

# Model-Based Estimation of Myeloid Hematopoietic Progenitor Cells in Ex Vivo Cultures for Cell and Gene Therapies

H. Yang, E.T. Papoutsakis, W.M. Miller

Department of Chemical Engineering, Northwestern University, 2145 Sheridan Road, Evanston, IL 60208-3120; telephone (847) 491-4828; fax: (847) 491-3728; E-mail: wmmiller@northwestern.edu

Received 8 February 2000; accepted 20 August 2000

**Abstract:** Ex vivo production of hematopoietic progenitor cells has potential applications for cell therapy to alleviate cytopenias associated with chemotherapy and for gene therapy. In both therapies, progenitor and stem cells are considered crucial factors for therapeutic success. Assays for progenitor cells, however, take 2 weeks to complete, which is similar to the length of a typical culture. Therefore, a real-time estimation of the percentage or number of progenitor cells, based on rapid measurements, would be useful for optimization of feeding and harvest decisions. In this study, metabolic activity assays and flow cytometric analysis were used to estimate the content of progenitor cells. The measured metabolic activities are a collective contribution from all types of cells. Cells in granulomonocytic cultures have been lumped into six cell types and metabolic rates have been modeled as a linear function of cell composition and growth rate and as a nonlinear function of cell density. Data from 24 experiments were utilized to determine the model parameters in a calibration step. These data include flow cytometric analysis of more mature hematopoietic cells, progenitor cell colony assays, total cell content, and metabolite concentrations, and cover a wide range of cell composition, cell density, and growth rate. After calibration, the model is able to deliver good predictions of progenitor cell content for cultures with higher percentages of progenitor cells, as well as the peak progenitor cell content, based only on parameters that can be rapidly measured. With the aid of those predictions a harvest strategy was developed that will allow optimizing the harvest time based on the culture kinetics of each patient or donor inoculum, rather than using retrospective analysis to determine a uniform harvest time. © 2000 John Wiley & Sons, Inc. *Biotechnol Bioeng* 72: 144–155, 2001.

**Key words:** prediction of %CFC; metabolic activities; ex vivo hematopoietic culture; chemometric modeling; harvest strategy

## INTRODUCTION

Ex vivo hematopoietic cultures offer great promise in cell therapies for restoring in vivo hematopoiesis in patients fol-

lowing chemotherapy or radiation therapy (Koller and Pals-son, 1993; McAdams et al., 1996). Since chemotherapy and radiation therapy are designed to kill rapidly growing cancer cells, proliferating hematopoietic progenitor cells also suffer considerable damage, leading to, e.g., severe neutropenia (considerable reduction in the number of infection-fighting neutrophils). Ex vivo expansion can be a very useful tool to abrogate neutropenia, depending on the transfused amounts of granulocytic progenitor and postprogenitor cells (Brugger et al., 1995; Colter et al., 1996; Nielsen et al., 1998; Reiffers et al., 1999; Scheduling et al., 1999). Another potential application of hematopoietic cell culture is for gene therapy for a wide range of disorders and to provide drug resistance (Crystal, 1995; Havenga et al., 1997). A large number of cycling stem and progenitor cells are required for retroviral transduction, which is the most commonly used method.

Under different cytokine combinations and culture conditions, hematopoietic cultures expand differently (Koller et al., 1992; Piacibello et al., 1997). Besides those environmental factors, it has also been observed (Koller et al., 1996; Sandstrom et al., 1995) that ex vivo hematopoietic cultures initiated with different patient samples can exhibit large variations in the rate and extent of expansion of total cells and progenitor cells, especially at low inoculum density and at the beginning of cultures. In addition, the contents of different cell types are constantly changing during an experiment as a result of cell proliferation, differentiation, and death. The current practice of using a fixed harvest day (e.g., Day 10; Reiffers et al., 1999), regardless of sample variations, may therefore lead to culture harvests with sub-optimal contents of progenitor and postprogenitor cells that are transfused to patients, so that the effectiveness of cell or gene therapy is diminished.

Optimization of ex vivo hematopoietic cultures would improve the effectiveness of cell or gene therapy. However, an engineered system can only be optimized when the relevant system parameters are known or available. Hematopoietic cultures are heterogeneous in nature, containing different cell types at various stages of differentiation. Using flow cytometric analysis, more differentiated cell types in a

Correspondence to: W.M. Miller

Contract grant sponsor: National Science Foundation

Contract grant number: BES-9809730

mixed hematopoietic culture can be measured. This method relies on the observation that hematopoietic cells express different surface antigens (and levels) as they differentiate along different lineages. However, it is not possible to measure the progenitor cell content using flow cytometry. Progenitor cells are often present at low frequencies in a culture and cannot be uniquely identified using only a few markers. Progenitor cells are commonly evaluated based on their proliferative capacity to form colonies in semisolid medium, and hence are often called colony-forming cells (CFC). Such a colony assay requires a 2-week cultivation before evaluation. That assay time is about as long as a typical *ex vivo* hematopoietic culture, which makes colony assays of progenitor cells useless for real-time feeding/harvest optimization of a hematopoietic culture.

Despite the significance of nutrient uptake and by-product secretion in mammalian cell cultures, relatively little attention has been given to metabolism in hematopoietic cultures. Collins et al. (1997) have shown that the maximum percentage of cells that are CFC (%CFC) coincides with the peaks of specific lactate production rate and glucose consumption rate for cultures of cord blood (CB) mononuclear cells (MNC), peripheral blood (PB) MNC, and PB CD34+ cells carried out in spinner flasks and in T flasks. It has been shown that the correlation between %CFC and specific lactate production rate can be well described by a two-population (CFC and other cells) model for experiments with similar/identical inoculum density, cytokine combination, and culture conditions (Collins et al., 1997). However, the model parameters can vary substantially at different inoculum densities, cytokine combinations, initial %CFC contents, and culture conditions. Hematopoietic cell cultures are heterogeneous, containing as many as 100 distinguishable cell types, and the cell distribution can vary greatly with the cytokines and culture conditions used. It is unreasonable to evaluate the metabolic contributions of so many cell types. However, the number of cell types can be reduced to some extent by using cytokines that promote differentiation along particular lineages. In this article we focus on two myeloid lineages: granulocytes (neutrophils) and monocytes. Expanded granulocytic cells were examined for transfusion to abrogate severe neutropenia, and significant numbers of monocytes are often produced as by-products in granulocytic cultures.

In this work, we extend the two-population model (Collins et al., 1997) to a multipopulation model, and hypothesize that the parameter variation for the two-population model could be eliminated when using a more comprehensive model that considers variations of several relevant cell types, total cell density, and growth rate. Based on flow cytometric analysis using surface markers CD15 and CD11b, the relevant cell types for the granulocytic and monocytic lineages can be lumped into six groups (Terstappen et al., 1990) for characterization of cell metabolism. It is also important to consider the effects of growth rate and cell density on cell metabolism. Thus, the measurable consumption/formation rate of a metabolite in granulomonocytic cell cultures can be modeled as a function of six cell types (groups), growth rate, and cell density. When extending the two-population model (Collins et al., 1997), several possible models that differ in the number of model parameters and the degree of nonlinearity can be proposed. The objective of this work is to select a mathematical model (with a minimal set of model parameters) that can reasonably predict %CFC. Prediction of %CFC or CFC density in a myeloid hematopoietic culture follows two steps: 1) evaluation of the specific metabolic rates of CFC and other cell types by deconvolution of the measured cell composition, cell density, growth rate, and overall metabolic rates using multivariate regression methods; and 2) prediction of the CFC density or %CFC by incorporating these specific metabolic rates (model parameters) into the model. Based on the predicted (estimated) CFC density or %CFC, a harvest strategy was developed for *ex vivo* hematopoietic cultures expanded for cell or gene therapy purposes.

cytic cell cultures can be modeled as a function of six cell types (groups), growth rate, and cell density. When extending the two-population model (Collins et al., 1997), several possible models that differ in the number of model parameters and the degree of nonlinearity can be proposed. The objective of this work is to select a mathematical model (with a minimal set of model parameters) that can reasonably predict %CFC. Prediction of %CFC or CFC density in a myeloid hematopoietic culture follows two steps: 1) evaluation of the specific metabolic rates of CFC and other cell types by deconvolution of the measured cell composition, cell density, growth rate, and overall metabolic rates using multivariate regression methods; and 2) prediction of the CFC density or %CFC by incorporating these specific metabolic rates (model parameters) into the model. Based on the predicted (estimated) CFC density or %CFC, a harvest strategy was developed for *ex vivo* hematopoietic cultures expanded for cell or gene therapy purposes.

## EXPERIMENTAL CONSIDERATIONS

One hundred seventy-seven datasets (individual time points) from 24 experiments of 12 granulocytic (G), seven monocytic (M), and five granulomonocytic (G/M) cultures were collected to investigate the relationship between the percentages of different cell types and the metabolic activities by means of multivariate linear regression. All cultures were performed in T-flasks at the same pO<sub>2</sub> (5%) and temperature (37°C) and were inoculated with either CD34+ cells or PB MNC at different CD34+ cell contents. Different feeding strategies were employed in different experiments to maintain cell density at different levels. Detailed descriptions of the materials and methods can be found in Collins et al. (1997) and Hevehan et al. (2000a).

A typical experiment lasted an average of 14 days. Each dataset at a given time point included metabolic (lactate and glucose) activities, viable cell density, and percentages of different cell types, as assessed by flow cytometry and colony assay. Estimation of metabolic activities was based on daily metabolite concentration measurements. In most experiments, flow cytometric analysis was conducted every 2 days. Thus, about seven datasets were available per experiment. However, only the datasets between Days 3 and 10 were used for further study, reducing the total available datasets from 177 to 101. This is because in the first 2 days the culture exhibits a lag phase, and the proposed mathematical model may not be able to describe cell metabolism during this transient period. Also, after 10 days the CFC content is generally very low and often the change in metabolite concentrations is also small. Thus, the measurement of %CFC and the evaluation of metabolic rates often have large errors for later culture days and it would not make sense to include those measurements.

### Measurement of Metabolite Concentrations and Evaluation of Metabolic Rates

Glucose and lactate are the metabolites considered in this study. A glucose/lactate analyzer (YSI; Yellow Springs,

Youngstown, OH) was used to measure the concentrations of glucose and lactate, with respective error ranges of 5–10% and 2–5%. The concentration measurements can be performed in less than 1 min and were used to calculate the metabolic rates.

Evaluation of a metabolic rate is based on the concentration changes of that metabolite. The overall specific metabolic rate  $q_S$  at a time point  $t$  is defined as:

$$q_S(t) = \frac{1}{C_X(t)} \frac{dC_S(t)}{dt} \quad (1)$$

where  $C_X$  is the viable cell density and  $C_S$  is the concentration of metabolite  $S$  at the time point  $t$ . Since the measurements of metabolite concentrations are conducted in a discrete manner, the time derivative in the above equation can only be approximated based on the metabolite concentration differences, e.g., by using two-point or three-point methods.

A two-point method is a first-order estimation, based on the difference between the current and previous time steps, such as:

$$q_S(t) = \frac{1}{C_X(t)} \frac{dC_S(t)}{dt} \approx \frac{\ln \frac{C_X(t)}{C_X(t_-)}}{C_X(t) - C_X(t_-)} \frac{C_S(t) - C_S(t_-)}{t - t_-} \quad (2)$$

where  $t_-$  means the previous sampling time point and the log-mean cell density is used to represent the variable cell density over the interval. A time-weighted three-point method, which can be stated as follows:

$$q_S(t) \approx \frac{t_+ - t}{t_+ - t_-} \frac{\ln \frac{C_X(t)}{C_X(t_-)}}{C_X(t) - C_X(t_-)} \frac{C_S(t) - C_S(t_-)}{t - t_-} + \frac{t - t_-}{t_+ - t_-} \frac{\ln \frac{C_X(t)}{C_X(t_+)}}{C_X(t) - C_X(t_+)} \frac{C_S(t_+) - C_S(t)}{t_+ - t} \quad (3)$$

is a second-order estimation, allowing more accurate estimation of the metabolic rate, where  $t_+$  is the next sampling time point. However, the measurement results for the next time step are required for evaluation of the current metabolic rates, which will delay the whole chemometric analysis and thus is not as suitable for real-time CFC estimation.

### Measurements of Cell Density/Composition, Evaluation of Growth Rate, and Grouping of Cell Types

Viable cell density assays were conducted in two steps: total cell density measurement, and cell viability assay. Cell density was measured using a cell counter (Collins et al., 1997) and cell viability was evaluated by Trypan blue exclusion. The time to complete these two steps is about 0.5 h. By replacing the metabolite concentration in Eqs. (2) and (3) with the viable cell density, an estimate of the specific popu-

lation growth rate can be obtained, based on the two-point or three-point method.

To evaluate the percentages of different cell types, two measuring methods were required. The number of CFC—that is, the sum of granulocyte (CFU-G), monocyte (CFU-M), erythrocyte (BFU-E), and mixed (CFU-GM and CFU-Mix) progenitor cells—was determined using methylcellulose colony assays (Koller et al., 1992). Evaluation of the number of cells at more developed stages was performed with the aid of flow cytometric analysis (Terstappen et al., 1990; Hevehan et al., 2000a). Estimation of progenitor cells based on colony assays takes 2 weeks to complete, while flow cytometric analysis is a more rapid method, usually requiring in the range of 3 h.

Figure 1 shows a sample sequence of flow cytometric analysis on different days in a granulocytic hematopoietic culture initiated with CD34+ (CD15- and CD11b-; progenitor and quiescent) cells on Day 0 (Hevehan et al., 2000a). As the cells become more mature, the level of expression of CD15 and CD11b moves from CD15-/CD11b- (zone A) to CD15+/CD11b- (zone B), CD15++/CD11b- (zone C), and finally to CD15++/CD11b+ (zone D). During monocytic differentiation (not shown), cells first express CD11b (zone E) and then pick up low levels of CD15 (zone F). The number of events in different zones can be used to evaluate the cell numbers of different lineages at each differentiation stage. These six zones form the six cell groups for which the specific metabolic rates will be investigated at the calibration step for CFC estimation. Zones A–F correspond to: 1) progenitor/quiescent cells, 2) myeloblast/promyelocyte, 3) myelocyte, 4) metamyelocyte/band/neutrophil, 5) monoblast/promonocyte/monocyte, and 6) metamonocyte/macrophage, respectively (Terstappen et al., 1990). It should be noted that the cell number in zone A includes progenitor cells and quiescent (nonactive) cells of different types, as well as cells of other lineages that do not proliferate under the culture conditions due to the lack of suitable cytokines.

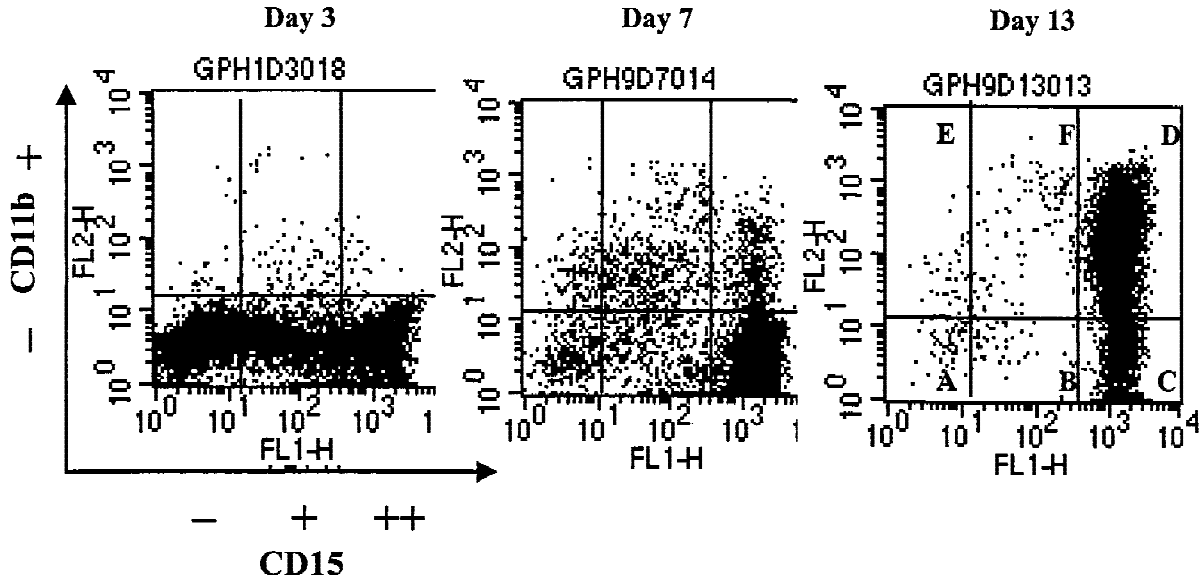
## MODELING CONSIDERATIONS

### Linear Chemometric Model

The two-population (CFC and other cells) model describing the correlation of %CFC with metabolic activity (Collins et al., 1997):

$$q_S(t) = \frac{\%CFC}{100} \alpha + \left(1 - \frac{\%CFC}{100}\right) \beta \quad (4)$$

with  $\alpha$  and  $\beta$  as the model parameters can first be extended to a linear multipopulation model by specifying other cells in the five granulocytic and monocytic cell groups noted above (Fig. 1). Assuming that quiescent cells and nonproliferative cells of other lineages have a negligible contribution to cell metabolism, the metabolic activity of cell group 1 (Fig. 1, zone A) comes mainly from progenitor cells.



**Figure 1.** Flow cytometric diagrams used for cell type grouping. -, +, and ++ mean no, moderate, and high surface marker levels, respectively (reproduced with permission from Hevehan et al., 2000a, Elsevier Science).

Thus, the total specific metabolic rate is the sum of individual metabolic activities of the six cell groups (Fig. 1), except that group 1 is replaced by only progenitor cells determined via colony assays. For a metabolite  $S$ , the overall (measured) specific metabolic rate,  $q_S$ , can be expressed as follows:

$$q_S(t) = \sum_{j=1}^p x_j(t) \beta_{S,j} \quad (5)$$

where  $x_j(t)$  and  $\beta_{S,j}$  are the respective cell fraction and specific metabolic rate of cell type  $j$ , and  $p$  ( $= 6$ ) is the number of cell types (groups) considered. The main assumption made here is that the specific metabolic rates for each cell type are constant at any culture time and in any cell composition. Metabolic activities of hematopoietic cultures are known to be dependent on environmental conditions, such as pH and the availability of cytokines, amino acids, dissolved oxygen, and glucose. Through regular feeding and saturation of these compounds, it is reasonable to assume that each individual specific rate remains almost constant during a culture.

When all  $n$  sets of measurements, which are gathered from different experiments and at different time points, are assumed to be independent of each other and to be uniformly distributed in the whole parameter space of G, M, and G/M cultures, we can then include and utilize all these datasets to determine the individual specific metabolic rates  $\beta_{S,j}$  from:

$$\begin{pmatrix} q_S(t_1) \\ \vdots \\ q_S(t_i) \\ \vdots \\ q_S(t_n) \end{pmatrix} = \begin{pmatrix} x_1(t_1) & \dots & x_j(t_1) & \dots & x_p(t_1) \\ \vdots & & \vdots & & \vdots \\ x_1(t_i) & \dots & x_j(t_i) & \dots & x_p(t_i) \\ \vdots & & \vdots & & \vdots \\ x_1(t_n) & \dots & x_j(t_n) & \dots & X_p(t_n) \end{pmatrix} \begin{pmatrix} \beta_{S,1} \\ \vdots \\ \beta_{S,j} \\ \vdots \\ \beta_{S,p} \end{pmatrix} \quad (6)$$

Chemometric methods used for estimation of  $\beta_{S,j}$  based on Eq. (6) will be discussed later. Once the values of  $\beta_{S,j}$  ( $j = 1, \dots, p$ ) become available, we can predict the %CFC by rewriting Eq. (5):

$$\% \hat{\text{CFC}}(t) = 100 \hat{x}_1(t) = \frac{1}{\beta_{S,1}} q_S(t) - \sum_{j=2}^p x_j(t) \frac{\beta_{S,j}}{\beta_{S,1}} \quad (7)$$

where %CFC is the prediction of %CFC and  $\hat{x}_1$  is the predicted value of  $x_1$ .

### Growth Rate Influence

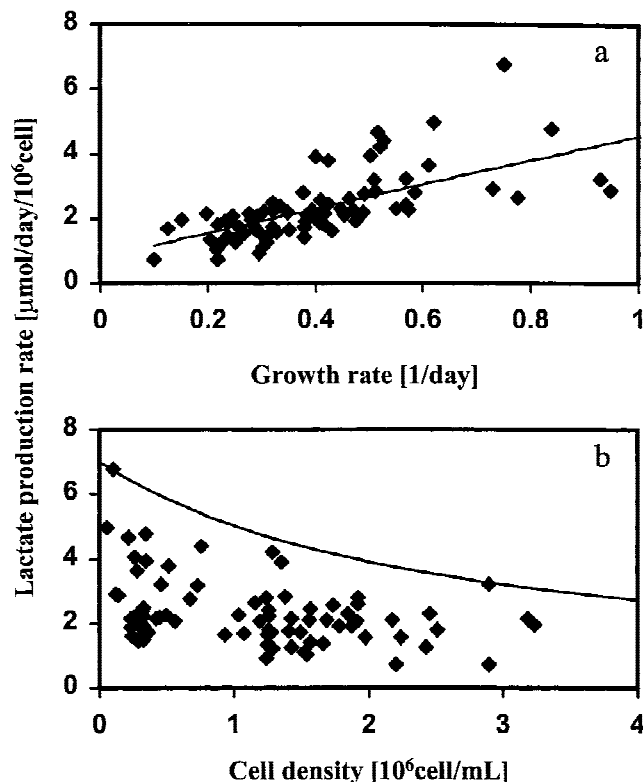
In the linear model (Eq. 5), the effects of growth rate have been neglected. However, it is well known that microbial metabolic formation/consumption rates are dependent on growth rate,  $\mu$ . For describing such dependence, mathematical models developed by Pirt (1965):

$$q_S = - \left( \frac{\mu}{Y_{X/S}} + m_S \right) \quad (8)$$

and by Ludeking and Piret (1962):

$$q_S = a_S \mu + b_S \quad (9)$$

have been widely used to describe microbial metabolic uptake and formation rates, respectively, where  $Y_{X/S}$  is the yield coefficient,  $m_S$  is the maintenance rate,  $a_S$  the growth-associated product formation coefficient, and  $b_S$  is the non-growth-associated product formation rate. We evaluated the effect of growth rate on specific lactate production and glucose consumption rates for 24 G, M, and G/M experiments. Although there is a lot of scatter in the data, there is a tendency for higher lactate production rate with increasing growth rate (Fig. 2a). The scatter in the data is probably due to the aforementioned culture heterogeneity, whereby other variables such as the fractions of different cell types are also



**Figure 2.** Effect of growth rate (a) and cell density (b) on lactate production rate. The metabolic rates were obtained by the three-point method. Each symbol represents a separate dataset (time point). The solid line in (a) is the regression line ( $R^2 = 0.46$ ), while the solid line in (b) represents the best fit of Eq. (10) to the maximum metabolic rate.

important for describing metabolic rates of neutrophils and monocytes.

### “Crowding” Effect

As for other mammalian cell cultures (Ozturk and Palsson, 1990), a crowding effect has been reported in the cultivation of hematopoietic cells (Talstad, 1971; Sand et al., 1977). The “crowding” effect means that cell metabolic activities diminish with increasing cell density. Figure 2b shows the effect of cell density on specific lactate production rate. Again, a scattered distribution of data is evident. However, it can also be clearly observed that the range of scattering shrinks, and the maximum metabolic rate declines, as cell density increases. The following inhibition kinetics are employed to describe the dependence of the maximum metabolic activities on cell density:

$$f_S(C_X) = \frac{K_S}{K_S + C_X} \quad K_S > 0 \quad (10)$$

where  $K_S$  is the model parameter to be evaluated.

### Modified Metabolic Model

To account for the growth rate influence, the growth-related term(s) could be added to the linear model (Eq. 5). To

incorporate the “crowding” effect into the model, one can assume that cell density and other factors (growth rate and cell types) affect cell metabolism independently. Thus, the total metabolic activity is made up of two parts. The first part describes metabolic activity at very low (“zero”) cell density, while the second part represents the dependence of the maximum metabolic activity on cell density. Based on the above discussion, the linear multipopulation model can be modified as follows:

$$q_S(t) = \sum_{j=1}^p x_j(t)(\alpha_{S,j}\mu_j(t) + \beta_{S,j}) \frac{K_{S,j}}{K_{S,j} + C_X(t)} \quad (11)$$

Comparing this model with Eq. (4), the model parameters ( $\alpha$  and  $\beta$ ) in the two-population model can be expressed as functions of cell densities, growth rates, and fractions of different cell types, as follows:

$$\beta(C_X, \mu, x_2, \dots, x_6) = \sum_{j=2}^p x_j(\alpha_{S,j}\mu_j + \beta_{S,j}) \frac{K_{S,j}}{K_{S,j} + C_X} \quad (12)$$

$$\alpha(C_X, \mu, x_2, \dots, x_6) = (\alpha_{S,1}\mu_1 + \beta_{S,1}) \frac{K_{S,1}}{K_{S,1} + C_X} + \beta(C_X, \mu, x_2, \dots, x_6)$$

When actually using the complex model (Eq. 11) to estimate the CFC density or %CFC, the parameter number triples compared to the model shown in Eq. (5), which requires a considerable increase in the number of experiments. Successful modeling often does not result in a complex model, which is able to represent every feature of a process perfectly in every aspect, but rather in a model that describes, on the one hand, the process reasonably well and has, on the other hand, a structure that is as simple as possible, with a minimal set of parameters. Thus, a reduced version should be considered.

In a reduced version, it is reasonable to assume that different cell types have similar metabolic responses to cell density variations. Hence, the function  $f_S(C_X)$  that describes the influence of cell density on metabolic activities will have the same parameter  $K_S$  for all cell types. A practical problem for using individual  $K_{S,j}$  is that evaluation of model parameters will require the use of nonlinear regression methods, since the model is nonlinear in terms of model parameters and cannot be linearized easily. Such nonlinear regression requires powerful optimization software and is out of the scope of this project.

Different cell types may proliferate at different rates. Using different  $\mu_j$  for different cell types means that growth rates for CFC and other cell types are needed as the model inputs when predicting %CFC. Real-time estimation of the growth rate for CFC, however, requires again real-time CFC measurement, and thus this kind of model using individual  $\mu_j$  is not useful. Furthermore, recent modeling efforts (Hevhan et al., 2000b) indicate that a single growth rate can be used for granulocytic cells ranging from CFC (CFU-G) to CD15+/CD11b+ cells, and that this composite growth rate varies with the culture conditions. In this work, the specific

cell growth rate  $\mu$  of the population is therefore used to replace the individual growth rates  $\mu_j$ .

The model can be further reduced by using either individual  $\alpha_{S,j}$  and a common  $\beta_S$  or individual  $\beta_{S,j}$  and a common  $\alpha_S$ . As will be shown later, %CFC prediction based on individual  $\beta_{S,j}$  is slightly better than that based on individual  $\alpha_{S,j}$ . Hence, a reduced version using individual  $\beta_{S,j}$ , a single  $\mu$ , a single  $\alpha_S$ , and a single  $K_S$  is proposed in this study. The model using individual  $\alpha_{S,j}$  and a single  $\beta_S$  is in the following referred to as the alternative model.

In order to eliminate the nonlinearity caused by the cell density term, a new variable  $y_S$  is introduced to describe  $q_S/f_S(C_X)$ . Considering the simplifications noted above, the proposed version of the metabolic model can then be rewritten as a linear model in terms of model parameters, namely:

$$y_S(t) = q_S(t) \frac{K_S + C_X(t)}{K_S} = \sum_{j=1}^p x_j(t) \beta_{S,j} + \alpha_S \mu(t) \sum_{j=1}^p x_j(t) \quad (13)$$

Analogous to Eq. (7), the prediction of %CFC can be carried out using the following equation:

$$\begin{aligned} \% \hat{\text{CFC}}(t) = 100 \hat{x}_1(t) &= \frac{1}{\beta_{S,1} + \alpha_S \mu(t)} y_S(t) \\ &- \sum_{j=2}^p x_j(t) \frac{\beta_{S,j} + \alpha_S \mu(t)}{\beta_{S,1} + \alpha_S \mu(t)} \end{aligned} \quad (14)$$

Furthermore, the CFC density (cells/mL; [CFC]) can be calculated from:

$$[\hat{\text{CFC}}(t)] = C_X(t) \times \frac{\% \hat{\text{CFC}}(t)}{100} \quad (15)$$

## Model Calibration and Data Processing

The key step in the prediction of %CFC or [CFC] is the calibration step for evaluation of  $\beta_{S,j}$  and  $\alpha_S$  by deconvolution of experimental data. For data deconvolution we employed multivariate linear regression methods, namely, least squares estimation (LS), ridge regression (RR), and principal component analysis (PCA) (see Appendix for more details).

An observation (measurement) is not only subject to measurement errors inherent in the analytical assay. Sometimes, human errors or method errors can also be added to these errors, such that misleading data or less representative (inconsistent) data have been collected. As method errors, we can include the errors caused by the approximation methods used to estimate the metabolic rates. PCA using the first two principal components was employed to detect outliers (measurements with large errors). In a biplot of these two components, each dataset is reduced to a single point. Outlier detection is based on the distance from a point to the central line that represents the general trend of all datapoints. Since %CFC is not available in real time, a matrix containing only metabolic activity and flow cytometric results is used for

PCA, corresponding to Eq. (14). The transfer matrix obtained in this way can be later employed to decide whether a set of real-time measurements is an outlier.

Parameter estimation is based on minimizing the sum of prediction errors over all datasets. Ideally, experimental datasets should be uniformly distributed in the parameter space. In this study, available datasets are, however, predominately of lower %CFC. Overrepresentation in a subspace means that the subgroup is weighted more heavily in the optimization process, which can bias the parameter estimation to some extent. Thus, some datasets have been removed from the databank, not because they are outliers, but simply because there are already enough similar datasets in the databank. Detection of overrepresentations is based on cluster analysis using the distance between two sets of data (Sharaf et al., 1986). The smaller the distance, the more similar the datasets are. When the distance from one dataset to another is smaller than a critical value, one of them will be removed from the list of the datasets used for calibration; a random number generator is employed to decide which dataset to remove.

Datasets that pass data processing (outlier and overrepresentation detection) are pooled together to form a databank. Only those datasets in the databank have been employed for model calibration.

## RESULTS AND DISCUSSION

### Off-Line Model Calibration

Model calibration can be carried out by means of LS, RR, or PCA, based on either glucose consumption rate or lactate production rate, using either the two-point or three-point approximation method (Eqs. 2 or 3, respectively). To compare different methods, the norm of %CFC prediction, which is also called the overall prediction error (PE) in this work, is introduced and defined as:

$$\begin{aligned} \text{PE} = \\ \|\% \hat{\text{CFC}} - \% \text{CFC}\| = \sqrt{\frac{\sum_{i=1}^n (\% \hat{\text{CFC}}(t_i) - \% \text{CFC}(t_i))^2}{n - (p + 1)}} \end{aligned} \quad (16)$$

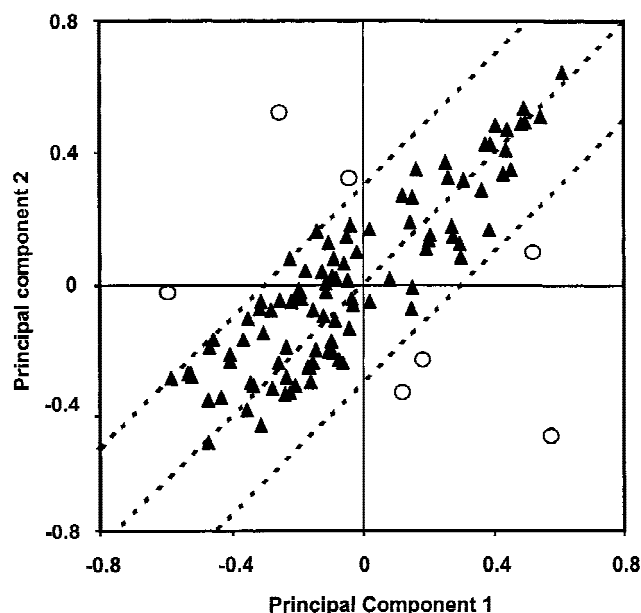
where  $p = 6$ . The smaller the PE value, the better the %CFC prediction and the more successful the model calibration.

Introducing  $y_S$  (Eq. 13) eliminates the nonlinearity caused by considering the ‘‘crowding’’ effect, and thus enables us to use multivariate linear regression for data deconvolution. For a given  $K_S$ , a set of  $\beta_{S,j}$  and  $\alpha_S$  can be determined by means of LS, RR, or PCA. Variation of  $K_S$  will lead to a new set of  $\beta_{S,j}$  and  $\alpha_S$ . Thus,  $K_S$  is determined via an iterative process by minimizing the PE (Eq. 16). The reason for considering different regression methods is to obtain an optimal set of  $\beta_{S,j}$  and  $\alpha_S$ , and to minimize the prediction errors of these parameters (see Appendix). Since the true

values of  $\beta_{S,j}$  and  $\alpha_S$  are unknown, the minimization of the prediction errors for  $\beta_{S,j}$  and  $\alpha_S$  is again based on minimizing the PE (Eq. 16). The best regression method was found to be the RR with  $k \approx 0.08$  (see Appendix), when model calibration is based on using the lactate production rate approximated by the three-point method. The improvement in reducing PE (Eq. 16) by means of RR in comparison to LS is, in this case, about 10%. All of the following studies employed the proposed model (Eq. 13) unless otherwise indicated.

When using the three-point approximation to estimate the lactate production rate, seven datasets were detected as outliers and were not put into the databank (Fig. 3). Table I shows that %CFC predictions involving different metabolic rates approximated by different methods yielded different number of outliers. In general, the two-point approximation of metabolic rates is less accurate than the three-point approximation, and thus there are more outliers. Since the lactate production rate is greater than the glucose consumption rate, there is a larger change in the lactate concentration. Together with the lower percent error for lactate measurement, this means that the approximation of the lactate production rate is more accurate, leading to fewer outliers (Table I). The %CFC for most datasets is below 5%, and thus almost all overrepresentations are for lower %CFC. All of the following predictions were carried out based on model calibration obtained after removal of outliers and overrepresentations. Removal of outliers decreased the prediction error (Eq. 16) by about 35%, while improvement of the data deconvolution based on removal of overrepresentations was negligible (<3% decrease in PE).

Different results are expected for the different models.



**Figure 3.** Outlier detection, using the three-point approximation for lactate production rate, based on the first two principal components. The middle (central) line shows the trend of all points and the top and bottom lines define the tolerance range (0.32; selected to minimize the overall prediction error). Outliers (○) are those outside this tolerance range.

The two-population model (Eq. 4) lumps all immature and mature G and M cells together, and the simple linear chemometric model (Eq. 5) considers only differentiation by specifying the 5 G and M cell groups, while the proposed model (Eq. 13) and the alternative model with individual  $\alpha_{S,j}$  and a common  $\beta_S$  also take into account growth rate and cell density. The performance of the proposed model vs. those of three other models are compared in Figure 4. All predictions are based on lactate production rate approximated by the three-point method (Eq. 3). It can be seen from Figure 4 that the proposed model delivers better results than the other three models. The PE values for the two-population model (Eq. 4), the linear chemometric model (Eq. 5), the alternative model, and the proposed model are 5.9, 4.7, 3.4, and 2.9, respectively. The improvement in %CFC prediction verifies our hypothesis that the parameter variations in a two-population model can be eliminated by a more comprehensive model (Fig. 4a). The prediction improvement also indicates that not only differentiation, but also the crowding effect and growth rate are important for characterizing hematopoietic cell metabolism (Fig. 4b). Furthermore, Figure 4c shows that the proposed model (using individual  $\beta_{S,j}$  and a single  $\alpha_S$ ) is slightly superior to the alternative model using individual  $\alpha_{S,j}$  and a single  $\beta_S$ .

The corresponding parameters for the proposed model are listed in Table II. The highest specific lactate production rate was found for CFC (7.73  $\mu\text{mol/h}/10^7\text{cell}$ ), while early G cells (CD15+/CD11b-) have a small lactate consumption rate (-0.46  $\mu\text{mol/h}/10^7\text{cell}$ ). These two distinctive values may account for the coincidental peaks for specific lactate production rate and %CFC (Collins et al., 1997). Immature (CD15+/CD11b-) and mature (CD15+/CD11b+) G cells (Table II) have similar metabolic rates (1.36 and 1.61  $\mu\text{mol/h}/10^7\text{cell}$ , respectively). These values are consistent with the observation (unpublished results from Y. Kuang and D. Pascoe) that the specific lactate production rate remains at a moderate level at the end of G cultures when %CFC is almost zero. It can be further observed from Table II that a higher lactate production rate was found for immature monocytes (CD15-/CD11b+), while mature monocytes (CD15+/CD11b+) have a small consumption rate. This corresponds well with the observation that in M cultures a second peak of specific lactate production rate appears about 8–12 days after the peak associated with the peak in CFC (unpublished results from Y. Kuang and D. Pascoe).

Figure 5 shows the %CFC predictions based on different metabolites and approximation methods for their rates. Although the less accurate two-point approximation for metabolic rates results in more outliers, the prediction error based on this approximation method is not much greater than that based on the three-point approximation (Table I). Thus, estimation of the individual specific metabolic rates and subsequent %CFC prediction need not be delayed until the next available measurements. From Figure 5 and Table I, it can be seen that the %CFC predictions based on lactate are superior to those based on glucose, indicating that more accurate measurements yield more accurate predictions. It

**Table I.** Number of datasets used in calibration of Eq. (13), number of outliers, and number of overrepresentations based on different metabolites and approximation methods

Metabolite	Approximation method	Number of outliers	Number of overrepresentations	Number of datasets used in calibration ( $n$ )	PE (Eq. 16)
Glucose	3-point	12	4	85	3.5
Lactate	3-point	7	5	89	2.9
Glucose	2-point	23	3	75	4.8
Lactate	2-point	14	6	81	3.4

can also be seen that prediction is reasonable for higher %CFC values—especially when using the lactate production rate, while predictions of lower %CFC can be far outside a 25% error range. The discrepancy for lower %CFC is likely to be partly due to the greater inaccuracy on a percentage basis of colony assays with lower CFC content. Model inaccuracy due to using single  $\mu$ ,  $K_S$ , and  $\alpha_S$  values for different cell types may also partly contribute to the prediction inaccuracy for lower %CFC. However, despite the greater percent error for lower %CFC, Figure 5 also shows that even at lower %CFC values, the model generally predicts higher %CFC for higher values and lower %CFC for lower values. Thus, if the prediction of %CFC values greater than a threshold value (~5%) is of primary interest, then the prediction method proposed in this article provides a useful tool.

### Real-time Estimation and Harvest Decision

Reasonable real-time predictions of %CFC and [CFC] could be useful in deciding when to manipulate a hematopoietic culture. In gene therapy, it might be better to initiate gene transfer when %CFC is highest, while the transplantation of an ex vivo culture with the maximal [CFC] may be more effective to abrogate neutropenia. Thus, in order to improve the harvest decision for cell or gene therapy, the peak values of %CFC and [CFC] should be reasonably estimated. By means of our proposed model, the time profiles of %CFC and [CFC] during an experiment can be estimated in real time. In Figure 6, predictions of %CFC and [CFC] for five representative experiments are depicted. The datasets from Exp. 1–4 were already embedded in the databank used for model calibration (Fig. 6a–h), while the datasets from Exp. 5 were not included in this databank (Fig. 6i,j).

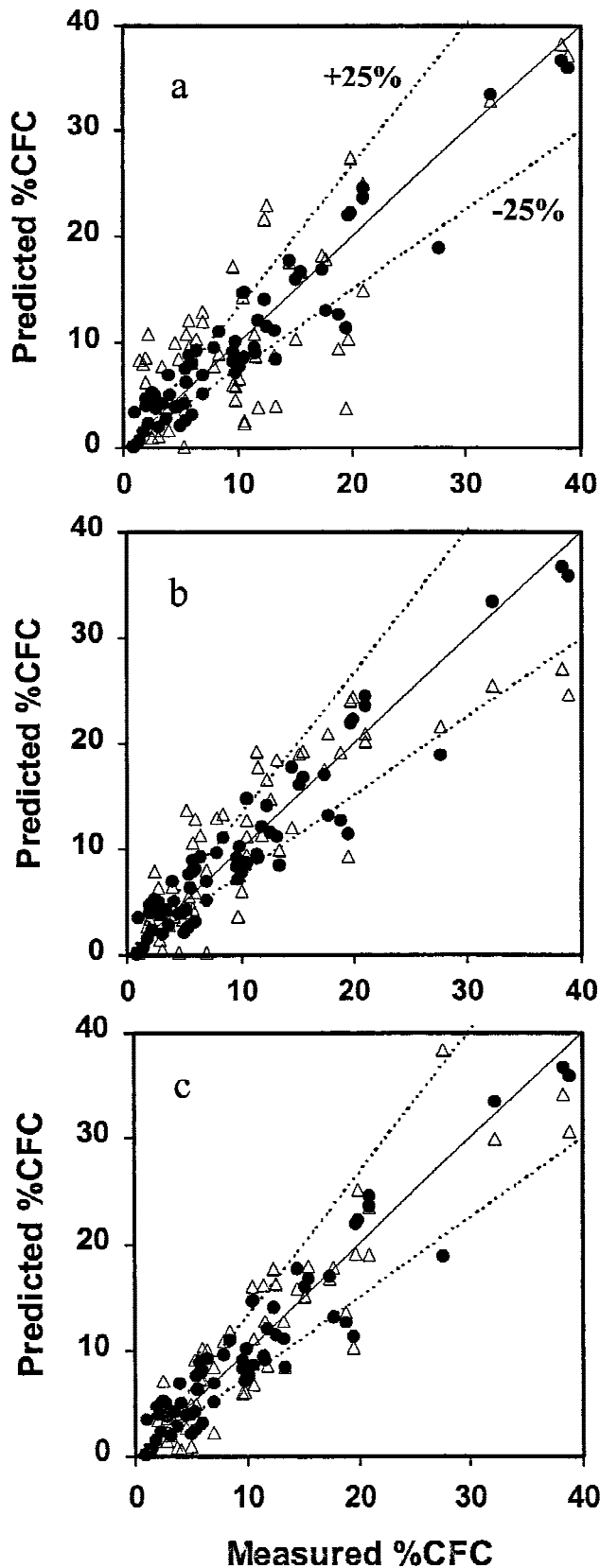
In 14 of 24 experiments available for this study, only three datasets per experiment have been embedded in the databank due to outlier or overrepresentation removal or the lack of one or more necessary measurements. Three experiments are represented in the databank with only two datasets. Representative %CFC and [CFC] predictions for those experiments with few datasets are depicted in Figure 6a,b.

There are seven experiments with four or more datasets in the databank. Three representative predictions are shown in Figure 6c–h. Except for Day 5 of Exp. 4 (Fig. 6g–h), good predictions for the %CFC and [CFC] peaks were obtained

for all these experiments. Due to a high value for the estimated lactate production rate, the model overpredicts the %CFC value on Day 5 (Exp. 4) to a certain degree (about 40% off). This error has been amplified in the [CFC] prediction because of a very high total cell density at this time point. It should be noted that the real-time measurements for this particular day were not recognized as an outlier. Thus, the unsuccessful predictions for %CFC and [CFC] at this particular time point illustrate a possible failure of the proposed model or the outlier detection method. However, it is also possible that there were errors in the colony assays for that day, especially since higher [CFC] values were obtained on Days 4 and 6. Unfortunately, we can only speculate, since repeating a colony assay at a later time is impossible.

A proposed harvest strategy is depicted in Figure 7, and its application is illustrated using Exp. 5 (Fig. 6i,j). Before performing the %CFC and [CFC] estimations for Exp. 5, the real-time measurements and rate approximations (cell density, growth rate, flow cytometry results, and metabolic rates) at a given time point were first transferred to a point on the PCA biplot (Fig. 3) for possible outlier detection (Fig. 7). The transfer matrix was obtained based on the datasets in the databank. If a new dataset is detected as an outlier, the measurements at this time point should be repeated to exclude human mistakes (Fig. 7). The original real-time measurements on Day 7 of Exp. 5 were detected as an outlier. Repeating the measurements showed that the previous lactate concentration reading on Day 7 before feeding was too low, leading to a lower calculated lactate production rate and considerable underestimation of %CFC and [CFC] (Fig. 6i,j). By correcting the lactate concentration, reasonable predictions of %CFC and [CFC] were achieved (Fig. 6i,j). The real-time measurements and rate approximations from new experiments, e.g., Exp. 5 in Figure 6, can be integrated into the databank after screening for overrepresentation once the colony assay results (%CFC measurements) are available. Such an accumulation of datasets would raise the accuracy and confidence level of the estimated model parameters, as well as the quality of %CFC and [CFC] predictions.

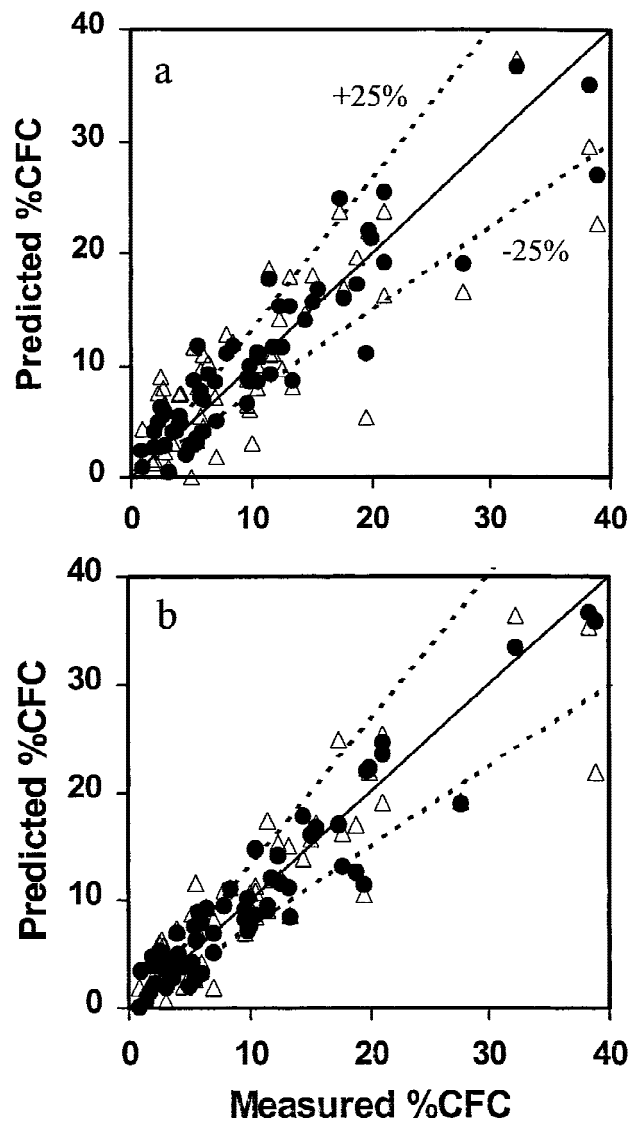
Based on real-time estimation, optimal harvest decisions can be made by using harvest rules developed for a particular application (Fig. 7). One possible harvest rule is that the culture will be harvested as soon as one observes a decreasing tendency of estimated %CFC or [CFC]. Another pos-



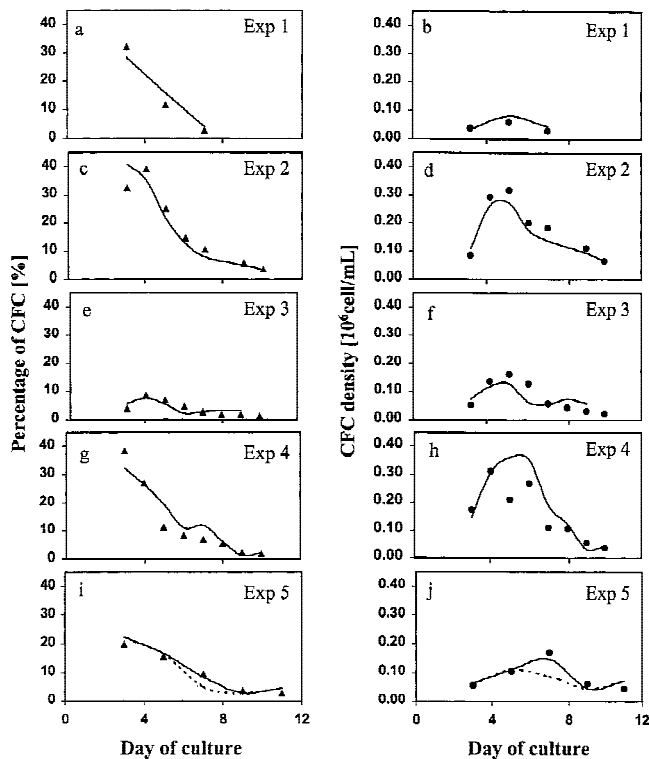
**Figure 4.** Comparison of the proposed multipopulation model (Eq. 13) (●) with three other models (Δ) for prediction of %CFC: (a) the two-population model (Eq. 4), (b) the linear multipopulation model (Eq. 5), and (c) the alternative multipopulation model with individual  $\alpha_{S,j}$  and a common  $\beta_S$ . The solid lines describe perfect predictions, while the dashed lines represent 25% over- or underestimations.

**Table II.** Parameters values for the proposed model (Eq. 13)

Parameter	Cell types	Unit	Value
$\beta_{S,1}$	CFC	$\mu\text{Mol/h}/10^7$ cell	7.73
$\beta_{S,2}$	CD15+/CD11b-	$\mu\text{Mol/h}/10^7$ cell	-0.46
$\beta_{S,3}$	CD15+/CD11b-	$\mu\text{Mol/h}/10^7$ cell	1.36
$\beta_{S,4}$	CD15+/CD11b+	$\mu\text{Mol/h}/10^7$ cell	1.61
$\beta_{S,5}$	CD15-/CD11b+	$\mu\text{Mol/h}/10^7$ cell	2.71
$\beta_{S,6}$	CD15+/CD11b+	$\mu\text{Mol/h}/10^7$ cell	-0.49
$\alpha_S$	All	$\mu\text{Mol}/10^7$ cell	9.6
$K_S$	All	$10^6$ cell/mL	2.5



**Figure 5.** Prediction of %CFC using the proposed model (Eq. 13) based on glucose consumption (Δ) and lactate production (●) using the (a) two-point (Eq. 2) and (b) three-point (Eq. 3) approximation methods. The solid lines describe perfect predictions, while the dashed lines represent 25% over- or underestimations.

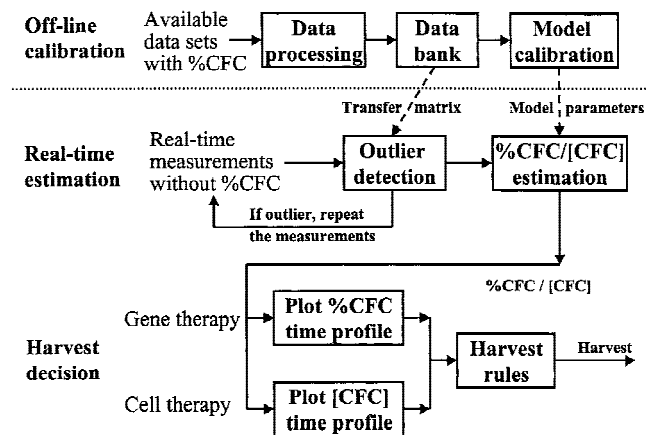


**Figure 6.** Measured (symbols) and predicted (solid lines) values for %CFC ( $\blacktriangle$ ) and [CFC] ( $\bullet$ ) for five experiments: Exp. 1 (a,b), 2 (c,d), 3 (e,f), 4 (g,h), and 5 (i,j). Dashed lines for Exp. 5 in (i) and (j) represent predicted values that would have been obtained without outlier detection and correction (see text for details).

sible harvest rule involves harvesting a culture as soon as the estimated %CFC or [CFC] exceeds a certain threshold value.

## CONCLUSIONS

A harvest strategy was developed for ex vivo expanded hematopoietic cultures to be used for cell or gene therapies



**Figure 7.** Proposed harvest strategy for ex vivo hematopoietic cultures for clinical application based on real-time measurements.

(Fig. 7). It involves developing a mathematical model that properly describes the relationship between %CFC and metabolic activity, evaluating the model parameters (model calibration), estimating %CFC or [CFC] in real-time, and making an optimal harvest decision based on predetermined harvest rules. The success of this strategy lies in reasonably predicting the %CFC and [CFC] peaks based only on real-time measurements. The proposed nonlinear multipopulation model, which was extended from a previous two-population model (Collins et al. 1997), is able to reasonably predict higher values of %CFC, and thus the %CFC or [CFC] peaks, for ex vivo hematopoietic cultures carried out under different cytokine combinations with different feeding strategies and inoculated with different patient samples, inoculum densities, and initial CD34+ cell contents. Prediction accuracy of %CFC or [CFC] is related to the model formulation (Fig. 4), measurement accuracy (Fig. 5), regression methods for parameter evaluation, and data processing (e.g., outlier detection). In order to improve the predictions, more complex models that may include additional metabolic rates should be tested. It would also be helpful to increase the rate approximation accuracy by using more sensitive metabolites, e.g., dissolved oxygen, which can be measured on-line. Furthermore, more advanced/complex regression methods could be employed, such as multivariate nonlinear regression. Besides the regression methods, database mining with cluster analysis and pattern recognition (Stephanopoulos et al., 1997) could be used to determine which parameters are most significant for %CFC estimation and to eliminate the outliers more effectively, thereby improving data deconvolution and model calibration. Despite these possible improvements, which could be further studied and addressed in the future, this study presents an overall strategy to approach the harvest problems of ex vivo hematopoietic cultures and outlines a detailed harvesting procedure that could be implemented in clinical application.

The experimental data used were provided by Paul Collins, Brian Finch, Diane Hevehan, Yu Kuang, Gerrie Liaw, Deborah Pascoe and Sanjay Patel. We thank Professor Gregory Stephanopoulos (MIT) for helpful discussions.

## NOMENCLATURE

$a_S$	growth-related metabolic coefficient in Eq. (9)
$b_S$	nongrowth-related metabolic rate in Eq. (9)
CB	cord blood
CFC	progenitor cells
$\hat{CFC}(t)$	estimated CFC content
%CFC	percentage of CFC
$\% \hat{CFC}(t)$	estimated percentage of CFC
[CFC]	CFC density
$[ \hat{CFC}(t) ]$	estimated CFC density
$C_S$	concentration of metabolite $S$
$C_X$	cell density
$E\{*\}$	average of $*$
$f_S(C_X)$	cell density influence factor on $S$ metabolic rate
$I_n$	$n \times n$ unit matrix
$L$	distance from $\beta_S$ to $\beta_S$
$K_S$	parameter used in the function $f_S(C_X)$

LS	least squares estimation
MNC	mononuclear cells
$m$	number of principal components
$m_S$	maintenance consumption on metabolite $S$
$n$	number of dataset points (sample size)
$p$	number of cell groups
PB	peripheral blood
PCA	principal component analysis
PE	overall prediction error
$q_S$	overall specific consumption/production rate of metabolite $S$
RR	ridge regression
$t$	time point of sampling
$U$	orthonormal matrix whose columns are the eigenvectors of $XX'$
$v_j$	$j$ -th column vector of $V$
$V$	orthonormal matrix whose columns are the eigenvectors of $X'X$
$W$	inverse matrix used in principal component analysis
$x_j(t)$	fraction of cell group $j$
$\hat{x}_1$	predicted value of $x_1$
$X$	cell composition matrix
$y_S$	ratio of metabolic rate to crowding effect correction
$Y_{X/S}$	yield coefficient
$Z$	diagonal matrix containing square root of eigenvalues

### Greek symbols

$\alpha, \beta$	parameters in the two-population model
$\alpha_S$	growth-related metabolic coefficient
$\beta_S$	vector containing all $\beta_{S,j}$
$\beta_{S,j}$	individual specific metabolic rate coefficients
$\hat{\beta}_{S,j}$	estimation of $\beta_{S,j}$
$\varepsilon_S$	measurement and modeling error
$\lambda$	eigenvalue
$\sigma$	standard deviation

### APPENDIX

The multivariate linear regression methods used to estimate  $\beta_{S,j}$  and  $\alpha_S$  in Eq. (13) are LS, RR, and PCA. Since multivariate linear regression is a statistical method, a term describing measurement or modeling error is added to the metabolic model. When all  $n$  experimental data are summarized in a matrix notation analogous to Eq. (5), we get

$$y_S = X\beta_S + \varepsilon_S \quad (17)$$

where  $X$  is a  $(n \times (p + 1))$  matrix with  $x_j(t_i)$  as its  $(i,j)$ -th element,  $x_{p+1}(t_i) = \mu(t_i)\sum_{j=1}^p x_j(t_i)$  (Eq. 11),  $y_S$  is a  $(n \times 1)$  vector with  $y_S(t_i)$  as the  $i$ -th element, and  $\beta_S$  is the unknown  $((p + 1) \times 1)$  parameter vector containing all  $\beta_{S,j}$  with  $\alpha_S$  as the last element. Finally,  $\varepsilon_S$  is the  $(n \times 1)$  measurement and modeling error vector. The mean of  $\varepsilon_S$  is usually assumed to be zero, mathematically expressed as  $E\{\varepsilon_S\} = 0$ . Furthermore, it is also assumed that the errors made at each sample are uncorrelated and thus the covariance matrix of  $\varepsilon_S$  is diagonal, i.e.,  $E\{\varepsilon_S\varepsilon_S'\} = \sigma_\varepsilon^2 I_n$  where  $\sigma_\varepsilon$  is the standard derivation and  $I_n$  is the  $(n \times n)$  unit matrix. It should be noted that the number of experimental data sets  $n$  is much larger than the value of  $(p + 1)$ , and the rank of  $X$  is assumed be  $(p + 1)$ .

### Least Squares Estimation

All linear regression methods are more or less based on the LS estimation that gives:

$$\hat{\beta}_S = (X'X)^{-1} X'y_S \quad (18)$$

as the estimation of  $\beta_S$ . The LS estimation is unbiased, i.e.  $E\{\hat{\beta}_S\} = \beta_S$  and its variance-covariance matrix can be given by

$$\text{VAR}(\hat{\beta}_S) = \sigma_\varepsilon^2 (X'X)^{-1} \quad (19)$$

Let  $L$  be the distance from the estimation  $\hat{\beta}_S$  to the real  $\beta_S$ . Then, the average of the squared distance can be given by

$$\begin{aligned} E\{L^2\} &= E\{(\hat{\beta}_S - \beta_S)'(\hat{\beta}_S - \beta_S)\} \\ &= \sigma_\varepsilon^2 \text{Trace}(X'X)^{-1} = \sigma_\varepsilon^2 \sum_{j=1}^{p+1} \frac{1}{\lambda_j} \end{aligned} \quad (20)$$

where  $\lambda_j$  is the  $j$ -th largest eigenvalue of  $X'X$ . When the matrix  $X'X$  is "ill-conditioned," i.e., it has one or more small eigenvalues, the prediction errors for  $\beta_{S,j}$  and  $\alpha_S$  tend to be large (Hocking et al., 1976), often making the estimation of  $\beta_{S,j}$  and  $\alpha_S$  unacceptable. RR and PCA mainly aim to correct the matrix  $X'X$  when it is "ill-conditioned." The former method adds a positive term to the matrix  $X'X$  diagonal elements to prevent any small eigenvalues from occurring (Hoerl and Kennard, 1970), while the latter eliminates the small eigenvalues in the matrix  $X'X$  (Jolliffe, 1986; Wold et al., 1987).

### Ridge Regression Method

Estimation of  $\beta_S$  based on RR gives:

$$\hat{\beta}_S = (X'X + kI)^{-1} X'y_S \quad k > 0 \quad (21)$$

In this case, the average of the squared distance from  $\hat{\beta}_S$  to  $\beta_S$  can be given by

$$\begin{aligned} E\{L^2(k)\} &= E\{(\hat{\beta}_S - \beta_S)'(\hat{\beta}_S - \beta_S)\} \\ &= \sigma_\varepsilon^2 \sum_{j=1}^{p+1} \frac{\lambda_j}{(\lambda_j + k)^2} + k^2 \beta_S'(X'X + kI)^{-2} \beta_S \end{aligned} \quad (22)$$

The parameter  $k$  is determined in such a way that the mean value of the squared distance is minimized.

### Principal Component Analysis

In PCA, the singular value decomposition is often used to decompose the matrix  $X$  in the form

$$X = UZV' \quad (23)$$

where:

- 1)  $U$  is a  $(n \times (p + 1))$  matrix and has orthonormal columns so that  $U'U = I_{p+1}$ . Furthermore, the columns of  $U$  are those eigenvectors of  $XX'$  which correspond to the non-zero  $(p + 1)$  eigenvalues.
- 2)  $V$  is a  $((p + 1) \times (p + 1))$  matrix and has orthonormal columns so that  $V'V = I_{p+1}$ . Furthermore, the columns of  $V$  are the eigenvectors of  $X'X$ .

3)  $Z$  is a  $((p + 1) \times (p + 1))$  diagonal matrix with  $\lambda_j^{1/2}$  as its  $(j,j)$ -th element.

The inverse matrix of  $X'X$  can then be given by:

$$(X'X)^{-1} = (VZ'U'UZV')^{-1} = V(Z'Z)^{-1}V' = \sum_{j=1}^{p+1} \frac{1}{\lambda_j} v_j v_j' \quad (24)$$

Assuming that the number of principal components is  $m$  ( $0 < m \leq p + 1$ ), a new matrix  $W$  is introduced to replace the matrix  $(X'X)^{-1}$  such that all of the smallest  $(p + 1 - m)$  eigenvalues are eliminated:

$$W = \sum_{j=1}^m \frac{1}{\lambda_j} v_j v_j' \quad (25)$$

where  $v_j$  is the  $j$ -th column vector of  $V$ . The principal component estimation then gives the estimation of  $\beta_S$  as follows

$$\hat{\beta}_S = WX'y_S \quad (26)$$

The selection of  $m$  is more or less arbitrary, depending on the distribution of all eigenvalues of  $X'X$ . However, minimizing the distance from  $\hat{\beta}_S$  to  $\beta_S$  is often a good criterion.

## References

- Brugger W, Heimfeld S, Berenson RJ, Mertelsmann R, Kanz L. 1995. Reconstitution of hematopoiesis after high-dose chemotherapy by autologous progenitor cells generated ex vivo. *N Engl J Med* 333: 283–287.
- Collins PC, Nielsen LK, Wong CK, Papoutsakis ET, Miller WM. 1997. Real-time method for determining the colony-forming cell content of human hematopoietic cell cultures. *Biotechnol Bioeng* 55:693–700.
- Colter M, Jones M, Heimfeld S. 1996. CD34+ Progenitor cell selection: clinical transplantation, tumor cell purging, gene therapy, ex vivo expansion, and cord blood processing. *J Hematother* 5:179–184.
- Crystal RG. 1995. Transfer of genes to humans: early lessons and obstacle to success. *Science* 270:404–410.
- Havenga M, Hoogerbrugge P, Valerio D, van Es HHG. 1997. Retroviral stem cell gene therapy. *Stem Cells* 15:162–179.
- Hevehan DL, Papoutsakis ET, Miller WM. 2000a. Physiologically significant effects of pH and oxygen tension on granulopoiesis. *Exp Hematol* 28:267–275.
- Hevehan DL, Wenning L, Miller WM, Papoutsakis ET. 2000b. A dynamic model of ex vivo granulocytic kinetics to examine the effects of oxygen tension, pH, and IL-3. *Exp Hematol* 28:1016–1028.
- Hocking RR, Speed FM, Lynn MJ. 1976. Class of biased estimators in linear regression. *Technometrics* 18:425–437.
- Hoerl AE, Kennard RW. 1970. Ridge regression: biased estimation for no orthogonal problems. *Technometrics* 12:55–67.
- Jolliffe IT. 1986. *Principal component analysis*. New York: Springer-Verlag.
- Koller MR, Palsson BO. 1993. Tissue engineering: reconstitution of human hematopoiesis ex vivo. *Biotechnol Bioeng* 42:909–930.
- Koller MR, Bender JG, Papoutsakis ET, Miller WM. 1992. Effects of synergistic cytokine combinations, low oxygen and irradiated stroma on the expansion of human cord blood progenitors. *Blood* 80:403–411.
- Koller MR, Manchel I, Palsson MA, Maher RJ, Palsson BO. 1996. Different measures of ex vivo human hematopoietic culture performance are optimized under vastly different conditions. *Biotechnol Bioeng* 50:505–513.
- Ludeking R, Piret EL. 1959. A kinetic study of the lactic acid fermentation. *J Biochem Microbiol Technol Eng* 1:393–412.
- McAdams TA, Winter JN, Miller WM, Papoutsakis ET. 1996. Hematopoietic cell culture therapies. II. Clinical aspects and applications. *Trends Biotechnol* 14:388–396.
- Nielsen LK, Bender JG, Miller WM, Papoutsakis ET. 1998. Population balance model of in vivo neutrophil formation following bone marrow rescue therapy. *Cytotechnology* 28:157–162.
- Ozturk S, Palsson BO. 1990. Effect of initial cell density on hybridoma growth, metabolism, and monoclonal antibody production. *J Biotechnol* 16:259–278.
- Pirt SJ. 1965. The maintenance energy of bacteria in growing culture. *Proc R Soc Lond B* 163:224–231.
- Reiffers J, Calliot C, Dazey B, Attal M, Caraux J, Boiron JM. 1999. Abrogation of post-myeloablative chemotherapy neutropenia by ex vivo expanded autologous CD34-positive cells. *Lancet* 354: 1092–1093.
- Sand T, Condie R, Rosenberg A. 1977. Metabolic crowding effect in suspension of cultured lymphocytes. *Blood* 50:337–346.
- Sandstrom CE, Bender JG, Papoutsakis ET, Miller WM. 1995. Effects of CD34+ cell selection and perfusion on mobilized blood mononuclear cell expansion ex vivo. *Blood* 86:958–970.
- Scheding S, Franke H, Diehl V, Wichmann HE, Brugger W, Kanz L, Schmitz S. 1999. How many myeloid post-progenitor cells have to be transplanted to completely abrogate neutropenia after peripheral blood progenitor cell transplantation? Results of a computer simulation. *Exp Hematol* 27:956–965.
- Sharaf MA, Illman D, Kowalski BR. 1986. *Chemometrics*. New York: John Wiley & Sons.
- Stephanopoulos G, Locher G, Duff MJ, Kamimura R, Stephanopoulos G. 1997. Fermentation database mining by pattern recognition. *Biotechnol Bioeng* 53:443–452.
- Talstad I. 1972. The influence of the crowding phenomenon on the oxygen consumption of blood cells as determined by Cartesian diver technique. *Acta Physiol Scand* 84:332–337.
- Terstappen LW, Safford M, Loken MR. 1990. Flow cytometric analysis of human bone marrow. III. Neutrophil maturation. *Leukemia* 4: 657–663.
- Wold S, Bsenensen K, Geladi, P. 1987. *Principal component analysis*. *Chemomet Intell Lab Sys* 2:37–52.