximized or minimized. The problem looks somewhat similar from the

process modeling point of view for a wide variety of biotechnological processes. From the process modeling point of view, the product properties are consequences of feed characteristics and process variables, as summarized in Figure 1.

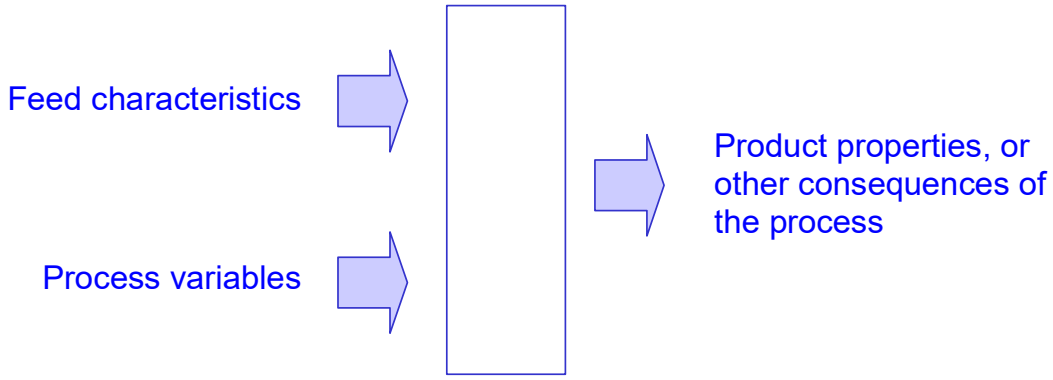


Figure 1. A typical model configuration for process development

There have been several, mostly academic, attempts at modeling biotechnological processes for different purposes. These have been either based on simulated data, or are attempts to illustrate how one approach is better than another. A larger fraction of them are based on linear statistical techniques. Some are meant to be illustrations of how certain functions can be performed better by certain techniques. These articles do not quite have solve industrial problems, one of which is process development. That process development should be carried out with as few experiments as possible aiming at achieving the desired product properties and optimising a production economic variable.

Nonlinear modeling

Nonlinear modeling has been in industrial use for more than ten years [8]. This new technology benefits you in several ways. It has been successfully utilized by various industries for a variety of purposes, particularly for process development. Nonlinear modeling has successfully been used for a large number of processes and materials in several sectors of process industries including biotechnology, polymers, plastics processing, ceramics, concrete, pulp, paper and board, power generation, semiconductor processing, water treatment, chemical production, food processing, etc. The awareness of these techniques in biotechnology is very limited.

Nonlinear modeling can roughly be defined as empirical or semi-empirical modeling which takes at least some nonlinearities into account. Nonlinear models can be static or dynamic. Nonlinear modeling can be performed in many ways. The simpler ways include polynomial regression and linear regression with nonlinear terms. Nonlinear regression is useful in some situations. The form of the nonlinearities, however, has to be specified in these older techniques. The new techniques of nonlinear modeling are based on free-form nonlinearities. They include series of basis functions, splines, kernel regression, feed-forward neural networks, etc. Feed-forward neural networks are a set of efficient tools for nonlinear modeling, particularly because of their universal approximation capability [9].

Why nonlinear modeling?

Mathematical models represent knowledge of quantitative effects of relevant variables in a concise and precise form. They can be used instead of experimentation if they are reliable enough. Mathematical models also permit the user to carry out various kinds of calculations, like optimization, which can be used to determine suitable values of process variables. Mathematical modeling can be performed in various ways, and different ways are suitable in different situations.

It is not possible to use physical modeling in many situations. Even if it is possible, physical models tend to compute the output more slowly than empirical or semi-empirical models. Development of physical models is time consuming. Nonlinear modeling tends to be expensive, but physical modeling usually costs even more. Physical models involve assumptions and simplifications. Thus empirical modeling is often a better alternative.

Traditional empirical modeling is based on linear statistical techniques. Nothing in nature is absolutely linear. So it helps to take nonlinearities into account rather than ignore them. If the range of variables is small, linear techniques are sometimes sufficient. New techniques of nonlinear modeling based on artificial neural networks allow us to approximate nonlinearities without specifying in detail the nonlinearties to be accounted for. They allow for free form nonlinearities, unlike linear and nonlinear regression methods.

New technologies open up new possibilities

There are many different types of neural networks, and some of them have practical uses in process industries [8, 10]. Neural networks have been in use in process industries for more than ten years. The multilayer perceptron is a kind of a feed-forward neural network. Most neural network applications in industries [11-20] ranging from polymers [11] and concrete [16] to optical fibre cables [17] are based on them.

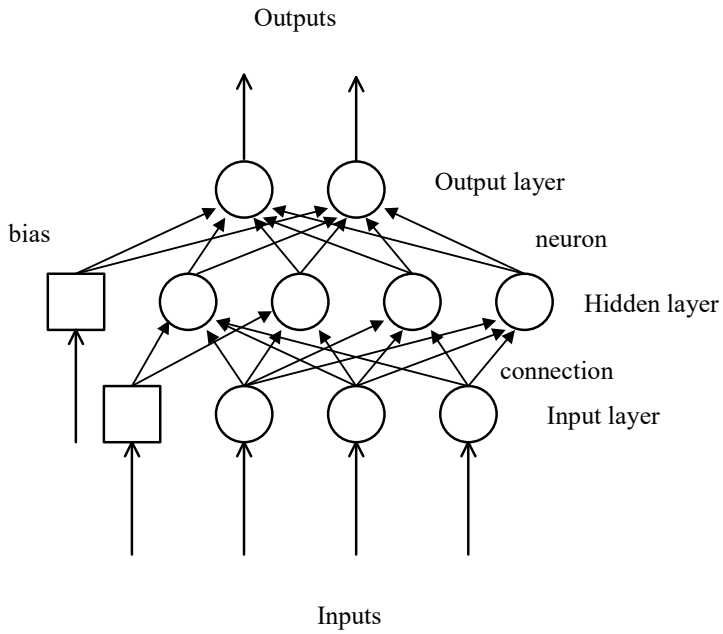


Figure 2. A typical feed-forward neural network

The output of each neuron *i* in a feed-forward neural network is given by

$$z_i = \sigma\left(\sum_{j=0}^{N} w_{ij} x_j\right)$$

where the activation function is often the logistic sigmoid, given by

$$\sigma(a) = \frac{1}{1+e^{-a}}$$

The incoming signals to the neuron are x_j, and w_{ij} are the weights for each connection from the incoming signals to the i^{th} neuron. The w_{i0} terms are called biases. This results in a set of algebraic equations which relate the input variables to the output variables. Thus, for each observation (a set of input and output variables), the outputs can be predicted from these equations based on a given set of weights. The training procedure aims at determining the weights which result in the smallest sum of squares of prediction errors. There are a variety of training methods in use today. It is also possible to combine neural networks with physical models or other empirical models, which often lead to better solutions.

Quality of the nonlinear models

A large number of people today claim to be able to develop neural network models. A large number of people can offer you impressive user interfaces that hide the details of the neural network models inside. However, not many people can solve real-world problems. Very few are able to produce industrially usable systems of good quality. Fewer are consistently successful in every project they agree to take. The result is that there is a wide variation in the quality of nonlinear models [8]. How do you know which models are good and which models are not? In other words, what are the characteristics that you look for in a good nonlinear model?

Characteristics of good nonlinear models

The simple answer to the first question is that the proof of the pudding is in the eating: A good model has to work. There is of course more to it than that. A good model for industrial purposes has to be reliable. Accuracy is secondary. A good industrial model has to be robust. You want a model that can be updated relatively easily. You want the model to have some transparency, and the simpler the better. At the same time, a good model efficiently treats all the important nonlinearities. Reliability, robustness, simplicity, maintainability often come at the cost of accuracy and efficient treatment of nonlinearities. These conflicting demands make nonlinear modeling a harder task. That does not leave too many people who will offer you all these attributes. Most are satisfied with claiming accuracy. It is important to insist on highly reliable and robust models because several decisions might be made based on the answers from these models. Considering the possible effects these decisions can have, it becomes quite obvious that the models should be of as good quality as possible.

In other words, the quality of a nonlinear model is much more than accuracy, and there is a scarcity of quality of nonlinear modeling today. There are no simple ways of measuring reliability and robustness of nonlinear models. How does one ensure these and other features? Experience and expertise are essential for neural network model development. However, a good software tool goes a long way by offering you a variety of measures that tell you of possible undesirable features in the models.

Enantioselective enzymatic process

In a recent process development work of PCAS Finland, a racemic mixture of two enantiomers was to be treated by an enzyme to destroy the undesired enantiomer. The difference between the concentrations of the two enantiomers is referred to as the enantiomer excess, commonly abbreviated to EE. The fraction of the two enantiomers which undergo reaction is referred to as conversion here. In this terminology, the ideal situation would be to obtain an EE of 100% and a conversion of 50%, which would leave the desired enantiomer unreacted. There are plenty of variables which determine the consequences of this enzyme catalysed reaction, of which some are more important than others. The dynamics of the process could be described in the form of

the following two equations, where C_D, C_L, C_S, C_E are concentrations of the two enantiomers, substrate and enzyme respectively.

$$\frac{dC_D}{dt} = f_1(C_D, C_L, C_S, C_E, T, pH)$$
$$\frac{dC_L}{dt} = f_2(C_D, C_L, C_S, C_E, T, pH)$$

It is not feasible to write the right hand sides of the two equations without making large and questionable assumptions. However, it is possible to determine these functions explicitly or implicitly from suitably planned experiments. In the sequel, we will consider only temperature, substrate concentration, enzyme concentration and of course, the reaction time as the inputs of Figure 1. A suitable production economic objective for this work was productivity divided by the amount of enzyme consumed. This objective function could have been different if the enzyme recovery were easier, or if the enzyme cost was much lower.

Experimentation

As mentioned in the earlier sections, nonlinear modeling requires experimental or production data. An experimental approach was selected, and experiments were carried out at laboratory scale in 250 ml reactors. A total of 19 experiments were planned and carried out for this work. It would have been possible to manage with fewer experiments and still produce equally good nonlinear models but this was already a large reduction from the typical number of experiments they were used to carrying out for process development. A couple of reactions were carried out for as long as two weeks, while a few were carried out for just 48 hours. It may be noted that most of the statistical theory on experiment design in literature is meant for development of linear empirical models, and is not very efficient for development of nonlinear models. It is therefore desirable to keep in mind while planning experiments that the results of the experiments should be suitable for nonlinear empirical or semi-empirical modeling.

Before model development is attempted, usually the data is analyzed and preprocessed. It is more important for production data, but experimental data should also be analyzed and preprocessed. The data was preprocessed and analysed by proprietroy software, which has several facilities, including simpler things like calculating the basis statistics of the data set, filtering observations with missing measurements or variables beyond the range of interest, calculation of correlation matrices, showing the plots of every variable against every other, and more advanced features like clustering, calculating sets of observations with maximum or minimum similarity, and dividing the data into training, test and validation sets needed for model development with desired forms of imbalance.

Figure 3 shows a plot of enantiomer excess against reaction time for different values of substrate concentrations, with other variables constant. Figure 4 shows enantiomer excess vs time for different values of temperature.

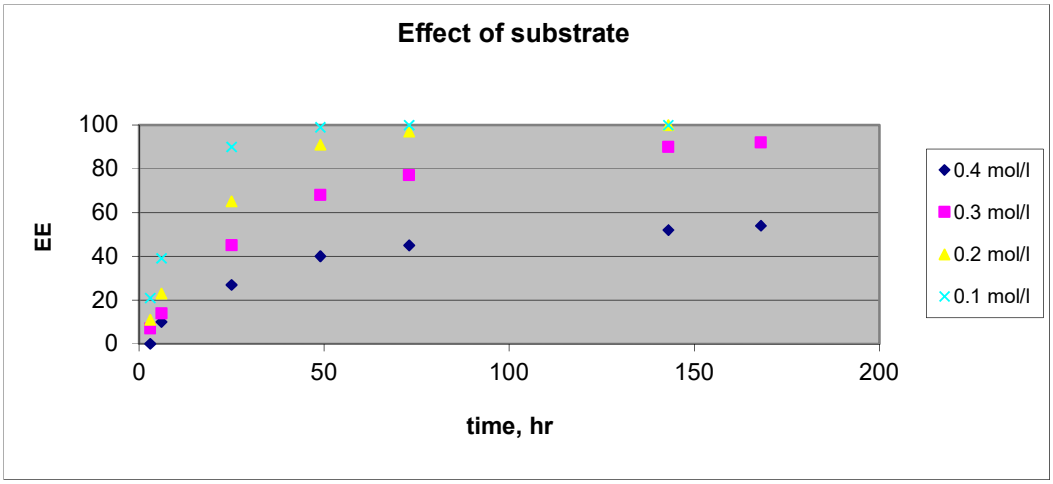


Figure 3. Enantiomer excess vs time plots indicate that the reaction proceeds faster when the substrate concentration is lower

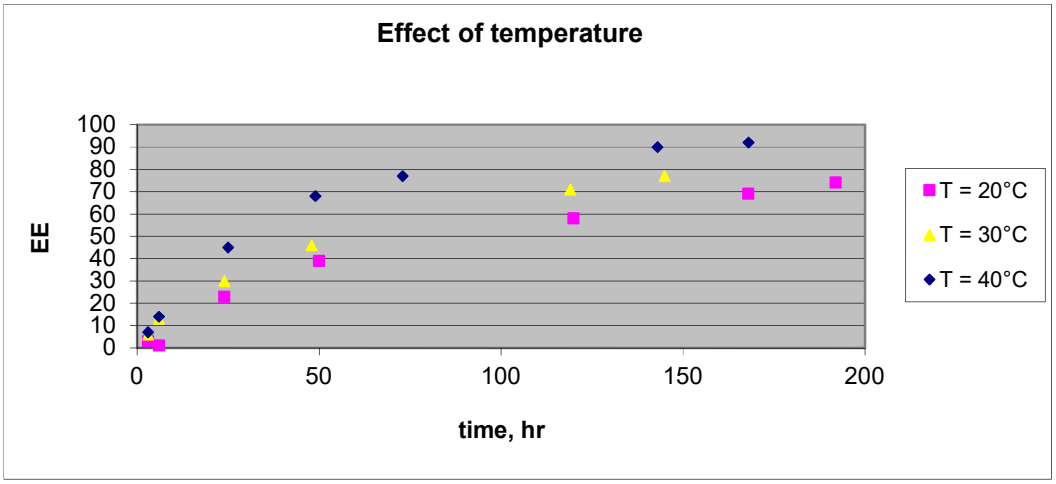


Figure 4. Enantiomer excess vs time plots indicate that the reaction proceeds faster at higher temperatures

Results

A large number of models were attempted with different configurations of feed-forward neural networks, with a single hidden layer, with different activation functions. One or more of the free parameters or weights of many of those models were restricted. It is often advantageous to transform the input or the output variables, from simpler transforms like logarithms of concentrations to more complicated transforms involving even derivatives of variables. Some of the models were trained to predict transformed outputs from the feed-forward neural network using NLS 020 software. Most of the better models showed the same qualitative features, with a high degree of correlation. The nonlinear models taken into use had the following characteristics.

Output variable 1: Enantiomer Excess, [%]
rms err : 4.459
mean |err| : 3.389
max |err| : 12.42
Correlation : 0.9802

Output variable 2: Conversion, [%]
rms err : 3.211
mean |err| : 2.507
max |err| : 8.752
Correlation : 0.9593

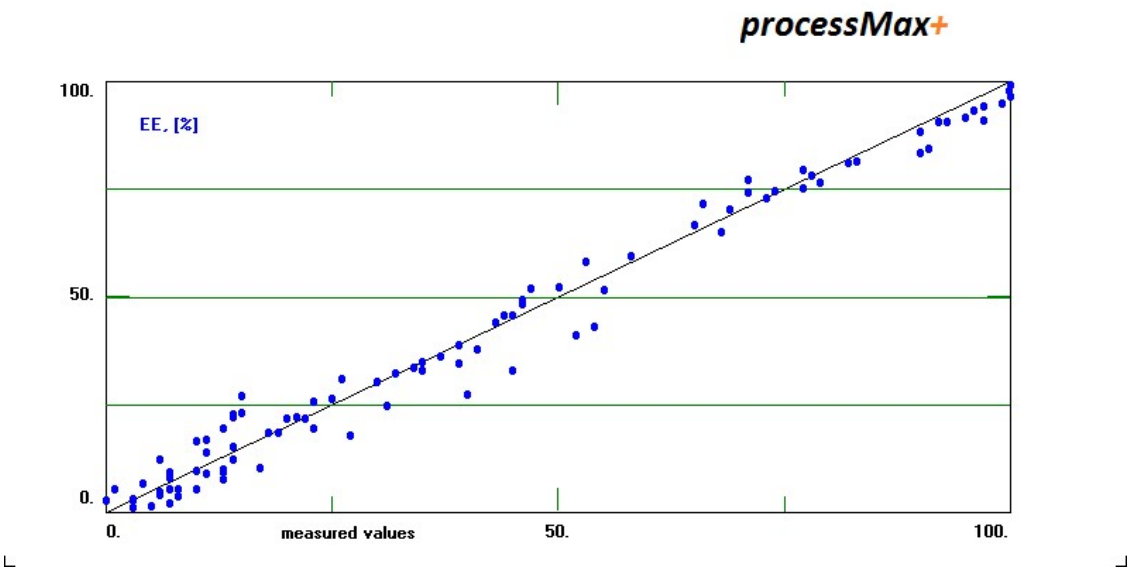


Figure 5. A comparison of the measured and predicted values from the nonlinear model for enantiomer excess

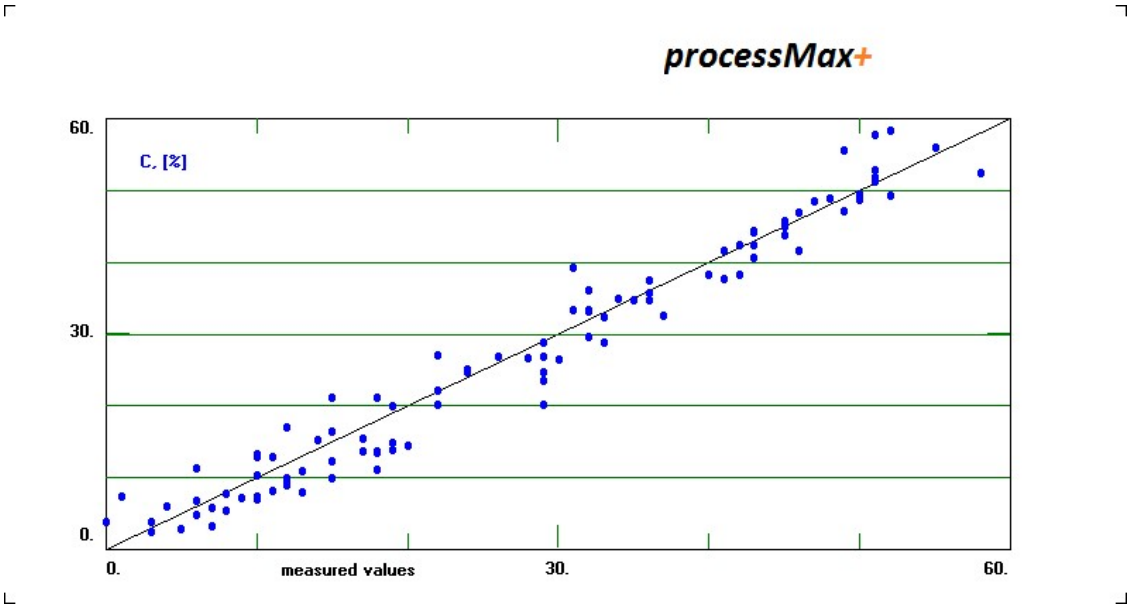


Figure 6. A comparison of measured and predicted values from the nonlinear model for conversion

The statistics of the prediction errors looks good, which of course can be expected since the data was very good. The rms error (root mean square error, roughly speaking, the standard deviation of the prediction errors) for enantiomer excess was calculated to be 4.459. That indicates a correlation coefficient of about 98%, which is really very good for biochemical processes. As can be seen from Figures 5 and 6, the observations have been predicted very well by the model. The rms error for conversion is 3.211, which amounts to a correlation coefficient of 96%, which too is very good.

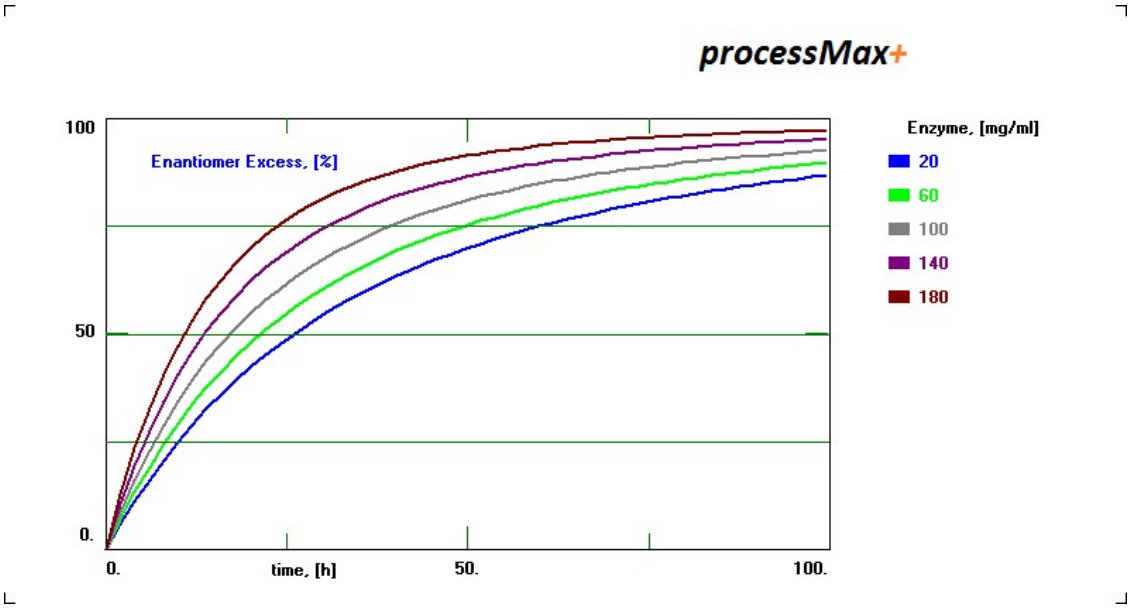


Figure 7. Effect of enzyme concentration on enantiomer excess according to the model

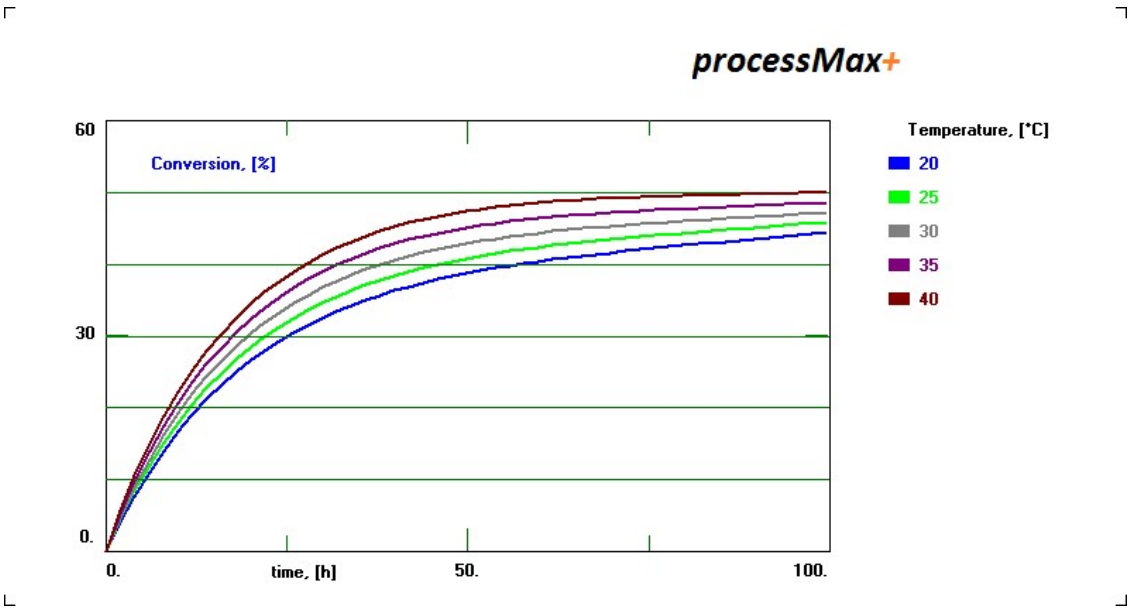


Figure 8. Effect of temperature on conversion according to the model

The models were then implemented in a *PROCESSMAX+* system, a software that allows facile use of nonlinear models which otherwise look unwieldy and are cumbersome to use. The plots in

Figures 7 and 8 are shown for durations under 100 hours because the duration of interest in determining the optimal conditions is about 4 days or less. The models are not perfect. For large values of substrate concentrations in combination with small concentrations of enzyme, the prediction accuracy is lower compared to the range of actual interest for determining optimal conditions.

After the model development was completed, four more experiments were carried out, partly to validate the model, and partly to confirm that the values of process variables and concentrations we were considering did result in desired EE while not getting a conversion much higher than 50%. The prediction error variance on these new observations from the nonlinear model for EE turned out to be less than half of that on the data it was developed from.

Comparison with linear models

It is more than obvious that linear models are hardly applicable to this process where there are strong nonlinearities in the effects of all the input variables. Linear models do not hesitate to predict negative values or values higher than 100%. Still, the biotechnology, pharmaceutical and medical sector today uses primarily linear statistical techniques for several purposes. The aim of this article is to improve the awareness of the new techniques of nonlinear modeling and their utility in biotechnological process development. The linear models for EE and conversion show the following characteristics.

Output variable 1: Enantiomer Excess, [%]
rms err : 20.3045
mean |err| : 16.7142
max |err| : 48.7633
Correlation : 0.5617

Output variable 2: Conversion, [%]
rms err : 10.4456
mean |err| : 8.5970
max |err| : 23.3703
Correlation : 0.5446
In terms of variance, the nonlinear model for EE is 20.74 times better than the linear model for EE, and the nonlinear model for conversion is 10.58 times better than the linear model for conversion.

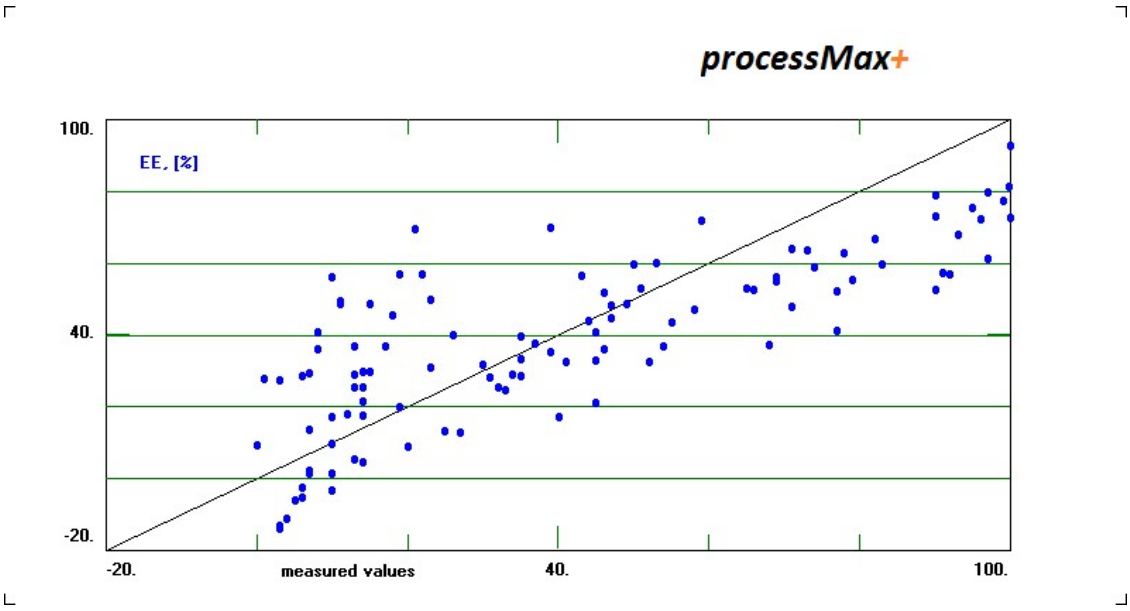


Figure 9. A comparison of the measured and predicted values from the linear model for enantiomer excess

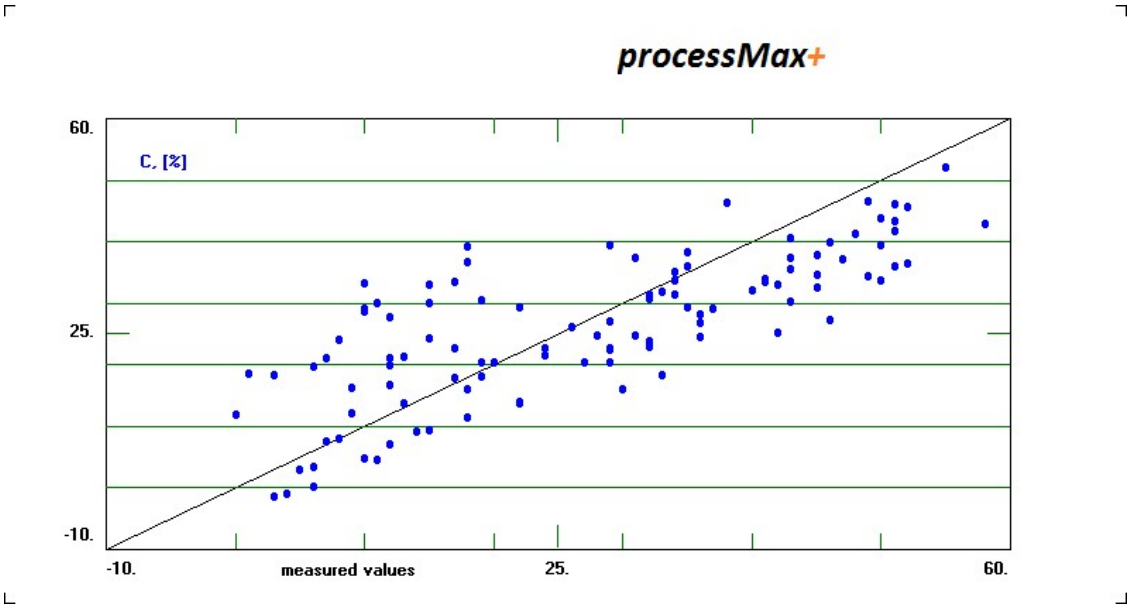
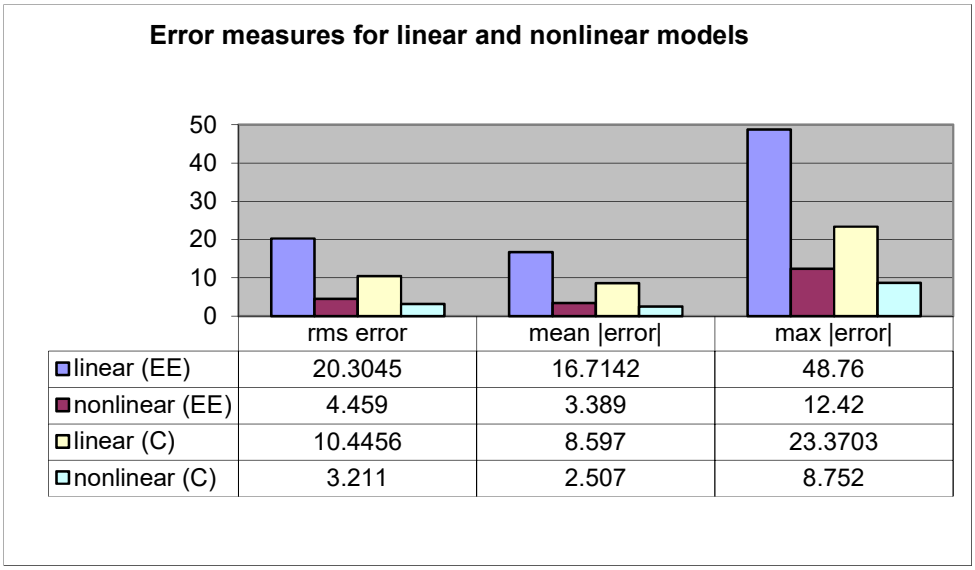


Figure 10. A comparison of measured and predicted values from the linear model for conversion



	rms error	mean |error|	max |error|
linear (EE)	20.3045	16.7142	48.76
nonlinear (EE)	4.459	3.389	12.42
linear (C)	10.4456	8.597	23.3703
nonlinear (C)	3.211	2.507	8.752

Figure 11. Error measures for linear and nonlinear models of EE and conversion show glaring differences

Figures 9 and 10 show comparisons of measured values and predicted values from the linear models. Figure 11 summarises the statistical characteristics of the nonlinear models and the linear models for EE and conversion.

Optimization helps you derive the extra mileage

One advantage of having process knowledge in the form of mathematical models is that it can be utilized for optimization also. Maximizing quality, maximizing production, maximizing profitability are all optimization problems. They usually come with constraints of two kinds – equalities and inequalities. All the process variables have to stay within operable limits, which are usually inequality constraints, and several results (like product properties) of operating conditions are defined by models which are equality or inequality constraints. In mathematical terms, this kind of optimization problems are written as

$$\underset{x \in \boldsymbol{R}^n}{\text{maximize}} \quad F(x)$$

$$\text{subject to} \quad c_i(x)=0, \text{ for } i = 1 \text{ to } m$$
$$\text{and} \quad c_k(x) \geq 0, \text{ for } k = 1 \text{ to } p$$

In process optimization, the inequality constraints are usually the limits on process variables and possibly also product properties, and are therefore simple inequalities in single variables. The problem can be rewritten more specifically as

$$\underset{x \in \boldsymbol{R}^n}{\text{maximize}} \quad F(x)$$

$$\text{subject to} \quad x_k \geq a_k, \text{ for } k = 1 \text{ to } n+n'$$
$$\text{and} \quad x_k \leq b_k, \text{ for } k = 1 \text{ to } n+n'$$

where n is the number of process variables. The 2n' constraints pertaining to product properties come from typically n' product properties. It is in principle possible to maximize or minimize two or

more variables at a time, but such multi-objective optimization does not yield straightforward answers, and is not used often in practice. Many books including [21] describe various techniques for this kind of optimization problems in varying degrees of detail. Many of the techniques are based on gradient descent. The mathematical details of these techniques are not suitable for this article, which primarily aims at making people working with biochemical process development aware of the new possibilities that have emerged with nonlinear modeling.

Making optimization accessible to biotechnology and pharmaceutical laboratories

It is good to have the quantitative knowledge in the form of models, but it is important to be able to use that knowledge. As mentioned in the previous section, a variety of techniques exist for optimization [21]. Engineers and scientists in pharmaceutical laboratories cannot be expected to be familiar with them. A software tool has therefore been developed that it is easy to use without knowing the details of the methods used in it. This brings the new technology to their fingertips. They only need to know what they want to maximize or minimize, and the limits on the variables.

Sometimes, one may want to determine the process conditions which will result in given product characteristics. This is essentially solving a few equations in several unknowns, which has under normal circumstances, an infinite number of solutions. One may instead specify the minimum and maximum acceptable limits for the product properties and optionally, also for process variables. Figure 12 shows an example of that kind.

What is often of greater interest to production people is to determine the conditions under which one of the production economic objectives can be minimized or maximized while keeping the product properties within acceptable limits. The limits for the process conditions can also be specified and one click on the mouse finds the optimum in a few seconds. Figure 13 shows an example where productivity per enzyme consumed is maximized while keeping EE limited to above 95% and conversion below 51%. However, it is also possible to make impossible demands on the set of equations and inequalities, in which case, even constraint satisfaction does not succeed, and the software then tries to look for the best compromise.

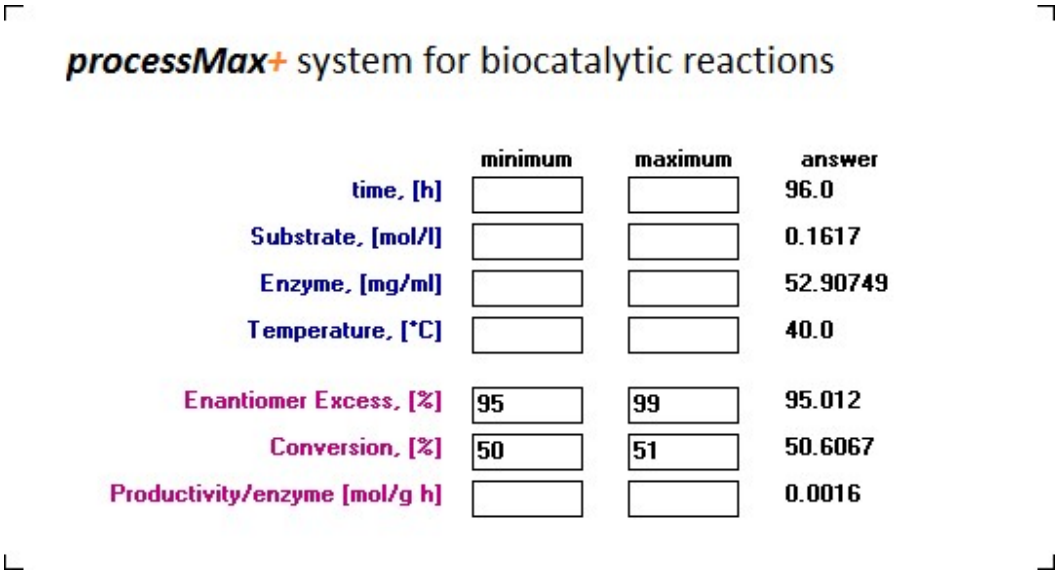


Figure 12. Determining suitable values of process variables and feed characteristics to obtain product properties within desired ranges

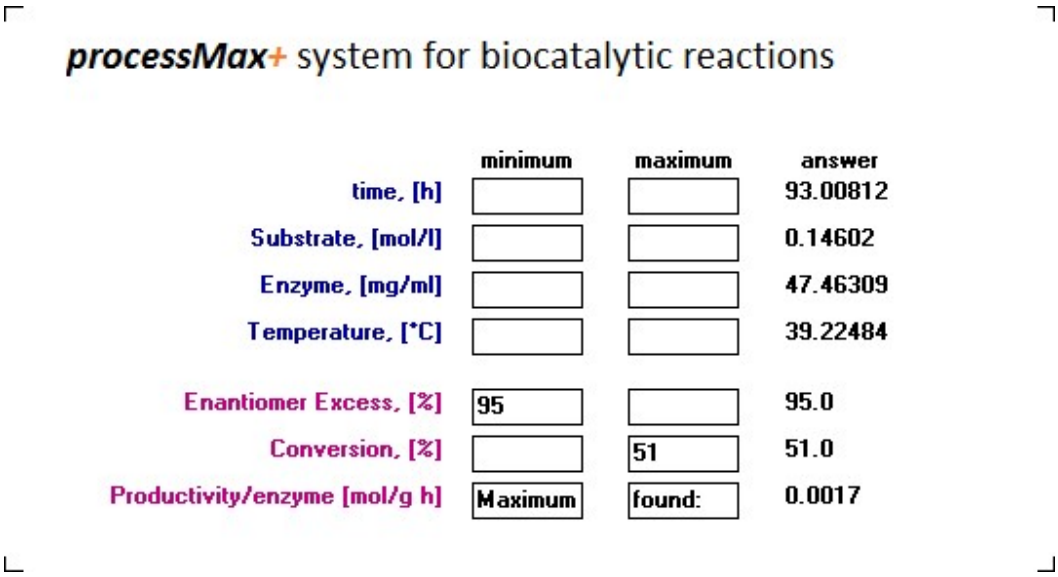


Figure 13. Determining suitable values of process variables and feed characteristics to obtain product properties within desired ranges, and maximising a production economic variable

Conclusions

Nonlinear modeling can contribute a lot in biotechnology sector industries. New technologies come up from time to time which affect the production economics, and open up new possibilities. Those companies which utilize new technologies effectively have an edge over those which do not. Nonlinear modeling is a new technology which helps production units derive more out of their equipment while also improving product properties. If used effectively, it can add to a company's competitiveness. Nonlinear modeling is expensive, but as experience shows time and again, the benefits clearly outweigh the costs.

The best of the scientists and engineers can perform better process development with less experimentation effort by using nonlinear modeling. Deriving maximum mileage out of the reactor or the fermenter or a separation equipment is an optimization task, for which, it is necessary to have detailed quantitative knowledge of its operation. Such knowledge can be summarized in the form on nonlinear empirical or semi-empirical models describing the effects of process variables and feed characteristics. It has become feasible to utilise this new technology in practice.

Development of mathematical models fit for industrial use, of biochemical processes in general, including bio-catalytic reactions, has been considered to be very difficult. However, new techniques of nonlinear modeling have clearly improved this situation, as results from this work demonstrate. In the case described in this article, the nonlinear models were an order of magnitude better than linear models for the same two variables. The nonlinear models show good statistical characteristics, covering about 98% of the variance for enantiomer excess and 96% for conversion. The nonlinear models also show qualitative features as are expected.

Since these nonlinear models can be quite complicated, and engineers and scientists cannot be expected to be experts on optimization techniques, a simple tool has been devised bringing these new possibilities within the reach of engineers and chemists.