REVIEW

Quantification of Microbial Productivity Via Multi-Angle Light Scattering and Supervised Learning

Alun Jones,[1,2] Danielle Young,[1] Janet Taylor,[1,2] Douglas B. Kell,[1] Jem J. Rowland[2]

[1]Institute of Biological Sciences, University of Wales, ABERYSTWYTH, Ceredigion SY23 3DD, Wales, United Kingdom; telephone: +44 (0) 1970 622353; fax: +44 (0) 1970 622350; e-mail: auj/diy/jjt95/dbk/jjr@aber.ac.uk
[2]Department of Computer Science, University of Wales, ABERYSTWYTH, Ceredigion SY23 3DB, Wales, United Kingdom

Received 12 July 1997; accepted 24 November 1997

Abstract: This article describes the use of chemometric methods for prediction of biological parameters of cell suspensions on the basis of their light scattering profiles. Laser light is directed into a vial or flow cell containing media from the suspension. The intensity of the scattered light is recorded at 18 angles. Supervised learning methods are then used to calibrate a model relating the parameter of interest to the intensity values. Using such models opens up the possibility of estimating the biological properties of fermentor broths extremely rapidly (typically every 4 sec), and, using the flow cell, without user interaction. Our work has demonstrated the usefulness of this approach for estimation of yeast cell counts over a wide range of values (10^5–10^9 cells mL^{-1}), although it was less successful in predicting cell viability in such suspensions. © 1998 John Wiley & Sons, Inc. *Biotechnol Bioeng* **59**: 131–143, 1998.
Keywords: chemometrics; light scattering; microbial productivity

Background

> *''On-line measurement of the biomass concentration is no longer a matter of comfort, it is essential for a functional controller.''* (Sonnleitner et al., 1992, P. 5)

There is a continuing need for more and better methods for estimating the biologically related properties of fermentor broths and of other cell suspensions (e.g., Clarke et al., 1985; Clarke et al., 1982; Junker et al., 1994; Kell, 1980; Kell et al., 1990; Kell and Sonnleitner, 1995; Konstantinov et al., 1994; Locher et al., 1992b; Singh et al., 1994; Sonnleitner et al., 1992). For its rapidity, convenience, and simplicity, the routine assessment of the biomass content of a cell suspension in almost every microbiological laboratory is normally carried out by determining its turbidity or optical density (Harris and Kell, 1985; Mallette, 1969). This is typically performed by taking a sample, placing it in a conventional spectrophotometer, illuminating with visible light (whose wavelength is normally chosen for arbitrary or historical reasons), and taking an absorbance reading, which (in the absence of true chromophores) is numerically equal to, and properly referred to, as the optical density. Calibration curves typically indicate that the OD (in a cuvette with a 1 cm path length) is linear with the concentration of a given type of biomass below an OD of 0.6 or so, whereupon multiple scattering results in a breakdown of the Beer-Lambert law, leading eventually to a complete independence of OD from cell concentration. Because an OD (1 cm) of 1 typically corresponds to cell concentrations of ca. 0.5 mg dry wt/mL, it is almost always necessary to dilute the sample when it comes from a fermentation of biotechnological interest.

The scattering of light by microbial and similarly sized particles depends (nonlinearly) on a number of factors, including the illuminating wavelength, the relative size of the scatterer, and the difference in refractive index (RI) between the scatterer and the medium (e.g., Bohren and Huffman, 1983; Carr et al., 1987; Davey and Kell, 1996; Harding, 1986; Kerker, 1983; Koch, 1968; Koch, 1984; Koch, 1986; Latimer, 1982; Salzman, 1982; Sharpless et al., 1977; Van der Hulst, 1957; Wyatt, 1968; Wyatt, 1973). In brief, Rayleigh scattering occurs when the particle sizes are significantly less than the wavelength of the light, while a modified form of Rayleigh scattering, usually referred to as Rayleigh-Debye-Gans scattering, occurs when, as in the case of many bacteria, the particle size and wavelength of light are of the same order (Koch, 1968; Koch, 1984; Wyatt, 1968). One result of these considerations is that at lower cell concentrations *direct* measurement of the scattered light (nephelometry), typically at right angles to the illuminating beam,

Correspondence to: Alun Jones
Contract grant sponsors: Chemicals and Pharmaceuticals Directorate of the UK BBSRC; Higher Education Funding Council for Wales

© 1998 John Wiley & Sons, Inc.
CCC 0006-3592/98/020131-13

provides a much more accurate measure of particle density than does the OD (Harding, 1986; Harris and Kell, 1985; Keilmann et al., 1980; Mallette, 1969; Wyatt, 1973). However, given the dependence on the above factors of the scattering at different angles, it would appear that even more accurate and informative analyses could be obtained by measuring light scattering at many angles simultaneously.

An excellent example of the potential of multiangle light scattering measurements is provided by the study of individual spores of *Bacillus sphaericus* by Ulanowski et al. (1987), using a scanning laser diffractometer in which the scattering was logged serially at every 1° between 4 and 176°, although in practice this was too often (Miller, 1990) and caused problems of collinearity when using multivariate calibration techniques. Despite the morphological simplicity of the target biological system, and the lack of complications from studying heterogeneous suspensions containing many different cells, the light-scattering data obtained (see Fig. 1 of Miller, 1990) were not at all well fitted by theoretical (Lorenz-Mie) scattering curves containing four free parameters (two radii, two refractive indices). Carr and colleagues also pursued the differential light-scattering approach (Jepras et al., 1991), and obtained very reproducible data from a variety of fermentations using *Escherichia coli*, *Saccharomyces cerevisiae* and *Pseudomonas aeruginosa*, but while the same group successfully exploited a variety of other modern optical methods for the extraction of very useful information from microbial fermentations (e.g., Carr, 1990; Carr et al., 1987; Chow et al., 1988; Clarke et al., 1986; Clarke et al., 1985; Jepras et al., 1991; Perkins et al., 1993), they were unable to account for the differential light scattering data on the basis of theoretical scattering models.

The generalized approach of multi-angle light scattering was developed within microbiology under the term, ''differential light scattering'' by Wyatt (Wyatt, 1968, 1973, Wyatt and Jackson, 1989) and has been embodied in a commercial instrument, the DAWN (Wyatt, 1993a,b). This instrument has the singular ability to acquire data through the simultaneous action of 18 separate detectors. However, it has been reported (Wyatt, 1993a,b) that the general approach has not been promoted, probably because of its inability, contingent on the complexity of biological systems involved, to adequately account for the differential light scattering in terms of the physical theories alluded to above (Wyatt et al., 1972).

Over the last decade there have emerged many powerful *chemometric* methods for *supervised learning* (see e.g., Brereton, 1990, 1992; Mark, 1991; Martens and Næs, 1989; Massart et al., 1988). These can be exploited to form calibration models which accurately relate the rather featureless *multivariate* multiangle light-scattering data to the biological, chemical, or physical properties of interest. In this sense, a very suitable analogy is that of near-infrared spectroscopy, which also produces broad and featureless multivariate data that without chemometrics, are analytically useless, however, *with* chemometrics it is an outstandingly powerful, convenient, and noninvasive analytical technique of very wide applicability (see e.g., Corti and Dreassi, 1993; Drennen et al., 1991; Hildrum et al., 1992; Martens and Næs, 1989; Martin, 1992; McClure, 1994; Murray and Cowe, 1992; Næs et al., 1993; Osborne et al., 1993).

In this article, we report on an experimental study in which we have used multiangle light scattering measurements and multivariate calibration techniques to produce robust predictive models for the monitoring of cell suspensions over a wide concentration range in which the optical density attains values grossly in excess of those measurable by conventional means.

MATERIALS AND METHODS

Equipment

A DAWN DSP-F laser photometer (supplied by Optokem Instruments Limited, Pistyll Farm, Nercwys, Clwyd CH7 4EW, UK) was used to acquire the multiangle responses. This instrument passes light from a 5mW Helium-Neon laser source into a chamber which may contain either a scintillation vial or a flow cell. The flow cell is arranged so that the biological suspension may be passed through the cell continuously and without interrupting data acquisition. Most of the experiments reported here were carried out using scintillation vials. Figure 1 shows the basic layout of the DAWN DSP-F photometer. Eighteen detectors are arranged asymmetrically around the sample chamber to record light intensities at angles in the range 22.5°–147.0° relative to the laser input.

To remove the effects of variation in detector sensitivity, the DAWN must be calibrated prior to experimentation. This is achieved by first measuring the voltages registered

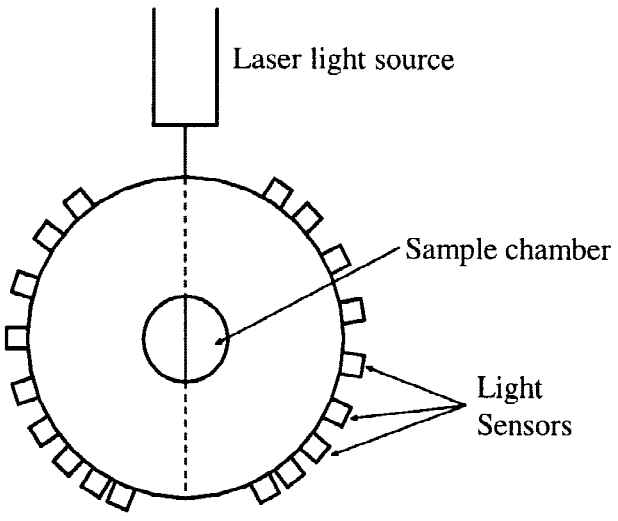

Figure 1. Basic layout of the DAWN DSP-F laser photometer. Light from the helium-neon laser source is passed through the sample chamber. The scattering intensity at 18 angles is recorded by sensors placed asymmetrically around the chamber.

by the photodiodes in the absence of any light, then measuring the voltages registered in the presence of light from an isotropic scattering agent (colloidal gold suspension, RMS radius 7.36nm). These voltages determine a zero-offset and gain value for each detector which are then applied as corrections via the instrument's software. The detectors are normalized automatically to have the same characteristics as the 90° detector, which is calibrated using a light standard with a known Rayleigh ratio (e.g., toluene). Using this methodology, the DAWN DSP-F photometer can be calibrated in an instrument-independent manner. The supplied DAWN data acquisition software performs all of the necessary transforms to the measured voltages, so that once calibration of the instrument has been performed, the final voltages recorded to disk are in a normalized form.

The DAWN DSP-F is connected to an IBM compatible PC via an RS423 serial link, and controlled using the supplied software. A digital signal processor, internal to the DAWN instrument, is responsible for acquiring the light sensor outputs and passing them through the link to the PC. The DAWN software provides a number of sophisticated data management and analysis methods. However, these methods are more suited to the analysis of light scattered by large *molecules*, and are of little use in the analysis of the large *particles* of interest in our study. In general, the data acquired from the DAWN were saved directly to disk in a textual form, and all data management was performed using our purpose-built database and chemometric software.

The DAWN DSP instrument was designed primarily for the analysis of macromolecules such as proteins (Wyatt, 1993a,b), and is exceptionally sensitive because the scattering by proteins is far less than that by cells. Our early experiments using suspensions of mixtures of latex beads of known concentration, and similarly characterized suspension of yeast, revealed that saturation of detectors was a significant problem. Although the DAWN permits external monitoring of the individual detector voltages, rather surprisingly it has no provision for generating a warning in the event of detector saturation. To overcome the problem of saturation, we placed a neutral density filter (Kodak N.D. 1.00, measured optical density 0.94) in the incident laser beam to provide the optical attenuation necessary to prevent detector saturation. The consequent loss of instrument sensitivity did not prove to be a problem in our subsequent experiments.

To form models that may be used to predict parameters of biological samples from their light scattering measurements, it is necessary to obtain data objects that contain the detector voltages and the corresponding biological parameters. For recording cell counts, our ''gold standard'' method was a Coulter Z1 counter (Coulter Electronics Limited, Northwell Drive, Luton, Bedfordshire LU3 3RH, England). This method is based on detecting the changes of conductivity observed when a cell suspended in a conductive liquid passes through a small aperture. Each cell produces a ''spike'' in resistance as it passes through the aperture, with the number and size of such spikes providing information on cell count and size distribution (Harris and Kell, 1985). The Z1 cell counter is specified as having an accuracy of ±1% of a ''reference Coulter system'' (Coulter Electronics Ltd., 1994) and a precision of better than 1% at metered volumes of 0.5 mL and 1 mL. This specification suggests that the Coulter Z1 would be ideal for cell count determination, and that using the DAWN DSP-F is unnecessary. However, there is the issue of sample preparation and measurement time. The Coulter system requires that the cells be diluted in an electrolyte solution and, to meet the above specification, the average of at least 20 replicate measurements must be taken. In addition, care must be taken to ensure that the aperture does not become blocked, and that the range of sizes counted is calibrated correctly (otherwise, cell fragments and other small particles may be included in the reported count).

In contrast, once suitably calibrated, the DAWN system should be capable of taking undiluted suspension samples and, by using a flow cell, a number of replicate measurements taken automatically. In fact, for our experiments, the DAWN integrated sampled light scattering profiles over a period of 20 sec for each recorded sample. The minimum sampling duration is 0.125 sec, but samples recorded at this rate would show a correspondingly higher noise level.

Chemometric Methods

One may perform a simplistic analysis of, for example, the cell concentration of a suspension merely by examining the light scattered by the cells (e.g., Fig. 2). While such an examination does not provide any quantitative information, it is easy to reason that the scattering profile provided at low cell concentrations will show higher intensities at the forward scattering angles than that at higher cell concentrations (Davey and Kell, 1996). The use of multivariate calibration methods to calibrate light scattering against cell count thus seems a reasonable approach to the problem of forming quantitative models.

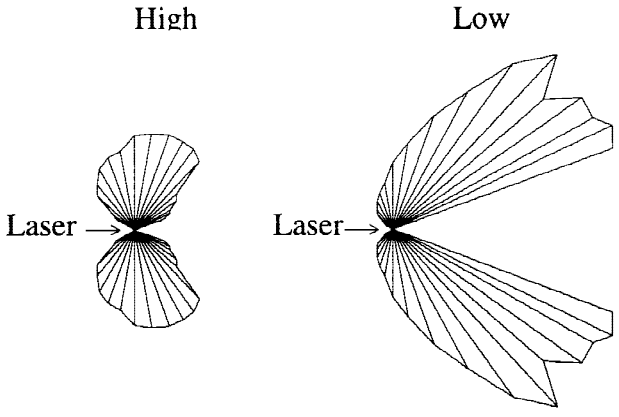

Figure 2. Example light scattering profiles for yeast cells (average diameter 6.5 mm) at high (7.4×10^7 cells mL^{-1}) and low (2.8×10^7 cells mL^{-1}) concentrations. Simple inferences may be made using these profiles (e.g., more light is scattered in the forward direction at low concentrations).

Multivariate calibration software was written in-house to provide portability between a number of machine architectures and operating systems and a flexible interface to other locally used software. It also allowed for the easy incorporation of many calibration methods into a single, coherent framework. The software was written in C++ (Stroustrup, 1994), and is based around a single ''modeling'' class, from which all the calibration methods are derived using the ''inheritance'' mechanism of C++. This allows new methods to be implemented and slotted into the framework supplied by the client software with little or no change to that client software. In particular, such general methods as model validation can be written without reference to the calibration method in use, and each new calibration method implemented inherits all validation methods from the top-level modeling class.

The client software consists of the set of simple ''wrapper'' programs which read matrices from file and invoke the model operations appropriate to their operation. The modular nature of the class structure means that, in general, exactly the same sequence of C++ statements is used to build models for each modeling type. Therefore, the modeling type can be specified by the user at program startup, and the model-specific code is isolated to one or two lines of C++. At present, the software can form and use models of the following types: MLR, PCA, PCR, PLS1, PLS2, (e.g., Brereton, 1990, 1992; Martens and Næs, 1989) and a variety of standard back-propagation artificial neural network types (Bishop, 1995; Chauvin and Rumelhart, 1995; Ripley, 1994; Rumelhart et al., 1986; Werbos, 1994).

To ensure portability, at this level the software is command-line driven. The input and output data, along with the generated model, are stored in text files. These are labeled by specifying their names on the command line, along with the modeling type used and any model-specific parameters. However, it is easy for other programs to generate such command lines and thus invoke the C++ software. Such programs can also ask the C++ software for details of model-specific operations, thus allowing dynamic alteration of the internal defaults of the models. This is particularly important in the case of back-propagation neural networks, where one may wish to adjust the learning rate, momentum, etc. The mechanism for such user-defined parameters is standardized, so that any modeling method added to the class hierarchy may supply its own parameters.

The DAWN specific application of this software is based around the Microsoft Access II Database system, running under Windows NT. The textual data files recorded by the DAWN software may be read into Access and stored using the experiment name and sample number as keys. Any subset of the stored data may be extracted and passed to the modeling methods. In addition, the Y data corresponding with any given sample can be retrieved, and samples can be retrieved on the basis of their Y values only. This gives the basis for a flexible chemometric modeling system. Models generated by the C++ software can be stored in the database, along with a description and the list of Y variables for which they generate data. Hence, a model, once formed, can be retrieved and used on any X data in the database. The predicted Y values generated by the model will be stored with their correct identification, both in terms of variable name and sample identity.

When a modeling method is selected by the user, the Access Basic code asks the C++ software whether there are any user-defined parameters applicable to that method. If so, the user is able to accept the default values or supply new parameters. This provides a user-friendly method for gaining access, where necessary, to the internals of modeling methods without having to build too much intelligence into the Access Basic code; this is an important factor, as it allows the addition of modeling methods to C++ with no changes to the database software.

Several methods are supplied to allow the user to validate the models produced. Validation is essential, because it is possible to generate a model which takes into account all *all* variations in the X data. In the real world, some parts of the X variation are noise. Therefore, it is important to know whether the model fits the random variations within the data. A model that fits the noisy part of its training set is unlikely to produce good predictions on unseen data.

Most chemometric methods are based around an iterative process which improves the fit of the model to the *training* data. At some point, the fit is *optimal*, and subsequent iterations start to fit the noise effects rather than the underlying effects of relevance. The validation methods supplied by the software allow the fitting process to be monitored, so that it can be terminated at the optimal training point. The following model validation methods are included:

1. Cross-validation (Geisser, 1975; Stone, 1974) splits the training data set into a number of subsets and repeatedly forms and evaluates models using disjoint collections of these subsets. The average error recorded by this process is assumed to be an estimate of the error generated by a model based on the entire training set.
2. Training/test set validation is based on an explicit statement of which points are to be used for forming the model and which are to be used for evaluating its performance. The software provides a number of methods for deciding which samples are to be used for training and testing. The first, and simplest, is user-selection. Responsibility lies with the user for deciding which points are for training and which are for testing. The danger of this approach lies in choosing an invalid test set for the training data. If the training data were to encompass a different range of values from that covered by the test set, then the results of model evaluation would most likely be entirely spurious. The user must ensure that such a situation does not occur. A method based on the Duplex algorithm (Snee, 1977) has been incorporated into the software, so that the computer itself can partition the set of candidate samples into training and test sets. This method ensures that no significant extrapolation of

the training data set takes place in the test set, and that the test and training points are distributed evenly throughout the data space. By using this method, one can be more confident that the model evaluation results are representative of future data *within the same region of data space*. The proviso that future data lie within the same region of space is important. When a model is formed, its performance is only determined over the range of values used for training the calibration method. We can say nothing about its performance for any other range of values without testing it for such values. If this problem is ignored, then one cannot be certain that the predictions made for new, unseen X data are correct.

A number of methods have been considered for the examination of new X data, to give a warning when the model is likely to produce invalid results (e.g., Euclidean, Penrose, and Mahalanobis distance measures (Manly, 1994), and the BEAST method (Lodder and Hieftje, 1988). Typically, these methods are based on evaluating the distance from the new points to the centroid of the set of training X data. As such, all are subject to problems when the cluster structure of the training data does not conform to certain assumptions. For example, Euclidean and Penrose distance metrics do not take into account correlations between the X variables, and the BEAST method fails when the X training data contain multiple clusters or curvature. As yet, we are not aware of a satisfactory solution to this problem. For the experiments presented in this article, we have had the luxury of being able to ensure that extrapolation does not occur. In an industrial context, one would strive to ensure that the models used encompass all expected data points so that significant faults can be identified.

We have also used artificial neural networks using the standard back-propagation algorithm (Rumelhart et al., 1986); these were found essential in forming calibration models over large ranges of cell concentration, where the variations in detector outputs are strongly nonlinear.

Genetic programming has been introduced as a comparative nonlinear approach to calibration model formation.

In addition, we have added the capability to pre-process the calibration Y data. This allows us, for example, to take the logarithms of the measured cell counts before forming the model, and has provided significant improvements in model performance over large ranges of cell concentration.

MATERIALS AND METHODS (BIOLOGICAL)

Here we consider measurements relating to a fermentation of commercial significance: *S. cerevisiae* (Davey et al., 1996; Locher et al., 1992a; Markx et al., 1991). To form models of cell count, we needed to record the light scattering data for a range of known cell counts. Suspensions of ''Mauri'' yeast were grown aseptically overnight at 25°C in ''YEP'' medium, consisting of 10 g/L yeast extract, 20 g/L peptone, 20 g/L glucose. The suspension was then spun down and resuspended in a buffer containing 10 m*M* citric acid, 50 m*M* KCl, 2.5 m*M* $MgCl_2$ at a pH of 5.0 and adjusted to the starting concentration of interest, as measured using the Coulter counter. Such suspensions were used within 2 d of growth. A serial dilution was then performed, diluting the suspension by 5% at each step using more of the buffer solution. The light scattering data were measured using the DAWN software, and recorded on disk. The corresponding cell counts were measured using the Coulter counter using a minimum size threshold of 3.5 μm, the mode cell size having been measured at 6.5 μm, using a Skatron Argus 100 flow cytometer (Davey et al., 1990; Davey and Kell, 1996; Kaprelyants and Kell, 1992).

For our experiments on cell viability, we required suspensions with a wide variety of viabilities. Because generating such suspensions by ''natural wastage'' was liable to take a long time, we worked with mixtures of live and dead cells. Yeast suspensions were grown as reported above. Half of each suspension was heated to 70°C for 60 min to kill the yeast. Mixtures of live and dead cells in differing proportions were then formed to give a range of viabilities. The cell counts and light scattering data were recorded as before.

For flow cell-based measurements, the suspensions were prepared as before, and injected into the cell. At least 1 mL of suspension was passed through the cell before each reading was taken, to ensure that no significant mixing of samples occurred in the sampled data.

EXPERIMENTS AND RESULTS

We investigated the relationship between light scattering and cell counts over a wide range of cell concentrations using yeast suspensions. Figure 3 shows a selection of normalized detector outputs for cell counts over four orders of magnitude. The outputs are strongly nonlinear. However, for each individual experiment, encompassing a small range of cell count values, a subset of the detector outputs can be found which shows a nearly linear relationship. For such small ranges then, multivariate linear calibration methods like PCR and PLS1 may be applicable. To cover the entire range, however, we had to resort to nonlinear calibration methods such as artificial neural networks, or to the use of piecewise linear approximation. We consider the latter method first.

The Piecewise Linear Approach

The aim of this approach is to form a number of quantitative linear models of cell concentration, each permitting prediction of counts over a small range of values. A classification model is then formed to choose which of the quantitative models should be used for a given sample. In this manner, we effectively handle the curvature in the light scattering profile by approximating it using a number of straight lines.

First, however, we need to ensure that linear models are sufficient to approximate the curves over reasonable ranges of cell concentration. Figure 3 suggests that this is the case. For each serial dilution experiment, a number of detector

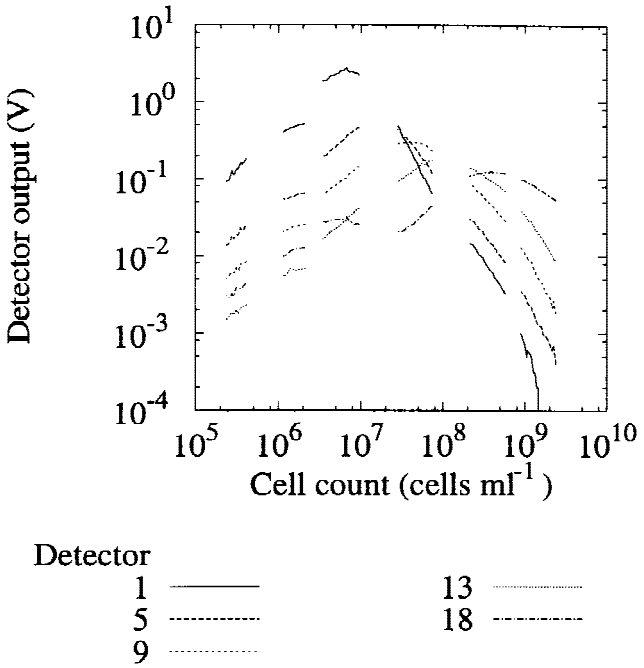

Figure 3. Normalized detector outputs for a range of cell concentrations. Each curve segment represents a detector output for a single serial dilution experiment. Over the whole range of cell count values, any given detector response is highly nonlinear.

outputs can be seen which are approximately linear over the range of the experiment. Therefore, a linear method should be reliable over such ranges. We now demonstrate this with an example from experiments in the 10^7 decade.

Three serial dilution experiments were used in this example, two for training and testing the model, the third for demonstration of the model performance. Each experiment recorded 20 samples at cell counts in the range 2.8×10^7–7.5×10^7 cells mL^{-1}. The data from the first two experiments were amalgamated to form a single data set, then partitioned using the Duplex method to give training and test sets in which the problems of extrapolation and uneven data spread are minimized. PLS1 regression was used to form models with up to 18 factors using the training set data. For each number of factors, the corresponding model was validated using the test set data.

Figure 4 shows the Root Mean Square Error of Prediction (RMSEP) for predicting the test set data for each PLS1 model. To obtain accurate predictions, we need to choose a number of factors which gives rise to a low RMSEP. Conversely, we need to choose a model in which each factor represents a relevant effect. Ideally, both of these requirements amount to the same thing. However, with these particular data, one could use anything from 2–5 factors with approximately the same performance from each. The ''simplicity'' requirement (Seasholtz and Kowalski, 1993) suggests that we use the 2-factor model. A useful check is to examine the weightings applied to the detector outputs in order to generate each factor. Figure 5 shows such weightings. We can see a strong degree of structure in the weightings for factors 1 and 2, while those for subsequent factors appear much more random.

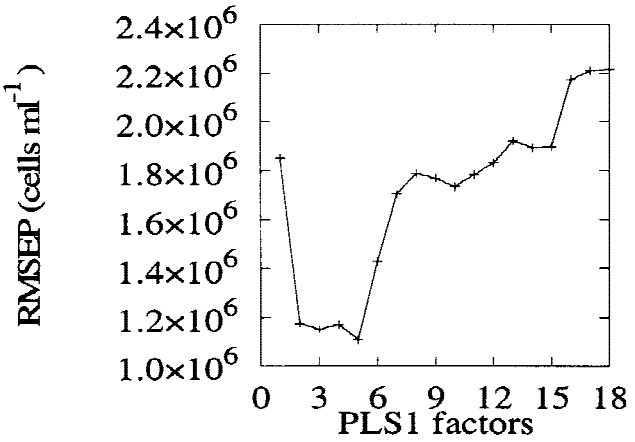

Figure 4. Prediction error vs. number of PLS1 factors used in model formation, test/training set validation on Duplex partitioned cell count data with cell counts in the range 2.8×10^7–7.5×10^7 cells mL^{-1}.

Another check of this effect is to examine the degree of variation of the weightings across different data sets. If a factor encodes a relevant effect, then we would expect it to be stable across (equivalent) training data sets. However, if the factor encodes noise effects, then it is likely to be strongly data dependent. Two 5-factor PLS1 models were formed using our previous ''training'' and ''test'' sets (note, there is no extrapolation consideration in forming models on

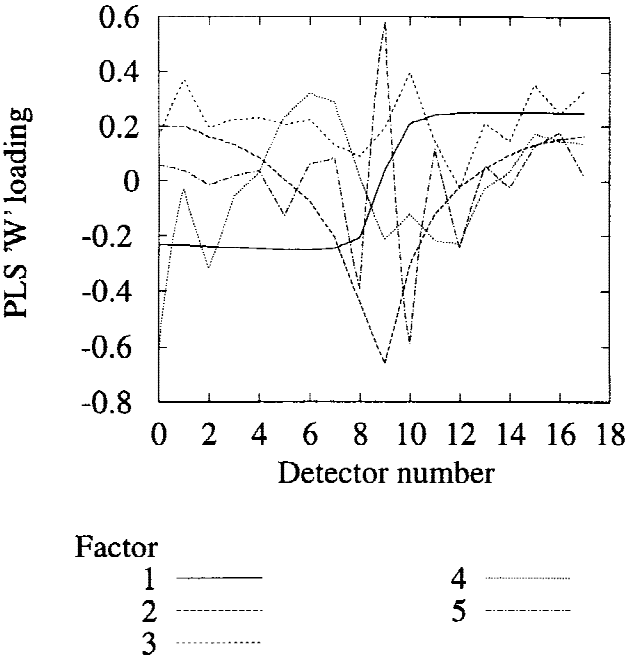

Figure 5. PLS ''W'' loadings for the first five factors in the cell count model. These encode the importance of each detector in the determination of each factor. For a continuous effect like light scattering angle, one would expect that the loadings should show a structured profile. This is true for the first two factors, but subsequent ones show disorder, which suggests that they are encoding noise effects.

the test set in this case, because we are not evaluating model error, merely looking at the structure of the model). The correlation coefficient between corresponding ''W'' loading vectors in each model was then calculated, giving values of 1.00, 1.00, 0.55, 0.87, and 0.51 for the first five factors. By using the RMSEP data from Figure 4 and these correlation coefficients, we can again assume that a 2-factor PLS1 model is optimal for this cell concentration range.

The results of applying the 2-factor PLS1 model to the training, test, and external validation data are show in Figure 6. The training set points are shown for reference only. The performance of the model for such points could be arbitrarily improved by including factors which are irrelevant to the cell count effect, and thereby modeling the noise. Such an action would remove the generality of the model and make it useless for predicting cell counts from future samples. The test set points are predicted more reliably than the external validation points. This is probably due to the fact that, while we use 20 points to represent cell counts in the training set, these points are taken from only two experiments. So, inter-experiment noise is poorly characterized by the model. In addition, we note that the performance for low concentrations is worse than that at the high concentration range. This could be an indication that the range of values used is close to the limit that can be used to approximate the nonlinear nature of the problem.

To implement the ''piecewise'' or hierarchical approach to fitting the nonlinearity with linear models, we need a controling quantification model to select the appropriate linear PLS1 model to use for any given X data. The method used for generating this model was PLS2. Light scattering data for a wide range of X values were classified into decades, 10^5, 10^6, etc. Each decade was represented by a single Y variable, having the value ''1'' if a value lay within that decade, ''0'' otherwise.

By using five Y variables in this manner, rather than a single linear variable, we can reduce the effect of nonlinearity. If we modeled a value in the range 5–9, we would require that the modeling process encompasses nonlinearity. Representing the ranges in this manner means that the modeling process needs only to identify an area of space using some form of threshold, something that does not necessarily require a nonlinear approach.

PLS2 uses an iterative method to form a model which can predict more than one Y variable. Using the Y variables described above, we can form a PLS2 model that outputs a value close to 1 for the Y variable number which corresponds to the appropriate decade for each light scattering sample. All other Y variables should be close to 0 for such a sample. By placing a simple threshold at 0.5, we should be able to decide in which decade a sample lies and hence, apply the appropriate linear model to form an accurate prediction of cell count. Figure 7 shows the learning curve for the PLS model based on a Duplex partitioned amalgamation of 12 experiments, containing 208 sample points. Using five factors, the classification error is optimal at 11.5%. Checking the W correlations shows that all five factors are stable.

Figure 8 shows the results of applying the classification model to the test data. The values plotted on gridlines indicate the decade in which the classification places each point. The remaining points show the value of $\log_{10}$(cell count), and are coded to indicate which points were misclassified by the model. Most such points are close to the border between classes, a problem which will be inherent in this arbitrary partitioning of the data. However, this form of misclassification will tend to push points from one model-

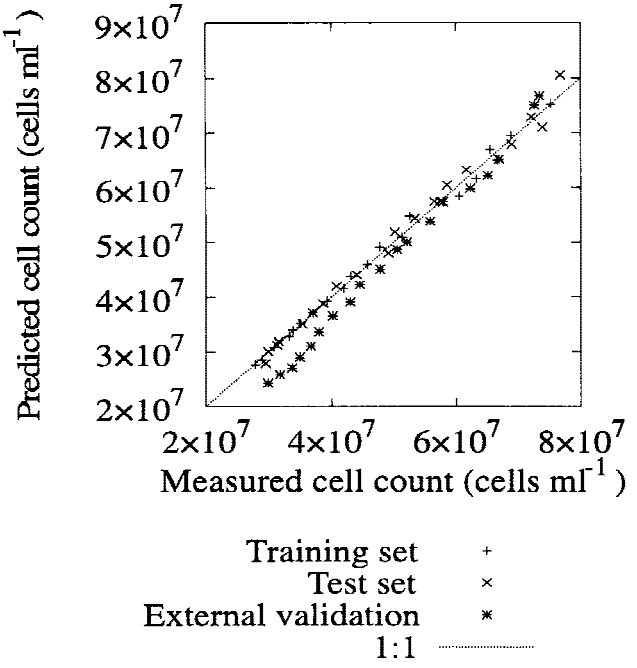

Figure 6. Predicted vs. measured cell counts using the two factor PLS1 model. The training set predictions are shown for comparison only. By adjusting the modeling method, one can make the training set self-prediction arbitrarily good (fitting the noise in the training set). Such an overfit model will have no practical use due to its dependence on the noise present.

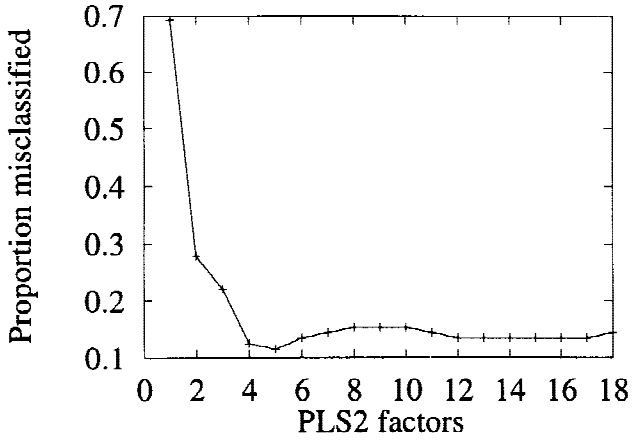

Figure 7. Learning curve for the classification PLS2 model used for determining which PLS1 model should be used to form predictions of cell count. The PLS2 model outputs a value close to 1 for the Y variable corresponding to a given modeling range, and close to 0 for other Y variables.

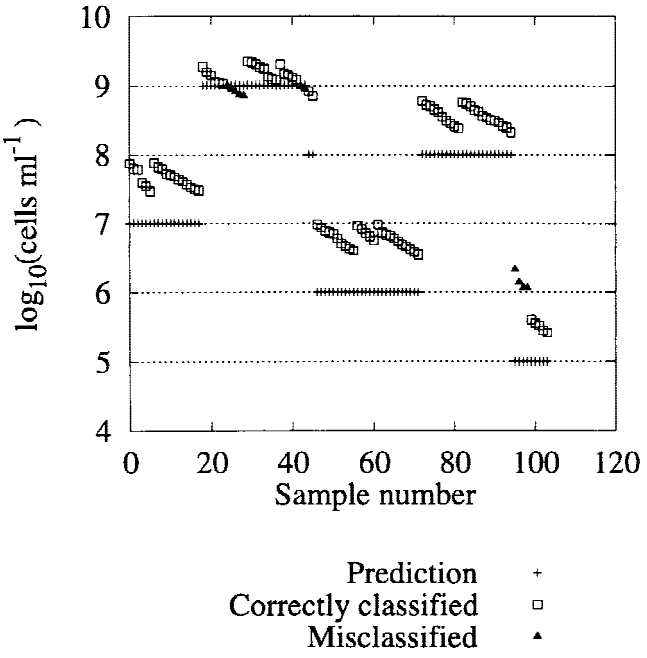

Figure 8. Performance of the classification model for test data over a wide range of cell counts. For those samples where the model selects the wrong decade, it selects an adjacent one.

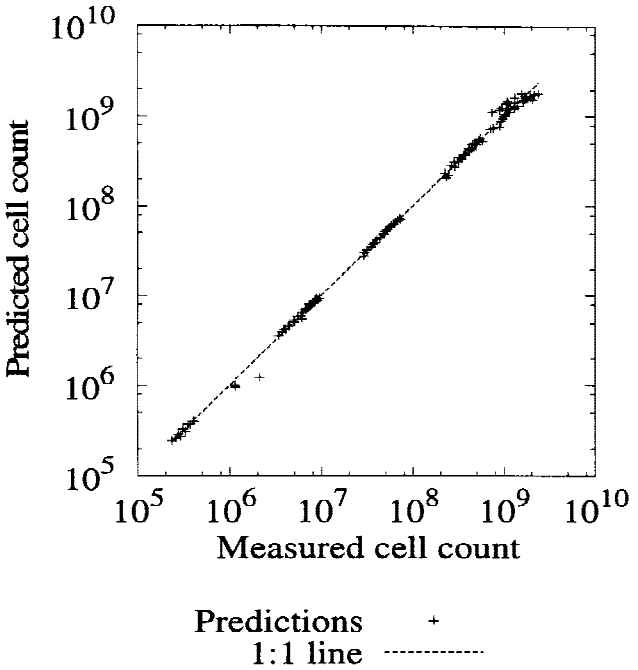

Figure 9. Predicted vs. measured cell counts for test data over a wide range of cell counts. Each decade of counts was modeled using a two-factor PLS1 model. For each light scattering sample, the appropriate PLS1 model was chosen using the output of the PLS2 classification model. This PLS1 model was then applied to the X data to form the cell count prediction.

ing region into the nearest adjacent one, so the linear models are likely to remain reasonably applicable, subject to the strength of the nonlinearity. A solution to this problem, given enough data, would be to ensure that the training set for each linear model contains a range of values which overlaps that of adjacent linear models by a few percent, thus ensuring that boundary points are well represented no matter which side of the boundary the classification process places them.

Having demonstrated that linear models can handle the nonlinearity in the light scattering data for small ranges of cell counts, and that a classification model is capable of partitioning the data into such small ranges, we can now demonstrate that the combination of such methods allows the prediction of cell count over a wide range of values. Figure 9 shows the results of applying such an arrangement to unseen test data.

This method provides an adequate method for predicting cell counts on the basis of the light scattering data. However, selecting suitable ranges of values for the PLS1 models is very much a rule of thumb. In addition, for each range, a large number of points must be recorded to form a reliable model. The prediction process is cumbersome, because five models must be stored, and their use requires two steps. While the effects of these problems could be reduced by writing dedicated software and sampling many more points, a better method would be to apply a nonlinear modeling method to the problem. The next section describes this approach.

The Nonlinear Approach

Artificial Neural Networks (ANNs) are, in many ways, similar to PLS, in terms of structure. While PLS forms a linear regression of the X variables onto a (usually smaller) set of factors, or underlying effects, then forms a linear regression from these onto the Y variables, ANNs form nonlinear relationships between the X variables (supplied to ''input nodes'') and the factors (''hidden nodes''), and between the hidden nodes and output nodes. While the PLS1 algorithm is simple, and can be calculated using a simple regression formula for each factor, ANN training methods typically use an iterative approach, ''training'' the network, evaluating how it performs, then adjusting its internal weightings to reduce the error. By repeating this training/evaluation/adjustment process many times, the internal weightings can be optimized. Because the internal relationships are nonlinear, the network can be trained to approximate an arbitrary nonlinear mapping.

We have used three layer back-propagation ANNs to form models of cell count over the entire range, with differing degrees of success. We now discuss the methods used and their results.

Modeling Total Cell Counts

This is our first, simplistic approach: Train the network using the cell concentration data, and use RMSE as the error

metric. Figure 10 shows the results of applying a two hidden-node network, previously trained in this manner, to the training data. The performance is very poor for the training data, and for all test-sets, except at high cell counts. The test set RMSEP of the optimally trained network is 8.76×10^7.

The problem with this training method becomes apparent when we consider the error value that is being minimized. At the top end of the cell count range (ca. 10^9 cells mL^{-1}), an RMSE of 8.76×10^7 cells mL^{-1} is under 10%. At the bottom end (ca. 10^5 cells mL^{-1}), this is nearly a 100,000% error. Conversely, to reduce the RMSE value to 10^4 cells mL^{-1} (i.e., 10% at the bottom end), for instance, the net would have to fit the top-end values to within 0.001%, a clearly unattainable goal, given the stated accuracy of the Coulter counter. So, when we minimize RMSE, the net is very likely to form a model which predicts the top-end values more effectively than the low-end ones. What we need is a different error metric to optimize.

Total Cell Counts, Proportional Error

Our second class of networks uses the back-propagation algorithm to minimize the mean squared proportional error,

$$\frac{1}{n}\sum_{i=1}^{n}\left(\frac{y_i-\hat{y}_i}{y_i}\right)^2$$

rather than mean squared error. Scaling by the measured y allows errors at all cell count ranges to contribute to the training process equally. Again, a number of models were formed, using between 2 and 8 hidden nodes. The proportional errors reported by these methods were in the region of 6. In other words, the average error on the predictions for models produced using this method was of the order of 6 times the magnitude of the measurements themselves. Clearly this approach is also insufficient, possibly because the ''sharpness'' of the detector output vs. cell count curve is making the global error minimum very difficult to find. Given enough time and effort, it may be possible to form a reliable network for these data. However, our approach has been to apply a transformation in order to ''smooth out'' the relationship.

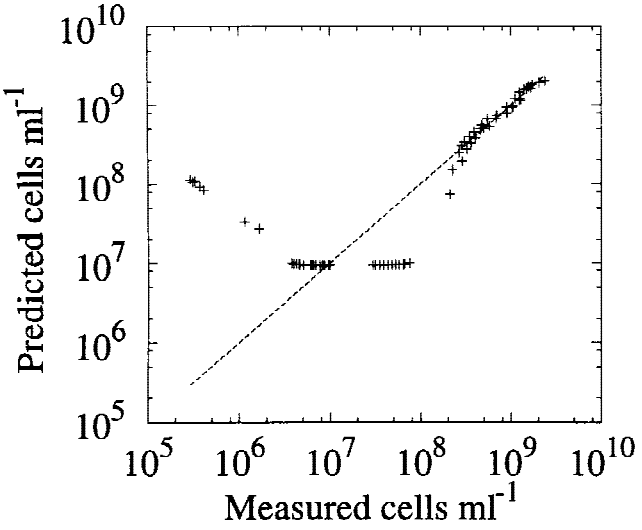

Figure 10. Self-prediction for an optimally trained two hidden-node neural network for yeast cell count prediction. The network performs very poorly at all except the highest cell counts. This result is seen for networks with a variety of different hidden-node counts, and also for networks with a linear output node.

Transformed Cell Counts, Proportional Error

Our final approach was to transform the cell count data and to train a neural network to predict the transformed numbers. A simple inverse transformation is then used to obtain the cell count predictions. The transformation used was to take the logarithm (base 10) of the cell counts. Using a five hidden node neural network using squared proportional error as the error metric, and a linear output node, we have generated a reliable model of $\log_{10}$(cell count).

Figure 11 shows the resulting predictions of $\log_{10}$(cell count) for unseen test data. The average absolute error for this range is 0.32%. Of course, what we are aiming to predict are the absolute cell counts, rather than their transformed values. A small percentage error in $\log_{10}$(cell count) will translate into a rather larger percentage error in the predicted cell count. The spread of errors shown in Figure 11 translates to an average absolute error in predicted cell count of 5.6%, with a maximum error of 25.5%. In 82% of test set cases, the error in predicted cell count is less than 10%.

Genetic Programming

The nonlinear nature of the data obtained from light scattering has been represented successfully by neural networks. One disadvantage in using this technique is that the internal

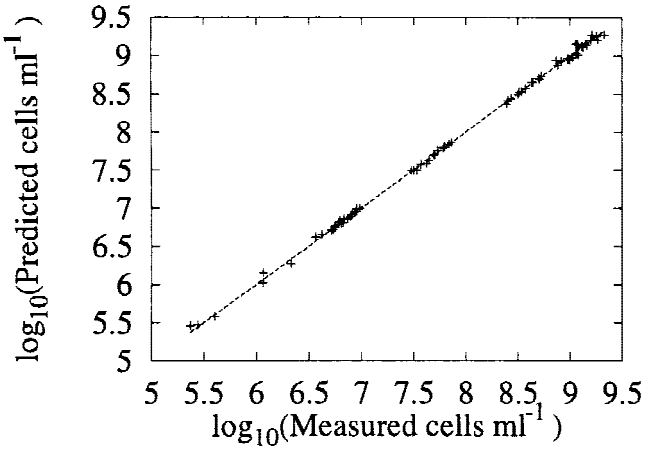

Figure 11. Predicted vs. measured transformed cell counts for validation cell data using a five hidden-node neural network.

representation of the relationship between X and Y variables is not easily interpreted. Genetic programming is a technique of software evolution developed by Koza (1992). The method attempts to provide problem solutions using automatically generated computer programs that require no explicit coding. The method effectively performs a directed search through the abstract space of possible computer programs from random starting points to find an optimal program. Genetic Programming follows the Darwinian theory of evolution and survival of the fittest to attain its goal. The fundamental principle of genetic programming is the creation of a population of computer programs, generated by the random selection of operators (such as +, −, log, etc.) and variables. These computer programs are then tested via a fitness function. Programs having a high ''fitness'' are preferentially selected to ''breed'' genetically, thus forming a second generation. This procedure is repeated until a program with a suitable fitness score has been evolved.

Before initiating an evolution process, several factors must be taken into consideration to control the development of suitable programs such as the population size, functions, and operators to be included in the instructions, the diversity of the training examples, and the size of the programs evolved. A larger, more complex function is less able to generalize, introducing the danger of overfitting.

Application of Genetic Programming to Light Scattering Data

It was hoped that by using genetic programming a simple function would be derived from the evolved program to provide an insight into the relationship between the input and output data. This experiment was carried out using a commercial GP package from SDI Products Ltd. (Fenton, MO) The software was installed and run on a 486DX66 PC under Windows NT 4.0. The evolution process was configured using the following parameters:

Number of ecosystems = 5. Each ecosystem begins its reproduction process as a distinct population following its own evolution course. The initial individuals are randomly generated to provide diverse populations. During the course of evolution however, migration may occur between populations at a probability of 0.1. This allows populations that have a low fitness to increase their diversity.

Ecology size = 20. This is the maximum number of individuals that make up a population.

The following operators were included as candidates in program formation: +, −, *, /, SQR (squaring), MIN(x, y), MAX(x, y), HI1 (Detector output having the greatest value for this sample), LO1 (Detector output having the smallest value for this sample), AVG (Average of all detector outputs for this sample).

In fact, a set of additional operators is used by the genetic programs internally, but these are never produced in the output programs.

Selection of individuals was fitness proportionate, i.e., each individual was weighted according to its fitness rank, increasing its probability of being chosen. The fitness function was the difference between the predicted Y value calculated by the individual, and the measured Y value of the training data.

The acceptable training set error threshold was set at 0.1 log units. When the score reached this level, or the training time had reached 99 h the evolutionary process halted. If the training time reaching 99 h without an acceptable training set error threshold, the population was deemed not to have solved the problem. The training data used were identical to the log-transformed training set used in the neural network analysis.

Ecology 0

RMSEP = 1.9% after 153,983 generations.

$$\log(\text{Cell Count}) = 8.45033 - \frac{\text{MIN}(1 - d_2 - d_4 - 3 * d_{18} - \text{AVG}, d_{14}) + 2 * d_1 + d_2 + d_4 - 2 * d_5 - d_{18}}{\text{HI1}} + d_6$$

Ecology 1

RMSEP = 2.4% after 2,393,177 generations.

$$\log(\text{Cell Count}) = \text{MAX}\left(1 - \text{MIN}(\text{HI1} - d_1 - d_2 * (1 + 2 * d_8) - d_5, \text{LO1}) - \frac{(d_3 - d_{11})}{d_9}, d_4\right) + 6.2743$$

Of the five ecologies, one failed to converge to a satisfactory solution in the given time. The remaining ecologies break down into pairs having very similar structures, i.e., ecologies 0 and 2; and ecologies 1 and 4. This similarity could be due to ''parallel evolution,'' where the same solution has been arrived at independently, or to the migration between ecologies of members of the populations. We, therefore, only present the results of applying ecologies 0 and 1.

As can be seen from Figures 12 and 13, the programs perform comparably well, despite their radically different structure. We attribute this fact to the structure within the data. All of the variables show a strong relationship to the cell concentration, and hence, any *reasonably sized* subset of the variables can be expected to form a model. Here, the model for ecology 0 has based its predictions on detectors 1, 2, 4, 5, 6, 14, and 18, while ecology 1 has used detectors 1, 2, 3, 4, 5, 8, 9, and 11.

Despite our assertion that the formulae resulting from genetic programs are more easily interpreted than those resulting from neural networks, in this case the strong structure seen means that any interpretation is more likely to be dependent on the particular ecology examined than on any information underlying the data.

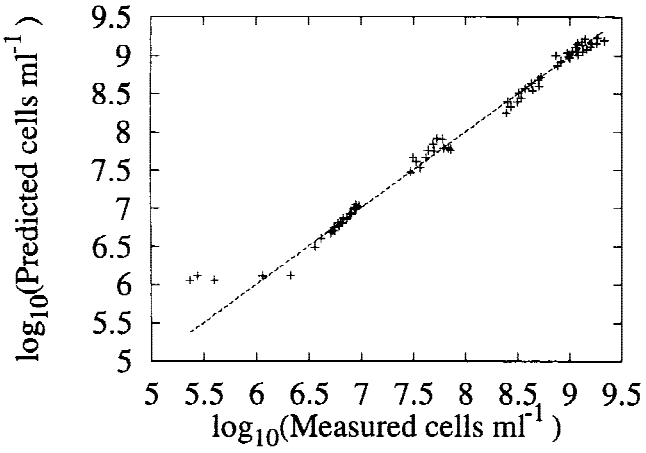

Figure 12. Validation set prediction for ecology 0 using a genetic program.

Using the DAWN ''Flow Cell'' Attachment

The DAWN ''flow cell'' attachment allows the instrument to record light scattering dynamically. It consists of a transparent cylinder with a fine-bored hole through it. The laser shines along the axis of this hole, through which media may be pumped. The DAWN software allows computer control of the pump, so that dynamic monitoring of light scattering in a bulk medium may be achieved without user interaction. It is, therefore, of interest to demonstrate the use of the flow cell in our application.

Figure 14 shows the detector outputs for a range of cell count values when the flow cell is used. The response is more nonlinear than the corresponding response when scintillation vials are used. For this reason, we would expect PLS to perform less well and this is, indeed, the case.

Using a neural network, the nonlinearity is fitted more closely, as shown in Figure 15. Over the range of values tested, the neural network model based on flow cell (RMSEP $= 1.63 \times 10^6$) data performs comparably with that based on scintillation vial data (RMSEP $= 1.62 \times 10^6$). The strong nonlinearity is highlighted once again at sample 19.

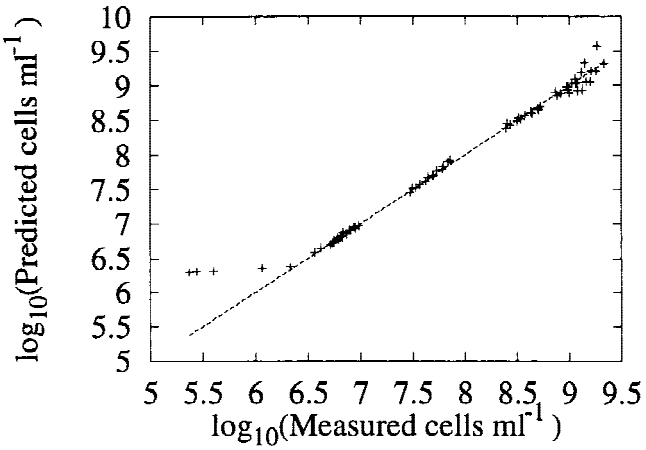

Figure 13. Validation set prediction for ecology 1.

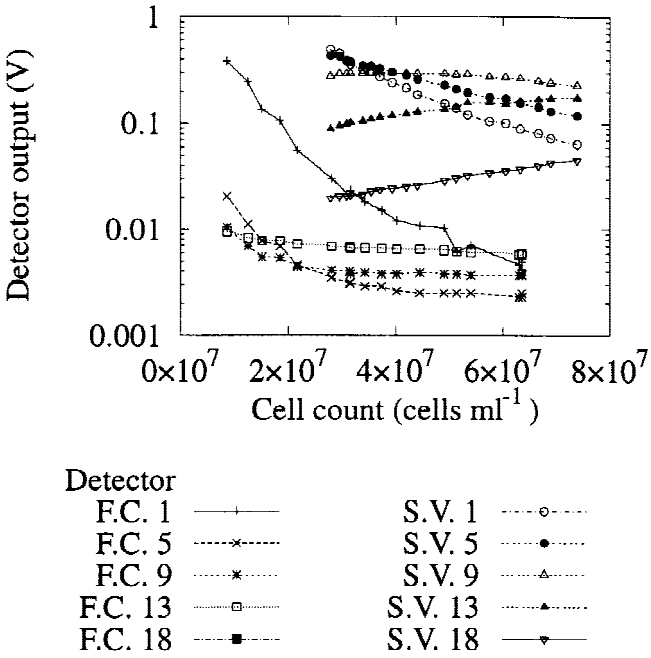

Figure 14. Detector outputs for a narrow range of cell counts using the flow cell and the scintillation vial. The nonlinearity in the light scattering profile is much more pronounced when the flow cell is used.

This sample is outside the range of the training data, and the neural network fails to predict its value adequately.

Viability

We have spent much time trying to use light scattering data for prediction of the percentage of cells which are ''viable''

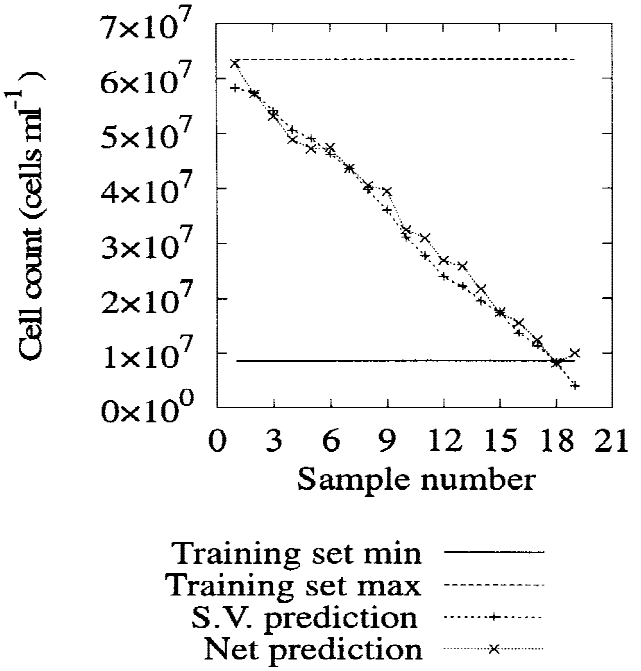

Figure 15. Using a two hidden-node neural network for prediction of cell counts from flow cell-generated light scattering data. The neural network performs comparably to one based on the scintillation vial data.

(as judged by the methods described above). We have, however, failed to form any reliable models. We feel that this is due to the gross effect of changes in concentration ''swamping'' the small effect due to viability. When the cell concentrations were allowed to vary over only a small range, trends could be seen in the predictions of viability during a ''dilution'' experiment, but these were accompanied by a significant offset, sometimes giving rise to large negative predicted viabilities! Of course, there are many possible changes which can lead to cell death (Davey and Kell, 1996), and we have investigated only a small number of methods. It is possible that further work may discover situations in which viability prediction by light scattering may become feasible.

CONCLUSIONS

We have demonstrated that the combination of multiangle light scattering and multivariate calibration techniques has proved a powerful means of determining the cell concentration in yeast suspensions over an exceptionally wide range of cell concentrations. Genetic programming may be used to form models of similar utility, but for this application, they do not provide any increased interpretability over neural networks. By using the flow cell for data acquisition, in conjunction with a neural network for prediction, we may acquire reliable estimates of cell count over an extremely wide range of values without user intervention. We also note that there are distinct qualitative differences between the light scattering profiles from suspensions of *E. coli* and yeast (unpublished data) of similar optical densities. There is, therefore, scope for future work on the determination of cell size, morphology, and/or species. It is important to note that our experiments have worked for a very well defined set of experimental conditions, and that they may not be so successful for biomass measurements of other biological particles of industrial importance. However, on the basis of our preliminary investigations of the scattering profiles of other suspensions, we would expect that this general methodology should have wide ranging use in a variety of industrial processes of biotechnological interest.

References

Bishop, C. M. 1995. Neural networks for pattern recognition. Clarendon Press, Oxford.

Bohren, C. F., Huffman, D. R. 1983. Absorption and scattering of light by small particles. Wiley, New York.

Brereton, R. G. 1990. Chemometrics: Applications of mathematics and statistics to laboratory systems. Ellis Horwood, Chichester.

Brereton, R. G. 1992. Multivariate pattern recognition in chemometrics. Elsevier, Amsterdam.

Carr, R. J. G. 1990. Fiber optic sensors for the characterization of particle-size and flow velocity. Sens. Actuat. **23**: 1111–1117.

Carr, R. J. G., Brown, R. G. W., Rarity, J. G., Clarke, D. J. 1987. Laser light-scattering and related techniques, pp. 679–701. In: A. P. F. Turner, I. Karube and G. S. Wilson (eds.), Biosensors: Fundamentals and applications. Oxford University Press, Oxford.

Chauvin, Y., Rumelhart, D. E. 1995. Backpropagation: Theory, architectures and applications. Lawrence Erlbaum, Hillsdale, NJ.

Chow, K. M., Stansfield, A. G., Carr, R. J. G., Rarity, J. G., Brown, R. G. W. 1988. Online photon-correlation spectroscopy using fibre-optic probes. J. Phys. E-Sci. Instr. **21**: 1186–1190.

Clarke, D. J., Blake-Coleman, B. C., Carr, R. J. G., Calder, M. R., Atkinson, T. 1986. Monitoring reactor biomass. TIBTech. **4**: 173–178.

Clarke, D. J., Calder, M. R., Carr, R. J. G., Blake-Coleman, B. C., Moody, S. C., Collinge, T. A. 1985. The development and application of biosensing devices for bioreactor monitoring and control. Biosensors **1**: 213–320.

Clarke, D. J., Kell, D. B., Morris, J. G., Burns, A. 1982. The role of ion-selective electrodes in microbial process control. Ion-Selective Electrode Rev. **4**: 75–131.

Corti, P., Dreassi, E. 1993. Near infrared reflectance analysis: Features and applications in pharmaceutical and biomedical analysis. Il Farmaco **48**: 3–20.

Coulter Electronics Ltd. 1994. Coulter Z1 user manual, B edition. Luton, England.

Davey, C. L., Dixon, N. M., Kell, D. B. 1990. FLOWTOVP: A spreadsheet method for linearizing flow cytometric light-scattering data used in cell sizing. Binary **2**: 119–125.

Davey, H. M., Davey, C. L., Woodward, A. M., Edmonds, A. N., Lee, A. W., Kell, D. B. 1996. Oscillatory, stochastic and chaotic growth rate fluctuations in permittistatically-controlled yeast cultures. Biosystems **39**: 43–61.

Davey, H. M., Kell, D. B. 1996. Flow cytometry and cell sorting of heterogeneous microbial populations: The importance of single-cell analysis. Microbiol. Rev. **60**: 641–696.

Drennen, J. K., Kraemer, E. G., Lodder, R. A. 1991. Advances and perspectives in near-infrared spectrophotometry. CRC Crit. Rev. Anal. Chem. **22**: 443–475.

Geisser, S. 1975. The predictive sample reuse method with applications. J. Amer. Stat. Assoc. **70**: 320–328.

Harding, S. E. 1986. Applications of light scattering in microbiology. Biotechnol. Appl. Biochem. **8**: 489–509.

Harris, C. M., Kell, D. B. 1985. The estimation of microbial biomass. Biosensors **1**: 17–84.

Hildrum, K. I., Isaksson, T., Næs, T., Tandberg, A. 1992. Near infra-red spectroscopy: Bridging the gap between data analysis and NIR applications. Ellis Horwood, New York.

Jepras, R. I., Perkins, E. A., Rarity, J., Carr, R. J. G., Clarke, D. J., Atkinson, T. 1991. Application of photon-correlation spectroscopy as a technique for detecting culture contamination. Biotechnol. Bioeng. **38**: 929–940.

Junker, B. H., Reddy, J., Gbewonyo, K., Greasham, R. 1994. Online and in-situ monitoring technology for cell-density measurement in microbial and animal-cell cultures. Bioprocess Eng. **10**: 195–207.

Kaprelyants, A. S., Kell, D. B. 1992. Rapid assessment of bacterial viability and vitality using rhodamine 123 and flow cytometry. J. Appl. Bacteriol. **72**: 410–422.

Keilmann, F., Bohme, D. Santo, L. 1980. Multichannel photometer nephelometer. Environ. Microbiol. **40**: 458–461.

Kell, D. B. 1980. The role of ion-selective electrodes in improving fermentation yields. Proc. Biochem. **15**: 18–23, 29.

Kell, D. B., Markx, G. H., Davey, C. L., Todd, R. W. 1990. Real-time monitoring of cellular biomass: Methods and applications. Trends Anal. Chem. **9**: 190–194.

Kell, D. B., Sonnleitner, B. 1995. GMP—Good modeling practice: An essential component of good manufacturing practice. Trends Biotechnol. **13**: 481–492.

Kerker, M. 1983. Elastic and inelastic light scattering in flow cytometry. Cytometry **4**: 1–10.

Koch, A. L. 1968. Theory of angular dependence of light scattered by bacteria and similar-sized biological objects. J. Theoret. Biol. **18**: 133–156.

Koch, A. L. 1984. Turbidity measurements in microbiology. ASM News **50**: 473–477.

Koch, A. L. 1986. Estimation of size of bacteria by low-angle light-scattering measurements: Theory. J. Microbiol. Meth. **5**: 221–235.
Konstantinov, K., Chuppa, S., Sajan, E., Tsai, Y., Yoon, S. J., Golini, F. 1994. Real-time biomass concentration monitoring in animal-cell cultures. Trends Biotechnol. **12**: 324–333.
Koza, J. R. 1992. Genetic programming: On the programming of computers by means of natural selection, 3rd edition. MIT Press, Cambridge, MA.
Latimer, P. 1982. Light scattering and absorption as methods of studying cell population parameters. Ann. Rev. Biophys. Bioeng. **11**: 129–150.
Locher, G., Hahnemann, U., Sonnleitner, B., Fiechter, A. 1992a. Automatic bioprocess control. 4. A prototype batch of *Saccharomyces cerevisiae.* Biotechnol. **29**: 57–74.
Locher, G., Sonnleitner, B., Fiechter, A. 1992b. Online measurement in biotechnology: Exploitation, objectives and benefits. J. Biotechnol. **25**: 23–53.
Lodder, R. A., Hieftje, G. M. 1988. Quantile BEAST attacks the false-sample problem in near-infrared reflectance analysis. Appl. Spectrosc. **42**: 1351–1365.
Mallette, M. F. 1969. Evaluation of growth by physical and chemical means, pp. 521–566. In: J. R. Morris and D. W. Ribbons (eds.), Methods in microbiology, vol. 1. Academic Press, London.
Manly, B. J. 1994. Multivariate statistical methods: A primer, 2nd edition. Chapman & Hall, London.
Mark, H. 1991. Principles and practice of spectroscopic calibration. Wiley, New York.
Markx, G. H., Davey, C. L., Kell, D. B. 1991. The permittistat: A novel type of turbidostat. J. Gen. Microbiol. **137**: 735–743.
Martens, H., Næs, T. 1989. Multivariate calibration. Wiley, Chichester.
Martin, K. A. 1992. Recent advances in near-infrared reflectance spectroscopy. Appl. Spectrosc. Rev. **27**: 325–383.
Massart, D. L., Vandeginste, B. G. M., Deming, S. N., Michotte, Y., Kaufmann, L. 1988. Chemometrics: A textbook. Elsevier, Amsterdam.
McClure, W. F. 1994. Near infrared spectroscopy: The giant is running strong. Anal. Chem. **66**: 43–53.
Miller, A. J. 1990. Subset selection in regression. Chapman and Hall, London.
Murray, I., Cowe, I. A. 1992. Making light work: Advances in near infrared spectroscopy. Verlag Chemie, Weinheim.
Næs, T., Kvaal, K., Isaksson, T., Miller, C. 1993. Artificial neural networks in multivariate calibration. J. Near Infrared Spectrosc. **1**: 1–11.
Osborne, B. G., Fearn, T., Hindle, P. H. 1993. Practical NIR spectroscopy with applications in food and beverage analysis. Longman, London.
Perkins, E. A., Carr, R. J. G., Rarity, J., Chow, K., Atkinson, T. 1993. A twin beam fibre optic laser light scattering system. Meas. Sci. Technol. **4**: 215–220.
Ripley, B. D. 1994. Neural networks and related methods for classification. J. Roy. Stat. Soc. Ser. B-Methodological **56**: 409–437.
Rumelhart, D. E., McClelland, J. L., The PDP Research Group 1986. Parallel distributed processing. Experiments in the microstructure of cognition, Vols I & II. M.I.T. Press, Cambridge, MA.
Salzman, G. C. 1982. Light scattering analysis of single cells. In: N. Castimpoolas (ed.), Cell Analysis. Plenum Press, New York.
Seasholtz, M. B., Kowalski, B. 1993. The parsimony principle applied to multivariate calibration. Anal. Chim. Acta. **277**: 165–177.
Sharpless, T. K., Bartholdi, M., Melamed, M. R. 1977. Size and refractive index dependence of simple forward angle scattering measurements in a flow system using sharply-focused illumination. J. Histochem. Cytochem **25**: 845–856.
Singh, A., Kuhad, R. C., Sahai, V., Ghosh, P. 1994. Evaluation of biomass. Adv. Biochem. Eng. **51**: 47–70.
Snee, R. D. 1977. Validation of regression models: Methods and examples. Technometrics **19**: 415–428.
Sonnleitner, B., Locher, G., Fiechter, A. 1992. Biomass determination. J. Biotechnol. **25**: 5–22.
Stone, B. M. 1974. Cross-validatory choice and assessment of statistical predictions. J. Roy. Stat. Soc. **36**: 111–133.
Stroustrup, B. 1994. The design and evolution of C++. Addison-Wesley, MA.
Ulanowski, Z., Ludlow, I. K., Waites, W. M. 1987. Water content and size of spore components determined by laser diffractometry. FEMS Microbiol. Lett. **40**: 229–232.
Van der Hulst, H. C. 1957. Light scattering by small particles. Wiley, New York.
Werbos, P. J. 1994. The roots of back-propagation: From ordered derivatives to neural networks and political forecasting. Wiley, Chichester.
Wyatt, P. J. 1968. Differential light scattering: A physical method for identifying living bacterial cells. Appl. Opt. **7**: 1879–1896.
Wyatt, P. J. 1973. Differential light-scattering techniques for microbiology, pp. 183–263. In: J. R. Norris and D. W. Ribbons (eds.), Methods in Microbiology, Vol. 8. Academic Press, London.
Wyatt, P. J. 1993a. The missing instrument: Part 1. Intl. Lab. **23**: 14–19.
Wyatt, P. J. 1993b. The missing instrument: Part 2. Intl. Lab. **23**: 21–28.
Wyatt, P. J., Berkman, R. M., Phillips, D. T. 1972. Osmotic sensitivity in *Staphylococcus aureus* induced by streptomycin. J. Bacteriol. **110**: 523–528.
Wyatt, P. J., Jackson, C. 1989. Discrimination of phytoplankton via light scattering properties. Limnol. Oceanogr. **34**: 96–112.