18 Brandow, S. L., Turner, D. C., Ratna, B. R. and Gaber, B. P. (1993) *Biophys. J.* 64, 989–902
19 Müller, W. T., Klein, D. L., Lee, T., Clarke, J., McEuen, P. L. and Schultz, P. G. (1995) *Science* 268, 272–273
20 Bourdillon, C., Demaille, C., Moiroux, J. and Savéant, J. M. (1994) *J. Am. Chem. Soc.* 116, 10328–10329
21 Martin, C. R. (1994) *Science* 266, 1961–1966
22 Braco, L., Dabulis, K. and Klibanov, A. M. (1990) *Proc. Natl Acad. Sci. USA* 87, 274–277
23 Ekberg, B. and Mosbach, K. (1989) *Trends Biotechnol.* 7, 92–96
24 Sagiv, J. (1979) *Isr. J. Chem.* 18, 346–353
25 Nowick, J. S., Feng, Q., Tjirikua, T., Ballester, P. and Rebek, J. (1991) *J. Am. Chem. Soc.* 113, 8831–8839
26 Oberholzer, T., Wick, R., Luisi, P. L. and Biebricker, C. K. (1995) *Biochem. Biophys. Res. Commun.* 207, 250–257
27 Leibsle, F. M., Murray, P. W., Francis, S. M., Thornton, G. and Bowker, M. (1993) *Nature* 363, 706–709
28 Radmacher, M., Fritz, M., Mansma, H. G. and Hansma, P. K. (1994) *Science* 265, 1577–1579
29 McIntyre, B., Salmeron, M. and Somorjai, G. A. (1994) *Science* 265, 1415–1418
30 Yanagida, T., Harada, Y. and Ishjima, A. (1993) *Trends Biochem. Sci.* 18, 319–324
31 Hagen, A. J., Hatton, T. A. and Wang, D. I. C. (1990) *Biotechnol. Bioeng.* 35, 955–965
32 Hagen, A. J., Hatton, T. A. and Wang, D. I. C. (1990) *Biotechnol. Bioeng.* 35, 966–975
33 Roberts, S., Domurado, D., Thomas, D. and Chopineau, J. (1993) *Biochem. Biophys. Res. Commun.* 196, 447–454
34 Kabakov, V. E., Merker, S., Klyachko, N. L., Martinek, K. and Levashov, A. V. (1992) *FEBS Lett.* 311, 209–212
35 Garza-Ramos, G., Tuena de Gomez-Puyou, M., Gomez-Puyou, A. and Gracy, R. W. (1992) *Eur. J. Biochem.* 208, 389–395
36 Chang, G. G., Huang, T. M., Huang, S. M. and Chou, W. Y. (1994) *Eur. J. Biochem.* 225, 1021–1027
37 Ryabov, A. D. *et al.* (1992) *Angew. Chem. Int. Ed.* 31, 789–790
38 Garza-Ramos, G., Tuena de Gomez-Puyou, M., Gomez-Puyou, A., Yuksel, K. U. and Gracy, R. W. (1994) *Biochemistry*, 33, 6960–6965
39 Levashov, A. V., Rariy, R. V., Martinek, K. and Klyachko, N. L. (1993) *FEBS Lett.* 336, 385–388
40 Drexler, K. E. (1994) *Annu. Rev. Biophys. Biomol. Struct.* 23, 377–405

GMP – Good Modelling Practice: an essential component of Good Manufacturing Practice

Douglas B. Kell and Bernhard Sonnleitner

There is much interest in the exploitation of more-or-less sophisticated mathematical methods combining the signals from different sensors for measured variables for estimating the present or future state of bioprocesses. In many cases, however, the application of these methods has failed to take into account many important principles. We therefore summarize what we consider to be some of the key issues, and present them in the form of a guide for assisting the development of these methods and their integration into mainstream biotechnology.

The growing interest in exploiting mathematical methods for estimating the present or future state of bioprocesses (Box 1) is largely contingent on the (albeit increasingly untenable[1–5]) belief that direct on-line sensors for many variables of interest are still lacking, and that estimation of these variables must be made by indirect 'soft sensing' methods[6]. An introduction to sensing for bioprocesses may be found in the books of Leigh[7] and of Bastin and Dochain[8], and in the article of Lübbert and Simutis[9]; Montague and Morris[10] have recently reviewed the artificial neural network (ANN) approach to these and related problems (see also Refs 11–13). Similar mathematical methods are widely used in analytical and computational chemistry, where the field is known as chemometrics[14], and in the related field of bioinformatics. However, many bioprocess models pay scant attention to determining valid inputs and outputs of such multivariate calibration models, and rarely attempt to relate the inputs and outputs to the biological processes involved, despite the central importance of this in cause–effect studies, and the performance of descriptive models applied to process control and state estimation. In addition, adequate criteria for evaluating model quality are often absent.

D. B. Kell is at the Institute of Biological Sciences, University of Wales, Aberystwyth, UK SY23 3DA. B. Sonnleitner is at the Institut für Biotechnologie, ETH Zürich Hönggerberg, CH-8093 Zürich, Switzerland.

 0167 – 7799/95/$9.50

Box 1. Why sense and model?

- Biomass or product yields vary between runs because of changes in the values of the parameters that control the process.
- All parameters have optimal values for the particular process.
- Optimal values may be time-dependent for nonstationary processes.
- To control the parameters at their optimal values, it is necessary to measure them, whether directly or otherwise.
- On-line and at-line sensors can give rapid, accurate and plentiful data, without artefacts related to sampling.
- The cost of on-line controllers is a very small fraction of the cost of an industrial production plant, the medium, the downstream-processing equipment and the personnel required to run the system.
- Historical data may be combined with mathematical methods to provide identification of the most important parameters and their optimal values, for potential use in control loops, for process modelling and thus for rational process improvement.

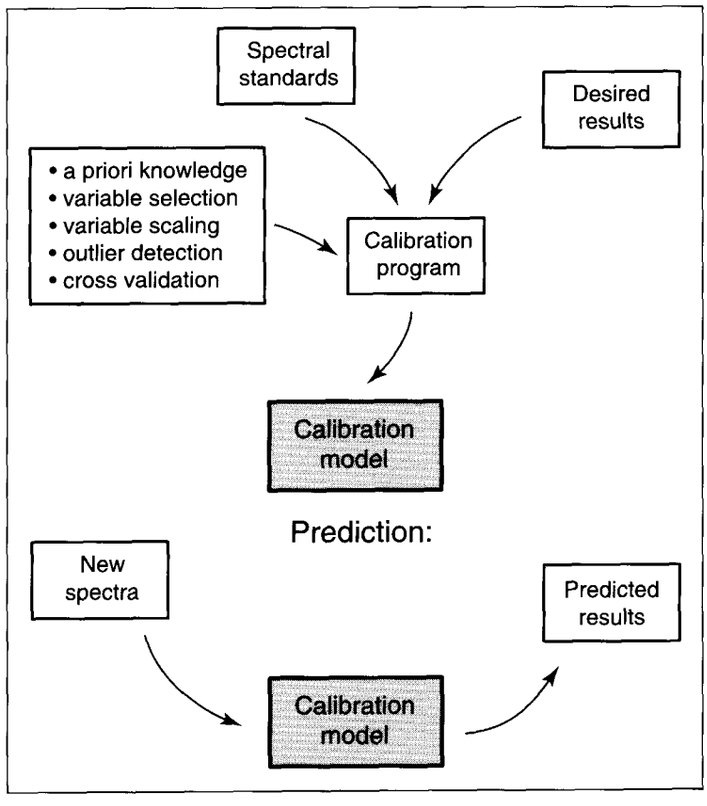


Figure 1

The process of multivariate calibration via supervised learning consists of two stages: (1) calibration involves establishing a calibration model for later prediction of results (e.g. the amount of a determinand) from spectra or other multivariate inputs by matching data from spectral or other standards (X-data) and the known results wanted (Y-data) from a set of calibration samples (the training set) in a supervised-learning calibration program; this is typically done using a neural network simulation, or with a program that performs multiple linear regression, partial least squares regression or principal components regression. It is sensible to exploit some deduced knowledge in the formation of the model, while variable selection and scaling, outlier detection and cross validation are all important steps in the development of a successful model. (2) Prediction involves converting comparable multivariate data for new samples into predictions of the output of interest (e.g. in terms of determinand concentration) using the previously established calibration model in the computer program.

Developments in this field suggest that it is no longer sufficient that methods 'work' (i.e. have good descriptive, if not necessarily predictive, ability) but optimal modelling approaches must now be selected, compared in a logical manner and reported more carefully. Careful consideration of modelling requirements will assist continuing development of these methods and their integration into mainstream biotechnology.

Figure 1 shows the basic principle of multivariate calibration[15], as applied to predictive modelling, for a case in which pyrolysis mass spectra are the inputs[16]. The principle is that it is possible to relate multiple, but often featureless, inputs that are relatively easy to measure (X-data) to something of interest (Y-data) that cannot easily be measured directly. Pairs of related X- and Y-data are subjected to an iterative algorithm, typically incorporating latent variables of some kind, while avoiding overfitting (fitting to noise) or the fitting of a model to outliers, to determine an optimal mathematical relationship between them. (Although the X-data cannot be assumed to be error-free, most models are constructed and used based on this assumption.) The multivariate calibration model may then be interrogated with new X-data to predict the Y-data of interest. These modelling approaches may be applied to systems in which all X-data for a given sample are obtained at the same time, or to cases in which the outputs of various sensors producing the X-data are used to predict the behaviour of the system at a later time.

Parameters and variables

Parameters are the properties of a system that are time-invariant over the period of observation, and which are inherent to the system. In some instances, one or other property is held constant by applying a closed-loop controller. Variables are those properties of the system that vary over time, and whose dynamic and steady-state properties are, therefore, determined by the values of the parameters. The distinction between parameters and variables is particularly clear cut in Metabolic Control Analysis (MCA; Refs 17–20). Here, enzymatic rate constants, the concentrations of enzymes and of external substrates and effectors that remain constant (and therefore permit the maintenance of a steady state) are examples of parameters. Variables typically include metabolite concentrations and the fluxes through metabolic pathways of interest.

For bioreactors, parameters include operational properties such as the pH, temperature or oxygen partial pressure (if controlled), and the composition of the medium (which can be known only if it does not contain any complex component), and its feed-rate (if in fed-batch or chemostat mode). Biological properties such as the genetic and enzymatic make-up of the inoculum are also parameters, although these are generally ill-defined[21], and can be kept reasonably unchanged only when pre-culture preparation follows a very strict standard operating procedure (SOP). Variables include levels of biomass (except in a turbidostat, where it is a parameter[22]) and metabolites, and the fluxes to metabolites of interest. Although culture

properties such as the specific or total oxygen-uptake rate (OUR) and CO_2-evolution rate are frequently referred to as culture parameters, they are, more accurately, culture variables. (Strictly speaking, they are dependent variables; only time is an independent variable.) Therefore, to understand system behaviour, it is of prime importance to make a correct distinction between parameters and variables. Our first principle (Box 2) is therefore:

> correctly identify the parameters and variables of the system under study (see Fig. 1 in Ref. 23), whether measurable or not. Regard every property as a variable and, only after definite identification of their (relevant) time-invariance, rename those fulfilling this criterion as parameters.

Forward and inverse problems

A forward problem is one in which the parameters and starting conditions of a system, and the kinetic or other equations that govern its behaviour, are known. Any of several simulation packages may be used to describe the transient and steady-state behaviour of such a system, taking the parameters and starting conditions as the inputs and the variables as the outputs. However, especially in a complex biological system, it is the variables, not the parameters, that are easiest to measure, and it is the variables that depend on the parameters.

In metabolic reactions, the usual parameters of interest are the enzymatic-rate and affinity constants, which are difficult to measure accurately *in vitro*, and nearly impossible so to do *in vivo*[24–27]. However, to describe, understand and simulate the system, the parameters must be determined; it is necessary to work back from variables such as the steady-state fluxes and metabolite concentrations, which are relatively easy to measure, to the parameters. Such problems, in which the inputs are the variables and the outputs the parameters, are known as inverse problems, or system-identification problems.

There is no cognate logical structure to describe a model in which parameters and variables appear as both inputs and outputs. Our second principle is:

> identify whether the problem is forward or inverse, and the validity of studying a mixed model in which variables are used to predict other variables.

However, such 'variable-predicts-variable' (covariance) models are widespread, and a model should also be judged by how well it serves its purpose, even if its theoretical foundations are weak.

Multivariate calibration models and artificial neural networks

While using such methods is straightforward, several issues must be addressed to produce a good predictive model that generalizes well. Such considerations apply to statistical chemometrics methods such as partial least-squares regression, as well as to approaches based on artificial neural networks (ANN; Refs 28–30; Fig. 2). Box 3 lists the issues that should be considered in ANN production. Multivariate calibration methods of this type, at least in the short term, are not likely to be fully 'plug-and-play' (unless designed for specific purposes), but it illustrates how many factors can be 'tweaked' during the development of a good predictive model. No single study seems to have adequately addressed all these issues for any real dataset, and most papers do not consider them at all. Therefore, our third principle is:

> in reporting experiments of this type, state, as far as possible, what the parameters used were, their values, and what the outcomes were in terms of speed of learning and precision of prediction.

There are three important points from Box 3:

- The ability of these methods to form models that can effect reasonably accurate mapping of multivariate inputs to outputs of interest is remarkable. It is possible to generate vectors of random data as the inputs, a series of 'sensible' numbers (0, 5, 10...100) as the outputs, and get a model that forecasts the outputs more-or-less perfectly. In this extreme form of overtraining, the model learns the training set, and the chief approach to its solution is to reserve some of the training data as test or cross-validation examples, stopping training when acceptably accurate learning of the test or cross-validation data ceases. In a random noise example, one would see that adequate learning does not occur. It is, therefore, extremely desirable to show

Box 2. Some recommendations for Good Modelling Practice

- Correctly identify the parameters and variables of the system under study, whether measurable or not.
- Identify whether a problem is true forward, or inverse, and to what extent it makes sense to study a mixed model.
- Describe which inputs and outputs were chosen for the model, and why.
- In reporting the outcome of predictive modelling experiments of this type one should attempt to state what parameters were used, their values, and what the outcomes were in terms of speed of learning and precision of prediction.
- Data should always be plotted to show the relationship between the predictions and the 'true' (gold standard) data.
- Show whether the predictive time-series model can do significantly better than that based on a first- or second-order trivial predictor.
- Any nonlinear predictive model should always be compared with the performance of the best types of linear multivariate calibration models for the same dataset.
- Provide quantitative data and performance criteria for predictions so that the effectiveness of models can be assessed.
- Do not claim general superiority of any particular algorithm or predictive modelling approach when it has been tested on only a limited dataset.

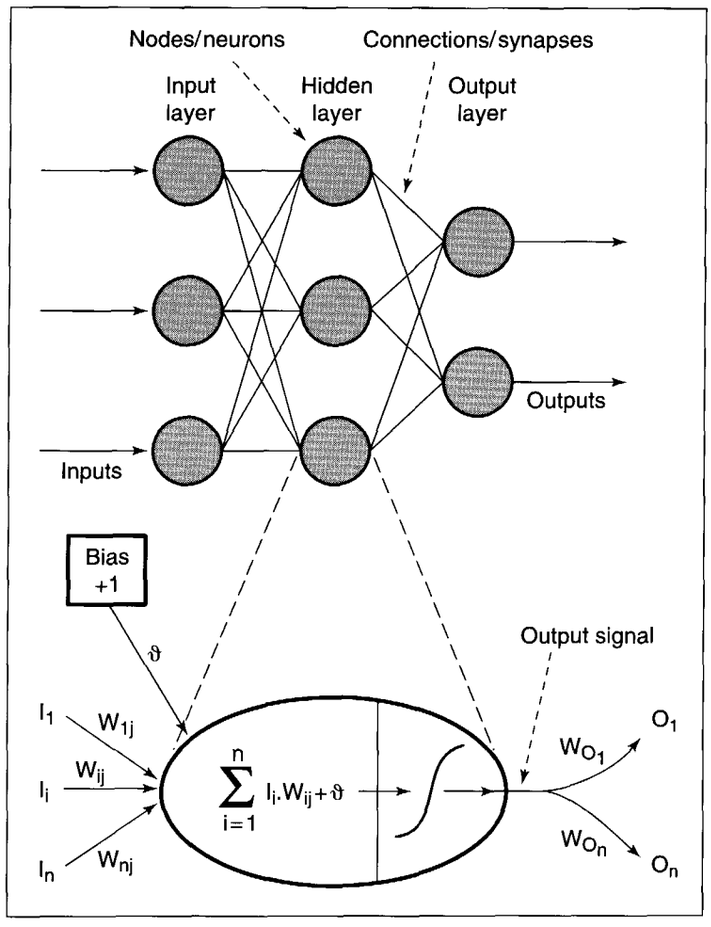


Figure 2

An artificial neural network (ANN) comprising three inputs and two outputs connected to each other by a hidden layer of three nodes. In the architecture shown, adjacent layers of the network are fully interconnected, although many other architectures are possible. One of the nodes in the hidden layer is given in more detail showing its method of information processing. An individual node sums its input (the Σ function) from nodes in the previous layer, including the bias, transforms them using a sigmoidal squashing function, and outputs them to the next node to which it is linked via a connection weight. The relevant principle of supervised learning in ANNs is that, as with the multivariate calibration described in Fig. 1, the ANNs take numerical inputs (the training data, which are usually multivariate) and transform them into desired (predetermined) outputs. The input and output nodes are connected to the 'external world' and to other nodes within the network. The way in which each node transforms its input depends on the so-called connection weights (or connection strength) and bias of the node, which are modifiable. The output of each node to another node or to the external world depends on its weight strength and bias, and on the weighted sum of all its inputs; these are then transformed by a, usually nonlinear, weighting function referred to as its activation, threshold or squashing function. As with other supervised learning methods, the great power of neural networks stems from the fact that it is possible to train them. ANNs can be trained by using sets of multivariate data from standard materials or systems of known identities or properties as the desired outputs. Training is effected by continually presenting the networks with the known inputs and outputs, and modifying the connection weights between the individual nodes and the biases. This is typically done according to some kind of backpropagation algorithm[84–86], until the output nodes of the network match the desired outputs to a stated degree of accuracy. The trained ANNs may then be exposed to unknown inputs (spectra), and they will 'immediately' provide the optimal best fit to the outputs. If these outputs are accurate, the network is said to have generalized. Therefore, the 'knowledge' of the neural network is contained in the values of the weights.

the learning curves [plots of the root mean square (RMS) errors of prediction (RMSEP) versus number of iterations] for training and test data. It is even more important to check the consistency of the data prior to feeding them into an ANN. Similarly, extrapolation of a model to use input values beyond the bounds of those encompassed in the training set is dangerous. Simutis *et al.* recently suggested a method to quantify this danger by determining an extrapolation measure[31]; they perform a cluster analysis of the training data and determine the centres of gravity of the individual clusters. A Gaussian distribution around these centres gives an estimate of the reliability of a forecast in the respective data space. Lodder and Hieftje used a similar approach for near infrared spectroscopy[32].

- It is hard to visualize and extract features from complex, multivariate data[33]. Relatively small changes in an algorithm or procedure can have marked effects on the performance of these supervised learning methods. In one study[34], we showed that appropriate scaling of the inputs to a neural network could improve its learning rate by more than two orders of magnitude, illustrating why it is so important to specify all parameters of models accurately and explicitly.
- It is widely considered that these methods have the power to deal with irrelevant inputs: i.e. if an input does not contribute to a good model, then the network, or program, will recognize this and set the relevant weights to zero; the model's forecasting ability will be unaffected and the only cost will be an increased computation time. This is incorrect; chance correlations always occur in high-dimensional space, and neural and multivariate calibration methods will always seek to fit to the chance correlations[35,36]. Cross-validation can only help to decrease the problem; chance correlations are still correlations, albeit poor ones, and the addition of input variables that are essentially irrelevant will always degrade the model[37], in our experience substantially. Such models may be expected to have poor forecasting power when there is no reasonable mechanistic background to their construction. The only solution is to test models of varying complexity to ensure that the parameters and variables used as inputs genuinely contribute to a good model[38].

However, it is important to include more input data if new, valuable, qualitative and/or quantitative information can then be derived. A variable that is useful for characterizing the physiological state, for example, is the respiratory quotient (RQ), which can be calculated simply from gas-balance data (O_2 and CO_2 in exhaust gas), and pH. Together, but not alone, these variables describe an important aspect of the physiological state (although one should be aware of potentially dramatic pitfalls[39,40]).

Understanding cause→effect relationships

High-performance bioreactors and on-line process analyses are essential tools for the identification and quantification of the fine regulation of metabolism for subsequent application in noninvasive measurement

Glossary

Artificial neural networks (ANNs) – are composed of a series of interconnected, simple elements operating in parallel and mimicking the biological nervous system. It is the connections between the individual elements that largely determine the network function. ANNs must be trained before they can perform their function.

Backpropagation – (more accurately known as backpropagation of error) is a method of altering the weights in an ANN in a manner that depends on the difference between the actual output of a layer of a neural network and the desired values for that output.

Black-box models – models in which the effects on the outputs of changing the inputs are studied, and equations that quantitatively describe these relationships can be derived, without any attempt to provide mechanistic explanations in terms of internal causative factors or properties of the system.

Endogenous metabolism – occurs where a fraction of the cell material is used as substrate, for instance to cover maintenance requirements for which no extracellular substrate is available.

Fuzzy reasoning – permits exploitation of verbally formulated information that is not crisp in a quantitative sense but allows classification of unsharp or fuzzy terms such as 'fairly dense', 'quite low', 'very rapid'. The rules and algorithms produce an output that is a given degree of certainty of membership in such a class; this output can either be used directly (e.g. for pattern recognition) or it must be de-fuzzified (converted into a crisp number) before the signal is acted upon. Fuzzy reasoning reflects the problem–solution approach 'what would you do if...' more than any numerical algorithm, and can therefore be developed much more rapidly, and less expensively.

Grey-box models – are intermediate between black-box and white-box models.

Inputs and outputs – inputs are observables that affect and determine the behaviour of a system that produces certain reactions, namely the outputs. Inputs and outputs are the external elements with which a system interacts. The system is a part of the 'real world' and has internal properties; all properties that are neither internal nor input nor output properties are external properties and irrelevant to the system. The definition of the system boundaries is crucial, but subjective. A model is a simplifying representation of a system; all factors, e.g. system definition, type, complexity and formulation of a model are subjective. Therefore, modellers should always give plausible reasoning for their models, as there cannot be a proof in a mathematical sense.

Maintenance metabolism – occurs in a substrate when a fraction of the extracellular substrate consumed is not directed to growth.

Monod-type models – these always relate a volumetric reaction rate to a specific rate (q) and the concentration of the biocatalyst. The specific rate depends on the concentration of a single limiting substrate (s) in a hyperbolic manner:

$$q_s = \frac{s \cdot q_{max}}{s + K_s}$$

where q_{max} is the maximal specific rate and K_s is a saturation parameter. This function passes through the origin of coordinates and becomes practically independent of s at high values of s (q tends towards q_{max} if $s >> K_s$). Therates of substrate conversion and product formation are linearly correlated by a yield coefficient (Y). This model is very simple as it needs only two state variables (s and x) and three parameters (q_{max}, K_s and Y); however, it does not accurately describe realistic systems behaviour and many modifications and extensions have therefore been proposed.

Multivariate calibration – involves establishing a calibration model for later prediction of results from multivariate inputs, by matching data (X-data) from known standards together with their values or properties (Y-data), using a set of calibration samples (the training set) in some sort of supervised learning calibration program.

Structured and segregated models – in classical models such as that of Monod, biomass concentration is a single state variable, and is thus in the form of a scalar. In structured models, an insight into the cells is claimed, and more detailed subcomponents of the cells are described as state variables; biomass concentration is a state variable vector with as many rows as the number of cellular components that are considered. The population is treated as homogeneous. In segregated models, it is recognized that not all cells are the same: segregated models distinguish subpopulations of cells, and biomass concentration is again a vector, with as many columns as there are subpopulations considered. In structured, segregated models, biomass becomes a matrix of state variables, reflecting the distribution of cellular components between the different subpopulations.

White-box models – are those in which the effects on the outputs of changing the inputs are studied, and measurements simultaneously made of the values of relevant internal parameters of the system. Equations that quantitatively describe the relationships between internal and external properties are established, so that mechanistic explanations are provided in terms of internal causative factors or properties of the system.

Box 3. Some parameters that may be varied to improve learning, convergence and generalization during the production of a feedforward backpropagation neural network calibration model

- Number of hidden layers
  One is thought sufficient for most problems; more increase the computational load.
- Number of nodes in hidden layer
  A simple rule of thumb says ln (natural log; number of inputs). More detailed analysis suggests an interaction between the number of input variables and the number of examples, and the actual complexity of the data.
- Architecture
  Fully interconnected feedforward nets are most common, but many others exist, including the Boltzmann machine, direct linear feedthrough and Hopfield networks.
- Number of exemplars in training set
  Sufficient are required to fill the parameter space and to allow generalization. When fewer are used, the network can store all the knowledge and overtrain. Examples should be selected on the basis of the quality of their data.
- Number of input variables
  Those that do not contribute positively to discrimination may impair generalization and are best removed by pruning the input data.
- Scaling of input and output variables
  Individual scaling on inputs can improve learning speed dramatically. There is a need to leave headroom, especially on the output layer. Non-numeric outputs may still be encoded numerically.
- Updating algorithm
  There are many variants on the original backpropagation (BP), most of which give small improvements, but standard backpropagation[84–86] is still the most popular.
  Various 'squashing' functions may be used for each node.
- Learning rate and momentum
  These may need to be carefully chosen so that the network does not get stuck in local minima or shoot off in the wrong direction when encountering small bumps on the error surface. For standard BP, a learning rate of 0.1 and a momentum of 0.9 are often best.
- Stability/generalization
  It is important to determine when to stop training, and which cross-validation scheme to use.
  Insensitivity of parameters may necessitate iterative process, e.g. including revision of network design. It is best to reserve some of the training data for cross-validation.

and process control. The best theoretical basis for such investigation includes structured and segregated models[41] – the only models able to predict outputs of biological systems on the basis of chemically and biologically reasonable cause–effect chains.

The dynamic behaviour of living organisms cannot be investigated satisfactorily with classical reductionist concepts alone: the *in vitro* kinetic analysis of reconstituted subsystems, incorporating genes, enzymes, substrates, co-substrates or products, represents an oversimplified system[23,27]. In part, this is why the majority of serious problems associated with metabolic engineering (see Ref. 42) are the unwanted deregulation of metabolism, poor genetic stability of the organisms, and a need for re-evaluation of optimal micro-environmental conditions. Metabolic networks are usually quite rigid[43], and the interrelated control mechanisms can cause an apparently straightforward modification to yield unexpected results[44]. The understanding of such networks cannot be improved by applying descriptive models with constrained forecasting power. However, mechanistic models can work well, as they are predictive. It is simply a question of good experimental planning to achieve reasonable falsification or verification of mechanisms. The underlying mechanisms must imply the 'right' inputs and outputs; in other words, the inputs affect the outputs in a straightforward way, although the outputs of one chain may well be inputs to others. Primary inputs are the operational and biological parameters of a system, whereas the outputs are the dependent variables.

An example that may be used to illustrate the importance of selecting the 'right' cause–effect mechanisms is an extension of the Monod growth model with a term correcting for what is generally called maintenance-energy requirements. This is widely documented (for example, see Ref. 45), but the formulation of the problem is not trivial. Consider two possibilities:

(1) Cells grow with a specific growth rate, given by

$$\mu = \frac{\mu_{max} s}{(s + K_s)}$$

in which μ_{max} is the maximum growth rate, s is the concentration of a single limiting substrate, and K_s is the saturation parameter. The specific substrate-consumption rate, (q_s), must then be calculated from μ, for instance, according to Pirt[45]:

$$q_s = \frac{\mu}{Y} + m$$

where Y is the 'true' growth yield for the 'limiting' substrates and m_s a growth rate-independent maintenance energy. As s approaches zero, μ approaches zero, but q_s must remain at the non-zero value of m_s. This means that substrate is being consumed, although there is no substrate available – a doubtful result.

(2) Substrate consumption is the cause and growth is the effect. It is reasonable to choose formulations such as:

$$q_s = \frac{q_{s,max} s}{(s + K_s)}$$

and calculate the specific growth rate as

$$\mu = (q_s - m_s)Y$$

As s approaches zero, q_s also approaches zero, and μ must take a negative value; in other words, cell-mass concentration is predicted to decrease according to the maintenance requirements as endogenous metabolism takes over – a plausible result.

Obviously, even for this simple example which

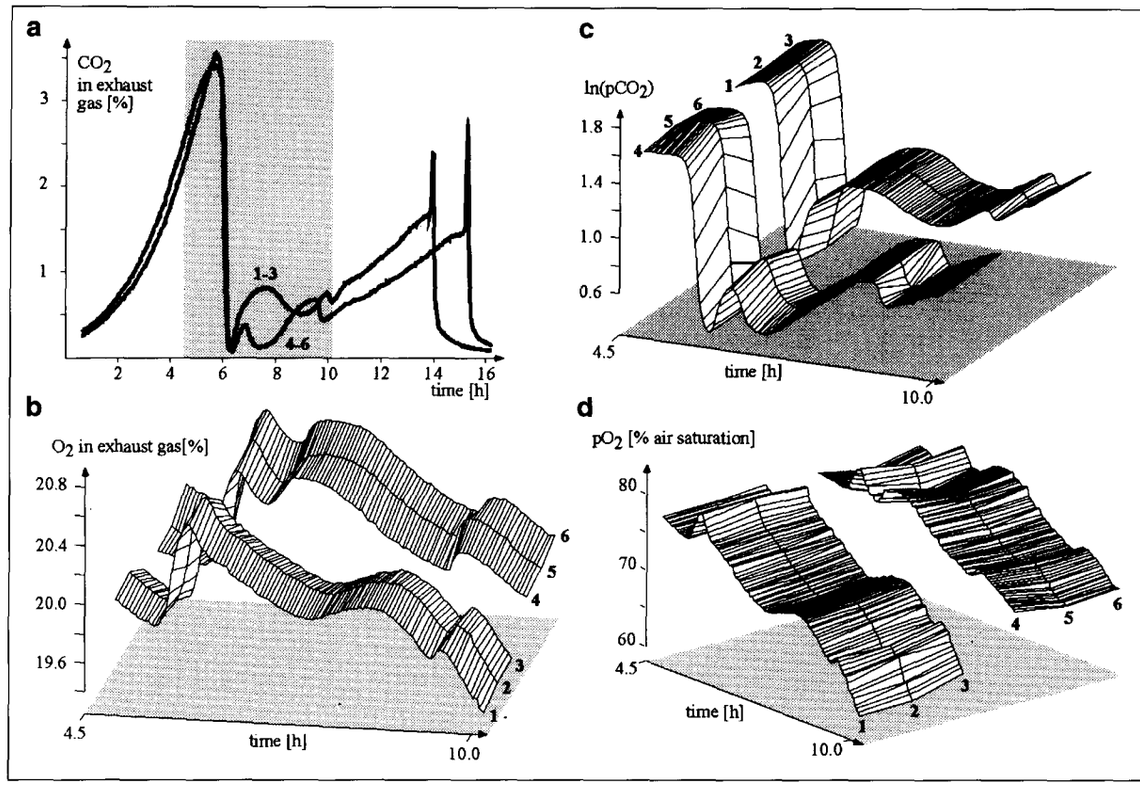


Figure 3

Time trajectories of the (**a**) CO_2- and (**b**) O_2-fraction in the exhaust gas, (**c**) logarithm of pCO_2 and (**d**) pO_2 during six cultivations of *Saccharomyces cerevisiae*. The zoomed period extends from the very last phase of growth on glucose (approximately 5 h) to the early phase of growth on ethanol (around 10 h). Glucose depletion causes a dramatic reduction in CO_2-formation and a less pronounced increase of the O_2-fraction and partial pressure. The various bumps, which differ significantly among the two sets of experiments, are due to the re-consumption of metabolites excreted during growth on glucose. Cultures 1–3 were grown on medium that was prepared by sterilizing salts and concentrated glucose solutions separately, while for cultures 4–6 the complete medium was heat-sterilized and acidified to pH 3 ± 0.1; in all cases, vitamins were filter-sterilized and added after cooling. All other operational conditions remained unchanged: [T: 30.0 ± 0.05°C, pH: 5.0 ± 0.02, aeration: 1.00 ± 0.01 vvm, pressure: 1.02 ± 0.01 bar(abs), stirrer and foam separator: 2000 ± 10 rpm each; NaOH and H_3PO_4 (for pH control): 10 ± 0.5%, starting contents: 5.00 ± 0.02 kg (= 80% filling of the compact loop reactor); inoculum: 5 ± 0.5% grown on the same defined medium transferred 30 ± 2 min after ethanol depletion (repetitive batch-culture technique), with media for each set of experiments prepared in a single batch and stored at 15 ± 1°C].

ignores transients, the second version is preferable.

It may be possible to circumvent this problem by cancelling out or eliminating all mechanistic-model ideas, including rate equations and yield coefficients, and working only with relationships. While there is still some work to be done before such white-box models can be formulated[46], all available background and process information should be collected and exploited, and there is more available than is generally believed.

Biology is sometimes considered to be an inexact science. It is thought that it lacks necessary reproducibility, and that the formulation and parametrization of detailed models for such complex systems is of little relevance for practical applications. Review of several examples has, however, shown that a satisfactory precision in the reproducible evolution over time of all the variables of interest in a series of cultivations could be obtained only after significantly improving the accuracy and precision of the measurements and of simple individual control loops[23]. Therefore, one should not aim to seek improvements by constructing a predictor–controller to cope with the fuzzy biology. Rather, an appropriate SOP should be devised, or better equipment used. It has become clear that major efforts are needed 'towards more measurement in biology' (Ref. 47).

Figure 3 shows the time trajectories of two state variables from two sets of experiments that were measured during the mid-phase of several runs of a simple aerobic cultivation of baker's yeast; namely, from the final depletion of glucose (the collapse of CO_2 evolution) to the start of ethanol utilization (in between, the cells reuse by-products that are excreted during growth on glucose; the first and last phases are identical and not shown[48]). Two distinct patterns are obvious. In this case, we can present the cause→effect conclusion: the only difference among these experiments was that in one set, the complete medium was sterilized, and in the other set, the glucose and salts fractions were combined after separate sterilization, a difference that might, a priori, have been considered trivial. Thus, the differences between the data sets are

a result of systematic methodological differences rather than irreproducible biology. Although it is possible that the usual type of algorithm could predict the actual behaviour, based on the similar traces in the first part of the cultivation, any predictive method will perform much better if it is given information on the preparation of the medium. Clearly, biological systems are nonlinear and complex, their time evolution is very sensitive to initial conditions, and one must find the real causes of any unexpected variation, and not just try to cure the symptoms. "The sins of the parents, one might say, extend even unto the third and fourth generations" (Ref. 49).

Select the 'right inputs'

As carbon and energy sources, sugars, especially glucose, fructose and their derivatives, have regulatory actions at two different levels: regulation of gene expression (induction, repression or derepression of transcription); and modification of enzymes (activation, inhibition, dis/association of subunits, covalent modification or digestion). The underlying reactions are basically well known, but not always well understood at the molecular level for organisms of industrial interest. Substrates are the primary effectors for the intact cell, representing the driving force for activating or arresting mass transformations. In addition, reaction rates or the rates of change of metabolite concentrations[50] are of paramount importance for determining the overall physiological state. Indeed, the dynamics of metabolism are often underestimated within the intact cell; glycolytic oscillatory control may well be of the order of fractions of seconds[51], and will normally be missed in nonsynchronous populations[49].

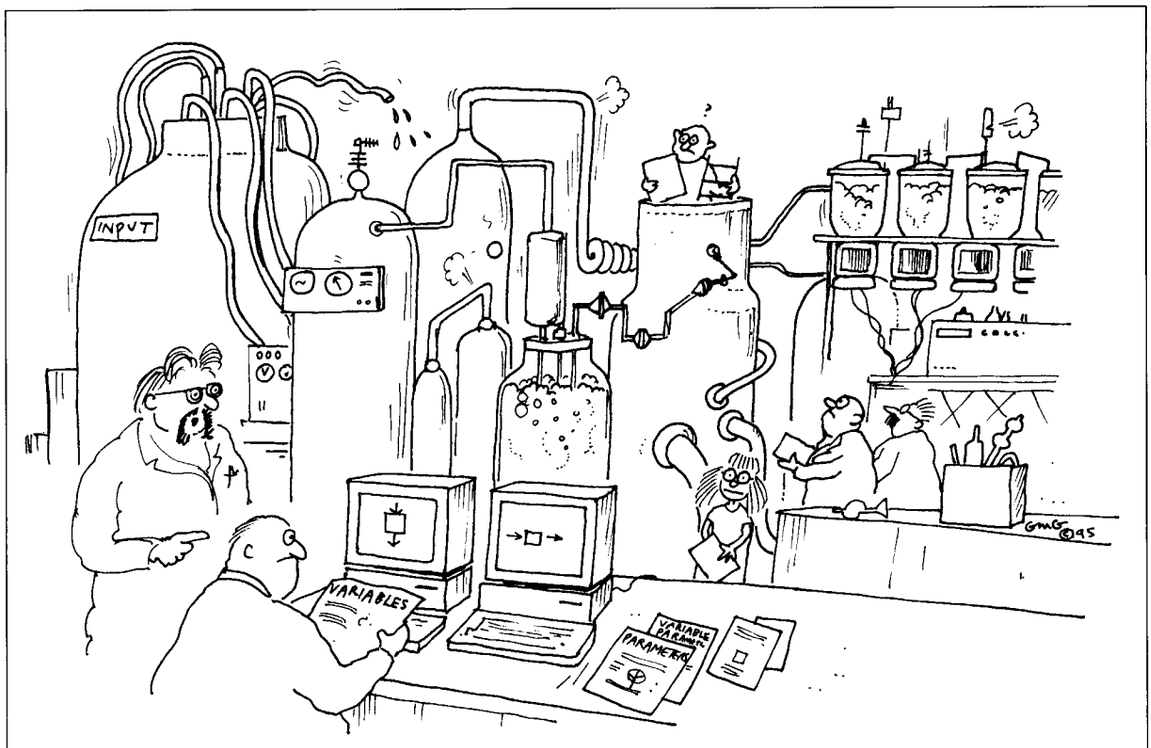
The pulsed addition of glucose to a continuous, carbon-limited culture of a wild-type baker's yeast results in a dramatic decrease in the intracellular concentration of ATP within only a few seconds. [This change was as predicted by Nielsen *et al.* (Ref. 52), but has only recently been demonstrated because experimental verification methods have only very recently been developed and applied[53,54].] On a longer time scale, within a few tens of seconds, the culture fluorescence (reflecting both biomass and the extent of pyridine nucleotide reduction), CO_2 fraction in the exhaust gas, and extracellular ethanol concentration also increase[55]. Only later, within the first few minutes after glucose addition, does the intracellular cAMP concentration increase[56]. Furthermore, it is not specifically glucose that can signal the presence of carbon and energy. Ethanol, acetic acid or pyruvic acid could also be successfully and reproducibly used to force a yeast subpopulation to enter a new round of the cell cycle[57]. Ethanol may act as an effector over a period of days (Refs 58,59; H. M. Davey, C. L. Davey, A. M. Woodward, A. N. Edmonds, A. W. Lee and D. B. Kell, unpublished). This example illustrates that there is ample mechanistic knowledge about biochemical causes and observable effects; however, the mechanisms of information transfer from other nutrients, metabolites, by-products and major products are often still unclear. For this, the careful study of the physiology of defined mutant strains is helpful.

It is not easy to define the most relevant variables. However, simulation of the behaviour of model variables allows the sensitivity of variables (towards parameters or other dependent variables) to be tested over a wide region of parameter space[60]. Such results must be checked experimentally, but prior knowledge derived from modelling can help experimental design[61]. However, it is important to measure as many parameters and variables as possible at an early stage of process development. After identification of the relevant variables (and of those cultivation parameters that are variables held constant by control loops), those that are irrelevant can be dismissed.

Biological parameters and model structure

It is essential to decide a priori which environmental and operating conditions should be made parameters (i.e. held constant by closed-loop control), and with what precision. The physical and chemical parameters should be wide-ranging, and should include the quality of media and pre-cultures. However, all measurable bioprocess data must be treated initially as variables: decide on-line whether they remain constant within a predefined window (i.e. behave as a parameter). The resulting documentation is both necessary and sufficient – even for Good Manufacturing Practice. Besides these operational parameters, biological parameters are determined by the choice of the 'most suitable' biological system.

A wide variety of bacteria (e.g. *Escherichia*, *Bacillus* or lactic acid bacteria) excrete acetic or lactic acid when grown on readily available carbon sources; many yeasts produce ethanol, glycerol and short-chain fatty acids; and mammalian cells excrete predominantly lactate but also amino acids. Such by-product formation is a very common phenomenon under growth conditions of high specific substrate-consumption rates that are caused by a high concentration of extracellular substrate – the right input, albeit at the wrong concentration – and minimizing this is essential for process or product quality[62]. Many cells can adapt their metabolism quite quickly in this way: as only a little of the reducing power generated during catabolism is needed for the synthesis of new cells, the reducing equivalents can be scavenged by reacting with accumulated carbon intermediates[63]. As a consequence, much of the carbon flux through the initial catabolic steps, when extracellular substrate is in excess, must be excreted, usually as C_2- or C_3-moieties; this phenomenon is usually referred to as metabolic overflow. In a bioreactor, the organism can reconsume the excreted molecules if, later, the supply of the primary substrate is limiting. Thus, a major cause of variations in the metabolic pattern is the external substrate concentration, which one might want to hold constant by appropriate control actions, as in the 'nutri-stat' (Refs 64,65). Such strategies are very useful for optimizing feeding regimes.



"So much for prediction but what about reality!"

Given that by-product formation can cause the macroscopically observable growth-yield coefficients to vary significantly, the identification of 'variable parameters' in simple unstructured models must be a problem: the yield coefficient may be a parameter of a kinetic model, but no constant value can describe all experimentally observed data. Auxiliary yield coefficients [such as P/O (the amount of ATP formed per O atom reduced during oxidative phosphorylation), or Y_{ATP}, (the amount of cellular material made per mole or gram of ATP produced), which cannot yet be measured] may sometimes be introduced[66]. The simpler model must, then, be replaced by a structured one, even though the structure's basis may be simple such as the distinction between actively growing and passive (not necessarily dead) biomass[52,67], or the distinction between major catabolic pathways, such as the oxidative and reductive processes in yeasts[68]. The predictive power of a model improves dramatically after such a simple upgrade[41], as it is based on a plausible mechanism. While we recognize that there is an urgency for more efficient and pragmatic (fuzzy) approaches to promote the progress of scientific research and technological applications, and would reiterate that one criterion of a good model is that it works, we nevertheless conclude that mechanistic models should be as simple as possible, but as complex as necessary to acquire a desirable degree of predictive power, particularly outside the data space used for construction and validation of the model.

Measurements of performance and their comparison

Although modelling is primarily concerned with making predictions, many papers appear to neglect to provide a quantitative assessment of predictive quality. Even error bars or confidence intervals appear rather infrequently. The essential difference between accuracy and precision is seldom addressed for either variables or parameters.

When assessing the ability of these methods to predict time series, it is common to see graphs in which data for the predicted and true (measured using a 'gold standard' method) values are plotted against time or against sample number. This has the effect of making the predictions appear far better than they actually are. Data should always be plotted so as to show the relationship between the predictions and the actual (gold standard) data. If data are plotted in this way, it becomes obvious that one can and should establish quantitative measures of the predictive quality of different models in terms of, for example, (non)-linearity, slope, intercept, correlation coefficient and RMSEP. [It should be noted that standard linear regression methods make a variety of assumptions (in particular of the accuracy and normality of the 'gold standard' data) that are rarely obeyed[69–71].]

Similarly, the fact that a particular predictive method gives a certain precision of prediction is not of itself particularly informative. The simplest, and usually fairly accurate, method for predicting the value of the next datapoint in a time series from a 'well-behaved'

system is to assume that it is the same as the present datapoint, a principle known as the first-order trivial predictor. Similarly, in the second-order trivial predictor the value of data point $n+2$ is equal to that for point n plus twice the signed difference between the values of points $n+1$ and n. If complex neural and nonlinear predictive models can do no better than the first- and second-order trivial predictors, then they are a waste of time. It is unfortunate that authors only rarely (H. M. Davey, C. L. Davey, A. M. Woodward, A. N. Edmonds, A. W. Lee and D. B. Kell, unpublished) provide the opportunity to assess whether this is the case. Thus, authors should always show whether their predictive time-series model can do significantly better than those based on a first- or second-order trivial predictor.

Even if the trivial predictors can be improved upon, it is desirable to give a comparison of models, and so provide a comparative benchmark; this is also very rarely done. This problem was highlighted in the results of a survey by Prechelt and colleagues[72] of 400 computer science papers; they found that in less than half of those containing software engineering results that required validation was any experimental validation actually carried out. Related disciplines, such as optical engineering and neural computation, were only somewhat better. It is always worth comparing at least one linear multivariate statistical method with the predictions from a nonlinear model, such as those based on neural networks[73–76]. If the linear methods perform adequately, they have the advantage that it is much easier to interrogate the models so as to establish how they are working, which inputs and datapoints are important to the model, and so on. This is known as the 'assignment problem'. It is, therefore, recommended that any nonlinear predictive model should always be compared with the performance of the best types of linear multivariate calibration models for the same dataset, to determine whether or not they outperform them significantly.

Commercial reasons partly dictate that many forecasting/modelling papers of time-series analysis in bioreactors appear in which the ordinate, and sometimes even the abscissa, lacks a scale (and even an indication of whether it is linear or logarithmic). While we recognize that authors using data obtained in industrial plant are not allowed to state the yields of product, this problem is adequately solved by, for example, normalizing the ordinate to the range. If this field is to progress as a science, then as much information as possible must be provided about the findings that authors are reporting. Authors should provide quantitative data and performance criteria so that it is possible to assess whether their models are any good.

Care should be taken not to overgeneralize any claim of the superiority of one algorithm over another when the comparison is made using a limited number of datasets; the methods are still not well understood from a fundamental statistical point of view, and approaches that work well on one dataset may not on another. For instance, a number of papers on bioprocess estimation have claimed superiority (in terms of speed of learning) for a certain algorithm over the standard backpropagation algorithm. The claim was based on only a limited set of data; and for many, and maybe most, datasets, the inventors pointed out that this algorithm performs worse than standard backpropagation.

What should one sense?

What one should sense or measure is governed by at least two factors: what sensors are available, and how cost-effective they are compared with alternative off-line or manual methods, or with doing nothing. Such cost-effectiveness can be measured in terms of the overall productivity of the process (and the product quality and purity) relative to the cost of implementing the measurements, which is normally small for on-line measurements. Most processes implement pH control, but it is commonplace that few processes are as highly instrumented as they could be.

An example – soft and hard sensing of microbial biomass

Ignoring distributions in the physiological state between organisms[51], it can be assumed that the rate of growth or of a biotransformation of interest is proportional to the biomass content[19]; automatic, undelayed biomass estimation is therefore a prerequisite for proper process control and on-line state estimation. The physiological state of a population, and transitions between various physiological states – some dependent variables – are characterized by *specific* consumption or production rates (as estimated by models or software sensors) that can be evaluated only if the actual biomass concentration is known. Only on-line methods can provide sufficiently high measurement frequency to investigate the dynamics of the physiology of microbial cultures in transient experiments. Thus, on-line biomass estimation is also an essential tool for basic scientific research. On-line and automatic determination is more accurate and precise than all-manual methods because:

(1) sampling and sample-pretreatment artefacts are systematically reproduced, minimized or excluded, so they are no longer variable and can be compensated for;

(2) the frequency of measurements is increased and no longer restricted to working times or personnel availability; and

(3) the dependence on personnel is eliminated.

The cellular biomass concentration is one of the most important process variables in a bioprocess (for reviews, see Refs 4,77–80), and a variety of on-line probes for microbial biomass estimation are now available. [Despite this, it is still possible to read such comments as 'unfortunately, the only method available to determine biomass is by off-line laboratory analysis' (Ref. 81), and 'the state of available sensors is not much different from a decade ago' (Ref. 82).] Although measurements from on-line probes could be

combined with estimates of O_2 uptake rate (OUR) and CO_2 evolution rate (CER) to give values for the specific rate of O_2 uptake or the specific rate of CO_2 evolution, for example, which are much more sensitive indicators of physiological state than the overall rates that are usually measured, many authors try to estimate biomass on the basis of external variables such as OUR and CER. Given that the values of OUR and CER per unit biomass can easily vary fivefold or more as a function of changing physiological state, substrate levels and so on[48], it should be apparent that OUR and CER alone are unlikely to be terribly good predictors of biomass, especially if such estimates are to be used for calculating specific rates

$$q_{O_2} = \frac{OUR}{x}$$

and

$$q_{CO_2} = \frac{CER}{x}.$$

A particular problem with biomass estimations of this type is that the precision of the dry weight measurements normally used as the 'gold standard' tends to be significantly lower that obtainable from on-line probes[8,78]. Indeed, as the estimation of the 'gold standard' accuracy is so important, it is essential to promote a variety of alternative on-line methods for monitoring biomass concomitantly based on independent, different principles, and to cross-evaluate them[78,83]. If soft sensing of microbial biomass is being carried out, optimal input selection should include a signal from at least one on-line biomass probe, so that any predictive model formed is improved. Biomass represents a particularly good example in which soft and hard sensing might be combined to advantage.

Concluding remarks

Experts – either humans or computers – have the data and the deterministic knowledge to trace observed behaviour back to the physical, chemical and physiological roots of a bioprocess, leading to a quantum leap in the improvement of bioprocess control. Process control can act on the causes of effects, rather than simply cure symptoms. The following SOP has proved particularly useful:

(1) measure everything that can be measured in the very beginning of process development;

(2) decide whether changes in the magnitude of each variable (parameter) are relevant to the desired performance of the process (model);

(3) choose the relevant variables to be controlled and/or documented;

(4) collect all raw data and distinguish on-line between variable and parameter behaviour, organize an archive of all these data accordingly, and keep even seemingly useless data, as they contribute to the treasure of expertise.

We believe that the adoption of these principles will lead to Good Modelling Practice.

Acknowledgements

DBK thanks the Chemicals and Pharmaceuticals Directorate of the UK Biotechnology and Biological Sciences Research Council for financial support, and Roy Goodacre and Adrian Shaw for comments on a draft. BS thanks the Swiss Priority Program BioTech for support. We also thank the participants at the symposia on Measurement & Control at ECB7, who gave us the final incentive to write this article, and the referees for a number of very helpful comments.

References

1 Locher, G., Sonnleitner, B. and Fiechter, A. (1992) *J. Biotechnol.* 25, 23–53
2 Locher, G., Sonnleitner, B. and Fiechter, A. (1992) *J. Biotechnol.* 25, 55–73
3 Nielsen, J. (1992) *Proc. Control Qual.* 2, 371–384
4 Sonnleitner, B., Locher, G. and Fiechter, A. (1992) *J. Biotechnol.* 25, 5–22
5 Villadsen, J. (1995) *Paper MAC161, ECB7*, Nice, France
6 Tham, M. T., Morris, A. J. and Montague, G. A. (1989) *Chem. Eng. Res. Des.* 67, 547–554
7 Leigh, J. R., ed. (1987) *Modelling and Control of Fermentation Processes*, Peter Peregrinus
8 Bastin, G. and Dochain, D. (1990) *On-Line Estimation and Adaptive Control of Bioreactors*, Elsevier
9 Lübbert, A. and Simutis, R. (1994) *Trends Biotechnol.* 12, 304–311
10 Montague, G. and Morris, A. J. (1994) *Trends Biotechnol.* 12, 312–324
11 Konstantinov, K. B. and Yoshida, T. (1992) *Biotechnol. Bioeng.* 39, 479–486
12 Konstantinov, K. B. and Yoshida, T. (1992) *J. Biotechnol.* 24, 33–51
13 Ishizaki, A. (1992) *J. Biotechnol.* 24, R7–R8
14 Massart, D. L., Vandeginste, B. G. M., Deming, S. N., Michotte, Y. and Kaufmann, L. (1988) *Chemometrics: A Textbook*, Elsevier
15 Martens, H. and Næs, T. (1989) *Multivariate Calibration*, Wiley
16 Goodacre, R., Neal, M. J. and Kell, D. B. *Zentralbl. Bakteriol. Mikrobiol.*, Reihe C (in press)
17 Kell, D. B. and Westerhoff, H. V. (1986) *FEMS Microbiol. Rev.* 39, 305–320
18 Kell, D. B. and Westerhoff, H. V. (1986) *Trends Biotechnol.* 4, 137–142
19 Kell, D. B., van Dam, K. and Westerhoff, H. V. (1989) *Symp. Soc. Gen. Microbiol.* 44, 61–93
20 Fell, D. A. (1992) *Biochem. J.* 286, 313–330
21 Nakajima, M. *et al.* (1992) *Biochemical Engineering for 2001* (Furusaki, S., Endo, I. and Matsuno, R., eds), pp. 681–684, Springer–Verlag
22 Markx, G. H., Davey, C. L. and Kell, D. B. (1991) *J. Gen. Microbiol.* 137, 735–743
23 Fiechter, A. and Sonnleitner, B. (1994) *Adv. Microb. Physiol.* 36, 145–180
24 Duggleby, R. G. (1991) *Trends Biochem. Sci.* 16, 51–52
25 Fuhrmann, G. F. and Völker, B. (1992) *J. Biotechnol.* 27, 1–15
26 Shiraishi, F. (1994) *Enzyme Microb. Technol.* 16, 349–350
27 Mendes, P., Kell, D. B. and Welch, G. R. (1995) in *Enzymology* in vivo (Brindle, K., ed.), pp. 1–19, JAI Press
28 Næs, T., Kvaal, K., Isaksson, T. and Miller, C. (1993) *J. Near Infrared Spectrosc.* 1, 1–11
29 Cheng, B. and Titterington, D. M. (1994) *Stat. Sci.* 9, 2–30
30 Ripley, B. D. (1994) *J. R. Stat. Soc.* 56, 409–437
31 Simutis, R., Havlik, I., Schneider, F., Dors, M. and Lübbert, A. (1995) Preprints 59–65, CAB6, Garmisch-Partenkirchen
32 Lodder, R. A. and Hieftje, G. M. (1988) *Appl. Spectrosc.* 42, 1351–1365
33 Tukey, J. W. (1977) *EDA – Exploratory Data Analysis*, Addison-Wesley
34 Neal, M. J., Goodacre, R. and Kell, D. B. (1994) *Proc. World Congr. Neural Networks 1994, San Diego*, I-318–I-323
35 Livingstone, D. J. and Manhallack, D. T. (1993) *J. Med. Chem.* 36, 1295–1297

36 Manhallack, D. T., Ellis, D. D. and Livingstone, D. J. (1993) *J. Med. Chem.* 37, 3758–3767
37 Seasholtz, M. B. and Kowalski, B. (1993) *Anal. Chim. Acta* 277, 165–177
38 Miller, A. J. (1990) *Subset Selection in Regression*, Chapman and Hall
39 Heinzle, E., Öggerli, A. and Dettwiler, B. (1990) *Anal. Chim. Acta* 238, 101–115
40 Noorman, H. J., Luijkx, G. C. A., Luyben, K. C. A. M. and Heijnen, J. J. (1992) *Biotechnol. Bioeng.* 39, 1069–1079
41 Nielsen, J. and Villadsen, J. (1994) *Bioreaction Engineering Principles*, Plenum
42 Bailey, J. E. (1991) *Science* 252, 1668–1675
43 Stephanopoulos, G. and Vallino, J. J. (1991) *Science* 252, 1675–1681
44 Kacser, H. and Acerenza, L. (1993) *Eur. J. Biochem.* 216, 361–367
45 Pirt, S. J. (1975) *Principles of Microbe and Cell Cultivation*, Blackwell
46 Liang, Y-Z., Kvalheim, O. M. and Manne, R. (1993) *Chemometrics Intell. Lab. Syst.* 18, 235–250
47 Maddox, J. (1994) *Nature* 368, 95
48 Locher, G., Hahnemann, U., Sonnleitner, B. and Fiechter, A. (1993) *J. Biotechnol.* 29, 57–74
49 Kell, D. B., Ryder, H. M., Kaprelyants, A. S. and Westerhoff, H. V. (1991) *Antonie van Leeuwenhoek J. Microbiol.* 60, 145–158
50 Li, J. H., McLellan, J. and Daugulis, A. J. (1995) *Biotechnol. Lett.* 17, 321–326
51 Aon, M. A., Cortassa, S., Westerhoff, H. V. and van Dam, K. (1992) *J. Gen. Microbiol.* 138, 2219–2227
52 Nielsen, J., Nikolajsen, K. and Villadsen, J. (1991) *Biotechnol. Bioeng.* 38, 1–10
53 de Koning, W. and van Dam, K. (1992) *Anal. Biochem.* 204, 118–123
54 Rizzi, M., Theobald, U., Baltes, M. and Reuss, M. (1993) in *Bioreactor and Bioprocess Fluid Dynamics* (Nienow, A. W., ed.), pp. 401–412, Mechanical Engineering Publishers
55 Beyeler, W., Einsele, A. and Fiechter, A. (1981) *Eur. J. Appl. Microbiol. Biotechnol.* 13, 10–14
56 Thevelein, J. M. (1992) *Antonie van Leeuwenhoek J. Microbiol.* 62, 109–130
57 Münch, T., Sonnleitner, B. and Fiechter, A. (1992) *J. Biotechnol.* 24, 299–314
58 Barford, J. P., Jeffery, P. M. and Hall, R. J. (1981) *Proc. 6th Int. Ferm. Symp.* (Moo-Young, M., Robinson, C. W. and Vezina, C., eds), pp. 255–260, Pergamon Press
59 Sonnleitner, B. and Hahnemann, U. (1994) *J. Biotechnol.* 38, 63–79
60 Mendes, P. (1993) *Comput. Appl. Biosci.* 9, 563–571
61 Rohner, M., Sonnleitner, B. and Fiechter, A. (1992) *J. Biotechnol.* 22, 129–144
62 Pascal, F., Dagot, C., Pingaud, H., Corriou, J. P., Pons, M. N. and Engasser, J. M. (1995) *Biotechnol. Bioeng.* 46, 202–217
63 Ratledge, C. (1991) *Bioprocess Eng.* 6, 195–203
64 Kleman, G. L., Chalmers, J. J., Luli, G. W. and Strohl, W. R. (1991) *Appl. Environ. Microbiol.* 57, 918–923
65 Gostomski, P., Mühlemann, M., Lin, Y. H., Mormino, R. and Bungay, H. (1994) *J. Biotechnol.* 37, 167–177
66 Andrews, G. F. (1993) *Biotechnol. Bioeng.* 42, 549–556
67 Kaprelyants, A. S., Gottschal, J. C. and Kell, D. B. (1993) *FEMS Microbiol. Rev.* 104, 271–286
68 Sonnleitner, B. and Käppeli, O. (1986) *Biotechnol. Bioeng.* 28, 927–937
69 Bland, J. M. and Altman, D. G. (1986) *Lancet* i, 307–310
70 Garber, C. C., Mallon, R. P. and Swern, A. S. (1994) *Anal. Chem.* 65, R480–R484
71 Matanguihan, R. M., Konstantinov, K. and Yoshida, T. (1994) *Bioprocess Eng.* 11, 213–222
72 Tichy, W. F., Lukowicz, P., Prechelt, L. and Heinz, E. A. (1995) *J. Syst. Software* 28, 9–18
73 Simutis, R., Havlik, I., Dors, M. and Lübbert, A. (1993) *Proc. Control Qual.* 4, 211–220
74 Simutis, R., Havlik, I. and Lübbert, A. (1993) *J. Biotechnol.* 27, 203–215
75 Goodacre, R., Neal, M. J. and Kell, D. B. (1994) *Anal. Chem.* 66, 1070–1085
76 Goodacre, R. *et al.* (1994) *Biotechnol. Bioeng.* 44, 1205–1216
77 Harris, C. M. and Kell, D. B. (1985) *Biosensors* 1, 17–84
78 Kell, D. B., Markx, G. H., Davey, C. L. and Todd, R. W. (1990) *Trends Anal. Chem.* 9, 190–194
79 Junker, B. H., Reddy, J., Gbewonyo, K. and Greasham, R. (1994) *Bioprocess Eng.* 10, 195–207
80 Konstantinov, K., Chuppa, S., Sajan, E., Tsai, Y., Yoon, S. J. and Golini, F. (1994) *Trends Biotechnol.* 12, 324–333
81 Morris, A. J., Montague, G. A. and Willis, M. J. (1994) *Chem. Eng. Res. Des.* 72, 3–19
82 Farza, M. and Chéruy, A. (1994) *Comput. Appl. Biosci.* 10, 477–488
83 Fehrenbach, R., Comberbach, M. and Pêtre, J. O. (1992) *J. Biotechnol.* 23, 303–314
84 Rumelhart, D. E., McClelland, J. L. and the PDP Research Group (1986) *Parallel Distributed Processing. Experiments in the Microstructure of Cognition*, MIT Press
85 Werbos, P. J. (1993) *The Roots of Back-Propagation: From Ordered Derivatives to Neural Networks and Political Forecasting*, Wiley
86 Chauvin, Y. and Rumelhart, D. E. (1995) *Backpropagation: Theory, Architectures and Applications*, Lawrence Erlbaum Associates