Utilizing a Genetic Algorithm to Elucidate Chemical Reaction Networks: An Experimental Case Study

Charles J. K. Hii, Allen R. Wright, and Mark J. Willis

***Abstract*—An artificial intelligence based on a genetic algorithm to build chemical reaction network (CRN) from chemical species concentration data from batch reaction is introduced. This is achieved through a two level optimization approach. The first level constructs the CRN through combinations of stoichiometric coefficients of all chemical species and optimized using genetic algorithm. Second level determines the best estimate for the reaction rate constants for each of the reactions using a standard non-linear optimization algorithm. The process is repeated through a number of generations where the genetic algorithm will successively reduce the number of possibilities through elimination of poor CRNs (based on how closely the CRN is able to predict concentration profiles) and retaining and re-optimizing better CRNs. This system's capability is demonstrated on an experimental data for the reaction between trimethyl orthoacetate and allyl alcohol. The results show that the system is able to develop a CRN that when simulated provides an accurate model (model predictions matching experimental measurements) with little human intervention.**

***Index Terms*—Reactor modeling, differential equations, system identification.**

I. Introduction

Detailed knowledge of a chemical reaction network (CRN) and a reaction mechanism can give a number of advantages in optimal plant and reactor design, including optimised process variables, higher quality product, more accurate process monitoring and safer, tighter and more accurate process control, lower amount of waste and by-products, more adaptation to variation in feed-stocks and measurable disturbances and improved production planning and scheduling [1]. Such knowledge is important within the fine chemical and pharmaceutical business that depends highly on short lead time between full-scale production and laboratory scale. Identification of chemical reactions occurring in a system usually involves first postulating a number of different reaction networks and then conducting relevant experiments to confirm or rule out the specific reactions. This can require substantial amount of time, resources and expertise in order to build the CRN.

Semi- automated elucidation of a chemical reaction network (CRN) can be achieved by analysing concentration profiles of chemical species from a reaction performed in a batch reactor. The analysis is usually via an inferential or deterministic approach. Inferential models such as S-systems [2] and the tendency model approach [3] attempt to fit the data into their respective models in order to achieve high level of model prediction accuracy. Although high accuracy can be achieved through such models, they can only be employed within the range of the training data. The models themselves also seldom provide any physical meaning and therefore do not provide any insight into the reactions that are occurring within the reactor.

The deterministic approach attempts to model the actual chemical reactions that are occurring through a basic understanding of the chemistry and underlying rate laws, e.g. using the law of mass action kinetics. An example of this approach is the use of target factor analysis [4] which focuses on testing postulated chemical reaction network stoichiometry. Another example is based on step-by-step statistical analysis of the concentration data and postulated models and is based on excluding reactions that do not conform to the process data [5]. Both of these earlier works require human input and consideration at every step of the CRN elucidation process and may prove cumbersome when the amount of data to analyse becomes significant.

A better solution is to automate the deterministic approach and attempts at this had been proposed through evolutionary algorithms such as differential evolution [6] and genetic programming [7]. These algorithms are able to produce successively better models from one generation to another with little human involvement. These automated approaches are advantageous in that they require less effort, time and expertise from the modeller and very minimal a priori information in order to deduce a CRN.

In this work, we aim to develop a fully automated identification system that requires little human intervention and works using minimal a priori information to elucidate a CRN based on the use of a Genetic Algorithm (GA).

II. Methodology

A. Genetic Algorithm

A GA is an evolutionary algorithm which is usually used as a numerical optimizer. Through the unique encoding used in this work, the GA can be used to build CRN structures which are mathematical models. This is achieved as shown in Fig 1, which shows three individuals and each individual represents a potential CRN (defined in terms of the stoichiometric matrix). Within each of the individuals are genotypes (rows in the matrix) which represent the reactions within the CRN. Each column in the matrix depicts the chemical species involved in the chemical reaction network. The values in the matrix describe the stoichiometric coefficients of each chemical species in a particular reaction (a negative number implying a reactant and a positive number implying a product of the reaction). Therefore, the two

Manuscript received January 15, 2014; revised April 20, 2014.

The authors are with the Newcastle University, Newcastle Upon Tyne, NE1 7RU United Kingdom (e-mail: charles.hii-jun-khiong@ncl.ac.uk, allen.wright@ncl.ac.uk, mark.willis@ncl.ac.uk).

DOI: 10.7763/IJCEA.2014.V5.439

individuals depicted in Fig 1, represent the following three CRNs.

Individual 1

$$\begin{bmatrix} -1 & 0 & 0 & 0 & -1 \\ -1 & 0 & 1 & 0 & 0 \\ 2 & 0 & 0 & -1 & 0 \\ 0 & -1 & 0 & 1 & 1 \end{bmatrix}$$

Individual 2

$$\begin{bmatrix} -2 & 0 & 0 & 1 & 0 \\ -1 & 0 & 0 & 0 & 1 \\ 0 & -2 & 1 & 1 & 0 \end{bmatrix}$$

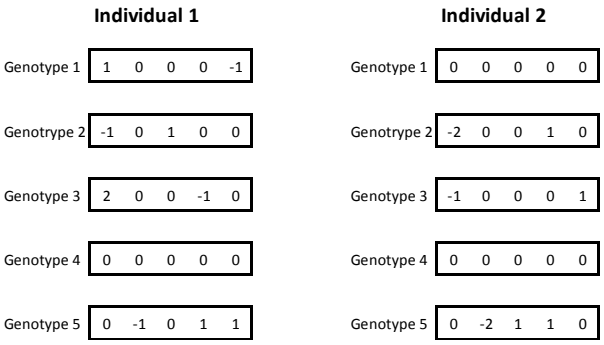

Fig. 1. Example of individuals or chemical reaction networks in the GA.

Using this type of encoding, the crossover operation has to be modified slightly from the classical version used in most GAs. When crossover is performed for classical GA, it is done to two integers in two different parent individuals. However, the encoding used in this GA uses crossover of an entire set of integers that represent the stoichiometry of the reaction. Fig 2 shows how the crossover operation is performed.

The encoding of the individuals are done with the set of following rules,

1) No two same reactions exist within the same individual.
2) All reactions must be mass balanced.
3) The highest order of reaction that can exist is second order.
4) Reactions using the same reactants but produce different products cannot exist within one individual.

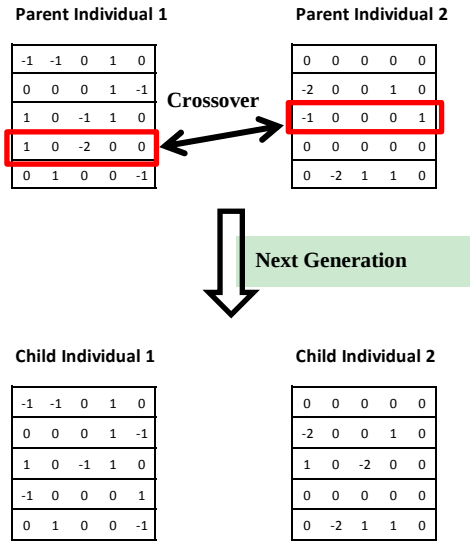

Fig. 2. Example of crossover in the GA.

B. Reaction Rate Constant Calculation

Each of the potential reactions generated by the GA must have their reaction rate defined. For example, the rate of reaction for each of the reactions in the reaction network

$$\mathrm{N} = \begin{bmatrix} -2 & 1 & 0 & 0 & 0 \\ -1 & 0 & 1 & 0 & 0 \\ 0 & 0 & -1 & 1 & 0 \\ 0 & -1 & 0 & -1 & 1 \end{bmatrix} \quad (1)$$

can be defined (based on the law of mass action) as

$$\begin{aligned} r_1 &= k_1[x_1]^2 \\ r_2 &= k_2[x_1] \\ r_3 &= k_3[x_3] \\ r_4 &= k_4[x_2][x_4] \end{aligned} \quad (2)$$

The rates of change of concentration (concentration derivatives) of the chemical species can be obtained by fitting the concentration data to a rational polynomial and then differentiating the rational polynomial. The concentration derivatives for the reaction network in (1) based on isothermal non-fed batch reaction can be defined as

$$\begin{aligned} \dot{x_1} &= -2r_1 - r_2 \\ \dot{x_2} &= r_1 - r_4 \\ \dot{x_3} &= r_2 - r_3 \\ \dot{x_4} &= r_3 - r_4 \\ \dot{x_5} &= r_4 \end{aligned} \quad (3)$$

With equation (2) and (3), the reaction rate constants k_1, k_2, k_3 and k_4 become the only unknowns. Using a least squares optimization algorithm, 'lsqnonlin' in MATLAB, the values for the reaction rate constants can be obtained.

Using the obtained reaction rate constants, the predicted concentration of the chemical species based on the reaction network in (1) can be calculated by solving the underlying ODEs. The predicted concentration can then be compared against the measured concentration using correlation coefficient or sum of squared error to evaluate the model's accuracy.

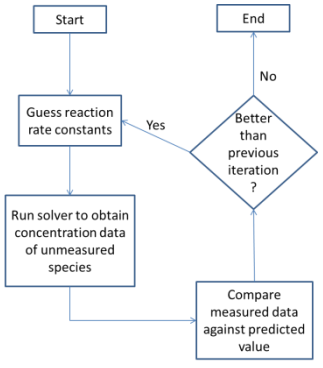

Fig. 3. Algorithm for reaction rate constants estimation.

C. Handling Unmeasured Chemical Species

The calculation in Section II-B requires the concentration profiles of all involved chemical species. When there is one or more unmeasured chemical species, this becomes infeasible. For example, without the information of a reactant, the system will be unable to predict the concentration profiles for the product of the reaction. To overcome this, the system

needs to build the concentration profiles of the unmeasured chemical species and this is achieved initially using 'guessed' values of reaction rate constants. Once built, the accuracy of the reaction rate constants can be determined through concentration data of the measured chemical species. The algorithm will then adjust the reaction rate constants in order to improve their accuracy. This is iterated until the best set of estimated reaction rate constants is obtained. Fig 3 provides the flowchart for such the algorithm.

D. Experimental Data of Reaction between Trimethyl Orthoacetate and Allyl Alcohol

For the purpose of demonstrating the capability of the automated system, experimental data for the reaction between trimethyl orthoacetate and allyl alcohol was used.

The experiment was conducted at 90 degrees Celsius in a 2 litre glass lined batch reactor for a period of 480 minutes with concentration of 1.6M for trimethyl orthoacetate and allyl alcohol. Throughout the experiment, samples were taken from the reactor and analysed using a GC-MS. The other chemical species that were detected and produced from the reaction are allyl dimethyl orthoacetate, diallyl methyl orthoacetate, triallyl orthoacetate and methanol. Due to the difficulty of measuring 'alcohol', allyl alcohol and methanol concentration profiles during the experiment were not measured.

III. Results and Discussions

Given the species concentration profiles, our GA was used to develop potential CRNs. The GA had the following parameters.

Number of generations = 200
Number of individuals = 200
Crossover probability = 20%
Mutation probability = 70%
Direct reproduction probability = 10%

The best CRN obtained from the run consisted of the following reactions,

$$R_1: \quad A + E \rightarrow B + F \quad k_1 = 0.0047\ dm^3 mol^{-1} min^{-1}$$
$$R_2: \quad C + F \rightarrow B + E \quad k_2 = 0.0032\ dm^3 mol^{-1} min^{-1}$$
$$R_3: \quad A + C \rightarrow 2B \quad k_3 = 0.0023\ dm^3 mol^{-1} min^{-1}$$
$$R_4: \quad 2B \rightarrow A + C \quad k_4 = 0.0032\ dm^3 mol^{-1} min^{-1}$$
$$R_5: \quad B + D \rightarrow 2C \quad k_5 = 0.0260\ dm^3 mol^{-1} min^{-1}$$
$$R_6: \quad 2C \rightarrow B + D \quad k_6 = 0.0116\ dm^3 mol^{-1} min^{-1}$$

where
A = Trimethyl orthoacetate
B = Allyl dimethyl orthoacetate
C = Diallyl methyl orthoacetate
D = Triallyl orthoacetate
E = Allyl alcohol
F = Methanol
R_i = reaction index
k_i = reaction rate constant

Fig. 4 shows the plot of the concentration data against time for the chemical species (predicted concentration profiles based on the CRN are shown as lines while experimental concentration data points are marked '×').

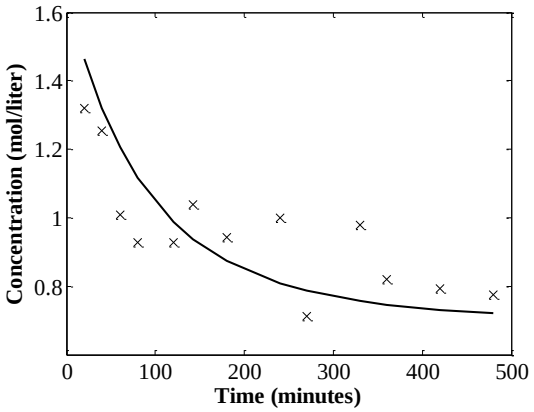

Fig. 4(a). Concentration of trimethyl orthoacetate against time for experimental and predicted data.

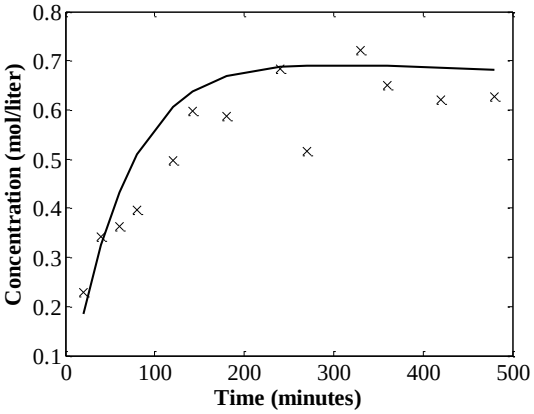

Fig. 4(b). Concentration of allyl dimethyl orthoacetate against time for experimental and predicted data.

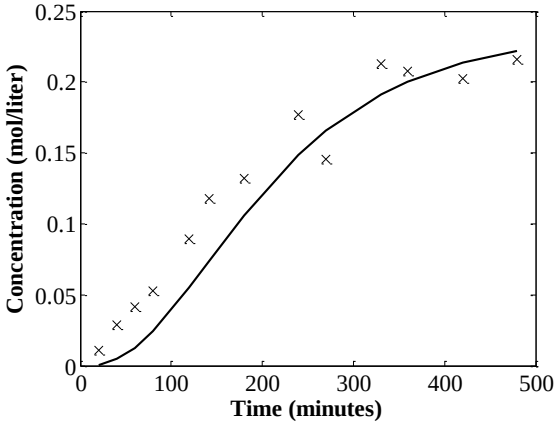

Fig. 4(c). Concentration of diallyl methyl orthoacetate against time for experimental and predicted data.

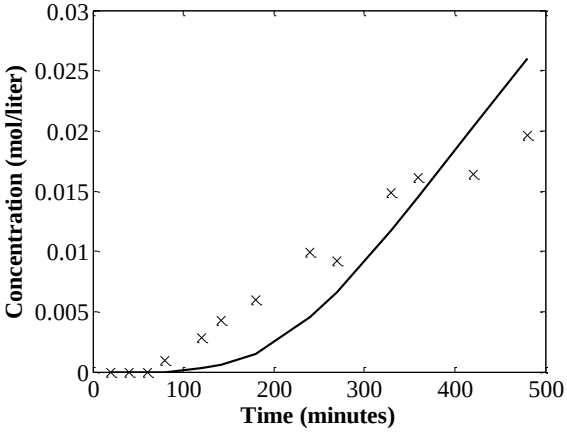

Fig. 4(d). Concentration of triallyl orthoacetate against time for experimental and predicted data.

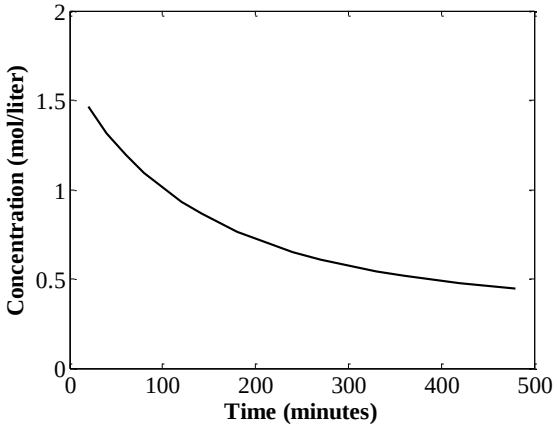

Fig. 4(e). Concentration of allyl alcohol against time for experimental and predicted data. Note: there is no experimental data.

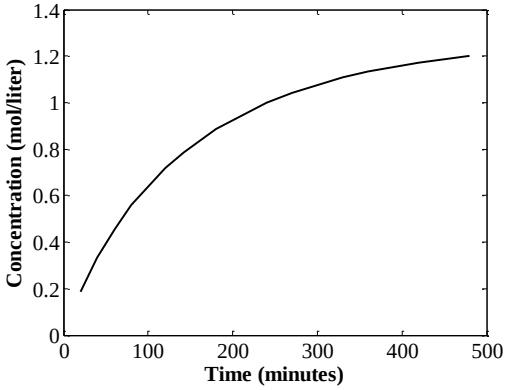

Fig. 4(f). Concentration of methanol against time for experimental and predicted data. Note: there is no experimental data.

Of the 6 reactions that have been elucidated by the GA, 2 pairs of them are reversible e.g. R_3 and R_4 describe the same reaction, one being the forward and the other the reverse reaction. The same case goes for R_5 and R_6 . This demonstrates the capability of the system to detect reversible reactions should they exist in the system.

The accuracy of the predicted concentration profiles can be seen in Fig 4. It can be observed that the CRN has modelled the concentration profiles accurately and even provides a prediction on the unmeasured chemical species (allyl alcohol and methanol).

Analysis on the elucidated CRN is shown to be similar to the ortho ester exchange between triethyl orthoacetate and allyl alcohol as discussed by Bollyn and Wright [8]. They had deduced that the CRN to be as followed:

$$R_{1\prime}\colon\ A' + E \rightarrow B' + F'$$
$$R_{2\prime}\colon\ B' + F \rightarrow A' + E$$
$$R_{3\prime}\colon\ B' + E \rightarrow C' + F'$$
$$R_{4\prime}\colon\ C' + F \rightarrow B' + E$$
$$R_{5\prime}\colon\ C' + E \rightarrow D + F'$$
$$R_{6\prime}\colon\ D + F' \rightarrow C' + E$$

where

A' = Triethyl orthoacetate
B' = Allyl diethyl orthoacetate
C' = Diallyl ethyl orthoacetate
D = Triallyl orthoacetate
E = Allyl alcohol
F' = Ethanol
R_i = reaction index

On the first look, R_1 and $R_{1\prime}$ is similar and the only difference is R_1 has methyl group exchanging within the reaction while $R_{1\prime}$ has ethyl group. The same can also be observed between R_2 and $R_{4\prime}$. Further investigation shows that R_3 to R_6 are actually combinations of $R_{1\prime}$ to $R_{6\prime}$ which is summarized below:

$$R_3 = R_{1\prime} + R_{4\prime}$$
$$R_4 = R_2 + R_{3\prime}$$
$$R_5 = R_{3\prime} + R_{6\prime}$$
$$R_6 = R_{4\prime} + R_{5\prime}$$

If the reaction between trimethyl orthoacetate and alyl alcohol do indeed undergo the same reaction scheme as deduced by Bollyn and Wright [8], then the elucidated CRN by the GA is able to show that such reactions do occur. The elucidated CRN failed to detect the in-between reactions (causing it to combine reactions) is due to the fact that the unmeasured chemical species plays a large part in the transition between the reactions. Without good data on the unmeasured chemical species, the system will assume that it is not required unless it affects the prediction of the concentration of other chemical species.

Although some of the elucidated reactions are shown to have skipped a reaction step, the possibility of such reaction occurring can be verified through experimental work. Furthermore, tacit knowledge of a chemist can also be used here to exclude certain reactions, for example the knowledge that reaction between two large molecules are less likely to occur than that between a large and smaller molecule.

IV. Conclusion

An automated system that employs GA has been designed to search and fit the best possible CRN from concentration profiles of involved chemical species in a batch reaction. The system has been demonstrated to be able to model a CRN even in the presence of unmeasured chemical species. Reversible reactions are also detected and can be evaluated by the system. Only minimal human intervention is required at the beginning of the GA run where parameters for the GA will need to be entered. The results from the elucidation of the CRN of the experimental data of the reaction between trimethyl orthoacetate and allyl alcohol showed a good level of accuracy.

Further work can still be done on improving the algorithm in terms of the objective function used by introducing multi-objective fitness function. Tacit knowledge such as relative molecular size between reactants or a preference for a more compact CRN can be entered as another objective.

References

[1] G. Maria, "A review of algorithms and trends in kinetic model identification for chemical and biochemical systems," *Chem. Biochem. Eng.*, vol. 18, no. 3, pp. 195–222, 2004.
[2] S. Kikuchi, D. Tominaga, M. Arita, K. Takahashi, and M. Tomita, "Dynamic modeling of genetic networks using genetic algorithm and s-system," *Bioinformatics*, vol. 19, no. 5, pp. 643-650, 2003
[3] M. V. le Lann, M. Cabassud, and G. Casamatta, "Modeling, optimization and control of batch chemical reactors in fine chemical production," *Annual Reviews in Control*, vol. 23, pp. 25-34, 1999.
[4] D. Bonvin and D. W. Rippin, "Target factor analysis for the identification of stoichiometric models," *Chemical Engineering Science*, vol. 45, pp. 3417–3426, 1990.

[5] S. C. Burnham, "Towards the automated determination of chemical reaction networks," Ph.D. dissertation, School of Chemical and Advanced Materials, Newcastle University, Newcastle Upon Tyne, United Kingdom, United Kingdom, 2007.
[6] D. P. Searson, M. J. Willis, and A. R. Wright, "Reverse enginering chemical reaction networks from time series data," in *Statistical Modelling of Molecular Descriptors in QSAR/QSPR*, M. Dehmer, K. Varmuza, and D. Bonchev, Eds. John Wiley & Sons Incorporated, 2012, pp. 327-348.
[7] J. R. Koza, W. Mydlowec, G. Lanza, J. Yu, and M. A. Keane, "Automatic computational discovery of chemical reaction networks using genetic programming," in *Computational Discovery of Scientific Knowledge*, D. Sašo and T. Ljupco, Eds. Springer-Verlag Berlin Heidelberg, 2007, pp. 205-227.
[8] M. P. Bollyn and A. R. Wright, "Development of a Process Model for a Batch Reactive Distillation – A case study," *Computers & Chemical Engineering Journal*, vol. 22, pp. 87-94, 1998.

Charles J. K. Hii is studying for his PhD in chemical engineering in School of Chemical Engineering and Advanced Materials under the tutelage of Dr. Willis. His research interest is focused mainly on utilizing the power of computing in the field of chemical engineering. The research goal of his PhD is to design an automated system using evolutionary algorithms that will enable elucidation of chemical reaction networks (chemical mechanisms).

Allen R. Wright is a professor of process development in the School of Chemical Engineering and Advanced Materials, Newcastle University, United Kingdom. He was the founder, MD and Technical Officer of BatchCad Ltd. He has also served as the Vice President of Advanced Technology and senior consultant during the initial planning of Avantium Technologies. His research interest lies in reaction engineering focusing on batch processing, understanding reaction kinetics, development of reactive distillation and process simulation. He is involved in the development of robotic workstations that carry out parallel synthesis and automated elucidation of chemical reaction network (chemical mechanism) using novel computer algorithms.

Mark J. Willis is a senior lecturer and the director of excellence in learning and teaching in the School of Chemical Engineering and Advanced Materials, Newcastle University, United Kingdom. The aim of his research is to develop enabling technologies for process development and he focuses on high throughput technologies. His primary research interest is automated elucidation of chemical reaction networks (chemical mechanism) using novel computer algorithms.