Modelling of mammalian cells and cell culture processes

F.R. Sidoli^{1,2}, A. Mantalaris^{2,*} and S.P. Asprey^{1,2,3}

¹Centre for Process Systems Engineering ²Department of Chemical Engineering and Chemical Technology, Imperial College London, South Kensington Campus, London SW7 2AZ, UK; ³Current address: Orbis Investment Advisory Ltd, Orbis House, 5 Mansfield Street W1G 9NG, UK; (*Author for correspondence; E-mail: a.mantalaris@imperial.ac.uk)

Received 10 January 2003; accepted in revised form 9 October 2003

Key words: Mammalian cells, Model discrimination, Model distinguishability, Model identifiability, Optimal experiment design, Parameter precision, Population balance model, Single cell model

Abstract

Mammalian cell cultures represent the major source for a number of very high-value biopharmaceutical products, including monoclonal antibodies (MAbs), viral vaccines, and hormones. These products are produced in relatively small quantities due to the highly specialised culture conditions and their susceptibility to either reduced productivity or cell death as a result of slight deviations in the culture conditions. The use of mathematical relationships to characterise distinct parts of the physiological behaviour of mammalian cells and the systematic integration of this information into a coherent, predictive model, which can be used for simulation, optimisation, and control purposes would contribute to efforts to increase productivity and control product quality. Models can also aid in the understanding and elucidation of underlying mechanisms and highlight the lack of accuracy or descriptive ability in parts of the model where experimental and simulated data cannot be reconciled. This paper reviews developments in the modelling of mammalian cell cultures in the last decade and proposes a future direction – the incorporation of genomic, proteomic, and metabolomic data, taking advantage of recent developments in these disciplines and thus improving model fidelity. Furthermore, with mammalian cell technology dependent on experiments for information, model-based experiment design is formally introduced, which when applied can result in the acquisition of more informative data from fewer experiments. This represents only part of a broader framework for model building and validation, which consists of three distinct stages: theoretical model assessment, model discrimination, and model precision, which provides a systematic strategy from assessing the identifiability and distinguishability of a set of competing models to improving the parameter precision of a final validated model.

Nomenclature

C_{ij} – component i concentration in j th model compartment; R_{ijk} – transport rate of component i into or out of compartment j from the k th source or sink compartment; N_{in} – number of source compartments; N_{out} – number of sink compartments; r_{ijl} – reaction rate of component i in compartment j in the l th i -generating or i -consuming reaction; N_{gen} – number of reactions generating component i ; N_{con} – number of reactions consuming component i ; μ – specific growth rate; μ_{max}^{app} – apparent maximum specific growth rate; K_d – specific death rate; N_v – viable cell number; N_t – total cell number; D – dilution rate; β – parameter; β_0 – parameter; α – parameter; m – cell mass; m' – mass of a dividing cell; t – time; N – cell number; r – single cell growth rate; S – nutrient concentration; s – nutrient concentration vector; Γ – cell transition rate; p – partition probability density function; Y – yield coefficient; $f(m)$ – the division probability density function; B – coefficient for the beta distribution incorporating the gamma function; $\mathbf{x}_{max}$ – vector of maximum

amounts of biochemical components of a newborn cell; $\mathbf{x}_{n,\min}$ – vector of minimum amounts of biochemical components of a newborn cell; $\mathbf{x}_{\min}$ – vector of minimum amounts of biochemical components of a dividing cell; $\mathbf{x}$ – vector of biochemical components contained within a cell (PBM), state variables (SCM); N_i – cell number of the i th stage; $\mathbf{r}_i$ – growth rate vector of stage i cells; Γ_i – cell transition rate from i th stage; $\mathbf{q}$ – nutrient consumption rate vector; ψ – density function of newborn cells; ϕ – density function of dividing cells; f – time-invariant density function of total cell population; $\mathbf{B}$ – domain of all physiologically feasible states; $N_t \equiv \partial N / \partial t$; $N_x \equiv \partial r_x N / \partial x$; $\dot{\mathbf{y}} \equiv d\mathbf{y} / dt$ where $\mathbf{y}$ is the vector of model outputs; $\mathbf{y}$ – vector of model outputs; $\dot{\mathbf{x}} \equiv d\mathbf{x} / dt$ where $\mathbf{x}$ is the vector of model state variables; θ – vector of model parameters (SCM); $\mathbf{V}$ – matrix of model variances and covariances; $\hat{\theta}$ – vector of best model parameter estimates; φ – vector of experiment decision variables; $\nabla_{\mathbf{x}} \equiv \sum_i \partial / \partial x_i$; $\sum_i$ – term summation over all possible values of i

Abbreviations: SCM – single cell model; PBM – population balance model

Introduction

The ability to predict the precise behaviour of all facets of mammalian cell cultures under different culture conditions is highly desirable for both commercial and scientific reasons. Monoclonal antibodies (MAbs) alone are expected to have sales exceeding \$10 billion globally by 2005 (Gray and Jasuja 2001). Although mammalian cells, unlike their bacterial counterparts, have the ability to produce these high-value products that require post-translational modification in order to become biologically active, there exists a commercial need to improve the efficiency and productivity of these processes. The drive for increased efficiency and productivity is becoming urgent not just to reduce high experimental costs but also to counter the emerging competition from non-cell culture technologies, such as *in vitro* ribosome display and transgenic animals (Duncan, 2002). These improvements can be either *biologically* based or *culture* based. The former can be achieved by the genetic manipulation of the cells. Optimising the culture process, while the intrinsic efficacy of the product remains unchanged, can also increase efficiency by increasing purity, quality, and productivity, aspects that require the modelling expertise of engineers.

Experiments are currently the major discovery method of underlying biological mechanisms and the only way to reliably validate relationships between important cellular variables. However, experiments are time consuming and expensive to perform, and unless carefully designed generate data that are typically very noisy. Alternatively, the use of a prototype model (or competing models) that is still in development can itself be utilised

to define quantitatively the future experimental schedule that will maximise the information obtained and subsequently the accuracy of the end model. Model-based experiment design is an area of modelling rarely utilised and yet one that addresses directly the problems of unnecessary experimental cost and insufficient model accuracy. Instead, models are developed retrospectively, after subjective experiments are performed, such that the information content of an experiment is not assessed quantitatively and thus unknown prior to being performed.

Once a model is complete and has been finalised it may be used for the optimisation of operational activities including online measurement and control, medium formulation, and feed strategy. If the aim is achieving an optimal value for some commercial objective, such as productivity maximisation, then a model is vital in achieving optimality of the operational aspects of culture experiments. In contrast, the use of quasi trial-and-error experiments is time-consuming and expensive since it may require a large number of experiments. Using a modelling framework, however, enables optimality and keeps experimental costs to a minimum. Experiments can be run *in silico*, reducing the number of experiments required and enabling the modeller to identify the operating conditions, medium compositions, and parameter spaces where experiments yield a better return on investment, that is, to obtain improved qualitative and quantitative information for the same time and financial expense incurred.

The accurate modelling of mammalian cell cultures is best approached by first modelling cellular function. Since mammalian cells have an extremely

complicated internal structure where a plethora of inter-linked biochemical processes occur in various spatial locations and where control of the biochemical processes is accomplished by a stupendously complex interconnected system of regulatory processes governed itself by the cell's genome, a single cell model (SCM) is best suited dealing with this level of detail. However, a cell-wide approach requiring significant investment is needed to develop such models and may explain why SCMs are not as abundant as models that only characterise parts of cell physiology, such as unstructured models. Nevertheless, significant progress has been made as has been detailed in a comparison study on the numerous literature models of hybridoma cell growth and metabolism (Pörtner et al. 1996), which helped to consolidate the work done at this level. Furthermore a new SCM with general applicability to mammalian cells has been recently developed (Sanderson 1997).

On the other hand, describing the behaviour of a single cell is only useful if it can be used to build a representation of a cell population, which is accomplished through the use of population balance models (PBMs). Recently, there has been renewed effort in the field of population balance modelling with work on multidimensional cell number distributions, the variable of interest in a PBM (Mantzaris et al. 2001a, b, c) and multi-staged PBMs (Hatzis et al. 1995); the dimensionality corresponding to the number of components that give a cell its internal structure and stages giving the cell culture more realistic structuring by accounting for sub-populations. The predictability of PBMs depends in part on important single-cell information about growth, when division occurs, and how dividing cells partition their components. Some of this information, typically incorporated in a SCM, can be used to supply the relevant information to a PBM and there has been some work in this direction (Martens et al. 1995).

The paper is divided into two major sections: models and modelling tools. We start with a careful discussion of model classification and their attributes. Since the last major modelling review was written almost a decade ago by Tziampazis and Sambanis (1994), we report on progress made and introduce population balance modelling in the *model* section. In the *modelling tools* section we introduce model identifiability, distinguishability,

discrimination, and adequacy as well as optimal design of experiments. Parameter estimation is also discussed, being an integral part of the validation process. Finally, in light of recent developments in genomics, proteomics and metabolomics, we conclude by discussing the urgent need for consideration of genes and gene expression, and their incorporation into SCMs.

Mathematical models

In cell process technology there are two levels of biological organisation, necessitating two categories of models. The first is the SCM, which considers the detailed characterisation of an individual cell. To describe a population of cells consisting of multiple populations with distributed properties within a culture, a PBM model is required. These two model types represent the extremes of the long established microbial model classification spectrum identified by Tsuchiya et al. (1966) (Figure 1), which can also be applied to mammalian cell models. In this classification system, a model can be structured or unstructured, segregated or unsegregated, deterministic or stochastic. Structured models attempt to explicitly describe intracellular processes in that they possess structure either in the physical sense, namely the organelles, cell shape or size, or in the biochemical sense where biomass is subdivided into its intracellular biochemical components. Classifying a model as being segregated or unsegregated pertains to the heterogeneity of the composition of the cell culture. If a cell culture is viewed as segregated, it is composed of cells in different stages of development and is heterogeneous. Conversely, an unsegregated model assumes an average cell such that the population can be treated as homogeneous. Lastly, while a deterministic model is one where cellular processes are not subject to variability, a stochastic model describes random processes. For example, two processes that warrant stochastic description are the partitioning of daughter cells at cytokinesis and the duration of cell cycle phases. Clearly a structured, segregated, stochastic model offers the most realistic representation of a cell; however there are important trade-offs to be made in terms of formulation time, model complexity, and solution time. Prioritisation of the model features is driven by the end purpose of the model.

Figure 1. Model classification for prokaryotic and eukaryotic cell culture systems (adapted from Ramkrishna 1979).

Single cell models

Cellular metabolism has been incorporated into existing SCMs as a composition of four interrelated component processes that consist of (i) transport, (ii) primary metabolism, (iii) product synthesis, and (iv) cell growth and death. In order to describe these biochemical processes, standard engineering formalisms are utilised. Intra- and extracellular transport considers transport kinetics of biochemical species across membranes and includes the compartmentalisation of the cell into distinct regions through which the transported species flow. Primary metabolism and product synthesis, in addition to transport, relate conceptually to two or more linked compartments within which either particular metabolic pathways and/or entire metabolic cycles operate. Conservation equations written in each compartment and for each biochemical species, result in a set of equations. These are ordinary differential equations once the usual assumption is made that each compartment is a lumped phase (i.e., concentrations are constant throughout the compartment). Equation (1) illustrates the generic form of these conservation equations. The accumulation of any given component i depends on its flow into or out of the compartment j as a result of transport across the compartment barrier (first and second terms on the RHS of Equation (1)), intra-compartmental reactions generating and consuming the component (third and

Figure 2. Single cell compartmentalisation.

fourth terms on the RHS) and its dilution due to growth of the cell and thus increasing compartment size (last term on RHS):

$$\begin{aligned}
 \underbrace{\frac{dC_{ij}}{dt}}_{\text{Accumulation}} &= \underbrace{\sum_{k=1}^{N_{in}} R_{ijk} - \sum_{k=1}^{N_{out}} R_{ijk}}_{\text{Transport}} + \underbrace{\sum_{l=1}^{N_{gen}} r_{ijl} - \sum_{l=1}^{N_{con}} r_{ijl}}_{\text{Reaction}} \\
 &\quad - \underbrace{\mu C_{ij}}_{\text{Dilution}}, \quad \forall i, j
 \end{aligned} \tag{1}$$

Conceptually, compartments correspond to cellular structures, such as organelles, pooled biochemical components or the cell environment (Figure 2). However, the ODEs actually represent distinct pooled concentrations at a single point and do not describe physical dimensions and therefore cell geometry. To remedy this, the lumped phase assumption must be relaxed, allowing concentration gradients to exist. Hence, the resulting conservation equations are now partial differential and

additional diffusion equations are required to compute the diffusive fluxes within the spatial geometry of each compartment. However, due to the additional mathematical and computational complexity involved in simulating the complex spatial variation of concentrations and the difficulty in validating experimentally the simulated intracellular concentrations, this relaxation is not normally implemented. Furthermore, as the cell is a reacting system, terms arising in the conservation equations relating to these compartment-specific reactions require characterisation via accompanying kinetic rate expressions, as do membrane transport rates also appearing in the conservation equations. Mathematically, the kinetics that describes metabolic reaction and membrane transport rates is written as a set of algebraic equations that relate rate to concentration, concentration gradient, or other rate. Reaction and transport mechanisms are classified as either enzymatic or non-enzymatic resulting in different structural forms of the algebraic equations. The most common form of expression used to describe rates is the Michaelis–Menten. Depending on how complex the reaction kinetics is in the model, the SCM will be parameterised to varying degrees. However, it is important not to overly parameterise models since this results in parameter estimability problems when considering model structural properties.

Stoichiometric modelling represents an alternative to the kinetic approach described earlier for the modelling of cellular metabolism. The mathematical formulation of a model, as either stoichiometric or kinetic, determines its optimal applicability. A kinetic model is represented by a system of differential algebraic equations (DAEs), which must be integrated over the domain of interest that is nearly always the time horizon and results in well-defined time trajectories for all model variables. It is therefore considered as a deterministic approach since the determined model variables physiologically can be interpreted as a ‘single phenotype’ description (Gombert and Nielsen 2000). A stoichiometric model is represented by a system of flux balance equations based on reaction stoichiometry of a metabolic network with accompanying constraints on flux values and solved as a constrained optimisation problem using some assumed cellular objective; there exists a mathematical and therefore physiological feasible region in

which a range of possible mathematical solutions or phenotypes are acceptable though not optimal. Kinetic models therefore can account for dynamic behaviour whereas stoichiometric models yield a static instance of metabolic activity and are unable to incorporate regulation and control of cellular activity, which requires dynamic simulation. However, a key drawback of kinetic models is that their additional predictive capability that is associated with incorporating complex dynamic expression normally results in non-linearity in both parameters and variables. This can make the models cumbersome and lead to severe numerical problems in their solution. One key advantage of stoichiometric models is that they can account for competing reactions, which is useful for studying the relative activity of certain pathways under various culture conditions and this has been their major use. For example, the existence of more than one physiological steady state, known as steady-state multiplicity, has been observed by a number of investigators using stoichiometric models and metabolic flux analysis (Follstad et al. 1999; Europa et al. 2000; Cruz et al. 1999; Zhou et al. 1997; Linz et al. 1997; Paredes et al. 1999).

In general, there will be parts of SCM kinetics that are not based on either known or proposed kinetic mechanisms but are inferred using mathematical functions that best fit the observed phenomenon. Clearly, this gives rise to a spectrum of models that range from the purely inferred to the completely mechanistic. Most existing SCMs are partially mechanistic models due to the fact that many aspects of cell behaviour that these models seek to describe have not yet been elucidated. The development of SCMs requires the incorporation and integration of either parts or the entirety of other smaller models each of which describe specific sub-processes. Utilising the most accurate sub-models that correspond to the components of cellular metabolism is, therefore, of key importance to the construction of a good SCM.

Pörtner and Schäfer (1996) presented a quantitative comparison between a number of unstructured hybridoma cell growth and death models commonly used in the literature and their associated data derived from chemostat cultures. Most of the correlations for the specific growth rate were Monod-type but differed in whether glucose or glutamine or both were used as the growth limiting substrate

and whether additional terms were included to account for growth inhibiting metabolites, such as lactate and ammonia. Some models allow for the absence of cell growth at non-zero substrate concentrations by incorporating a ‘threshold term’. This type of growth limitation is known as kinetic growth limitation, as opposed to stoichiometric growth limitation where the substrate is completely consumed. Some investigators include the serum concentration as either a limiting substrate or as a variable upon which the maximum specific growth rate is dependent. However, industrially, there is a general trend towards serum-free media, which may remove the need to consider serum effects in any future models. Furthermore, Pörtner et al. (1996) have stressed that, although most growth models are dependent on glucose and glutamine, actual growth may not be limited by these components and that despite the fact that they are the main carbon and nitrogen sources their extensive use is mainly due to the fact that they are easily measured experimentally. Despite the differences amongst the models, a plot of specific growth rate (μ) versus dilution rate (D) revealed the remarkable agreement the various models showed with each other, all within the analytical error for dilution rates above 0.015 h^{-1} . The study also showed, however, that data scatter increased with decreasing dilution rates, attributed to either cell disintegration or culture instability. Additionally, the specific cell growth rate in all models deviated from the expected trend ($\mu = D$) as a dilution rate of zero was approached, a phenomenon attributed to the potential existence of a non-zero growth rate at $D = 0$. Pörtner et al. (1996) correctly point out, however, that this is clearly not possible as steady-state at $D = 0$ necessarily means a steady-state batch culture, which requires the specific growth rate to be equal to zero. The growing deviation may be due to the difficulty in obtaining steady-state conditions at low dilution rates.

Some investigators highlight, through their use of ‘unknown’ species, the difficulties still remaining in understanding the causality of many factors. Zeng et al. (1998) proposed in a comparison of six hybridoma cell lines that the specific growth rate is limited by either an auto-inhibitor or an unknown limiting growth factor, which manifests itself through the strong dependence of μ on the ratio of cell number to dilution rate N_v/D . The growth rate correlates negatively and almost

linearly for all cultures, in both chemostat and perfusion modes, with the ratio of viable cell number to the dilution (perfusion) N_v/D , until cessation of growth is reached. The identification of the key determinant of growth, that is, the auto-inhibitor or limiting growth factor, remains unanswered. However, experimental data from several groups provide strong evidence for the general existence of auto-inhibitor(s). Following this, a growth equation of the form:

$$\mu = \mu_{\max}^{\text{app}} \left(1 - \alpha \frac{N_v}{D} \right) \quad (2)$$

was proposed where, in contrast to traditional growth equations, there are fewer parameters – two. Parameter α has physiological significance and takes just two values depending on whether the media contains serum or not. The work of Lee et al. (1995) proposed the dependence of specific cell death rate on a non-growth linked auto-inhibitor and suggested a positive linear correlation between K_d and N_v/D . Zeng et al. (1998) tested this relation on a number of hybridoma cell lines and found it to hold only for some chemostat cultures and only at relatively low values of N_v/D . For higher values, the relation becomes severely non-linear. Instead, based on evidence of a dependence of K_d on μ , Zeng et al. (1998) modified the relation of Lee et al. (1995) to incorporate this so that the same inhibitor that affects growth through μ also affects death. They also make a correlation with N_t not N_v thus finding a more robust linear relation between K_d/μ and N_t/D that is cell line non-specific, yet strongly media dependent. The final form of their relation, as with their growth correlation, has only two parameters and β , like α , is serum dependent:

$$K_d = (\beta_0 + \beta\mu) \frac{N_t}{D} \quad (3)$$

It is apparent from the sheer number of different mathematical forms used to describe the specific death rate that a lot of difficulty is encountered in modelling it (Pörtner et al. 1996). In addition to the work on cell growth, Pörtner et al. (1996), through a transformation of the form $K_d = f(\mu)$, performed a comparison of death rate correlations to discover that the scatter far exceeds the error margins attributed to differences in what count as dead cells for any of the death rates. The comparison usefully revealed that, in general, the specific

death rate falls in the range of $0.001\text{--}0.006\text{ h}^{-1}$ for high specific growth rates and that specific death rates increase sharply with decreasing specific growth rates.

Modelling the formation by the cells of their highly valuable protein products is of great importance, as ultimately the drive to improve cell culture modelling is due to the need for achieving increased productivity. Pörtner et al. (1996) compared a number of unstructured models for cell-specific antibody production by hybridomas. Specific growth rate, substrate concentration, serum and death rate have all been proposed as being the variables on which productivity depends, although which of these should be used and to what extent productivity is sensitive to them is very much cell line dependent. Substrate concentration, for example, has been shown to have no influence (Hiller et al. 1991; Pörtner et al. 1996), negatively (Kurokawa et al. 1994), or positively correlated with antibody production depending on the cell line used (Dalili et al. 1990; Linardos et al. 1991; Frame and Hu 1991). More recently Lüdemann et al. (1996) demonstrated how optimised serum-free media can yield the same productivity as serum containing media.

Jang and Barford (2000) developed an unstructured, unsegregated, deterministic model describing the MAb production of batch and fed-batch cultures by considering the anabolism of cellular macromolecules; specifically rates of DNA replication, RNA transcription, protein translation and lipid/carbohydrate creation. Protein concentration depended on mRNA availability, which depended on gene expression and so on, such that there was a cascade type dependency of one macromolecule to another. The specific rates of DNA replication, transcription and translation were assumed proportional to the specific growth rate and dependent on DNA and mRNA levels, respectively. The viable cell population was further subdivided into those cells that proceed through the cell cycle and those that do not (those arrested in the G_0 phase). DNA replication and viable cell growth was assumed to occur only in cycling viable cells. G_0 cells were also assumed to have higher antibody productivity. The model predicted both the batch and fed-batch experimental data satisfactorily.

A large, structured, unsegregated, deterministic SCM has been developed by Sanderson (1997). It is designed to have general applicability to any type

of suspension culture of mammalian cells. It incorporates glycolysis, glutaminolysis, the TCA cycle, the pentose-phosphate pathway, and fatty and amino acid metabolism in addition to cell growth and death and antibody production into a compartmental form. There are three cell-related compartments: the medium, cytoplasm and mitochondria. In what is a continuation of the model of Barford and Phillips (Barford et al. 1992; Phillips 1996), the concentrations of some 49 biochemical components are modelled using Monod-type kinetics. The model accounts for the effects of feedback inhibition in certain reactions and competitive reactions by modifying the Michaelis–Menten expression. Intracellular and medium concentrations are related by using transport equations to account for membrane transport and use a simplified concentration driving force to quantify flux. The model's practical accuracy is accessed by tuning it to experimental data at one set of conditions and then performing a second experiment under different conditions and comparing the data with the model's simulation. The ratio of the difference (error) between measured and predicted values of component concentration and measured values was used as a measure of accuracy. In the first experiment most errors were within the 15% expected scatter range and in the second averaged 10%. When a control experiment with no cells was conducted, it yielded an error of between 10% and 15% and the model seemed to perform satisfactorily. Glutamine and antibody concentration, however, were significantly inaccurate. In other experiments on a different cell line the model, when tuned, fitted the majority of component concentrations within acceptable error, but there were always some components that deviated significantly. It has also been used for optimum media design (Sanderson et al. 1995), where through the use of an economic objective function, a medium composition was obtained that resulted in the doubling of profitability.

Population balance models

The most mathematically concise way in which to describe the property variation of cells within a population is, currently, in the form of a PBM. In essence, PBMs are a number balance on a cell population. Unfortunately, they are also particularly

difficult to solve, being partial-integro-differential equations. Extending this type of model slightly to accommodate additional detail or structure leads very quickly to intractable models. In all but the simplest cases, which assume special conditions, such as that described by Liou et al. (1997) involving at most two cell properties, analytical solutions are not possible. Hence, in general, numerical methods must be used for their solution.

PBMs can be classified as follows: single- or multi-variable, single- or multi-staged, and mass or aged structured (Mantzaris et al. 2001a). By their very nature, these models recognise the individuality of cells within a population, and are considered to lie within the segregated (or corpuscular) class of models. A multi-variable model uses more than one cell property to distinguish between cells and can be used to account for any number of biochemical constituents within the cell whereas a single-variable model differentiates on the basis of one cell property only. A multi-staged model is required when more than one development phase is described, each additional phase requiring an extra stage, which is described by a single population balance equation resulting in separate population distributions. Finally, if mass or any property that adheres to mass conservation laws is the variable(s) used to differentiate cells, then the model is mass structured, whereas if cell maturity is used, the model is said to be age structured.

PBMs have been available since the 1960s and research has been spearheaded by Fredrickson and co-workers (Eakman et al. 1966; Tsuchiya et al. 1966). Looking beyond the mathematically complex nature of PBMs, the one crucial characterising difference that distinguishes them from SCMs is their unmatched ability to account for the corpuscular nature of a cell. They are a concise one-equation (for a single stage) description of what is a very complex process. Despite their unique capabilities, they have never been widely used by biochemical engineers or biotechnologists, mainly because of two major problematic features (Srienc 1999; Villadsen 1999). First, they are complicated to handle and solve. Convenient solution methods have not been available until recently and the more complex PBMs still pose serious numerical and computational problems. Nevertheless, this current constraint is not a serious problem as improved computer power becomes available; accurate

determination of model parameters being a more limiting restriction. Three parameters are required in a generalised PBM, each having physiological significance. They are: (i) the single cell growth rates; (ii) the transition rates between each of the cell cycle phases; and (iii) the partition function. Together, these three process parameters define the collective state of the cell population. However, they are unknown, which represents their main limitation. These functions are dependent on the environment or more precisely, the conditions of the surrounding medium. The effects of these on the cell are for the most part not well understood and sparse noisy experimental data are used to infer parameter values often using highly implicit methods (Villadsen 1999). Thus, the generation of relevant experimental data needed to validate PBMs has, in the past, been a significant problem. Although with the advent of flow cytometry some distribution data can be obtained, the lack of complete data remains an obstacle that must be overcome if the application of PBMs is to become accepted and more widespread. These problematic issues should not dissuade the use of PBMs; despite their abundant use for microbial cells, there are fewer examples for mammalian cell cultures. Accordingly, we will highlight recent developments in PBMs in the context of animal cell cultures, placing emphasis on the increasingly more complex structure of a model that can be formulated to account for more detailed cellular structure rather than the numerical aspects of their solution.

A single-stage single-variable PBM

A detailed derivation of the generic structure of a population balance equation can be found in Ramkrishna (2000). We introduce it using a relatively simple form presented by Mantzaris et al. (1999). The cells are considered heterogeneous on the basis of their mass, which is used to distinguish the physiological state of cells. The key variable of interest is the total number of cells or cell number distribution, $N(m, t)$. This quantity, which is assumed to vary with time and mass (LHS of Equation (4)), changes as a result of specific life-cycle events (captured by terms on the RHS of Equation (4)). The first derivative term accounts for accumulation while the second accounts for loss of cells of a given mass due to their transition into

larger cells – single cell growth. There are assumed to be two life-cycle processes that cause changes in the cell number. One is cell division, which results in the loss of dividing cells of mass m with a division rate $\Gamma(m, S)$. The second is cell birth, which gives rise to two daughter cells from the division of a single parent cell of mass m' . The product ΓN is the number of parent cells of a given mass. Information on how each of the parent cells partitions its mass is contained in the partition function, $p(m, m', S)$ and expresses a probability of a parent cell of mass m' giving birth to a daughter cell of mass m so that $2p\Gamma N$ quantifies the number of newborn cells originating from parent cells of a single mass. But as parent cells will take on a range of masses m' along the mass continuum, to include all parent cell divisions, integration is required for all possible parent cell masses ranging from the minimum, τ , to the theoretical maximum, infinity:

$$\begin{aligned} & \overbrace{\frac{\partial N(m, t)}{\partial t}}^{\text{Accumulation}} + \overbrace{\frac{\partial}{\partial m} [r(m, S)N(m, t)]}^{\text{Single cell growth}} \\ &= \overbrace{-\Gamma(m, S)N(m, t)}^{\text{Division}} \\ &+ \overbrace{2 \int_{\tau}^{\infty} \Gamma(m', S)p(m, m', S)N(m', t) dm'}^{\text{Birth}} \quad (4) \end{aligned}$$

To complete the model, the physiological functions r , Γ and p appearing in Equation (4) must be specified, as must the substrate variation S along with initial and boundary conditions. At the start of the culture process ($t = 0$), the number distribution of cells (N_0) is known and so this forms the initial condition for Equation (4):

$$N(m, 0) = N_0(m) \quad (5)$$

One or more boundaries may exist on the physiological state space (Fredrickson et al. 1967). The boundary condition represented by Equation (6) prohibits the existence of cells of zero mass.

$$N(0, t) = 0 \quad (6)$$

Currently the favoured characterisation of S is by relating it to biomass production rate via a yield coefficient, Y , as done by Mantzaris et al. (1999).

For a well-mixed system

$$\frac{dS}{dt} = \frac{1}{Y} \int_0^{\infty} r(m, S)N(m, t) dm \quad (7)$$

where, the initial substrate concentration S_0 is known

$$S(0) = S_0 \quad (8)$$

The partition function, p , which specifies how cellular material is divided amongst the two daughter cells, is taken to be a unimodal symmetric (i.e., the shape parameters are equal) beta distribution $B(q, q)$:

$$p(m, m') = \frac{1}{B(q, q)} \frac{1}{m'} \left(\frac{m}{m'}\right)^{q-1} \left(1 - \frac{m}{m'}\right)^{q-1} \quad (9)$$

Symmetry is required because cells obey the mass conservation law and therefore the probability of a parent cell m' giving birth to a daughter cell of mass m , necessitates also giving birth to a daughter cell of mass $m' - m$, that is, $p(m', m) = p(m' - m, m)$. The distribution is unimodal, centred on $m = m'/2$, reflecting the assumption that equal partitioning is the most likely outcome, although unequal partitioning is still possible. Other descriptions of cell partition have been used. Eakman et al. (1966) originally used a Gaussian-type distribution where, if the assumption is made that the partition function variance is very small, then the Gaussian distribution reduces to a Dirac delta function, which translates to equal partitioning always occurring. The division rate Γ is proportionally related to growth r by the expression:

$$\Gamma(m, S) = \frac{f(m)}{1 - \int_0^m f(m') dm'} r(m, S) \quad (10)$$

where, $f(m)$ is the division probability density function taken to be a LHS truncated normal distribution with a mean of μ_f and standard deviation σ_f . The normal is truncated to exclude cells of mass less than zero. The denominator of the density function represents the proportion of all cells that have reached a mass m and remain undivided. It is used to re-normalise $f(m)$ in order to yield the fraction of cells of mass m that will divide, given that no division of cells with mass less than m has yet taken place. For detailed derivation see Eakman et al. (1966).

Using a model consisting of Equations (4)–(10), Mantzaris et al. (1999) carried out a number of

simulations. Various forms were used to describe the growth rate, including constant, linear and quadratic growth forms. The effects of equal and unequal partitioning of parent cells into daughter cells was considered, as well as the effect of constant or changing substrate concentrations. Each of these issues affected the underlying intrinsic periodicity that is characteristic of population balance equations to some extent. In order to test the validity of the solutions from the numerical scheme, a comparison to the analytical solution of the successive generation approach presented by Liou et al. (1997) was performed. The problem considered by Liou et al. (1997) was a mass structured single variable constant substrate concentration assuming linear growth. Under these conditions the two methods, analytical and numerical, had almost precisely matching results for both equal and unequal partitioning. The slight mismatch was attributed to truncation and accumulation errors occurring during evaluation of the numerical scheme. Simulations showed that constant and quadratic forms of growth achieved balanced time-invariant growth states using equal and unequal partitioning. Under conditions of linear growth and equal partitioning, however, the population distribution appeared to have a periodic behaviour, a special case originally predicted in theory by Diekmann et al. (1984) described as a ‘multiplication machine’. In all the remaining cases, unequal division resulted in balanced growth being achieved faster and with broader distribution. It is suggested that this is due to the model’s ability to permit more than one form of partitioning hence cells could divide in any number of ways. The time-independent solution was achieved fastest with quadratic growth and slowest with linear growth, with constant growth in-between. The growth rate also affected the number density distribution, with quadratic growth leading to a taller, narrower distribution and constant growth a shorter, broader one again with linear growth in-between.

A single-stage multi-variable PBM

As we have mentioned already, cells can be differentiated on the basis of more than just one property (i.e., mass in the last example). In a series of papers Mantzaris et al. (2001a, b, c) present a multi-variable PBM. Cells distinguished on the basis of more than

one physiological state mathematically require the use of a ‘physiological state’ vector $\mathbf{x}$ in place of the mass m in Liou et al. (1997). The mode of reactor operation used was a CSTR, which requires an extra output term incorporating the dilution rate D of the culture. The form of this PBM was analogous to the one-dimensional form:

$$\begin{aligned} \frac{\partial N(\mathbf{x}, t)}{\partial t} + \nabla_{\mathbf{x}}[\mathbf{r}(\mathbf{x}, \mathbf{s})N(\mathbf{x}, t)] \\ = -\Gamma(\mathbf{x}, \mathbf{s})N(\mathbf{x}, t) - DN(\mathbf{x}, t) \\ + 2 \int_{\mathbf{x}}^{\mathbf{x}_{\max}} \Gamma(\mathbf{x}', \mathbf{s})p(\mathbf{x}, \mathbf{x}', \mathbf{s})N(\mathbf{x}', t) d\mathbf{x}' \end{aligned} \quad (11)$$

where, the initial conditions are correspondingly:

$$N(\mathbf{x}, 0) = N_0(\mathbf{x}) \quad (12)$$

The boundary conditions for Equation (11) ensuring that cell division does alter population mass and cell growth does not change population number are given by:

$$\mathbf{r}(\mathbf{x}, \mathbf{S})N(\mathbf{x}, t) = 0, \quad \forall \mathbf{x} \in \mathbf{B} \quad (13)$$

where, $\mathbf{B}$ defines the physiological state space boundary where at least one element of $\mathbf{x}$ is at its maximum or minimum value.

The cell environment is considered as multi-component represented by a vector form $\mathbf{s}$. The dividing cells have a physiological state vector denoted $\mathbf{x}$. Bounds are placed on the maximum ($\mathbf{x}_{\max}$) and minimum ($\mathbf{x}_{\min}$) states of daughter cells, and parent cells cannot divide unless in a minimum physiological state ($\mathbf{x}_{\min}$). The substrate characterisation takes on a vector form and has an extra term to account for the effect of continuously fed and removed substrate:

$$\frac{d\mathbf{S}}{dt} = D(\mathbf{S}_f - \mathbf{S}) - \int_{\mathbf{x}_{\min}}^{\mathbf{x}_{\max}} \mathbf{q}(\mathbf{x}, \mathbf{S})N(\mathbf{x}, t) d\mathbf{x} \quad (14)$$

where, $\mathbf{q}$ is a nutrient consumption vector. Equation (14) has a corresponding initial condition:

$$\mathbf{S}(0) = \mathbf{S}_0 \quad (15)$$

Mantzaris et al. (2001a), using this model, performed simulations using up to three physiological states, that is, $\mathbf{x} = [x_1, x_2, x_3]^T$, where x_i represented cellular material i .

A multi-stage multi-variable PBM

The model of Mantzaris et al. (2001a) is a multi-variable model that treats cell lifetime as one uneventful stage from birth to death, reflected through the use of one distribution to describe the entire population. The cell cycle gives rise to fractions of the overall population, existing as a sub-population in a particular cycle phase. It is more realistic to have a population balance equation for each subpopulation. Hatzis et al. (1995) illustrated the general structure and framework of a multi-variable multi-staged PBM. Equations (16)–(18) combined form the balances of a three-phase cell cycle model:

$$\begin{aligned} & \underbrace{\frac{\partial N_1(\mathbf{x}, t)}{\partial t}}_{\text{Accumulation 1}} + \underbrace{\nabla_{\mathbf{x}}[\mathbf{r}_1(\mathbf{x})N_1(\mathbf{x}, t)]}_{\text{Single cell growth 1}} \\ &= \underbrace{-\Gamma_1(\mathbf{x}, t)N_1(\mathbf{x}, t)}_{\text{Transition 1} \rightarrow 2} \\ &+ \underbrace{2 \int \Gamma_3(\mathbf{x}')p(\mathbf{x}', t)N_3(\mathbf{x}', t)d\mathbf{x}'}_{\text{Birth}} \quad (16) \end{aligned}$$

$$\begin{aligned} & \underbrace{\frac{\partial N_2(\mathbf{x}, t)}{\partial t}}_{\text{Accumulation 2}} + \underbrace{\nabla_{\mathbf{x}}[\mathbf{r}_2(\mathbf{x})N_2(\mathbf{x}, t)]}_{\text{Single cell growth 2}} \\ &= \underbrace{-\Gamma_2(\mathbf{x}, t)N_2(\mathbf{x}, t)}_{\text{Transition 2} \rightarrow 3} + \underbrace{\Gamma_1(\mathbf{x}, t)N_1(\mathbf{x}, t)}_{\text{Transition 1} \rightarrow 2} \quad (17) \end{aligned}$$

$$\begin{aligned} & \underbrace{\frac{\partial N_3(\mathbf{x}, t)}{\partial t}}_{\text{Accumulation 3}} + \underbrace{\nabla_{\mathbf{x}}[\mathbf{r}_3(\mathbf{x})N_3(\mathbf{x}, t)]}_{\text{Single cell growth 3}} \\ &= \underbrace{-\Gamma_3(\mathbf{x}, t)N_3(\mathbf{x}, t)}_{\text{Division}} + \underbrace{\Gamma_2(\mathbf{x}, t)N_2(\mathbf{x}, t)}_{\text{Transition 2} \rightarrow 3} \quad (18) \end{aligned}$$

Each equation provides a population balance for each cycle phase population N_i for $i = 1, \dots, 3$. This formulation is for a ‘constant environment’ since the substrate concentration s does not feature. Hence, it does not require supporting ordinary integro-differential equations characterising the dynamics of the substrate S . Phase-to-phase transitions (not necessarily division, which is one form of transition mathematically), described by Γ_i , can be characterised as either deterministic or stochastic (random transitions) and clearly, all phases will require these transitions for the populations

described to be coupled. A deterministic transition, for example, assumes that cells move to the next phase when they achieve a predetermined mass or age. A stochastic mechanism assumes that of all cells achieving the necessary physiological state, a proportion may make the transition to the next stage, albeit not with certainty. Each cell would have an equal chance of moving on to the next phase or remaining in the current phase. The distribution of the cellular properties at the phase transitions can be either, measured and used to generate a histogram that can be used directly to describe the transition rate, or be an approximation of the data using a parametric distribution. Irrespective of which approach is used the shape of the distribution will determine the transition rate.

Occasionally, instead of simulating the behaviour of a population based on assumed functions for single cell parameters, one may want to extract single cell information from known experimental population data, described as the inverse problem (Ramkrishna 2000). Ramkrishna et al. (1968) showed that, under special conditions, the physiological functions r , Γ , and p are related to experimentally obtainable cell properties and can thus be determined. For a batch culture under balanced exponential growth conditions, the number density function $N(m, t)$ can be decoupled and written as the product of two univariate functions: $f(m)$, the time-invariant density function for the total cell population, and $e^{\mu t}$, the exponential growth expression where μ is the specific growth rate:

$$N(m, t) = N_0 f(m) e^{\mu t} \quad (19)$$

Substitution of Equation (19) into a single variable mass structured balance equation, such as Equation (4), under batch culture conditions ($D = 0$) yields the following ordinary differential equation:

$$\begin{aligned} & \mu f(m) + \frac{d}{dm}[r(m)f(m)] \\ &= -\Gamma(m)f(m) + 2 \int_m^\infty \Gamma(m')p(m, m')f(m')dm' \quad (20) \end{aligned}$$

Ramkrishna et al. (1968) showed that the terms with the *a priori* division rate, Γ , and mass distributions, $f(m)$, can be rewritten using a *a posteriori* probability distributions for the state of dividing cells,

ϕ , and the newly formed daughter cells, ψ as shown:

$$\int_m^\infty p(m, m') \Gamma(m') f(m') dm' = \mu \psi(m) \quad (21)$$

$$\Gamma(m) f(m) = \mu \phi(m) \quad (22)$$

Integrating Equation (20) using Equations (21) and (22) leads to an expression that can be used to calculate the growth rate:

$$r(m) = \frac{\mu}{f(m)} \int_0^m 2\psi(m') - \phi(m') - f(m') dm' \quad (23)$$

Derivation of expressions for p , the partition function, and Γ , the division rate, can be found elsewhere (Ramkrishna 2000). Since these apply at balanced conditions (i.e., metabolism is at steady-state conditions) and substrate concentrations are thus constant, the dependence of all variables on substrate is removed and, of course, they are also time-invariant. For hybridoma cells, the relevant population data used to obtain the distributions ψ , ϕ and f were obtained by exploiting the light scattering property variations during mitosis (Kromenaker and Srienc 1991). Srienc (1999) notes that the physiological functions in Equations (20)–(22), despite the restrictive conditions under which the equations are derived, hold even under transient conditions. This is because they are functions of the physiological state of the cell. Therefore, for a given mass m , the functions are valid whether the environment is constant or dynamic, provided they have been evaluated at m .

Combined single cell and population balance models

The combination of SCMs and PBMs represents the next logical challenge. Despite the additional model fidelity, such a hybrid model is extremely computationally intensive; hence the solution of even the PBM component of an overall model becomes intractable. To overcome this problem, investigators have used finite-representation techniques to discretise populations avoiding the problems of continuous distributions and the integral differential features they bring (Domach and Shuler 1984; Kim and Shuler 1990). However, tractability can

be maintained for the coupling of simpler PBMs to SCMs (Martens et al. 1995; Morales 2001 unpublished). Equation (24) shows how such models could be coupled:

$$\begin{aligned} f(N_t, N_x N, r, p, \Gamma, s, m) &= 0 & \text{PBM} \\ \Downarrow r = x_i \equiv \mu & \\ \mathbf{f}(\dot{\mathbf{y}}, \mathbf{y}, \dot{\mathbf{x}}, \mathbf{x}, \theta, t) &= \mathbf{0} & \text{SCM} \end{aligned} \quad (24)$$

The SCM, with its detailed single-cell structure, can be used to reliably predict the single-cell information that is needed by the PBM in the form of the physiological functions (in this case Equation (24), the single-cell growth rate). As a result, this greatly enhances the theoretical fidelity of PBMs and makes use of ongoing developments in improving SCM structure.

Increasingly, investigators are using the cell cycle upon which to build models of cell growth and death. Martens et al. (1995) presented a model for steady-state continuous hybridoma cultures. The overall model consisted of two sub-models: a PBM describing the cycling populations and an unstructured model of metabolism. The cell cycle model was based on that of Cazzador and Mariani (1993) and is composed of a two-phased cycling population and an apoptotic population. Key assumptions relating to cell growth and death included the cells entering apoptosis from each cycling phase and apoptotic death occurring at a fixed time, with necrotic death occurring in the cycling and apoptotic phases at equal rates. The specific growth and death rates were calculated by the cell-cycle model and fed to the metabolic model. Even though parameter values from the literature were used, the cell-cycling part of the overall model appeared to predict cell viability and specific growth well, although variables relating to the metabolism were predicted less accurately.

Cain and Chau (1998) have extended a cell cycle population model to account for product synthesis and export. It assumed two stages: an A-state, which represented the immediate post mitotic G_1 phase, and a B-state representing late G_1 , S , G_2/M . Hence, the model consisted of two population balance equations. A-state cell density was distributed over a single mass property while the B-state was distributed over both a local B-state age and mass property. Both balances considered cell death using a death rate term in addition to usual transition

rate and dilution terms. The varying substrate environment was expressed as an ODE, with an additional term to account for substrate utilisation for cycling cell maintenance. The duration of the B-state was constant, yet transition from A to B was random. The transition rate, in contrast to the earlier treatment, was characterised as a Monod-like function with a substrate dependency. Two possible formulations were proposed to account for protein production, depending on the availability of mRNA and protein product data. The ‘direct protein’ model explicitly considered the product protein and consisted of two equations for the intracellular accumulation and secreted accumulation rates, while the ‘transcription’ model represented this property as the specific mRNA content of a cell.

Modelling tools

The development of modelling tools is as important as the development of models themselves. Models are gradually bringing order to biology and there is now an urgent need for new tools (Bailey 1998) to bring order and understanding to the resulting myriad of models; to screen them, scrutinise them, and refine them. It has always been the case in the past that mathematical modelling and its methodologies were utilised once culture experiments have been performed and the data obtained. However, these data will more than certainly not be informative enough to uniquely estimate the most appropriate model. The method, therefore, of obtaining useful data by these means has been one of trial and error. Another difficulty is that biological phenomena nearly always include relatively fast or slow dynamics and only a limited number of samples can be obtained due to either financial or practical limitations. This leads to data that are not equidistant in time, which results in some samples being taken at operating conditions or times when there are relatively few changes occurring, in turn yielding little additional information. In contrast, highly dynamic occurrences are tracked with relatively few samples and would benefit by additional highly targeted sampling in decisively validating a given model. A systematic framework exists that designs experiments in an optimal way so as to minimise the necessary

experimentation whilst simultaneously maximising information obtainable from the data, with the overall objective of achieving a uniquely validated model. In general, experiments are designed with one of two objectives in mind: discriminating between a number of models and identification (parameter precision) of a final model. In general, experimental design and subsequent parameter estimation form only two parts of a larger modelling strategy.

The abundance of available models necessitates the development of a systematic procedure to, first of all, compare models, validate the best model, and finally optimise its predictive precision using experimental design in an effective manner. This constitutes model building and validation, and a systematic quantitative framework for this process is presented. The model building and validation strategy can be considered to comprise of three sequential stages: preliminary analysis, model discrimination, and identification or parameter precision (Asprey and Machietto 2000) as shown in Figure 3.

Theoretical Model Analysis (Stage I)

In trying to describe the biological processes under investigation, several different inferred or mechanistic models may be conceived without knowledge of which one to use. At the outset, before any data collection, we wish to know whether we can, in principle, design experiments so that the model parameters can be uniquely identified from the mathematical structure of the model in a noise-free context. This is known as model structural identifiability (Walter 1987). Once the identifiability test has been applied, those models that pass are retained and those that fail are rejected and are no longer considered, as their failure of the test indicates mathematical singularity with respect to the model parameters. Resources therefore are not committed to performing experiments, in which model parameters are known in advance to be unidentifiable. Structural identifiability, however, requires symbolic manipulation of the model and can only be applied to very small models, which effectively excludes all models where the number of equations and parameters summed exceed 10. Consequently, less theoretically rigorous identifiability tests are used, which are non-symbolic.

A new optimisation-based approach to dynamic model identifiability was proposed by Asprey and

Figure 3. Model building and validation strategy.

Mantalaris (2001). The model used was a modified form of the unstructured hybridoma cell model employed by Jang and Barford (2000). Modifications made by Asprey and Mantalaris (2001) accounted for cell volume changes, dissolved oxygen, and fed-batch operation. Identifiability was posed as an optimisation problem in that the procedure sought to maximise the variability of each parameter in the model subject to the output trajectories of the model being invariant. Performing additional experiments clearly cannot uniquely identify those parameters that, at the end of the optimisation, can take values within a range greater than a predefined tolerance and still cause the model to make the same prediction. Using this method, Asprey and Mantalaris (2001) were able to determine, prior to any experiments being performed, that the model could not be uniquely identified based on its parameters, and which were the problematic parameters (i.e., those that could take any value in a range and still not affect the model output). Their stated conclusion was that it would be futile performing any experiments with this model.

Following model identifiability and having a collection of identifiable models, the next issue is

model distinguishability, which is determining whether competing models can be distinguished from each other in terms of their mathematical structural form. Asprey and Machietto (2000) applied an optimisation-based approach to model distinguishability as well as identifiability. Using a fed-batch, yeast fermentation example, two small models consisting of three equations were proposed to characterise substrate consumption rates: one assuming Monod-type kinetics and the other Contois-type. The models were said to be distinguishable when, for the same initial conditions, their response trajectories did not overlap. Through the minimisation of the divergence between the trajectories of the two models and the identification of the maximum of these minima over the entire time horizon of interest, again subject to the constraints that the outputs of both models were identical, distinguishability manifested itself as two unsuperimposed trajectories within a given tolerance. The scalar quantity resulting from the optimal objective of the optimisation problem, a measure of the separation of these trajectories was found to be larger than the tolerance and so the models were shown to be distinguishable.

Model discrimination (Stage II)

The task of the model discrimination stage is to select the best model from the available set of competing identifiable and distinguishable models, through the use of experiments. Experiments designed for this purpose are termed T-optimal. Using the experimental data, parameter estimation is performed for each model to maximise the consistency of each model relative to the data (not to improve model accuracy). The residuals (deviations between experimental data and simulated data) of each of the models resulting from the parameter estimation step are a result of two contributing factors: noise corrupted experimental data and lack-of-fit of the model itself. Model discrimination experiments aim to maximise the contribution due to lack-of-fit and hence the divergence between the model predictions, thus facilitating comparison. Once the residuals have been obtained after parameter estimation, the information contained in them is used to assess the degree of lack-of-fit in each model – a step known as model adequacy testing (Mason et al. 2003). Models can be rejected if their lack-of-fit is unacceptably high or retained if within statistically acceptable limits. Since after each adequacy test there may be more than one adequate model, new experiments must be designed to amplify the lack-of-fit of the remaining models so that after the next adequacy test is performed, one or more will fail and can be rejected. This process is continued until a single model remains, which represents the best model. There is, however, an inescapable flaw in this procedure. Given that the experiments performed have been used to yield the best model, the resulting model is only that – a best model and potentially not the *true* model. One can never know with certainty whether the best model predicted is in fact the true model.

Model precision (Stage III)

Following the selection of the best model (from Stage II), new experiments are designed to improve model accuracy by reducing the parameter uncertainty inherent in all parameter values. Decreasing the size of the joint parameters inference region increases parameter precision. Parameter uncertainty is expressed through the size of the

variance– covariance matrix:

$$\mathbf{V} = \mathbf{V}(\hat{\theta}, \varphi) \quad (25)$$

where, $\hat{\theta}$ is the vector of best model parameter estimates and φ is the vector of experiment decision variables and contains variances of individual parameters and covariances of pairs of parameters. Optimal experiment design criteria seek to reduce some metric of this matrix (or increase the same metric of the inverse of $\mathbf{V}$, known as the Fisher Information Matrix $\mathbf{M}$) – the fact that matrix size is quantifiable in a number of ways leads to the existence of several criteria: D-optimal designs minimise the volumes of the confidence ellipsoids (i.e., determinant of $\mathbf{V}$) and are independent of the parameter scales; A-optimal designs minimise the average variance of the parameter estimates (i.e., trace of $\mathbf{V}$) and are dependent on parameter scales; C-optimal designs minimise the average squared coefficient of variation (i.e., trace of $\mathbf{C}^{-1}\mathbf{V}$ where $\mathbf{C}^{-1}$ is a matrix of squares of parameter values at the point of minimum residuals) and are parameter scale independent; E-optimal designs minimise the length of the principle axis of the confidence ellipsoid (i.e., length of the largest eigenvalue of $\mathbf{V}$) in parameter space; M-optimal designs (Nathanson and Saidel 1985) are multi-objective, trading off the reduced correlation between parameter estimates by minimising the angle between indifference ellipsoid axes and the reference axis in the parameter space, with the volume reduction obtained from implementation of a D-optimal criterion. There has been extensive application of these techniques in microbial systems (Versyck et al. 1997, 1999; Bernaerts et al. 2000). These investigators have tended to use E-optimal designs, as these types of experiment designs aim at reducing correlations between parameters by trying to make the parameter inference region more circular. However, an important issue that these investigators fail to address is that the mathematical property of the E-optimal criterion (the ratio of eigenvalues) means that it is discontinuous and will therefore cause convergence problems when used with a gradient-based optimisation routine. In general, D-optimal designs are most widely used (Jacques 1998).

To overcome the shortcomings and the restrictive nature of the criteria mentioned earlier, two areas of research are being developed: (i) robust

design criteria and (ii) design of dynamic experiments. From the definition of $\mathbf{V}$, it is clear that the value of this matrix is dependent on the vector $\hat{\theta}$ and so will any experiment design. Robust designs seek to take into account the variability of $\hat{\theta}$ itself, resulting in designs that are relatively more insensitive to initial parameter estimates. In dynamic experiments, controlled inputs are free to take on different values during experiments. Hence, by changing controls dynamically, more freedom is introduced into the optimisation process of design allowing for potentially even more information-rich designs. Recently, Asprey and Machietto (2000) embedded the D-optimal design criteria into an optimal control framework but used a modified form of a dynamic information matrix first defined by Zullo (1991), Körkel et al. (1999) considered optimal design of experiments with models described by general DAE systems and Asprey and Machietto (2002) presented work on the design of robust dynamic experiments, with multi-response non-linear models where the new form accommodates point estimate uncertainty in parameter values.

In the future, one can potentially envisage designs that are not sequential but where more than one experiment can be designed and performed simultaneously, leading to a higher rate of information attainment (parallel designs). The increase in computing power, the improved optimisation algorithms, and the sufficiently reduced models would allow experiment design and model improvement processes to occur in minimal time, hence assisting this process to occur in a real-time closed-loop fashion.

Advancing model fidelity and complexity

Modelling

The recent explosion in developments and information emanating from, especially, the field of genomics has created a new state-of-the-art starting point for research in both model- and biology-based research. Bailey (1998) highlighted the ‘yawning chasm’ between the current state in genomics and mathematically and computationally orientated research. There is clearly a huge gulf that has emerged between having genomic information and using it to characterise a higher order

description, a problem that is at present dealt with just by functional genomics. Shuler (1999) hinted on the use of SCMs as a means of bridging the gap between ‘genetic knowledge and physiology’ and it has already been shown how stoichiometric models, using flux balance analysis, can be developed to analyse various aspects of genotype–phenotype behaviour that relate to *E. coli* cellular metabolic activity (Schilling et al. 1999), such as the effects on growth and metabolic shifts of gene knock-outs. Metabolic phenotyping is presently the only facet of ‘phenomics’, the prediction of genotype–phenotype relations from genomic information (Schilling et al. 1999), which can be modelled as it is supported by decades of research into metabolism. Limiting the modelled genome to those genes that are known to be significant to cell death, productivity, and cell cycle, would be a tractable extension to current dynamic models keeping the modeled gene set to within a manageable size. Additionally, the potential identification of a minimal set of genes required for cell survival may result in a small enough number of genes to make such modelling feasible. However, the realisation that certain cell functions are multigeneic, that is there is a many-to-one mapping between gene and function, giving rise to potentially very complex gene–gene regulatory interactions further complicates an already complex situation. Hence, the addition of proteomic and metabolomic information currently represents the natural upstream extension to existing process models since these correspond to one level up in the cell complexity (Figure 4).

Dynamic modelling has been applied to molecular interactions governing the cell cycle (Tyson and Novak 2001) and how it affects limited aspects of cellular physiology. However, there are very few dynamic models developed, which include treatment of genes. One example is the cell simulation model, known as E-cell (Tomita et al. 1999), an ongoing project at Keio University in Japan. It is a modelling and simulation environment for biochemical and genetic processes and its software package allows for a user-defined model of a single cell. The theoretical cell has a gene set of 127; most of the genes are those of the microbial organism *Mycoplasma genitalium*, which has the smallest known genome and has been fully sequenced. The model attempts to unify multiple levels of biological processes and time-scale hierarchies by

considering long-term processes occurring at the DNA and RNA level, right through to shorter timescale processes such as expression of proteins, protein complexes and metabolite interactions. In E-Cell, the user defines the substances in the cell, which gene(s) is knocked out of the gene set, the composition of the external medium, the reaction equations (i.e., reactants and products), the relevant kinetics such as enzymatic (Michaelis-Menton type equation assumed) or non-enzymatic (elementary type kinetics assumed), and the spatial and/or functional structure of the cell and the environment. The output of the model is viewed through tracer windows showing the concentration–time

profiles of the selected components. Depending on the user requirements, the model allows the use of different ‘off-the-shelf’ numerical integration methods, so that preference for either speed of calculation or accuracy of results can be catered for. Differential equations are automatically generated, creating a dynamic model from the user specified information. Major assumptions have been used in that the model does not account for genetic regulatory factors such as enhancers and repressors because gene switching is not included (gene expression is not described), the actual nucleotide and amino acid sequences are not used, and cell growth, DNA replication, chromosome segregation and cell division are not accounted for either.

It is becoming increasingly clear that structured models represent the best way forward as they are constructed on the basis of biological mechanisms as opposed to empirical relations. Since they attempt to replicate internal processes and not just match output data, they offer more confidence in their validity beyond limited ranges and conditions. Furthermore, they are easier to modify and extend as better and more detailed information becomes available, since incorrect mechanisms or sub-processes can be isolated in the form of modular computer code. Although many animal cell models exist in the literature, most focus on characterising only selected parts of the overall cell metabolism. For example, some describe only the product formation side of metabolism, while others the primary metabolism. These are essentially sectional models of cellular behaviour. Whole cell

Figure 4. The evolution process of bioprocess models.

Figure 5. Model development and design of experiment symbiosis.

models are necessarily much larger, comprising of large sets of equations and variables and therefore it takes considerably longer to formulate and solve. Batt and Kompala (1989), Wu et al. (1992), and Sanderson (1997) are examples of models in which attempts were made to simulate whole cell behaviour of hybridoma, CHO, and generic mammalian cells, dealing with single cell level metabolic and transport processes.

Currently, there are no models that accommodate the more complex regulatory and genetic processes. With the recent advances in genomics, along with proteomics and metabolomics, the time has undoubtedly come, at the very least, to consider addressing the incorporation of basic genetic and proteomic features in cell models as demonstrated in E-cell. However, single cell modelling should represent only one aspect of future efforts since, at a practical level, it is the knowledge of how to describe an entire cell culture that has any industrial value. It is imperative therefore that equal consideration is given to this task, which is best tackled through the use of PBMs despite being at this time restricted by numerical methods to using simple forms of this model type.

Tools

To be able to harness the leaps in genomics/proteomics/metabolomics, biochemical engineers must become aware and involved in the area of experiment design and its role into the model building process. Competing models can and should be screened prior to experimentation. Rigorous statistical validation should be performed on promising models and experiments should be model-based to enhance parameter precision. Through the rigorous application of these methods, progress can be made much more efficiently and effectively in terms of time and cost than is currently the case. Furthermore, should techniques be developed in the future to enable closed-loop real-time parallel design of dynamic experiments, the model building and validation timeline could be reduced drastically, by orders of magnitude (Figure 5). With the aid of better-proposed structured models that have withstood the scrutiny of tough validation methods, a better understanding of cellular physiology and the convergence of biochemical engineering

process models and genomics/proteomics/metabolomics could be achieved in the future.

Acknowledgements

The authors would like to thank the UK EPSRC for FRS' PhD studentship support.

References

- Asprey S.P. and Machietto S. 2000. Statistical tools for optimal dynamic model building. *Comput. Chem. Eng.* 24: 1261–1267.
- Asprey S.P. and Mantalaris A. 2001. Global parametric identifiability of a dynamic unstructured model of hybridoma cell culture. In: *Proceedings of the 8th International Conference on Computer Applications in Biotechnology*.
- Asprey S.P. and Machietto S. 2002. Designing robust optimal dynamic experiments. *J. Process Control* 12: 545–556.
- Bailey J.E. 1998. Mathematical modeling and analysis in biochemical engineering. *Biotech. Progr.* 14: 8–20.
- Barford J.P., Phillips P.J. and Harbour C. 1992. Simulation of animal cell metabolism. *Cytotechnology* 10: 63–74.
- Batt B.C. and Kompala D.S. 1989. A structured kinetic modeling framework for the dynamics of hybridoma growth and monoclonal antibody production in continuous suspension cultures. *Biotech. Bioeng.* 34: 515–531.
- Bernaerts K., Versyck K.J. and Van Impe J.F. 2000. On the design of optimal dynamic experiments for parameter estimation of a ratkowsky-type growth kinetics at suboptimal temperatures. *Int. J. Food Microbiol.* 54: 27–38.
- Cain S.J. and Chau C.C. 1998. Transition probability cell cycle model with product formation. *Biotech. Bioeng.* 58: 387–399.
- Cazzador L. and Mariani L. 1993. Growth and production modelling in hybridoma continuous culture. *Biotech. Bioeng.* 38: 781–787.
- Cruz H.J., Moreira J.L. and Carrondo M.J.T. 1999. Metabolic shifts by nutrient manipulation in continuous cultures of BHK cells. *Biotech. Bioeng.* 66: 104–113.
- Dalili M., Sayles G.D. and Ollis D.F. 1990. Glutamine-limited batch hybridoma growth and antibody production: experiment and model. *Biotech. Bioeng.* 36: 74–82.
- Diekmann O., Heijmans H.J.A.M. and Thieme H.R. 1984. On the stability of the cell size distribution. *J. Math. Biol.* 19: 227–248.
- Domach M.M. and Shuler M.L. 1984. A finite representation model for an asynchronous culture of *E. coli*. *Biotech. Bioeng.* 26: 877–884.
- Duncan A. 2002. Antibodies hold the key. *Chemistry & Industry*.
- Eakman J.M., Fredrickson A.G. and Tsuchiya H.M. 1966. Statistics and dynamics of microbial cell populations. *Chem. Eng. Progr.* 62: 37–49.
- Europa A.F., Gambhir A., Fu P.C. and Hu W.S. 2000. Multiple steady states with distinct cellular metabolism in continuous culture of mammalian cells. *Biotech. Bioeng.* 67: 25–34.

- Follstad B.D., Balcarcel R.R., Stephanopoulos G. and Wang D.I.C. 1999. Metabolic flux analysis of hybridoma continuous culture steady state multiplicity. *Biotech. Bioeng.* 63: 675–683.
- Frame K.K. and Hu W.-S. 1991. Kinetic study of hybridoma cell growth in continuous culture. ii. behavior of producers and comparison to nonproducers. *Biotech. Bioeng.* 38: 1020–1028.
- Fredrickson A.G., Ramkrishna D. and Tsuchiya H.M. 1967. Statistics and dynamics of procaryotic cell population. *Math. Biosci.* 1: 327–374.
- Gombert A.K. and Nielsen J. 2000. Mathematical modelling of metabolism. *Curr. Opin. Biotech.* 11: 180–186.
- Gray L. and Jasuja R. 2001. B-147 The New Future of Biotechnology: Enabling Technologies and Star Products. Business Communication Company, Inc.
- Hatzis C., Srien F. and Fredrickson A.G. 1995. Multistaged corpuscular models of microbiol growth: Monte carlo simulations. *Biosystems* 36: 19–35.
- Hiller G.W., Aeschlimann A.D., Clark D.S. and Blanch H.W. 1991. A kinetic analysis of hybridoma growth and metabolism in continuous suspension culture on serum-free medium. *Biotech. Bioeng.* 38: 733–741.
- Jacques J.A. 1998. Design of experiments. *J. Franklin Inst.* 335B: 259–279.
- Jang J.D. and Barford J.P. 2000. An unstructured kinetic model of macromolecular metabolism in batch and fed-batch cultures of hybridoma cells producing monoclonal antibody. *Biochem. Eng. J.* 4: 153–168.
- Kim B.-G. and Shuler M.L. 1990. A structured, segregated model for genetically modified *E. coli* cells and its use for prediction of plasmid stability. *Biotech. Bioeng.* 36: 581–592.
- Körkel S., Bauer I., Bock H.G. and Schloder J.P. 1999. A sequential approach for nonlinear optimum experimental design in dae systems. In: *Proceedings of the International Workshop on Scientific Computation in Chemical Engineering*.
- Kromenaker S. and Srien F. 1991. Cell-cycle-dependent protein accumulation by producer and nonproducer murine hybridoma cell lines: a population analysis. *Biotech. Bioeng.* 38: 665–677.
- Kurokawa H., Park Y.S., Iijima S. and Kobayashi T. 1994. Growth characteristics in fed-batch culture of hybridoma cells with control of glucose and glutamine concentration. *Biotech. Bioeng.* 44: 95–103.
- Lee Y.-K., Yap P.-K. and Teoh A.-P. 1995. Correlation between steady-state cell concentration and cell death of hybridoma cultures in chemostat. *Biotech. Bioeng.* 45: 18–26.
- Linardos T.I., Kalogerakis N. and Behie L.A. 1991. The effect of specific growth rate and death rate on monoclonal antibody production in hybridoma chemostat culture. *Can. J. Chem. Engin.* 69: 429–438.
- Linz M., Zeng A.-P., Wagner R. and Deckwer W.-D. 1997. Stoichiometry, kinetics, and regulation of glucose and amino acid metabolism of a recombinant BHK cell line in batch and continuous cultures. *Biotech. Progr.* 13: 453–463.
- Liou J.J., Srien F. and Fredrickson A.G. 1997. Solutions of population balance models based on a successive generations approach. *Chemical Eng. Sci.* 52: 1529–1540.
- Lüdemann I., Pörtner R., Schaefer C., Schick K., Sramkova K., Reher K., Neumaier M., Franek F. and Märkl H. 1996. Improvement of culture stability of non-anchorage-dependent cells grown in serum-free media through immobilization. *Cytotechnology* 19(2): 111–124.
- Mantzaris N.V., Liou J.J., Daoutidis P. and Srien F. 1999. Numerical solution of a mass structured cell population balance model in an environment of changing substrate concentration. *J. Biotech.* 71: 157–174.
- Mantzaris N.V., Daoutidis P. and Srien F. 2001a. Numerical solution of multi-variable cell population balance models. I. Finite difference methods. *Comput. Chem. Eng.* 25: 1411–1440.
- Mantzaris N.V., Daoutidis P. and Srien F. 2001b. Numerical solution of multi-variable cell population balance models. II. Spectral methods. *Comput. Chem. Eng.* 25: 1441–1462.
- Mantzaris N.V., Daoutidis P. and Srien F. 2001c. Numerical solution of multi-variable cell population balance models. III. Finite element methods. *Comput. Chem. Eng.* 25: 1463–1481.
- Martens D.E., Sipkema E.M., de Gooijer C.D., Beuvery E.C. and Tramper J. 1995. A combined cell-cycle and metabolic model for the growth of hybridoma cells in steady-state continuous culture. *Biotech. Bioeng.* 48: 49–65.
- Mason R.L., Gunst R.F. and Hess J.L. 2003. *Statistical Design and Analysis of Experiments: With Applications to Engineering and Science*, John Wiley & Sons, New York, USA.
- Morales J.A.A. 2001. Dynamic modelling of mammalian cell culture systems. MSc thesis, University of London.
- Nathanson M.H. and Saidel G.M. 1985. Multiple-objective criteria for optimal experimental design: application to ferrokinetics. *Am. J. Physiol.* 248: R378–R386.
- Paredes C., Prats E., Cairo J.J., Azorin F., Cornudella L. and Godia F. 1999. Modification of glucose and glutamine metabolism in hybridoma cells through metabolic engineering. *Cytotechnology* 30: 85–93.
- Phillips P.J. 1996. *The interaction of Experimentation and Computer Modelling for Animal Cell Culture*, University of Sydney.
- Pörtner R. and Schäfer T. 1996. Modelling hybridoma cell growth and metabolism – a comparison of selected models and data. *J. Biotech.* 49: 119–135.
- Pörtner R., Schilling A., Lüdemann I. and Märkl H. 1996. High density fed-batch cultures for hybridoma cells performed with the aid of a kinetic model. *Bioprocess Eng.* 00: 000–000.
- Ramkrishna D. 2000. *Population balances: Theory and Applications to Particulate Systems in Engineering*. Academic Press, New York.
- Ramkrishna D. 1979. Statistical Models of Cell Populations. *Adv. Biochem. Eng.* 11: 1–47.
- Ramkrishna D., Fredrickson A.G. and Tsuchiya H.M. 1968. On the relationships between various distribution functions in balanced unicellular growth. *Bull. Math. Biophys.* 30: 319–323.
- Sanderson C.S. 1997. *The Development and Application of a Structured Model for Animal Cell Metabolism*. Ph.D thesis, University of Sydney.
- Sanderson C.S., Barton G.W. and Barford J.P. 1995. Optimisation of animal cell culture media using dynamic simulation. *Comput. Chem. Eng.* 19: S681–S686.

- Schilling C.H., Edwards J.S. and Palsson B.O. 1999. Toward metabolic phenomics: analysis of genomic data using flux balances. *Biotech. Progr.* 15: 288–295.
- Shuler M. 1999. Single-cell models: promise and limitations. *Cytotechnology* 71: 225–228.
- Srienc F. 1999. Short communication: Cytometric data as the basis for rigorous models of cell population dynamics. *J. Biotech.* 71: 233–238.
- Tomita M., Hashimoto K., Takahashi K., Shimizu T.S., Matsuzaki Y., Miyoshi F., Saito K., Tanida S., Yugi K., Venter J.G. and Hutchison C.A.III 1999. E-cell: Software environment for whole-cell simulation. *Bioinformatics* 15: 72–84.
- Tsuchiya H.M., Fredrickson A.G. and Aris R. 1966. Dynamics of microbial cell populations. *Adv. Chem. Eng.* 6: 125–206.
- Tyson J.J. and Novak B. 2001. Regulation of the eukaryotic cell cycle: molecular antagonism, hysteresis, and irreversible transitions. *J. Theor. Biol.* 210: 249–263.
- Tziampazis E. and Sambanis A. 1994. Modeling of cell culture processes. *Cytotechnology* 14: 191–204.
- Versyck K.J., Claes J.E. and Van Impe J.F. 1997. Practical identification of unstructured growth kinetics by application of optimal experimental design. *Biotech. Progr.* 13: 524–531.
- Versyck K.J., Bernaerts K., Geeraerd A.H. and Van Impe J.F. 1999. Introducing optimal experimental design in predictive modeling: A motivating example. *Int. J. Food Microbiol.* 51: 39–51.
- Villadsen J. 1999. Short communication: On the use of population balances. *J. Biotech.* 71: 251–253.
- Walter E. 1987. *Identifiability of Parametric Models*. Pergamon Press, Oxford.
- Wu P., Ray N.G. and Shuler M.L. 1992. A single cell model of Chinese hamster ovary cells. *Ann NY Aca Sci* 665: 152–187.
- Zeng A.-P., Deckwer W.-D. and Hu W.-S. 1998. Determinants and rate laws of growth and death of hybridoma cells in continuous culture. *Biotech. Bioeng.* 57: 642–654.
- Zhou W., Rehm J., Europa A. and Hu W.-S. 1997. Alteration of mammalian cell metabolism by dynamic nutrient feeding. *Cytotechnology* 24: 99–108.
- Zullo L.C. 1991. *Computer-aided Design of Experiments: An Engineering Approach*. Ph.D thesis, University of London.