Available online at www.sciencedirect.com
ScienceDirect
International Journal of Pharmaceutics 351 (2008) 194–200
INTERNATIONAL JOURNAL OF PHARMACEUTICS
www.elsevier.com/locate/ijpharm

Modelling drug dissolution from controlled release products using genetic programming

Duong Q. Do, Raymond C. Rowe, Peter York *

Institute of Pharmaceutical Innovation, University of Bradford, Richmond Road, Bradford, West Yorkshire BD7 1DP, United Kingdom

Received 25 June 2007; received in revised form 25 September 2007; accepted 26 September 2007
Available online 12 October 2007

Abstract

This study has investigated and compared genetic programming (GP) – a method of automatically generating equations that describe the cause-and-effect relationships in a system – and statistical methods for modelling two controlled release formulations—a matrix tablet and microspheres. With the improved GP models exhibiting comparable predictive power, as well as simpler equations in some cases, the results obtained indicate that GP can be considered as an effective and efficient method for modelling controlled release formulations.

Keywords: Genetic programming; Statistical methods; Modelling; Controlled release; Formulation

1. Introduction

The design of pharmaceutical formulations involves complex interactions of ingredients and processing conditions. Genetic programming (GP) is known as a technique with the capability of generating mathematical equations, which are able to define models for data. Developed by Koza in 1992 (Koza, 1994, 1997, 1998), GP is founded on the basic principle of Darwinian theory and generates a mathematical equation based on experimental data. However, both the functional form and the numeric coefficients are found by an evolutionary mechanism. A population of possible solutions (mathematical equations) for a specific problem is randomly created and this is considered as the first generation. New generations are created by mutation and crossover. The fitness of each solution (mathematical equation) is evaluated using the fitness function. The final GP equation is considered as the optimum equation for the problem. A background to the concept and operation of GP is given below. The paper then reports the application of GP to two sets of published formulation data, one for a matrix tablet, the other for controlled release microspheres and compares the results obtained with statistical analyses.

2. Genetic programming concept

GP has been known primarily as a learning technique producing mathematical equations (Koza, 1994, 1997, 1998). It is founded on the basic principle of Darwin's theory of evolution in nature. In nature, biological structures that are more successful at operating in their given environment have a higher probability of surviving and reproducing. These structures are considered as the result of Darwinian natural selection operating in an environment over a period of time. With a similar principle of natural selection, Koza introduced GP that attempts and succeeds at applying this evolutionary theory in order to find the best (fittest) or the most appropriate equation (solution) for a problem domain.

2.1. Representation scheme

GP modifies individuals of the population after a number of runs in every cycle, called a generation, to find the best solution. The structure of individuals in GP is a tree (see Fig. 1(a)).

Possible structures in genetic programming are the set of all compositions of functions and the set of terminals.

The elements in the function set may include:

- arithmetic operations (+, −, *, etc.),
- mathematical functions (such as sin, cos, exp, and log),
- Boolean operations (such as AND, OR, NOT),

* Corresponding author. Tel.: +44 1274 233890; fax: +44 1274 234679.
E-mail address: p.york@bradford.ac.uk (P. York).

0378-5173/$ – see front matter © 2007 Elsevier B.V. All rights reserved.
doi:10.1016/j.ijpharm.2007.09.044

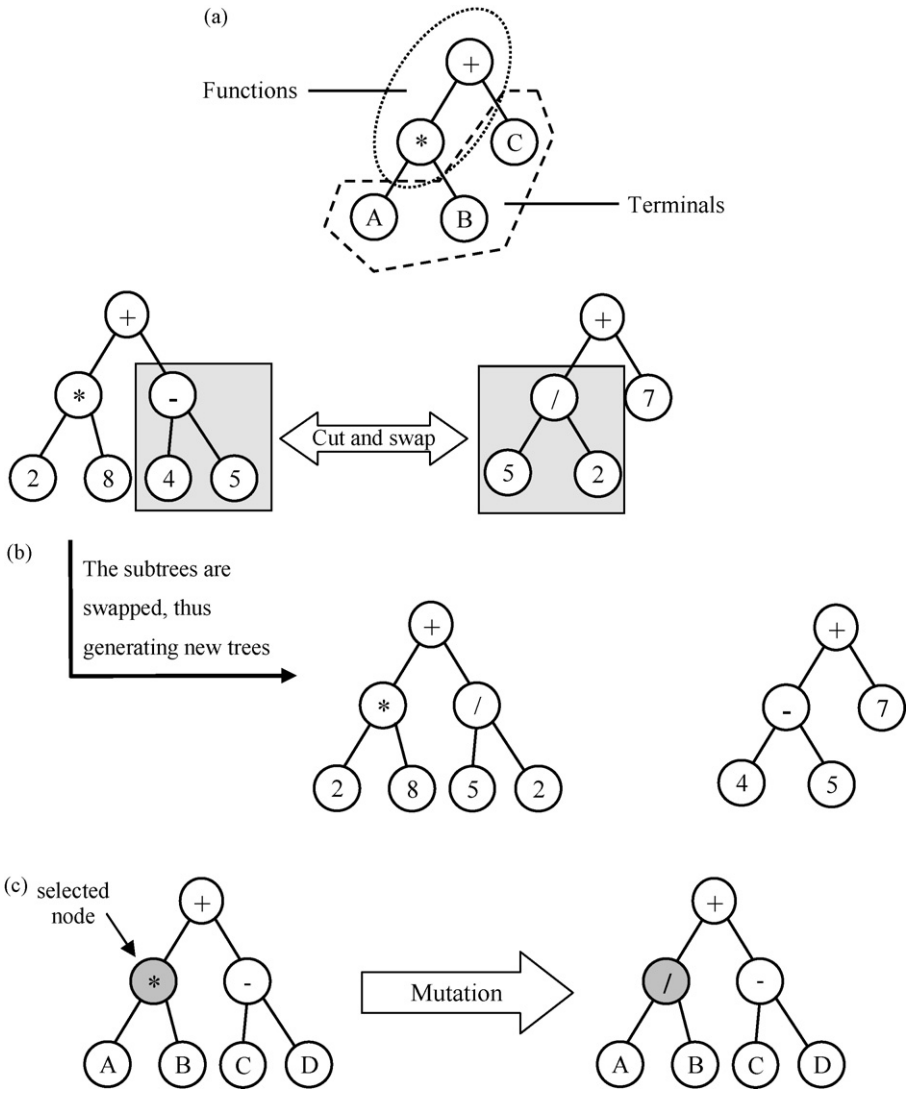

Fig. 1. (a) An example of a GP structure, (b) crossover operation and (c) process of mutation.

- conditional operators (such as If-Then-Else),
- functions causing iteration (such as Do-Until),
- functions causing recursion, and
- any other domain-specific functions that may be defined.

The terminals are typically either a variable or a constant value.

2.2. Genetic operations

In the Darwinian principle of reproduction and survival of the fittest, a population always has many generations and the individuals of new generations are the result of the combination of the individuals in the previous generations. In other words, there is a transfer of a set of individuals into a new generation of the population using genetic operations.

Crossover, reproduction, and mutation are genetic operations for creating the new individuals in each generation (see Fig. 1(b and c)). The crossover operation creates a new individual from two parental individuals selected from the population based on fitness. Within each parental tree, a random node is selected and the subtrees under the selected node are swapped to create two new trees. The simplest operation of genetic operations selects the individuals from a population based on their fitness and then copies them into the new population. The mutation operation creates a new individual from an existing tree in a population by deleting and replacing one node of that tree with another node from the same set. Note that a function only replaces a function and a terminal only replaces a terminal.

2.3. Fitness function

To select individuals for crossover, reproduction, and to determine how good the individuals are at solving the given problem, fitness functions are employed. The fitness function assesses how an individual is fitted to the environment of a domain problem, after calculating the fitness for all individuals. Every individual in a population is assigned a fitness

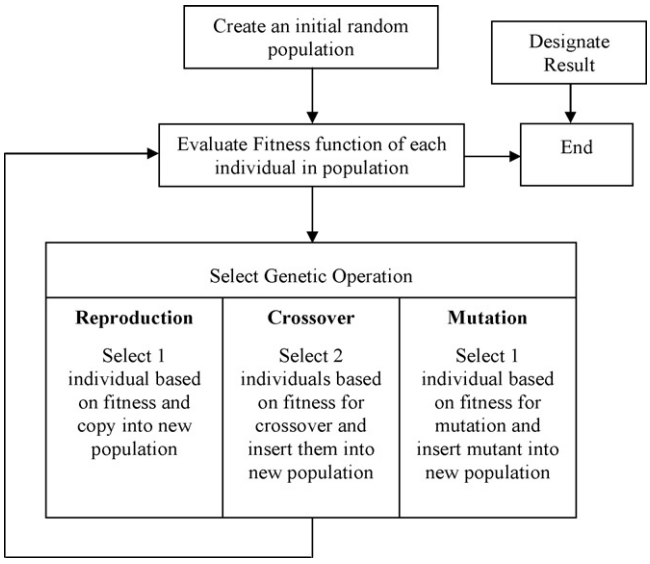

Fig. 2. The implementation of GP.

value and then, based on the adopted selection method, specific individuals are identified for crossing over, mutating, or reproducing. The considered fitness functions used in many systems for solving research and optimisation problems are MSE (Mean Square Error), SRM (Structural Risk Minimisation), AIC (Akaike's Information Criterion), MDL (Minimum Description Length), FPE (Final Prediction Error) (Weyer and Kavli, 1994).

2.4. Implementation scheme

The implementation of GP is performed by the following steps (see Fig. 2):

Step 1: One or more initial population of individuals is randomly generated with functions and terminals related to the problem domain.

Step 2: The implementation of GP iteratively performs the following steps until the termination criterion has been satisfied:

i. The fitness value of every individual is estimated according to a selected fitness measure.
ii. All individuals in the population are sorted based on their fitness values.
iii. The next generation is produced using the genetic operations.
iv. The termination criterion is checked. If it is not satisfied, the next iteration is performed; if satisfied go to step 3.

Step 3: The result may be a solution to the problem domain.

In particular, the quality of the final model is controlled by either the selection of training data or the values of the control parameters (crossover, mutation, reproduction and fitness function).

3. Materials and method

3.1. Formulation data

The formulation database of the matrix tablet taken from the literature (Bodea and Leucuta, 1997), consisted of 14 experimental records, and involved varying percentages of two hydrophilic polymers (hydroxypropylmethylcellulose, HPMC—X_1, sodium carboxymethylcellulose, CMCNa—X_2) and propranolol HCL—X_3. The measured outputs were the cumulative percentages of drug released after 1, 6, and 12 h sampling intervals (Y_1, Y_2, and Y_3, respectively). These data were modelled and optimised in the original study (Bodea and Leucuta, 1997) by statistical methods using a D-optimal quadratic model. In the present study, 10 records were used for training and records 5, 8, 13 and the observed optimum formulation from the original paper (Bodea and Leucuta, 1997) used as unseen data for testing the predictive power of GP equation.

A formulation database for controlled release diclofenac sodium microspheres containing 27 experimental records was taken from a published paper (Gohel and Amin, 1998). In this study, microspheres were prepared using sodium alginate as a polymer and $CaCl_2$ as a cross-linking agent. A 3^3 full factorial design was used to investigate the joint influences of three variables – the stirring speed during preparation of the microspheres (X_1), concentration of $CaCl_2$ (X_2) and % of heavy liquid paraffin in a blend of heavy and light liquid paraffin in the dispersion medium (X_3) – on the time for 80% drug dissolution (t_{80}). In addition, in the published study (Gohel and Amin, 1998), the % drug released after 60 (Y_{60}), 360 (Y_{360}), and 480 min (Y_{480}) was also considered as outputs that were analysed. 23 records were used as training data, and 4 records used as unseen data to test predictive power.

3.2. Software tool

The software used was a modified form of that described previously (Zongker et al., 1996), but with additional exponential and other mathematical functions as well as the fitness functions listed (Intelligensys, 2005). Table 1 lists the values of the control parameters used in the present study.

In order to evaluate the quality of a model generated by GP, the correlation coefficient R-squared (R^2) was computed, with

Table 1
The value of control parameters of GP used in this study

Population size: 1000–2000–5000
Generation: 100–500–1000
Mutation: 0.2–0.5
Crossover: 0.2–0.5
Regeneration: 0.5–0.7
Reproduction: 0.05–0.2–0.5
Number of nodes: 20–30–50–100
Constant mutation: 0.05–0.5–0.7
Fitness functions: SRM, MDL, MSE, AIC, FPE
Node weighting factor: 0.01
Addition function: exp

higher values of R^2 indicating the improved quality of the model.

$$R^2 = \left(1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}\right) \times 100 \quad (1)$$

where $\bar{y}$ is the mean of the dependent variable; $\hat{y}$ the predicted value from the model; n is the number of records.

4. Results and discussion

4.1. Validation of the method

The GP method was validated successfully, as reported elsewhere (Do, 2006), using mathematical data with 1% and 10% noise added. The results indicated that in the majority of cases the equations generated by GP were similar to the corresponding mathematical equations used to generate the mathematical data. This confirmed the validity of the GP approach to deal with real data, including the inevitable variation associated with individual data points.

Table 2
The predicted results for Eq. (2) by genetic programming

Added noise (%)	Equations predicted
0	$X_3X_2 - (X_1{}^* - 25.405)/101.62$
1	$X_3X_2 + (X_1/4.26) + (X_1/96.25)$
10	$X_3X_2 - (X_1/-4.97)$

For example from Table 2, although the equation constants in the GP derived equations were different from the original Eq. (2), the results were extremely close to the original Eq. (2). The different constants from the original equation were caused by the noise added to the data.

$$y = \frac{x_1}{4} + x_2x_3 \quad (2)$$

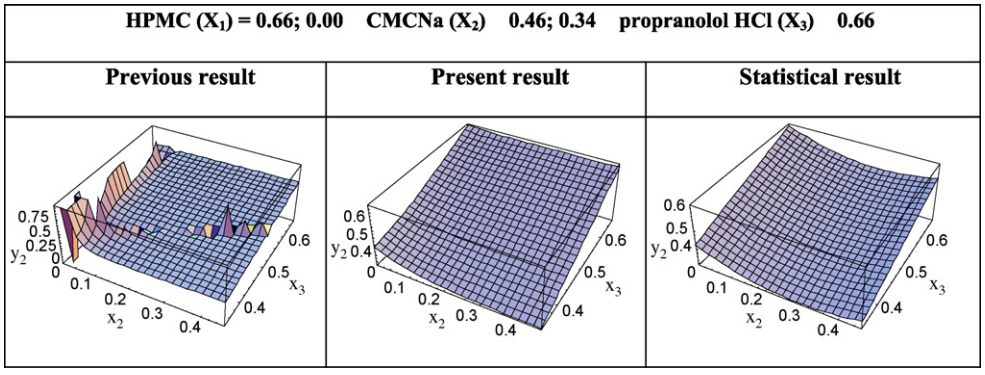

Fig. 3. Comparison of previous and present models from GP and the statistical model for Y_2.

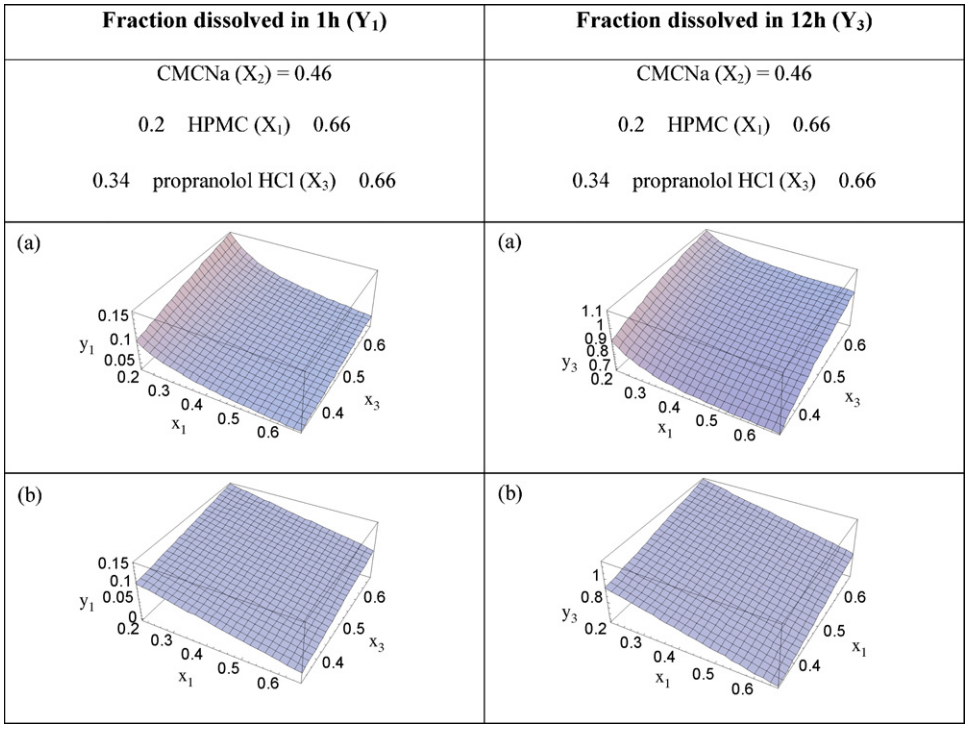

Fig. 4. 3D plots of (a) GP and (b) the statistical equations for Y_1 and Y_3.

4.2. Matrix tablet formulation

By choosing suitable values of control parameters, GP generated satisfactory models for all responses of the matrix tablet formulation. The equations generated from GP were as follows:

$$Y_1 = ((\exp((X_2/X_1)X_2))/((2.68X_2) + X_3)) \times (2.20 - X_1)(X_3/12.29) \quad (R^2 = 0.97) \tag{3}$$

$$Y_2 = X_3 + \exp(-27.94X_3(X_2 + 0.24)) + \exp(-53.39X_1X_3(X_3 + \exp(-59.68X_3X_3X_3))) \quad (R^2 = 0.94) \tag{4}$$

$$Y_3 = ((X_2 + X_2)/ - 27.78) + ((((X_1/26.34)/(\exp X_3)) \times ((75.01X_3) - (X_2/(((X_1/91.94)/ (X_2/ - 52.97))X_1)))))/X_1) \quad (R^2 = 0.97) \tag{5}$$

Compared with a previous report (Do et al., 2005), the present study gave improved models for all responses after further analyses. For the models of the cumulative percentage release after 1 h (Y_1) and 12 h (Y_3), the quality of the models was improved with higher R^2 values. In addition, for the Y_2 response, the equation was less complicated. The previous equation for Y_2 (Eq. (6)) was extremely complicated, because of overtraining or overfitting (Dias et al., 2006; Hwang et al., 1998).

$$Y_2 = (((((((X_3 - X_1)/X_2) + (((X_2 + (X_2/X_3)) \times (((X_1/ - 80.24) - X_1)(77.43X_1)))/ - 27.16))/ ((X_2 + (X_2 - ((X_3/27.38)/(20.77/40.35)))) - ((X_3 + X_3)X_2)))/(((X_2/X_2) + ((((X_3 - X_1) \times ((X_1/X_1)79.38)) + X_2)(X_1/42.05))) + ((((X_2 + X_2) + X_1)/((X_3 - X_1) + (X_1(X_3 - X_2)))) + (X_2 + 13.73)))))/ - 52.54) - ((X_2/X_1)/ ((-32.35)X_3))) + X_3), \quad (R^2 = 0.99) \tag{6}$$

Comparison of models for Y_2 in Fig. 3 indicates that whilst the present model had a lower R^2 value when compared to the previous result (Do et al., 2005), it was improved with good representation of the relationships between the Y_2 property and ingredients (X_2, X_3).

The statistical equations for the outputs Y_1, Y_2, Y_3 taken from the literature (Bodea and Leucuta, 1997) are shown below:

$$Y_1 = -0.015 + 0.145X_1 - 0.062X_2 + 0.168X_3 + 0.594X_2^2 - 0.691X_1X_2 \quad (R^2 = 0.96) \tag{7}$$

$$Y_2 = 0.279 - 0.1X_1 - 0.08X_2 + 0.626X_3 + 1.27X_2^2 - 0.841X_1X_2 \quad (R^2 = 0.88) \tag{8}$$

$$Y_3 = 0.629 - 0.246X_1 - 0.08X_2 + 0.653X_3 + 1.122X_2^2 - 0.841X_1X_2 \quad (R^2 = 0.91) \tag{9}$$

In comparison with the statistical result reported in the literature (Bodea and Leucuta, 1997), whilst producing more complex equations for many responses, GP generated equations with higher R^2 values. In addition, the 3D graphs from Figs. 3 and 4 demonstrate that comparable representations of the relationships between the ingredients and properties of the formulation can be seen from both the GP and statistical models. For example, the increasing levels of both hydrophilic polymers: hydroxypropylmethylcellulose (HPMC) and sodium carboxymethylcellulose (CMCNa) reduced amount of drug released after 1, 6 and 12 h.

Fig. 5 demonstrates the improved predictive power of the GP models for the unseen data. The linear R^2 values for all these responses were higher or very similar to those from the statistical models. For the outputs Y_2 and Y_3, the slope and the intercept coefficients from the GP models were much improved compared to those from the statistical models. For this database, it is apparent that overall GP produced improved models.

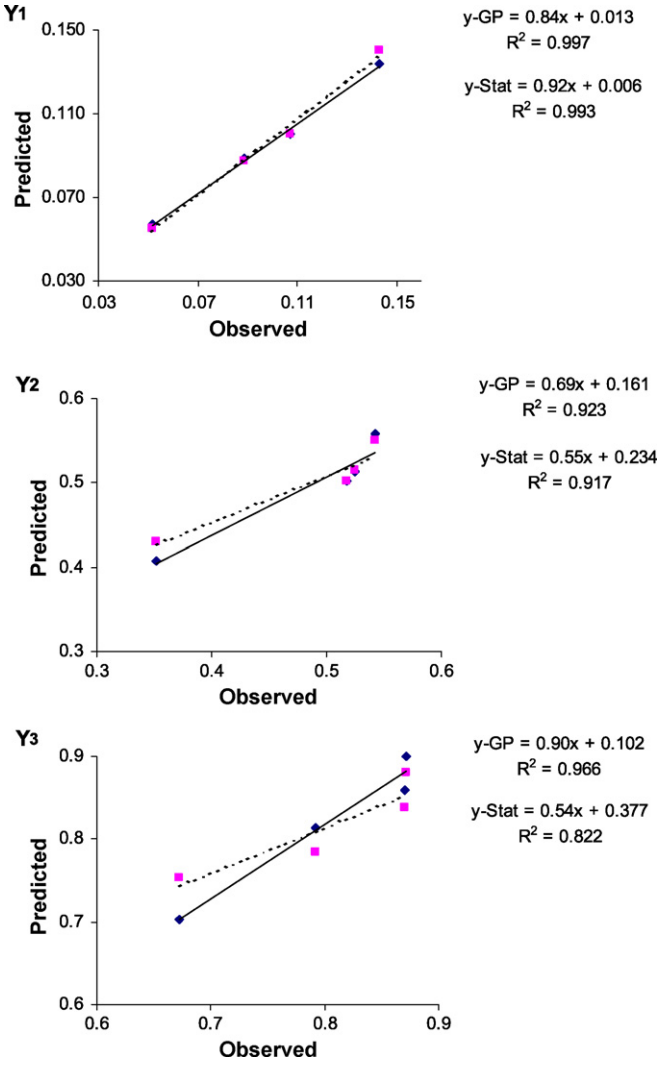

Fig. 5. Scatter plots, linear equations parameters and R^2 values for four unseen data points from GP (♦) and statistical (■) methods for Y_1, Y_2 and Y_3.

4.3. Controlled release diclofenac sodium microspheres formulation

The equations predicted by GP are as follows:

$$t_{80} = 426.22 - 51.15X_1 + 59.03X_2 + 27.60X_3 \quad (R^2 = 0.93) \tag{10}$$

$$Y_{60} = 40.15 + X_1 - X_2 - X_3 + \exp(X_1 - X_2) - \exp X_3 \quad (R^2 = 0.73) \tag{11}$$

$$Y_{360} = 73.95 + 4.00X_1 - 5.19X_2 - 2.00X_3 \quad (R^2 = 0.92) \tag{12}$$

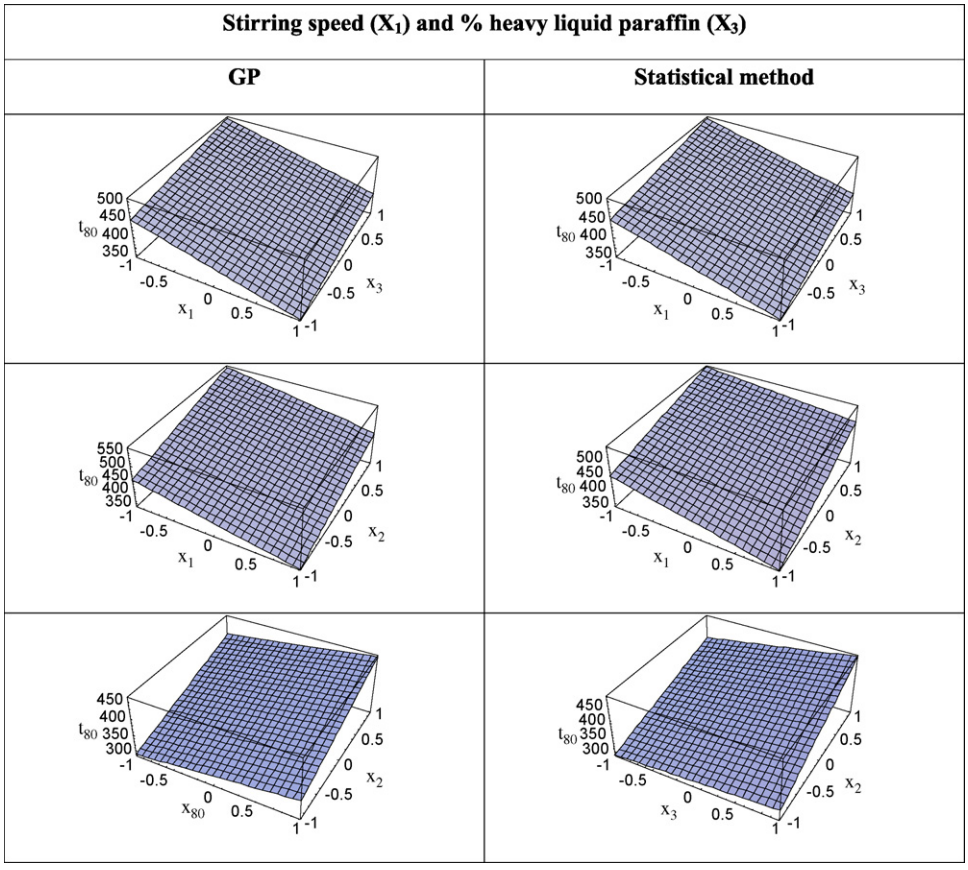

Fig. 6. 3D graphs of (a) GP and (b) the statistical equations for the response t_{80}.

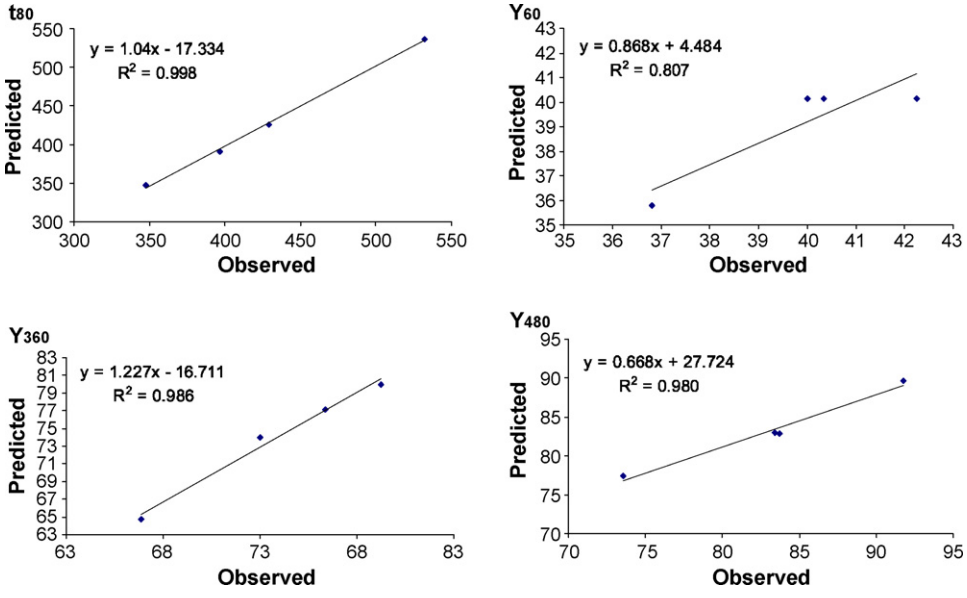

Fig. 7. Scatter plots, linear equations parameters and R^2 values for four unseen data points from GP for all responses.

$$Y_{480} = 83.02 + 3.82X_1 - 3.82X_2 - 2.82X_3 - 2.00X_1X_2 - X_2X_3 \quad (R^2 = 0.90) \tag{13}$$

It can be seen from Eq. (10), stirring speed X_1 has a positive effect on the drug release rate whilst increased amount of $CaCl_2$ (X_2) and percentage of heavy liquid paraffin in the dispersion medium (X_3) decrease the release rate. The time for 80% drug dissolution (t_{80}) is influenced simultaneously by all observed variables and increases with higher values of X_2 and X_3 as well as a low level of X_1.

The study of Gohel and Amin (1998) provided the following equation for t_{80} from statistical analysis:

$$t_{80} = 426.296 - 51.22X_1 + 58.28X_2 + 27.22X_3 + 13.08X_1X_2 + 11.92X_2X_3 \quad (R^2 = 0.96) \tag{14}$$

Eq. (10) generated by GP for t_{80} is considerably simpler than Eq. (14) with a similar R^2 value. When comparing these two equations, there are no major differences between the coefficients and the equation from GP can be considered as a simplified form of the statistical equation. Also as seen in Fig. 6, both models show that X_1 has a negative effect on the response t_{80} whilst an increased percentage of heavy liquid paraffin results in an increase of the time for 80% drug release.

For other responses, Gohel and Amin (1998) did not provided the statistical results using the same database as used for evaluating t_{80}. Thus in the present study, the predictive power of these responses was examined based on the regression analysis for unseen data.

From Fig. 7, it is clear that the satisfactory predictive power of the GP models for the unseen data can be seen. The linear R^2 values for all these responses were significantly high and the slope and the intercept coefficients from the GP models were acceptable. In general in comparison with the statistical method, GP produced satisfactory models for all responses. Moreover for the t_{80} response, the predictive equation of GP for this formulation is simpler compared to the equation generated from statistical analysis.

4.4. General comments

When validating the capability of GP using mathematical data (Do, 2006) and comparing the predictive power of GP and the statistical methods for both controlled release products, it was recognised that the basis of the statistical approach is to use standard equations and procedures based on statistical theory to obtain the final equation. This equation is unique for the input data and selected statistical procedures. The statistical output is fixed and if a formulator wants to improve the quality of the final statistical equation, he must carry out further experiments to obtain a higher quality data set. However with GP, a formulator can obtain alternative outputs, with a selection of an appropriate training model. For example, by changing values of control parameters or adding the "exp" function, the quality of the predictive equation can be improved. In other words, a formulator can perform GP in an iterative manner by directed change of control parameter values until the most appropriate and/or predictive model is obtained. However care must be taken to avoid overtraining as observed in the matrix tablet example.

5. Conclusion

Genetic programming, with its advantage of generating mathematical equations, has been shown to be an efficient method for modelling controlled release formulations. In contrast to statistical approaches, GP requires no assumption of the functional form (e.g. linear or quadratic) to be used to describe the cause-and-effect relationships.

References

Bodea, A., Leucuta, S.E., 1997. Optimization of hydrophilic matrix tablets using a D-optimal design. Int. J. Pharm. 153, 247–255.

Dias, F.M., Antunes, A., Vieira, J., Mota, A., 2006. A sliding window solution for the on-line implementation of the Levenberg–Marquardt algorithm. Eng. Appl. Artif. Intell. 19, 1–7.

Do, D.Q., 2006. Genetic programming for formulation design. Ph.D. Thesis. University of Bradford, United Kingdom.

Do, D.Q., Rowe, R.C., York, P., Colbourn, E.A., 2005. Modelling controlled release tablet formulation using Genetic programming. Presented as a poster presentation at the 32nd Annual Meeting and Exposition of the Controlled Release Society, Poster Session 2, No. 671.

Gohel, M.C., Amin, A.F., 1998. Formulation optimization of controlled release diclofenac sodium microspheres using factorial design. J. Controlled Release 51, 115–122.

Hwang, R.C., Huang, H.C., Chen, Y.J., Hsieh, J.G., 1998. Power load forecasting by neural network with a new learning process for considering overtraining problem. In: Proceedings of the International Conference on Energy Management and Power Delivery, pp. 317–322.

Intelligensys Ltd., 2005. Private communication.

Koza, J.R., 1994. Introduction to Genetic Programming—Advances in Genetic Programming. MIT Press, Cambridge (Chapter 2).

Koza, J.R., 1997. Future Work and Practical Applications of Genetic Programming—Handbook of Evolutionary Computation. Institute of Physics Publishing/Oxford University Press, Bristol, UK/New York, pp. 1–6.

Koza, J.R., 1998. Genetic Programming on the Programming of Computers by Means of Natural Selection. MIT Press, Cambridge, Massachusetts, London, England (Sixth printing).

Weyer, E., Kavli, T., 1994. The ASMOD algorithms some new theoretical and experimental results. In: Proceedings of IEEE Colloquium on Advances in Neural Networks for Control and Systems, vol. 3, pp. 1–7.

Zongker, D., Punch, B., Rand, B., 1996. Lil-gp Genetic Programming System. Michigan State University. (http://garage.cse.msu.edu/software/lil-gp/).