Hodgson, B.J., C.N. Taylor, U. Ushio, J.R. Leigh, T. Kalganova and F. Baganz (2004) Intelligent modelling of bioprocesses: a comparison of structured and unstructured approaches. *Bioprocess Biosystems Eng.,* Springer, 26 (6), 353-359

Intelligent Modelling of Bioprocesses: A Comparison of Structured and Unstructured Approaches

B J Hodgson, C N Taylor, M Ushio, J R Leigh, T Kalganova, F Baganz

Abstract

This contribution moves in the direction of answering some general questions about the most effective and useful ways of modelling bioprocesses. We investigate the characteristics of models that are good at extrapolating.

We trained 3 fully predictive models with different representational structures (diff eqns, inheritance of rates, network of reactions) on Saccharopolyspora erythraea shake flask fermentation data using genetic programming. The models were then tested on unseen data outside the range of the training data and the resulting performances compared.
It was found that constrained models with mathematical forms analogous to internal mass balancing and stoichiometric were superior to flexible unconstrained models even though no A priori knowledge of this fermentation was used.

1 Introduction

Artificial intelligence techniques have been used for several years in the modelling of bioprocesses. The majority of this work has been focussed on producing models for monitoring and control of manufacturing processes. Where the need is for robust models accurate within a narrow range of operating conditions. With data being abundantly available a number of techniques have been developed; from neural network based inferential sensors, to using GP to infer models in the form of process control diagrams[i] and hybrid models combining neural networks with a mass balance over the reactor [ii].

Models can also be used to guide the optimisation of process conditions and operating strategies during process development resulting in a saving in the number of experiments that need to be performed compared with fractional factorial statistical designs (IFED) or OFAT experiments. Kennedy and Spooner(1996)[iii] found that simple neural networks and fuzzy logic models could produce a saving of 63% in the number of experiments required for media optimisation.
Such simple statistical models only predict single response variables e.g. final DCW and so people have looked to dynamic models capable for predicting fermentation profiles to assist in optimisation of the feeding and control strategies which are intimately connected with the media and operating conditions. However the requirements and challenges of building models for process development are radically different to those of manufacturing and it is not immediately clear what modelling strategies are appropriate for this usage. Data is scarce, time is at a premium and because process conditions are changing significantly models are called upon to extrapolate far from their training data. However the models need only be indicative rather than totally accurate.

Hybrid models incorporating NN's[iv] have been used, however most work has tended to focus on parameter identification[v] in mechanistic models. Notably the work of Hans Roubos(2002)[vi] who developed a hybrid model of clavulanic acid production by streptomyces clavuligerus using an existing[vii] metabolic model as the base and inferring the kinetics using a genetic algorithm.

These approaches relying as they do on a large amount of a priori knowledge have proved to be both accurate and to have good extrapolative properties. However formulating the theoretical parts is time consuming and therefore in a development program an expensive task. This restricts their practical use to where adequate metabolic models exist.

Accepting that mechanistic models and hybrid models derived from biochemical knowledge and first principles are good at extrapolating the question we ask in this paper is: *"Is this because the theoretical part is true as a whole. Or because of certain features of mechanistic models"*
Such features may be: mass balancing, stoichiometric relations, classical enzyme kinetics or other proven subunits and concepts. Specifically we attempt to infer models(that make use of such concepts) from data with no a priori knowledge of that particular

fermentation. Our hypothesis is that such models (which we term structured models) will have superior extrapolative abilities when compared against unstructured black box methods.

This is similar to the philosophical episteme that theories that are internally consistent and built from accepted axioms as well as fitting observations have more truth-value than those that merely fit observations.

The 3 Models we compare are shown in figures 1-3.

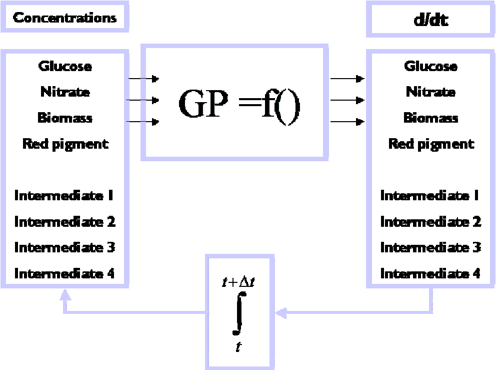

Figure 2 system of eqns with the form determined by GP

2 Materials and methods

2.1 optimisation/search method

The models were trained on a PIII computer using a general evolutionary system written in Borland C++ by the researchers. This system uses the principles of genetic programming[viii], a powerful global stochastic method inspired by Darwinian evolution capable of finding solutions in complex search spaces.

1. Randomly generate initial population
2. *Tune the values of constants in the models using simplex method*
3. Test the fitness of individuals within the population. Fitness is defined by the user essentially as being able to perform a specific task. e.g how good the model is
4. Individuals with poor fitness are killed
5. Individuals are allowed to breed by either cloning and mutation or by crossover – analogous to that in sexual reproduction.
6. These offspring and surviving individuals make up the new population

Genetic programming is very good at this global search; finding the form of models, however it is poor at finding the exact values of constants. For this reason a simplex method was employed for fast local search[ix].

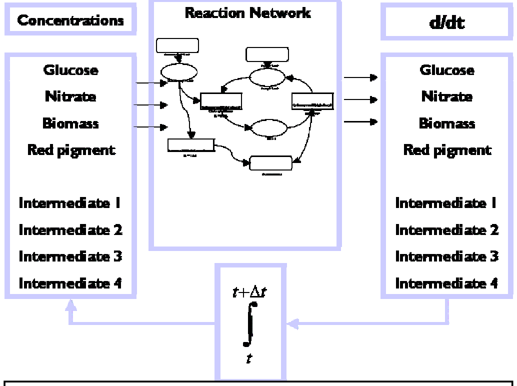

Figure 3 network of reactions with classical kinetics

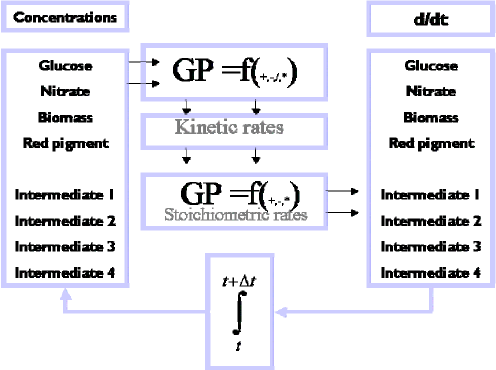

Figure 1 system of equations with inheritance of rates

This hybrid algorithm is a powerful search method and has successfully been used to find the forms of differential equations[x], and to gene regulatory networks using artificial data[xi], using a similar approach to the (GP=diff) representation used in this paper. However unlike our system they did not allow the use of internal intermediates not in the training data e.g. an internal variable representing and intermediate such as ATP or pyruvate.

2.1 fitness function[xii]

We use a fitness function that is a weighed average:

$$fitness = \frac{1}{1 + av\left(\mathrm{a.r}_{\mathrm{v}}^{\mathrm{t}}\ \mathrm{b.r}_{\mathrm{v}}^{\mathrm{b}}, \mathrm{c.}\frac{\mathrm{R}^2}{\mathrm{av(var)}}, \mathrm{d.\,N}_{\mathrm{nodes}}\right)}$$

where

$$r = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \mu_x)(y_i - \mu_y)}{\sigma_x \cdot \sigma_y}$$

x_i = array of actual values wrt(time or batch)

y_i = array of predicted values wrt(time or batch)

Highly weighted towards the correlation coefficient and with parsimony included to encourage simple models and so reduce computational burden.

We consider correlation in two dimensions:

- time with variable and batch fixed
- batch with time and variable fixed

This captures the model response with respect to the two independent variables time and initial conditions. The use of the correlation coefficient rather than R^2 error is critical in two respects, firstly because the purpose of models in optimisation is to predict that a batch run under condition x will do better than under condition y and therefore the exact yield is of lesser importance. The second feature is that since all models that have a given dynamic response will have the same correlation coefficient it performs an equivalence mapping on the search space effectively reducing the difficulty of the search.

2.1 Termination criteria

We terminate when fitness constant for n generations and the fit to training data is reasonably good.

3. Assessment of models

3.1 test system

Saccharopolyspora Erythraea (red variant wild type) was grown in defined media under different carbon and nitrogen source concentrations. The cultures were allowed to incubate concurrently at 28'C for 72 hours in a rotary shaker. The pH, biomass, nitrate, carbon and red pigment concentrations were monitored.[xiii]

The system was chosen since there is great variation in the data as the bacteria shifts from carbon to nitrate limited growth depending on the initial conditions. The production of red pigment is growth dependant under carbon-limiting conditions and is produced at the onset of the stationary phase in nitrogen-limited conditions.

3.1 goodness of fit criteria

Experiment	Training batches	Testing batches
1	1,3,4	2,5
2	1,2,3	4,5
3	3,4,5	1,2

The aim here is to produce models of the fermentation that can predict the concentration time profiles of DCW, Red pigment, Glucose, Nitrate from initial conditions only.
We will train on only 3 batches of data and test the models on 2 unseen batches of data outside the range of the training set. i.e. a glucose/nitrate concentration higher or lower than the nearest one in the training set. We do this for 3 different combinations of training and testing data so our results are independent of bias in the training data.

In order to quantify the relative performance of the models over all combinations of training and testing data we turn to 3 measures of goodness of fit.

- Scaled error on

$$\text{batch=} \sum_{1}^{nseries} \frac{\sum_{1}^{n} |error|}{n \cdot \max}$$

This gives a measure of fit corrected for the magnitude of the individual variables.

- Improvement v average=

$$\frac{\text{scaled error of model on batch}}{\text{scaled error of average profile of training data on batch}}$$

This gives us a measure of how different testing models are from the training data. Since profiles similar to the training data will be close to the average profile of the training data

- Correlation coefficient

$$= \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \mu_x)(y_i - \mu_y)}{\sigma_x \cdot \sigma_y}$$

Where x,y are the final red pigment concentrations. And the average is taken of the correlation in each training and testing set.
this provides a measure of how well the model would perform in practical use.

	Glucose (in sol'n), g/L	Nitrate (in sol'n), g/L	Pyruvic Acid, g/L	a-Ketoglutaric Acid, g/L	Red Pig, g/L	DCW, g/L
batch 1	**33.21**	**1.76**	0.028	0.152	0.000	1.371
batch 2	**35.33**	**2.35**	0.033	0.122	0.000	1.185
batch 3	**32.78**	**2.94**	0.031	0.136	0.000	1.087
batch 4	**30.05**	**4.22**	0.030	0.342	0.000	1.542
batch 5	**29.70**	**8.77**	0.034	0.001	0.000	1.839

Table 1 initial conditions of shake flask experiments

Results

Figures 4-5 show the fermentation profiles predicted by each model trained on batches 1,3,4 from initial conditions for an unseen batch(2) against actual measurements. It can be seen that the models all make "usefully accurate" predictions although they all seem to struggle with the dynamics of the red pigment. It can also be seen that the flexible pure GP structure performs worst, with the inheritance based model performing slightly better and the reaction model performing significantly better than that.

If we now turn our attention to the results of all testing and training combinations (table 3) The overall performance based on scaled errors shows that the models in order of superiority are reaction, inheritance, GP. This confirms our hypothesis that the more structured models will perform better. However looking at the data in more detail we see firstly that the difference between models is not massive. We also see that the superiority is not necessarily evident on any individual batch. This would suggest that future work should consider far more data in order to produce more statistically significant results

If we divide by the error between the average of the training data and that batch we get a scale according to how different a batch is from the training data and therefore a measure of extrapolation rather than interpolation. (1-imp v av) Again according to this measure Gp based models are significantly worse than Inherit and reaction based ones.

Looking at the correlation coefficient with respect to final product concentration which is a better indicator of how useful the models would be in practice [figure 6] we see a more pronounced difference between the models.

All these measures indicate that structured models perform better although they do not allow us to say which of the two more structured models are superior.

training	testing		GP	Inherit	reaction	average
3,4,5	batch 1	scaled error	0.44	0.21	0.63	0.85
3,4,5	batch 2	scaled error	0.46	0.23	0.60	0.77
1,3,4	batch 2	scaled error	0.39	0.26	0.15	0.29
1,3,4	batch 5	scaled error	0.56	0.52	0.43	1.16
1,2,3	batch 4	scaled error	0.17	0.21	0.17	0.68
1,2,3	batch 5	scaled error	0.56	0.92	0.26	1.36
Average	overall	scaled error	0.43	0.39	0.38	0.853
		1- imp v av	62%	50%	49%	
Table 2		correlation coeff	0.84	0.89	0.92	

Conclusions

In this paper we showed that it was possible to produce usefully predictive models of fermentation. The results showed that in this particular fermentation that more structured models had better extrapolative power than unstructured models. Since many fermentations share similar features to this one we may be tempted to generalise and extrapolate from this result to other systems. However caution is required in making this leap from a single test system. It is a critical feature of the methodology outlined in this paper that we see experiments to determine rationally what modelling approaches will work in a given situation as the future rather than a one size fits all approach. Future work is required to extend this methodology to other test systems, illuminate any critical differences in the approach required for different types of fermentation and thus find representations ideally suited to process development.

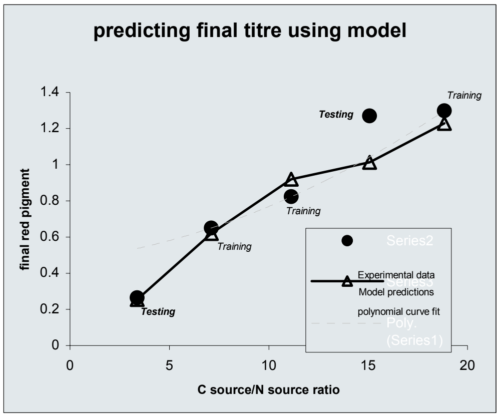

Figure 7

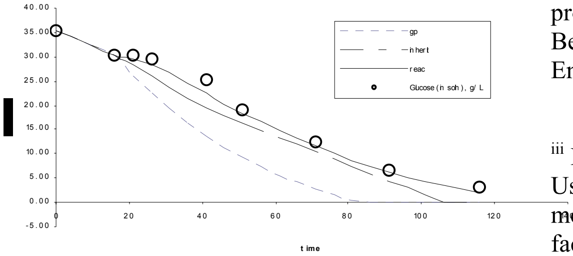

Figure 8

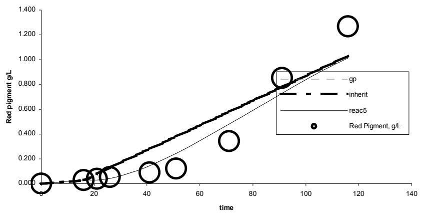

Figure 9

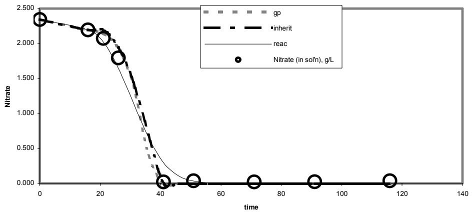

Figure 10

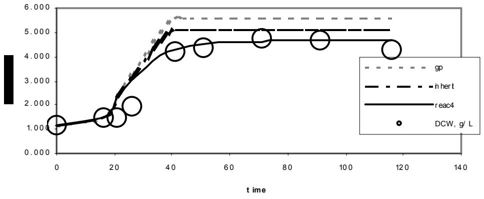

Figure 11

[i] Data driven structured modelling of a biotechnological fed-batch fermentation by means of genetic programming Pmarenbach, K D Betterhausen, S Freyer, U Nieken and H Retttenmaier Proc Instn Mech Engrs 1997 vol 211 part I

[ii] Hybrid modelling of biotechnological processes using neural networks L. Chen!, O. Bernard", G. Bastin!,*, P. Angelov Control Engineering Practice 8 (2000) 821}827

[iii] Max J. Kennedy and Natalie R. Spooner Using fuzzy logic to design fermentation media: a comparison to neural networks and factorial design. Biotechnology Techniques vol. 10 No1 January 1996 pp 47-52

[iv] Optimisation of fed-batch fermentations using hybrid models M. Ignova, G.A. Montague, G.C. Paul[+], C.A. Kent[+], C.R. Thomas[+], J. Glassey, and A.C. Ward

[v] (Isermann 1981 ? and hybridoma paper)

[vi] A semi- stoichiometric model for a streptomyces fed-batch cultivation with multiple feeds J.A. Roubos, P. Krabben, R. Luiten, R. Babuska and J.J. Heijnen 2001, Quebec City, Canada

[vii] Metabolic flux analysis of the growth of s.cavuligerus in batch-cultivations with different nitrogen sources J.A. Roubos, P. Krabben, R. Luiten, R. Babuska and J.J. Heijnen

[viii] Koza, J.R.(1992): Genetic Programming: On the Programming of Computers by Means of Natural Selection. MIT Press.

[ix] *downhill simplex method* Nelder, J.A., and Mead, R. 1965, Computer Journal, vol. 7, pp. 308–313. [1]

[x] Cao, H., Kang, L., Chen, Y., Yu, J., Evolutionary Modeling of Systems of Ordinary Differential
Equations with Genetic Programming, Genetic Programming and Evolvable Machines, 1, pp.309-337, 2000.

[xi] Erina Sakamoto Hitoshi Iba Inferring a system of differential equations for a gene regulatory network by using genetic programming proc. congress on evolutionary computation 01, 720-726, 2001

[xii] Evaluating Goodness-of-Fit in Comparison of Models to Data Christian D. Schunn Dieter Wallach

[xiii] [ref misti]