Elucidation of Chemical Reaction Networks through Genetic Algorithm

Charles J. K. Hii

School of Chemical Engineering and Advanced Materials

Newcastle University

A thesis submitted to the Faculty of Science, Agricultural and Engineering, Newcastle University, in partial fulfilment of the requirements for the degree of Doctor of Philosophy

Abstract

Obtaining chemical reaction network experimentally is a time consuming and expensive method. It requires a lot of specialised equipment and expertise in order to achieve concrete results. Using data mining method on available quantitative information such as concentration data of chemical species can help build the chemical reaction network faster, cheaper and with less expertise.

The aim of this work is to design an automated system to determine the chemical reaction network (CRN) from the concentration data of participating chemical species in an isothermal chemical batch reactor. Evolutionary algorithm ability to evolve optimum results for a non-linear problem is chosen as the method to go forward. Genetic algorithm's simplicity is modified such that it can be used to model the CRN with just integers.

The developed automated system has shown it can elucidate the CRN of two fictitious CRNs requiring only a few a priori information such as initial chemical species concentration and molecular weight of chemical species. Robustness of the automated system is tested multiple times with different level of noise in system and introduction of unmeasured chemical species and uninvolved chemical species. The automated system is also tested against an experimental data from the reaction of trimethyl orthoacetate and allyl alcohol which had shown mixed results. This had prompted for the inclusion of NSGA-II algorithm in the automated system to increase its ability to discover multiple reactions.

At the end of the work, a final form of the automated system is presented which can process datasets from different initial conditions and different operating temperature which shows a good performance in elucidating the CRNs.

It is concluded that automated system is susceptible to 'overfitting' where it designs its CRN structure to fit the measured chemical species but with enough variation in the data, it had shown it is capable of elucidating the true CRN even in the presence of unmeasured chemical species, noise and unrelated chemical species.

Acknowledgement

I would like to take the opportunity to express my gratitude to supervisor, Dr. Mark Willis for his excellent guidance and advice. Next, I will like to thank Dr. Searson for sharing his knowledge on evolutionary algorithm. Also, thanks to Prof. Allen Wright for sharing his experimental data with me and his suggestions on the work. I would like to also thank Julie Parker for her help in understanding the experimental data. Last but not least, I would like to acknowledge all the staff members in school CEAM for their help and support.

Table of Contents

Nomenclature

Symbols	**Description**
i	Index for chemical species
j	Index for reactions within a chemical reaction network
$k_{j,t}$	Reaction rate constant of j^{th} reaction at time t
m_{x_i}	Molecular mass of chemical species i
N_c	Total number of chemical species
N_r	Total number of chemical species
N_t	Reaction end time
$r_{j,t}$	Rate of reaction of j^{th} reaction at time t
r_{x_i}	Rate of consumption or production of chemical species i
t	Batch reaction time
V	Volume of reactor
F	Volumetric flow rate
ν_{x_i}	Stoichiometric coefficient of chemical species i
x_i	Chemical species i
$[x_i]_t$	Input concentration data of chemical species i at time t
$[\hat{x}_i]_t$	Predicted concentration data of chemical species i at time t
$y_{i,t}$ or $[\dot{x}_i]_t$	Experimental rate of concentration change of chemical species i at time t
$\hat{y}_{i,t}$	Predicted rate of concentration change of chemical species i at time t
Z_{x_i}	Number of moles of chemical species i
μ	Mean
σ	Standard deviation

E_a	Acitivation energy
R	Universal gas constant
T	Absolute temperature
N	Stoichiometric coefficient matrix of a single reaction

Matrix	**Description**
$\boldsymbol{K}$	Matrix of all reaction rate constants
$\mathbf{M}$	Matrix of molecular mass of involved chemical species
$\boldsymbol{V}$	Stoichiometric matrix of chemical reaction network
$\widetilde{\boldsymbol{N}}$	Transpose of $\lvert N \rvert$ after positive elements are made 0
$\mathbf{R}$	Matrix of all rate of reaction within chemical reaction network
$[\boldsymbol{X}]$	Matrix of all experimental concentration data of involved chemical species
$\left[\widehat{\boldsymbol{X}}\right]$	Matrix of all predicted concentration data of involved chemical species
$\boldsymbol{Y}$	Matrix of all experimental rate of concentration change of involved chemical species
$\widehat{\boldsymbol{Y}}$	Matrix of all predicted rate of concentration change of involved chemical species

List of Figures

List of Tables

Chapter 1. Introduction

1.1 Aims and objectives

The main of this thesis is the development of an automated system that elucidates chemical reaction network from the concentration data of participating chemical species from an isothermal chemical batch reactor. For the system to be entirely automated, evolutionary algorithm can be employed in order to build the chemical reaction networks and MATLAB can be used to design it. To develop the automated system, the underlying mathematics to solve for the reaction rate constants based on available data must be understood. Through a good estimation of the reaction rate constants, concentration data be reconstructed from the chemical reaction network and used for comparison against simulated data. The evolutionary algorithm can then evolve the structure based on the comparison, aiming to achieve the best estimation between the simulated/experimental data and predicted data.

1.2 Motivation

Information on chemical reaction network and its reaction kinetics are among the most important information required in order to scale-up any chemical process for industrial scale (Rostrup-Nielsen, 2000). Among the benefits (Maria, 2004) of having an accurate chemical reaction network are

a. Optimal plant and reactor design
b. Optimised process variables
c. Higher quality product
d. More accurate process monitoring and safer
e. Tighter and more accurate process control
f. Lower amount of waste and by-products
g. More adaptable to variation in feed-stocks and measurable disturbances
h. Improved production planning and scheduling.

Therefore, it is important that such information to be developed quickly to reduce the amount of time required to scale up production from laboratory scale to an industrial scale with highly efficient plant (Le Lann, Cabassud, & Casamatta, 1999). This is especially true among fine chemicals and pharmaceutical companies where their products are complex and depended on development of new products to succeed.

1.3 Overview of the thesis structure

The thesis begins with an introduction to the chemical kinetics principles and the different types of chemical reactions in Chapter 2. The terminology used for chemical reactions are explained and also includes the basic integral and differential technique for estimation of the reaction rate constants which would provide a basic background to the topic presented in this thesis. This is followed on by a literature review on work done in the field of elucidation of chemical reaction networks which is divided into inference, deterministic and automated system. The pros and cons of each methods are discussed which leads to the design of an automated system to discover the chemical reaction network.

Chapter 3 gives an introduction to the genetic algorithm and reveals the function of each parts of the evolutionary algorithm. Genetic algorithm is the base for the automated system developed through this thesis and in this chapter, the first form of the automated system is introduced. A method to calculate the reaction rate constants through the solution of multiple linear regression is demonstrated.

Chapter 4 through to Chapter 8 are about the application and development of the automated system. In Chapter 4, the basic automated system for chemical reaction networks elucidation is tested against two fictitious chemical reaction network. The weaknesses of the automated system are revealed to be unable to deal with reversible reaction, unmeasured chemical species and suffer from ‘overfitting’ of data.

In Chapter 5, the weaknesses of the automated system are being addressed, namely its inability to work in the presence of unmeasured chemical species. A new upgraded automated system, named automated system (version 2) is introduced which uses a different approach to reaction rate constants calculation. Automated system (version 2) approaches the problem by splitting the optimisation problem into two tiers, where the first tier optimises the chemical reaction networks and the second tier on the reaction rate constants. The modification has the added benefit of the automated system able to handle reversible reactions now. Further testing shows that it does work on reversible reactions and unmeasured chemical species, but suffer from ‘overfitting’ as well.

A slight overview on multi-objective optimisation is presented in Chapter 6 which leads into the enhancement of the automated system to become multi-objective optimisation capable. The new enhanced automated system is named automated system (NSGA-

II). Diversity in the chemical reaction networks became one of the objectives of the automated system in the bid to reduce the effect of ‘overfitting’. The implementation of automated system (NSGA-II) provides user with a choice of reactions to choose rather than a single chemical reaction network. Tests conducted revealed that although diversity in the population in the genetic algorithm had increased significantly, success rate for the elucidating the entire CRN is still low.

Chapter 7 takes automated system (version 2) and automated system (NSGA-II) to test on a laboratory experimental data which involves the reaction between trimethyl orthoacetate and allyl alcohol. The poor state of the experimental data meant that it needs to be adjusted before it can be used to further testing both of the automated systems. Their behaviour in the presence of chemical species that is totally unrelated to the chemical reaction networks that is being elucidated is tested and results show that the unrelated chemical species does affect their performance.

The automated system is modified further in Chapter 8 where it can now evaluates data from different batches running with different initial conditions and at different temperature. The results had shown that the modified automated system is able to cope better with noise, increased ability to determine the all the reactions in the CRN and ability to discern from involved and uninvolved chemical species. All in all, a better performance as compared to the previous iterations of the automated system. A final form of the automated system is proposed which is a combination of both of the modifications but the final form was not tested due to the lack of suitable datasets.

The thesis ends with Chapter 9 which summarised all the findings from this work and discussed on the direction on future work that will be more involved in developing the system further.

Chapter 2. Literature Review

2.1 Overview

In this chapter, an introduction to the chemical kinetics, terminology and the calculation involved in reaction rate constants are presented. This is followed by a survey on the work done in the field of elucidation of chemical species which is divided into inference, deterministic and automated systems. Inference modelling refers to modelling the chemical reaction network by fitting mathematical models that have no physical meaning. Deterministic models are model that are based on the reaction kinetics and provides insight on the workings of the reactions. Automated systems are system that are based on evolutionary algorithms that determine the chemical reaction network automatically. The benefits of automated systems are discussed and the reason it is chosen for this thesis.

2.2 Introduction

Every industrial chemical reactors is designed to produce one or more intended chemical products which can serve as raw materials for another chemical process or as the final desired output. The design of the chemical reactors are mainly determined by two factors (Davis & Davis, 2003).

a. Expected changes in the environmental conditions within the reactor
b. Chemical reactions that are expected to occur

The chemical reactors are expected to first be able to withstand the change in temperature, pressure and volume that is caused by the chemical reactions within it. It can then be further enhanced through additional installation such as heating elements to deal with endothermic reactions or cooling system for exothermic reactions.

In order to determine the final design of the chemical reactor, the knowledge of the chemical reactions will be crucial. The type of chemical reactions that are expected to occur within the reactor will determine the thermodynamics of the process which in turn decide the changes in the environmental conditions within the reactor. Kinetic analysis of the reactions that will be occurring and the chemical reaction pathways that form the chemical reaction network then must be done.

The first step towards understanding the underlying kinetics of the chemical reactions is to understand what a chemical reaction is.

2.3 Chemical Reaction Basics

The process of converting chemical reactants to desired chemical products is considered as a chemical reaction. This conversion involves the rearrangement of the structures of the chemical reactants at the molecular level to produce new distinct chemical structures. For instance given,

$$x_1 + x_2 \rightarrow x_3 \qquad \text{(Equation 2-1)}$$

In the above example, x_1 reacts with x_2 to produce x_3. x_1 and x_2 serve as the reactants in this reaction and are consumed in the process of the creation of x_3. The product of the reaction is x_3 and is produced through the reaction. Often, more than one chemical reaction can take place. Consider the following,

$$x_1 + x_2 \rightarrow x_3 \rightarrow x_4 \qquad \text{(Equation 2-2)}$$

In this example, the role of x_1 and x_2 remain unchanged. However, x_3 now serves as the reaction intermediate in the production of the product x_4 from the reactants, x_1 and x_2. As a reaction intermediate, x_3 may increase in quantity at first from the reaction of x_1 and x_2 and subsequently decreased as x_3 is consumed to produce the final product, x_4.

In both of the examples shown in Equation 2-1 and Equation 2-2, only forward reactions are presented. Forward reactions are chemical reaction that only proceeds in one direction, from reactants to products. However, it should be noted that all chemical reactions are inherently reversible and it is theoretically possible to obtain the reactants from the products through the reverse reaction. Revisiting the first example shown in Equation 2-1, if the reaction is reversible it can be written as,

$$x_1 + x_2 \rightleftharpoons x_3 \qquad \text{(Equation 2-3)}$$

or in two separate forward reactions (the nature of the automated system developed in this work will present reversible reactions in this way),

$$\begin{gathered} x_1 + x_2 \rightarrow x_3 \\ x_3 \rightarrow x_1 + x_2 \end{gathered} \qquad \text{(Equation 2-4)}$$

All chemical reactions, given time will achieve a state of chemical equilibrium at which point the forward reaction and the reverse reaction will be at the same reaction rate. The concentration level of the reactants and products at this stage will not change any further.

In practice, a chemical reaction where the chemical equilibrium lies strongly on the side of the products and have reactants that will react until only minute amount of them are left at the end of the reaction can be considered as an irreversible reaction. On the other hand, a reversible reaction is a reaction that will achieve chemical equilibrium and both reactants and products will be present in the system when the equilibrium is achieved.

2.4 Chemical Reaction Network (CRN)

The most basic reaction one can encounter is the elementary reaction where the reaction proceed from reactant to product in a single step without any intermediates or transition steps involved in between (McNaught & Wilkinson, 1997). Using the hydrogenation of dibromine (Davis & Davis, 2003) as an example,

$$\mathrm{H_2} + \mathrm{Br_2} \rightarrow 2\mathrm{HBr} \qquad \text{(Equation 2-5)}$$

This reaction may be simple but it does not proceed in a single action. It requires light or photon as initiator to first sever $\mathrm{Br_2}$ into two Br radicals which will then attack and replace the hydrogen atom in the hydrogen molecule. The replaced hydrogen atom will then be radicalised and will continue on the reaction. The chain reaction will stop when two radicals meet and recombine. In term of chemical equations, the reactions are normally presented as

$$\begin{gathered}
\mathrm{Br_2} \rightarrow \mathrm{Br}\cdot + \mathrm{Br}\cdot \\
\mathrm{Br}\cdot + \mathrm{H_2} \rightarrow \mathrm{HBr} + \mathrm{H}\cdot \\
\mathrm{H}\cdot + \mathrm{Br_2} \rightarrow \mathrm{HBr} + \mathrm{Br}\cdot \\
\mathrm{Br}\cdot + \mathrm{H}\cdot \rightarrow \mathrm{HBr} \\
\mathrm{Br}\cdot + \mathrm{Br}\cdot \rightarrow \mathrm{Br_2} \\
\mathrm{H}\cdot + \mathrm{H}\cdot \rightarrow \mathrm{H_2}
\end{gathered} \qquad \text{(Equation 2-6)}$$

Each of these reactions occur in a single step and can be deemed as elementary reaction. It is clear that even a simple reaction cannot be assumed to occur in one single step but through multiple steps with reaction intermediates and interdependent relationships between each of the reaction steps. This is one of the many examples of a CRN which can be defined as a network that connects all the initial reactants, reaction intermediates and products through elementary reactions (Crampin, Schnell, & McSharry, 2004).

The reactions within a CRN can occur simultaneously and they can be classified into different types. A series reaction is a set of consecutive reactions where the reactants will produce reaction intermediates which will then be used to produce the product. It is also possible for more than one transition steps before the product is produced.

$$x_1 + x_2 \rightarrow x_3 \rightarrow x_4 \rightarrow x_5 \qquad \text{(Equation 2-7)}$$

The example above shows a series reaction where x_5 is formed through x_4 which is in turn formed from x_3 which again is formed by the reactants x_1 and x_2. The reaction intermediates are generally short-lived and highly reactive and their detection can sometimes be difficult as they are sometimes consumed as soon as they are produced.

Another type of reaction that can be found within a CRN is the parallel reaction. In this type of reaction, two or more competing reactions will have at least one common reactant.

$$\begin{gathered} x_1 \rightarrow x_2 \\ x_1 \rightarrow x_3 \end{gathered} \qquad \text{(Equation 2-8)}$$

Here, x_1 produces x_2 and x_3 simultaneously albeit possibly at a different rate. The reaction kinetics of the reactions will determine which of the reaction is more favourable.

A more complicated scheme which involves both series and parallel reaction is also possible.

$$\begin{gathered} x_1 + x_2 \rightarrow x_3 \\ x_3 + x_2 \rightarrow x_4 \end{gathered} \qquad \text{(Equation 2-9)}$$

In this case, x_2 is the target of competition from both of the reactions and x_3 is used as reaction intermediate in the series reaction to produce x_4.

The last type is the independent reaction which are just reactions that are not dependent on the reactant or product of other reactions.

$$\begin{gathered} x_1 \rightarrow x_2 \\ x_3 + x_4 \rightarrow x_5 \end{gathered} \qquad \text{(Equation 2-10)}$$

Both of the reactions may occur at the same time but will not participate in each other's reaction.

2.5 Molecularity and order of reaction

Molecularity or the order of reaction of an elementary chemical reaction describes the number of molecules involved in the reaction. A chemical reaction that uses only one molecule is called unimolecular reaction or first order reaction. If two molecules are involved, it is called bimolecular reaction or second order reaction. Termolecular reaction or third order reaction is uncommon because the level of difficulty to achieve it. It requires three different molecules to collide at the same time, at sufficient kinetic energy and at the precise orientation.

2.6 Stoichiometry

Stoichiometry is quantitative balancing of number of participating molecules of reactants and products in a reaction through conservation of the number of atoms of chemical elements in the reaction. The obtained values are generally referred as the stoichiometric coefficients. As per convention, reactants' stoichiometric coefficients are affixed with negative sign while the products' stoichiometric coefficients are positive. For example, the chemical reaction,

$$x_1 + x_2 \rightarrow x_3 \qquad \text{(Equation 2-11)}$$

will have a stoichiometric coefficient of -1 for the chemical species x_1 and x_2 and 1 for chemical species x_3. This can also be written in a matrix form:

$$\begin{bmatrix} -1 & -1 & 1 \end{bmatrix} \qquad \text{(Equation 2-12)}$$

Where the columns in the matrix refers to the chemical species x_1, x_2 and x_3 from left to right. In the case of a CRN, the stoichiometric matrix of each of the reactions can be combined. For example, the CRN,

$$x_1 \rightarrow x_2$$
$$x_1 \rightarrow x_3 \qquad \text{(Equation 2-13)}$$
$$x_1 + x_2 \rightarrow x_3$$

will have the stoichiometric matrix:

$$\begin{bmatrix} -1 & 1 & 0 \\ -1 & 0 & 1 \\ -1 & -1 & 1 \end{bmatrix} \qquad \text{(Equation 2-14)}$$

The rows in this stoichiometric matrix refers to the reactions. First row describes the first reaction, the second row the second reaction and so on. This in general will gives the stoichiometric matrix, v the matrix dimensions of $N_r \times N_c$.

2.7 Reaction Kinetics

Reaction kinetics is also known as chemical kinetics is the study of speed of chemical reactions and how factors such as concentration, pressure and temperature affects a chemical reaction. One of the goals of the study is to derive mathematical model to explain the phenomenon that surround the chemical reaction.

At the heart of the reaction kinetics is the law of mass action (Guldberg & Waage, 1879). The law states that the rate of reaction of an elementary reaction is proportional to the product of the reactants' concentrations and each of the concentration is raised to the power of its stoichiometric coefficient in the reaction. Below is an example to better understand the law:

$$\alpha\, x_1 + \beta\, x_2 \overset{k}{\rightarrow} \gamma\, x_3 \qquad \text{(Equation 2-15)}$$

For this reaction, based on the law of mass action, the reaction rate of this reaction can be written as

$$r \propto [x_1]^{\alpha}[x_2]^{\beta}$$
$$r = k[x_1]^{\alpha}[x_2]^{\beta} \qquad \text{(Equation 2-16)}$$

The kinetic rate constant, k will determine the speed of which the reaction will occur. α, β and γ are to the stoichiometric coefficients of x_1, x_2 and x_3.

The reaction rate of chemical reactions based on law of mass action in a multiple reactions environment can be written into a more concise form,

$$r_j = k_j \prod_{i=1}^{N_c} [x_i]^{-v_{i,j}}, \forall\, v_{i,j} < 0 \qquad \text{(Equation 2-17)}$$

Wheres

r_j is the reaction rate of the j^{th} reaction

k_j is the kinetic rate constants of the j^{th} reaction

N_c is the total number of chemical species

$[x_i]$ is the molar concentration of the i^{th} chemical species

$v_{i,j}$ is the stoichiometric matrix of the i^{th} chemical species in the j^{th} reaction

The change of concentration of each of the participating chemical species based on the reaction rate of the chemical reactions in an isothermal homogenous constant density batch reactor can then be written as follows:

$$\frac{d[x_i]}{dt} = \sum_{j=1}^{N_r} v_{i,j}\, r_j \qquad \text{(Equation 2-18)}$$

Where

N_r is the total number of reactions in the system

2.8 Kinetic rate constants

A parameter that arises from the law of mass action is the kinetic rate constant, k. The constant is the coefficient of proportionality from the law and it quantifies the rate of reaction of the chemical reaction. Although it is termed a constant, it is only a constant for a reaction occurring in an isothermal condition and is actually dependent on temperature.

Svante Arrhenius proposed the Arrhenius equation that relates the temperature to the changes in the reaction rate constant. The Arrhenius equation is as shown below:

$$k = Ae^{-E_a/(RT)} \qquad \text{(Equation 2-19)}$$

Where A is the pre-exponential factor, E_a is the activation energy, R is the universal gas constant and T is the absolute temperature.

The pre-exponential factor, A is determined empirically relating temperature, T to the reaction rate constant, k. It is considered to be a measure of the frequency of successful collisions (collisions that will generate reaction given enough energy) between the reactants.

Activation energy, E_a is the minimum energy required for the chemical reaction to occur. The presence of the right reaction catalyst can help in reducing the value by providing an alternative transition state or formation of reaction intermediate that requires a lower energy to form. The figure below will provide a better understanding on the effect of catalyst on activation energy,

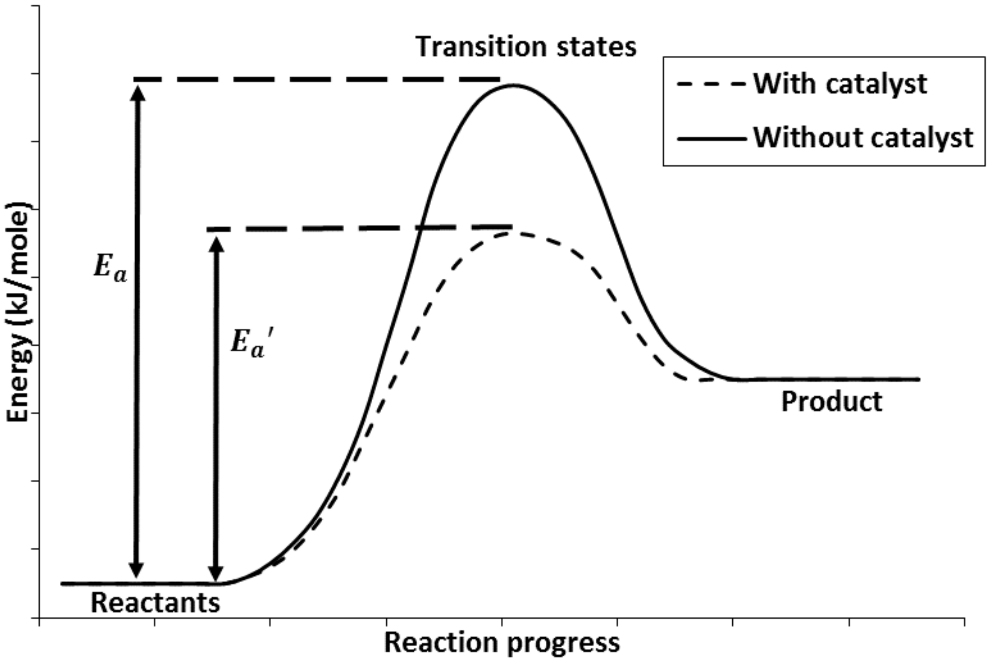


Figure 2.8-1 Effect of catalysts on activation energy

The activation energy in the presence of catalyst, E_a' is lower than that of the original activation energy, E_a. A lower activation energy means that more reactants can reach the transition state in order to progress the reaction towards the production of the product. In other short, a lower activation energy will increase the reaction rate and this can be easily be seen through the Arrhenius equation shown in Equation 2-18. The frequency of collision between reactants is also affected by the temperature, T. The higher the temperature, the more kinetic energy each molecules possesses which leads to more collisions that surpasses the required activation energy, E_a for chemical reaction to occur. The relationship between the molecules speed and temperature can be viewed through the Maxwell-Boltzmann distribution.

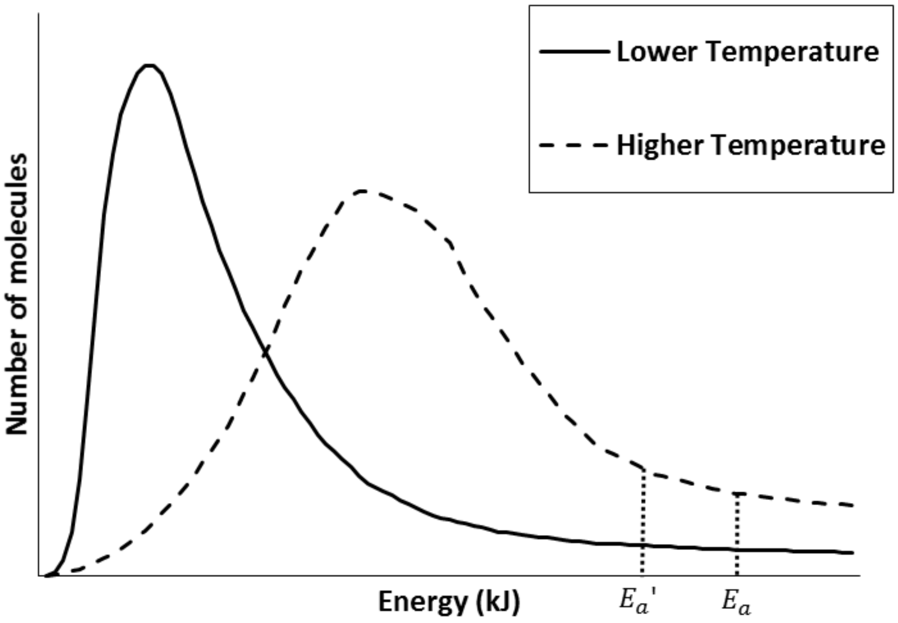


Figure 2.8-2 Maxwell-Boltzmann distribution plot

With a higher temperature, the Maxwell-Boltzmann distribution skewed more to the right, showing more molecules possess higher kinetic energy and thus more of them will be able to reach the activation energy, E_a compared to those at lower temperature. Reducing the activation energy using catalyst also shift the required energy for reaction, $E_a{}'$ to the left. This increases the number of molecules that are able to achieve the required energy and this applies to both higher and lower temperature molecules.

2.9 Batch reactor material balance

The change of concentration of a chemical species involved in the reaction can be derived through molar balance of the chemical species. Below is an example of molar balance done for chemical species x_i.

$$\begin{array}{ccccccc} \text{Accumulation of } x_i & = & x_i \text{ entering reactor} & - & x_i \text{ exiting reactor} & + & \text{Production/ consumption of } x_i \end{array}$$

$$\frac{1}{dt}\int [x_i]\, dV \quad = \quad F_{in}\left[x_{i,in}\right] \quad - \quad F_{out}\left[x_{i,out}\right] \quad + \quad \int r_{x_i}\, dV$$

Where V is the volume of the reactor, F_{in} is volumetric inflow into the reactor, F_{out} is the volumetric outflow from the reactor, $\left[x_{i,in}\right]$ is the concentration of x_i in the inflow and $\left[x_{i,out}\right]$ is the concentration of x_i in the outflow and r_{x_i} is the production or consumption rate of x_i due to chemical reaction within the reactor.

For a batch reactor, there is no inflow or outflow of material. Therefore, $F_{in} = F_{out} = 0$. The equation is then reduced to

$$\frac{1}{dt}\int [x_i]\, dV = \int r_{x_i}\, dV \qquad \text{(Equation 2-20)}$$

For a well-mixed constant density batch reactor, $\int dV = V$. The equation is then simplified to

$$V\frac{d[x_i]}{dt} = Vr_{x_i}$$
$$\frac{d[x_i]}{dt} = r_{x_i} \qquad \text{(Equation 2-21)}$$

The final form of the equation basically shows that for a well-mixed constant density batch reactor any changes in $[x_i]$ is entirely due to the reaction that occurs within the reactor.

2.10 Determination of reaction rate constants

2.10.1 *Integral method*

Consider a first order reaction:

$$x_1 \overset{k_1}{\rightarrow} x_2 \qquad \text{(Equation 2-22)}$$

Based on the law of mass action, the reaction rate, r_1 can be determined as:

$$r_1 = k_1[x_1] \qquad \text{(Equation 2-23)}$$

As a first order reaction where x_1 is the sole reactant, the stoichiometric coefficient for x_1 will be -1. It then follows that the rate of concentration change in x_1 is:

$$\frac{d[x_1]}{dt} = r_{x_1} = -(1)r_1$$
$$\frac{d[x_1]}{dt} = -\,k_1[x_1] \qquad \text{(Equation 2-24)}$$

Integrating both sides of the equation gives:

$$-\int_{[x_{1,0}]}^{[x_1]} \frac{1}{[x_1]}d[x_1] = \int_0^t k_1 dt$$
$$\ln\left([x_{1,0}]\right) - \ln([x_1]) = k_1 t \qquad \text{(Equation 2-25)}$$

Plotting $\ln\left([x_{1,0}]\right) - \ln([x_1])$ against t will gives a linear graph. The gradient of the graph will be the reaction rate constant, k_1.

This is only a simple demonstration on how integral method can be used to determine the reaction rate constant. The difficulty of the method increases when more chemical species is involved in the reaction. Consider the following 2nd order reaction:

$$x_1 + x_2 \overset{k_2}{\rightarrow} x_3 \qquad \text{(Equation 2-26)}$$

Working through the integral method, one will arrive at:

$$-\int_{[x_{1,0}]}^{[x_1]} \frac{1}{[x_1][x_2]}d[x_1] = \int_0^t k_2 dt \qquad \text{(Equation 2-27)}$$

The presence of x_2 will complicate the solution. The method of solution (Levenspiel, 1999) includes introducing an additional variable, X_{x_1} which describes the fraction of x_1 converted and the use of partial fractions.

The method becomes highly complex when applied to a CRN. Consider the CRN that includes both of the reactions above:

$$x_1 \overset{k_1}{\rightarrow} x_2$$
$$x_1 + x_2 \overset{k_2}{\rightarrow} x_3 \qquad \text{(Equation 2-28)}$$

The rate of concentration change for x_1 and x_2 becomes

$$\frac{d[x_1]}{dt} = -\,k_1[x_1] - k_2[x_1][x_2]$$
$$\frac{d[x_2]}{dt} = \ k_1[x_1] - k_2[x_1][x_2] \qquad \text{(Equation 2-29)}$$

Additional work will need to decouple k_1 and k_2 from the equations before the integral method can proceed.

2.10.2 *Differential method*

An alternative method to the integral method is to evaluate the rate of change of concentration of chemical species directly. The rate of concentration change can be obtained through the tangent of the graph of the chemical species' concentration against time. For example, for the reaction

$$x_1 \overset{k_1}{\rightarrow} x_2 \qquad \text{(Equation 2-30)}$$

The graph of concentration of x_1 against time can then be plotted:

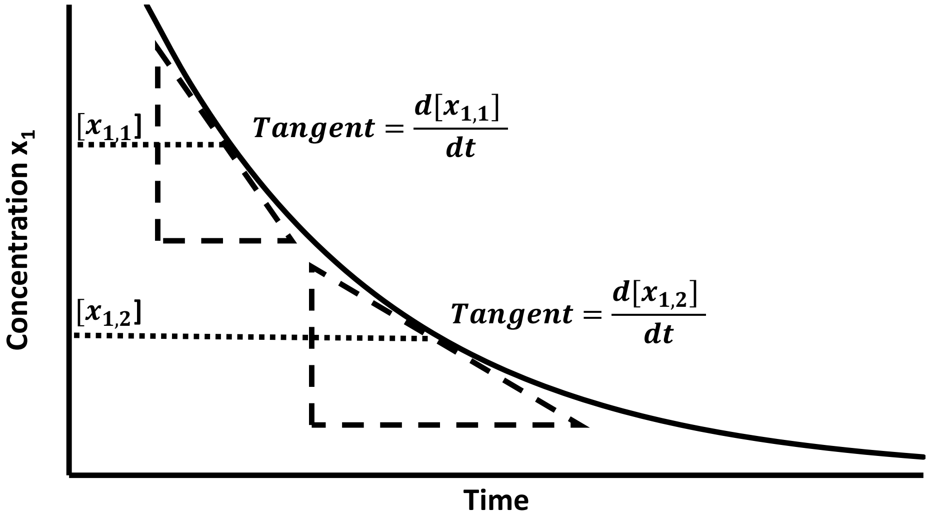

Concentration x$_1$
$[x_{1,1}]$
$Tangent = \frac{d[x_{1,1}]}{dt}$
$[x_{1,2}]$
$Tangent = \frac{d[x_{1,2}]}{dt}$
Time

Figure 2.10-1 Differential method to solve for reaction rate constants

The tangent at the point $[x_{1,1}]$ and $[x_{1,2}]$ is evaluated and then the obtained values can be substituted into the rate of concentration change of x_1 as derived in the integral method section.

$$\frac{d[x_1]}{dt} = -\,k_1[x_1] \qquad \text{(Equation 2-31)}$$

The rate constant, k_1 can then be calculated from the equations. Getting more values of k_1 and then taking the average is definitely more beneficial in reducing the effect of noise in the system.

In the case where the order of reaction with respect to x_1 is unknown, natural logarithm can be applied to the equation:

$$\frac{d[x_1]}{dt} = -\,k_1[x_1]^{\alpha}$$

$$\ln\left(-\frac{d[x_1]}{dt}\right) = \alpha \ln[x_1] + \ln k_1 \qquad \text{(Equation 2-32)}$$

Where α is the unknown reaction order. Using the tangent values obtained on the previous step, the graph $\ln\left(-\frac{d[x_1]}{dt}\right)$ can be plotted against $\ln[x_1]$.

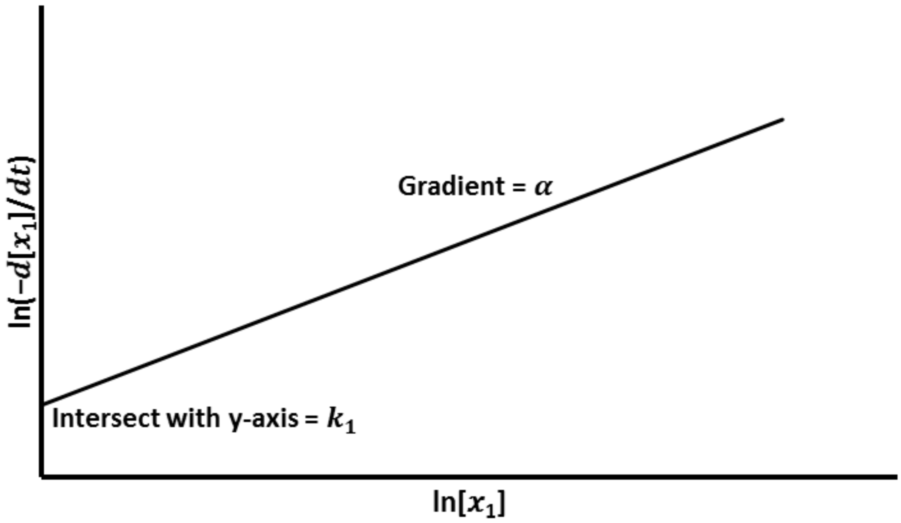


Figure 2.10-2 Plot of $ln\left(-\frac{d[x_1]}{dt}\right)$ *against* $ln[x_1]$.

With this method, k_1 can be determined at the same time as the order of reaction, α. It has to be noted that this method is limited to reactions that only uses one chemical species as reactant and cannot be used for complex CRNs.

Although the graphical method cannot handle more complex systems, the application of natural logarithm to the equation will be further discussed in a later chapter to calculate reaction rate constants for more complex CRNs.

2.11 Chemical Reaction Network Elucidation Introduction

Traditionally, trial-and-error methods are used to identify the chemical reaction networks. This is done by first hypothesising the possible reactions and testing out each of the hypothesised reactions one by one. The outcomes are then analysed and the chemical reaction network can be built based on the result of the analysis (Lin, 2004). This can become a very tedious task when the amount of involved chemical species is a huge number and the possibilities become numerous. A high level of expertise will also be required and specialised equipment may also be needed in order to conduct these reactions test. Needless to say, this will also require a huge amount of time and is not desirable.

Thus, a faster method to elucidate the chemical reaction network is needed in order to expedite the scaling-up process. Data driven techniques can be used to analyse the time-series concentration data of involved chemical species in batch reactors. If the structures of the chemical reaction network can be reasonably postulated, it becomes a matter of solving a set of ordinary differential equations in order to obtain the chemical rate constants or kinetics. The accuracy of the chemical reaction network can then be tested using the obtained chemical rate constants.

2.12 Inference Model

One of the methods to obtain a model of chemical reaction network is to infer it rather than trying to model it deterministically. This will usually result in a model that are able to fit the experimental data well but will have no meaning in its expressions. S-systems is one of the more popular method that is used to infer chemical reaction network and it originated from Savageau (1976). The S-systems used in chemical reaction network modeling formulate the ordinary differential equations almost similar to the one derived in equation 2.18. The difference that S-systems has is that the power term used are not restricted to integers like the law of mass action. The terms used can also differ widely as it can be a product of concentration of any chemical species (to the power of a non-integer value) and not restricted to only the reactants. The parameters in the S-systems, coefficient of each term (loosely relate to the reation rate constant in equation 2.17) and the power value of each concentration term are obtained by trying to fit the model to the experimental data. The effectiveness of this method in producing a good chemical reaction network model has been shown by Kikuchi et al. (2003), Voit and Almeida (2004), Searson et al. (2007).

Another technique that attempt to describe chemical reaction network is the tendency model by Filippi et al. (1986). The technique is a set of algorithm that consists of a set of rules and procedures to build the model step by step. It starts with assuming the model has only one reaction and fit it into equation 2.17. The stoichiometric coefficients are then obtained through minimisation of error between the model and the experimental rate of concentration change. The obtained coefficients are then rounded to the nearest integers or simple fraction and the error is calculated again. If the error is not satisfactorily low (depends on the required level of accuracy of the model), the steps are repeated again with addition of another reaction and so on. However, this method of modeling will only infer the reactions and will not provide the actual stoichiometric network (Le Lann et al., 1999). When the aim is not to elucidate the actual reaction network, tendency model has been shown to provide a good inference of the reactions as shown by work done by Rastogi et al. (1990 & 1992), Fotopoulou et al. (1994a) and Le Lann et al. (1999).

Both the S-system and the tendency model have the ability to infer the reactions in the chemical reaction network in order to provide a good model to predict the concentration profile of the chemical species without a priori information. Their common weakness is that the model will not have any significant information on the actual chemical reaction network and the models will only be accurate within the operating conditions of the experimental data that are used to produce them. A deterministic model will be able to model the reactions better especially in the scaling up of process when the operating conditions may differ from that of the laboratory. It can also help in obtaining an optimal operating condition (Maria, 2004).

2.13 Deterministic Model

One of the earliest methods to discover the actual chemical reaction network deterministically was proposed by Bonvin and Rippin (1990) which employs mathematical technique term as target testing or target factor analysis (TFA). The technique requires first to decompose the measured experimental data using singular value decomposition and from it obtained the number of independent reactions. With the number of independent reactions known, a postulated chemical reaction network is built and then tested using TFA. TFA will be able to test out the reactions within the chemical reaction network to determine whether they can be reasonably accepted or rejected. This method was tested on simulated models and had been shown successful (Bonvin & Rippin, 1990). It also provides a good starting point for other works that deals with tendency model such as works done by Rastogi et al. (1990 & 1992) and incremental chemical reaction network modelling such as work by Brendal et al. (2006).

However, the TFA technique requires a chemical reaction network to first be postulated before it can be used remains a problem. The possibilities can be numerous especially dealing with a system with large number of chemical species. A priori information will be required in order to lower down the number of these possibilities (Fotopoulos et al., 1994a). A step by step method in order to reduce the possibility mathematically was proposed in the form of structured target factor analysis (Fotopoulos et al., 1994b). The method is to systematically test out all possible reactions set by set. Each set of reactions consist different number of reactants and products for example Set I consist of 1 reactant 1 product, Set II consist of 2 reactants and 1 product and so on. The sets are evaluated and reactions gets eliminated if they did not satisfy the chosen criterion. This method however, can get very cumbersome as well when dealing with large number of chemical species and may eliminate correct reactions as it evaluate the reactions by the set and not globally.

Burnham (2007) demonstrated an incremental method using systematic mathematical and statistical analysis of experimental data. The method starts by listing all possible reaction combinations from the list of involved chemical species while ensuring they are mass-balanced. The reaction combinations will provide the reaction terms for the rate of change for each of the chemical species. Each of the reaction terms will have a constant (reaction rate constant) which can be determined through multiple linear regression using the experimental data. t-statistics and p-value are then calculated for each of the reaction terms and insignificant terms are removed. The process is repeated until the number of reaction had reduced to the estimated number of reactions obtained from singular value decomposition of the experimental data. The reaction network is then deduced from the final form of the reaction terms. The method is shown to work in Burnham (2007) work but suffered from becoming too unwieldy when the number of involved chemical species increases. Rationalisation will also be required as the statistical tests do not give absolute answer and consistency of the model with experimental data has to be checked at every step.

These techniques are generally cumbersome in nature and require a lot of work in order to eliminate possibilities. Statistical tests may help in the process but a lot of decisions and rationalisation will need to be taken in order to use the techniques. It also meant that the chemical reaction network that is built will determine highly on the expertise of the person who employ the technique. Using automated system will be able to remove the need for human intervention and even if the system is cumbersome, it is done automatically by computers.

2.14 Automated System

A global search method called differential evolution (Storn & Price, 1997) that is fully automated had been proposed by Searson et al. (2012) to reconstruct chemical reaction network from experimental concentration data. Differential evolution (DE) is an evolutionary algorithm which attempts to solve the problem iteratively by improving on its solutions from one generation to another. The algorithm begins by randomly building a set number of chemical reaction networks using integers to describe the stoichiometric matrix. The content of the networks (in this case individual stoichiometric coefficient) are then exchanged with each other randomly (mutation) in order to build a new chemical reaction network. If the new chemical reaction network is deemedbetter than the previous network, it is passed on the next generation else the original network will be passed on. After a set amount of generations determined at the start of the run, the algorithm will stop and the fittest chemical reaction network will be extracted. The method however has not been tested extensively and Searson et al. (2012) is only able achieve a considerably good result on simpler chemical reaction networks but struggled slightly with a more complex one. The paper by Searson et al. (2012) is also used as a comparison to test out the system that is proposed by this work.

Another global search method is by Koza et al. (2007) using genetic programming. Genetic programming (GP) is another branch of evolutionary algorithm that originated from Koza et al. (1999). The algorithm is similar to that of DE that it started with randomly building a set of chemical reaction networks. What is noticeably different here is that the GP builds the chemical reaction network as a tree with branches. Each branch of the tree will consist of a number of functions. These functions are pre-defined such as first order reaction with one product, second order reaction with one product and so on. The functions are then connected to chemical species which can serve as reactant or product. Every connected function will be considered as reaction and so if a tree consists of 3 functions, the tree represents a chemical reaction network with 3 reactions. The evolution of these ‘trees’ or networks are done through the exchange of these functions and chemical species to build new ‘trees’ or network. The network can also undergo mutation where any random function or chemical species within it can be change into something else randomly. Similar to DE, this process is repeated for a few generations and the best result is extracted at the end of the run. It has been shown to work for one example shown in Koza et al. (2007) but has not been tested extensively as well.

Both DE and GP are methods that have the benefit of operating without any prior information of the chemical reaction and require only the concentration data and molecular mass of the involved chemical species. They are also fully automated and the parameters needed to set the algorithm running is minimal. These two examples set as precursor of the decision to use evolutionary algorithm namely genetic algorithm for the basis of this work to produce a more complete automated system.

2.15 Summary

This chapter touches on the building blocks of chemical reaction networks which consists of combinations of multiple reactions. Different reaction types are discussed and the terminology for the topic is presented. The factors that can affect the reaction rate constant of a reaction are presented and methods of calculating reaction rate constants from available concentration data is shown.

The second part of the chapter discusses the advancement made in the field of elucidation of chemical reaction networks. Three methods are discussed namely, the inference modelling, deterministic modelling and the use of automated system through evolutionary algorithm. The next chapter will delve into the design of an automated system using genetic algorithm.

Chapter 3. Design of Automated System for Chemical Reaction Elucidation

3.1 Overview

This chapter starts with an introduction to genetic algorithm, explaining the different terminology used that goes with it. Each functions of the genetic algorithm are explained in detail and examples are given for easier understanding. This follows on to the design of the automated system for elucidation of chemical reaction network that will be used in this thesis which will be based on genetic algorithm. As part of the design of the automated system, a solution for reaction rate constants through the use of multiple linear regression is shown.

3.2 Introduction

A genetic algorithm (GA) is a global optimisation technique that utilises stochastic search method in order to find optimised parameters. It is first introduced by Holland (1975) as an algorithm that is inspired by nature. The algorithm is based loosely on the natural change in genetic material in individuals in a population through the generations (De Jong, 1988). De Jong (1988) explained the basic elements of GA as

a. Follows the Darwinian notion that the "fitness" of individuals of current generation will affect future generations. Generally, it means that the tough survives and the weak get weeded off.
b. Requires a "mating" process to create new population in the new generation using current individuals.
c. Each of the individuals is made up of "genes" that are used to describe each individual and these are the things that are passed down from one generation to another.

The GA neither requires the problem to be continuous nor differentiable making it very versatile and applicable to a lot of different type of problems. The following figure is a flowchart describing GA search process.

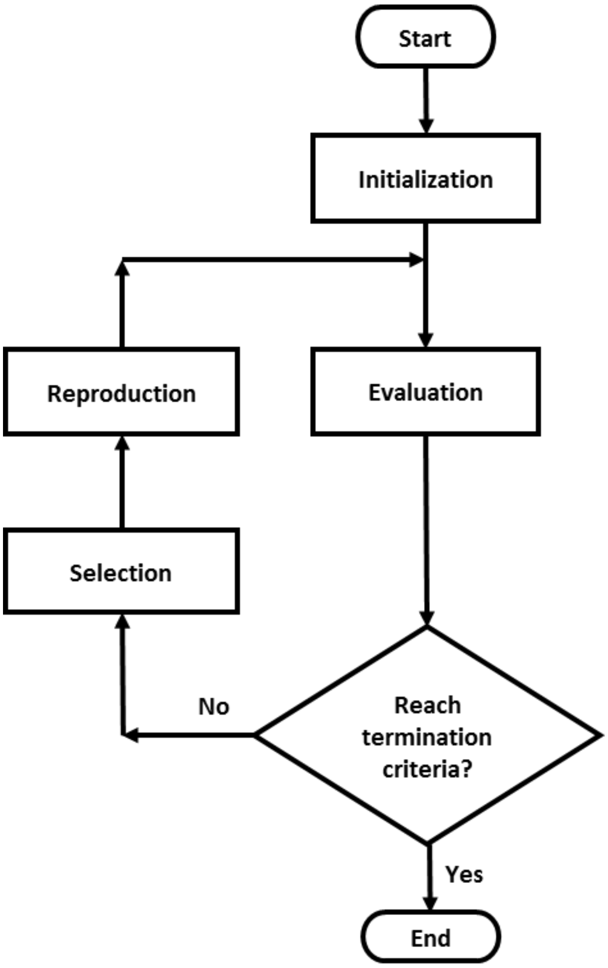


Figure 3.2-1 Flowchart for a basic genetic algorithm

3.3 Basic definition of terms

As mentioned, GA follows loosely on the selection process that exist in the biological world. The terms used in GA are also based on the biological process and their interpretation when used for GA are as follows:

a. Gene is the smallest possible unit in GA that when combined within a chromosome will define the chromosome. For example, for a binary encoding system, this can be 0 or 1.
b. Individual or chromosome contains a number of genes and the genes combinations will define the parameter/parameters the chromosome holds.
c. Population refers to all the individuals that belong to a single GA generation. The next generation of population will be formed through the reproduction process between the individuals within the same population.

d. Generation is the term used to describe the number of iterations that GA has undergone. The first generation refers to the initial population which will be used to reproduce the second generation and so on.

3.3.1 *Initialisation*

GA starts with initialisation of an initial population where each individual are encoded with. The classical method used by Holland (1975) is to encode using binary code strings. Examples of simple binary code string encoded individuals are shown below.

Individual 1	: 101010101010
Individual 2	: 010101010101
Individual 3	: 100100100100
Individual 4	: 111000111000

These can then be decoded as required by the variables sought for by the optimisation problem (Haupt & Haupt, 2004). For example, if the variable that is sought for contains 4 possibilities, then only 2 binary values are required to define it (00, 01, 10 and 11). Note that each of the binary value here is considered a gene. In the above example, there are 12 binary values for each individual and these can be separated as required. So if there are 6 variables and each of them has 4 possible results, the 12 binary values will be sufficient with every 2 values for each variable. The length of the individual and the required number of binary values for each variable can be change according to need.

The encoding of these individuals can affect the outcome of the GA significantly (Yin, Wei, & Meng, 2005) and the choice is dependent on the optimisation problem. Several examples of other type encoding are non-binary encoding (Gen, Cheng, & Wang, 1997), floating point number encoding (Budin, Golub, & Budin, 1996), variable string length (Goldberg, Korb, & Deb, 1989) and weight-coded GA (Raidl, 1999). The possibilities is numerous and it is up to the user to decide what is best to use for the particular problem.

3.3.2 *Evaluation*

This is the process where each of the individual genes content is translated into useful mathematical parameters and then evaluated against the problem's "fitness function". The fitness function can be based on the objective function of the optimisation problem. For an example, sums of squared error between the predicted values from GA and experimental values. The choice of fitness function used is crucial in determining the success of GA because it will determine whether GA will converge to a global solution, local solution or failure to achieve any result. This step is also the most computative intensive in GA and a fitness function that is simple to evaluate will help tremendously in reducing the amount of time required to run the GA.

3.3.3 *Selection*

Next, the individuals undergo the selection process which chooses the individuals which will participate in the next step, reproduction. In general, the “fitter” the individuals are (more desirable fitness function) the better chance they have in being selected by the selection process to be used for the reproduction step. A few examples of selection process that are the roulette-wheel selection, stochastic universal sampling, tournament system, truncation selection and elitism.

Roulette-wheel selection is done by first sorting the individuals according to their fitness functions. Each of sorted individuals’ accumulated fitness function are calculated which is done by adding up current individual and all the previous individuals’ (based on the order of the sorting) fitness functions. Then, adding up all the individuals’ fitness functions and generating a random number between 0 and the summed fitness functions value. The selected individual will be one that possess the accumulated fitness function right after the generated random number. The process is repeated until enough individuals are chosen.

A variation of the roulette-wheel selection is the stochastic universal sampling. The sorting of individuals, calculation of accumulated fitness function and obtaining random number are done the same way. The difference is that it divides the sorted population into equal intervals (based on number of required individuals that need to be selected) with the random number as the starting point. The individuals that lie right after the intervals will be selected.

A tournament system is done by randomly choosing a set number of individuals from the population. The individual with the most desired fitness function within the set will be selected. Repeat until the required number of individuals have been obtained.

Truncation system is simply the retention of a determined top percentages of the population. For example, if it chosen to have 50% of the population truncated, only the top 50% of the population will be selected for the next stage. If more individuals are required, duplicates will be made from the selected individuals.

Elitism helps to preserve a small percentages of the top performing individuals. This selection method helps to reduce the workload for GA because it will remove the need to rediscover these individuals in later generations. It can be used in conjunction of the other selection methods and the selected individuals bypasses the reproduction step.

3.3.4 *Reproduction*

The step where the next generation of population is created is called the reproduction step. The selected individuals from the selection process will participate in this step as parent individuals. These parent individuals will be used to produce child individuals which will be used to populate the next generation. The reproduction is generally done through two processes namely, crossover and mutation (Haupt & Haupt, 2004).

Crossover is done by randomly choosing a point within two parent individuals and having them exchange their genes. For example:

Parent Individual 1 : 1010101***01010***

Parent Individual 2 : ***0101010***10101

↓

Child Individual 1: 101010110101

Child Individual 2: ***010101001010***

A point is chosen in between the 7th and 8th binary variables for the parent individuals in the above example. The genes are then exchanged or crossover-ed between the parent individuals to produce two new child individuals.

Mutation is done by randomly selecting a point within a single parent individual and have it changed randomly or mutated (Haupt & Haupt, 2004). For example, Parent Individual 1 : 1***0***1010101010

↓

Child Individual 1 : 1***1***010101010

In this case, the second binary variable is chosen and gets mutated into a new value. For binary variable this is straightforward as it is either 0 or 1 so any mutation will have a definite outcome. It is also the choice of the user to have multiple point mutation or single point mutation as shown above. In other type of encodings, such floating point encoding, the result of the mutation will have to be randomly determined (Haupt & Haupt, 2004).

3.3.5 *Termination*

Once all the new child individuals for the new population are created, GA will proceed to the next generation. All the child individuals will be evaluated for their fitness just as their parents did and will participate in creation of future generations. This iteration will continue until the set number of maximum generations is completed or the GA gets terminated by other form of criterion such as reaching certain value of fitness (Hedar, Ong, & Fukushima, 2007). Other works such as done by Aytug & Koehler (1996) describes a theoretical bound on number of iterations to prevent overkill (Aytug & Koehler, 2000), stopping criterion based on the variance of fitness among the population (Tsoulos, 2008) and terminating if difference between best individual and worst individual reached a certain confidence value (Kaelo & Ali, 2007). Different termination criterion have different disadvantages (Hedar, Ong, & Fukushima, 2007) and there is no absolute choice in which to use. Once the GA run is terminated, the final result can be extracted from the GA’s final generation.

3.4 Application in Chemical Reaction Mechanisms

Genetic algorithm use in the area of chemical reaction mechanisms is not new. It has been used in obtaining optimised values of kinetic constants of known chemical reaction networks such as works done by Harris et al. (2000), Elliott et al. (2004) and Maeder et al. (2004). GA has also been used in reduction of highly complex reaction mechanisms to aid in reducing the amount of resources require to simulate it (Hernandez et al., 2010 and Perini et al., 2012). Keyvanloo et al. (2012) used GA to optimise the parameters of a polynomial model thermal cracking of naphtha. Cao et al.

(1999) built the model structure using genetic programming while Wang et al. (2007) employed fuzzy neural network and both optimised the parameters using GA to infer chemical mechanisms

There is a lack of work done to use GA to model the chemical reaction network itself and this is going to be investigated in this work where a novel GA is introduced to show its ability to elucidate the chemical reaction network.

3.5 Automated system design

GA is primarily a numerical optimisation method but it will be used as a modelling tool in this work. This is done mainly through a novel modification on the encoding method used on a classical GA and evaluation system that will be able to cope with the encoding system for it to be suitable for the elucidation of chemical reaction network.

For this encoding method, each genes will be used to represent a single chemical reaction. A set number of genes can then be grouped together to form a potential chemical reaction network or an individual in the custom GA. The population in the custom GA will be populated by these potential chemical reaction networks and will be used to evolve through the generations in the GA.

To do this, each of the genes will be represented by a set number of integers unlike classical GA where it is only represented by a single integer. These set of integers comes as a set and need to be materially balanced and cannot be changed partially and therefore becomes the smallest unit in the system, which is a gene. Figure 3.5-1 shows an example of three individuals (potential chemical reaction networks) in the population and each of the individual consists of five genes (reactions).

	Individual 1		**Individual 2**		**Individual 3**
Genotype 1	1 0 0 0 -1	Genotype 1	0 0 0 0 0	Genotype 1	-1 -1 0 1 0
Genotrype 2	-1 0 1 0 0	Genotrype 2	-2 0 0 1 0	Genotrype 2	0 0 0 1 -1
Genotype 3	2 0 0 -1 0	Genotype 3	-1 0 0 0 1	Genotype 3	1 0 -1 1 0
Genotype 4	0 0 0 0 0	Genotype 4	0 0 0 0 0	Genotype 4	1 0 -2 0 0
Genotype 5	0 -1 0 1 1	Genotype 5	0 -2 1 1 0	Genotype 5	0 1 0 0 -1

Figure 3.5-1: Example of individuals or chemical reaction networks in GA

Each gene shown in Figure 3.5-1 consisted of 5 integers and each of these integers represent the stoichiometric coefficient of a chemical species in the chemical reaction. For example, Gene 1 from Individual 1 is $[1 \quad 0 \quad 0 \quad 0 \quad -1]$ refers to the reaction:

$$x_5 \rightarrow x_1 \qquad \text{(Equation 3-1)}$$

Combining all the genes in Individual 1 will form the stoichiometric matrix:

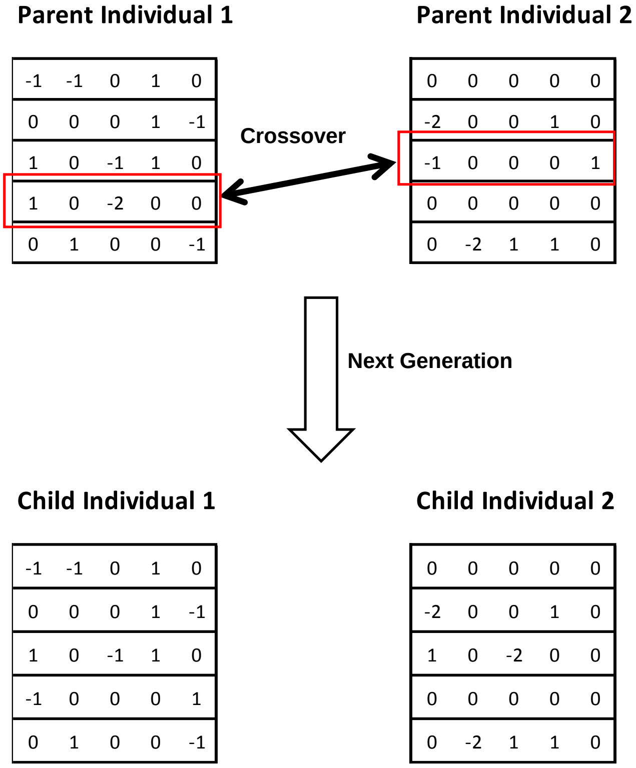
$$\begin{bmatrix} 1 & 0 & 0 & 0 & -1 \\ -1 & 0 & 1 & 0 & 0 \\ 2 & 0 & 0 & -1 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & -1 & 0 & 1 & 1 \end{bmatrix} \qquad \text{(Equation 3-2)}$$

which can be translated into the chemical reaction network:

$$\begin{gathered} x_5 \rightarrow x_1 \\ x_1 \rightarrow x_3 \\ x_4 \rightarrow 2x_1 \\ x_2 \rightarrow x_4 + x_5 \end{gathered} \qquad \text{(Equation 3-3)}$$

Note that, the gene $[0 \quad 0 \quad 0 \quad 0 \quad 0]$ is used to represent absence of any additional reactions in the chemical reaction network.

With this type of encoding, the crossover and mutation operations during the reproduction stage of the custom GA have to be modified from the classical GA's crossover and mutation. For crossover, rather than choosing a point where crossover between parent individuals will occur, the GA is programmed to choose a gene from each parent individuals to crossover to create two new child individuals. This exchanges chemical reactions from two potentially good chemical reaction network in the hope that the child individuals produced from the crossover operations will be stronger and better than their parents. Figure 3.5-2 gives a better picture of this.



Parent Individual 1

-1	-1	0	1	0
0	0	0	1	-1
1	0	-1	1	0
1	0	-2	0	0
0	1	0	0	-1

Parent Individual 2

0	0	0	0	0
-2	0	0	1	0
-1	0	0	0	1
0	0	0	0	0
0	-2	1	1	0

Child Individual 1

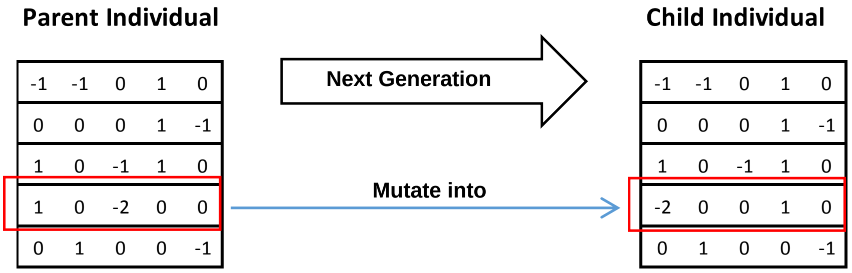
-1	-1	0	1	0
0	0	0	1	-1
1	0	-1	1	0
-1	0	0	0	1
0	1	0	0	-1

Child Individual 2

0	0	0	0	0
-2	0	0	1	0
1	0	-2	0	0
0	0	0	0	0
0	-2	1	1	0

Figure 3.5-2: Example of crossover in GA

For mutation, the custom GA will choose a gene randomly within the parent individual to mutate as compared to the classical GA where only a single integer will be mutated. The process will eliminate the chosen chemical reaction and randomly create a new chemical reaction within the child individual. This is shown in Figure 3.



Parent Individual

-1	-1	0	1	0
0	0	0	1	-1
1	0	-1	1	0
1	0	-2	0	0
0	1	0	0	-1

Child Individual

-1	-1	0	1	0
0	0	0	1	-1
1	0	-1	1	0
-2	0	0	1	0
0	1	0	0	-1

Figure 3.5-3: Example of mutation in GA

The encoding is done under a number of rules and these rules are applied not only during the initialisation step but also during the reproduction (crossover and mutation). The rules are as followed:

1. No two same reactions will exist within the same reaction network as they serve no purpose apart from complicating the estimation of the reaction rate constants.

2. All the chemical reactions within the chemical reaction network must be mass balanced.
3. The highest reaction degree is set as two. This is based on assumption that elementary reaction generally does not involve more than two molecules (Jackson, 2004) of reactants.
4. The reactions within the reaction network are checked against each other to avoid inconsistencies such as two reactions that use the exact same reactants producing different products.

Using these rules, the probability of generating infeasible chemical reaction network can be avoided and increases the utilisation of computing power in determining the correct chemical reaction network.

3.5.1 *Algorithm*

The automated system follows similar path as the classical GA but on the evaluation step, an additional step to estimate the reaction rate constants for each of the chemical reactions within the potential chemical reaction networks is required.

3.5.2 *Pre-run Parameters*

Before the algorithm can commence, a few pre-run parameters for the GA will need to be set up. Below are the few user defined parameters that need to be entered into the system:

a. Number of individuals per generation
b. Maximum number of generation
c. Maximum number of genes per individual
d. Mutation probability
e. Crossover probability
f. Elitism probability

Having more individuals per generation will enlarge the pool of available genes for use in the reproduction step. This typically help reduce the number of generation required to converge to a good result. However, too large a number of individuals per generation will bogged down the computing power and increases run time.

Running the GA for more generations may produce better results as the goal of the GA is to produce better and better individuals at every generation but at the risk of redundant generations because the results have already converged much earlier or faced the possibility of over-fitting to the experimental data.

Having more genes per individual is also another good way to increase the pool of available genes but at the same time also causes more computing power to be required.

Higher crossovers probability will help to recombine good parent individuals to build better child individuals but may be limited to the gene pool available in the parent individuals. This may lead to convergence to a local solution.

Mutation helps injecting new gene or re-introduce genes that are eliminated in previous generations into the gene pool and increases the diversity. It will lead to a creation of more diverse child individuals which can help in reaching a global solution but at the same time, it will increase runtime of the GA because a lot of poor individuals may be created.

Elitism will help preserve a small percentage of individuals that have the best performance to the next generation but too much elitism will cause convergence to a local solution.

These parameters are user defined and they vary case by case. In general, in a system with more chemical species involved will require a larger number of individuals, genes and generations. Crossover and mutation are mainly based on the experience of the user with the system. Elitism is best kept at a small percentage such as at 5%.

3.5.3 *Initialisation of the custom GA*

At the initialisation step, the initial population (first generation) is created. Each of the individuals will have their genes randomly created.

Each gene is created by first determining the order of reaction which can be first order, second order or no reaction. For first order reaction, a single reactant will be assigned randomly and for second order two reactants will be assigned. For no reaction, the gene will be filled with zeroes. Those assigned as reactants will have stoichiometric coefficient, $\upsilon_{x_i} < 0$. Next, the product of the chemical reaction is randomly determined as well but must not be the same chemical species as the reactant. The product will be assigned positive stoichiometric coefficient that is limited to not more than two for the scope of this work. The integers can easily be increased if need be, to any value.

Next, the reaction's mass balance will be checked to ensure no infeasible reaction is created. This is done by the following equality

$$\mathbf{v} \times \boldsymbol{M} = 0 \qquad \text{(Equation 3-4)}$$

Where

$\mathbf{v} = \begin{bmatrix} v_{x_1} & \cdots & v_{x_{N_c}} \end{bmatrix}$ and $\boldsymbol{M} = \begin{bmatrix} m_{x_1} \\ \vdots \\ m_{x_{N_c}} \end{bmatrix}$ for chemical reaction network that only has N_c chemical species.

v_{x_i} is the stoichiometric coefficient of the i^{th} chemical species

m_{x_i} is the molecular mass for the i^{th} chemical species

If the gene that is created fails to adhere to the mass balance, the gene will be re-created again. The GA currently is set up to repeat this up to 20 tries and a no reaction gene will be created instead if it cannot create a feasible reaction by then.

The process is repeated until all the genes in all the individuals in the population are filled.

3.5.4 *Reaction rate constants estimation*

The reaction rate constants for each of the individuals that have been created are estimated by solving the ordinary differential equations below:

$$\frac{d[x_i]}{dt} = [\dot{x_\imath}] = \sum_{j=1}^{N_r} v_{i,j}\, r_j \qquad \text{(Equation 3-5)}$$

The differential equations is only for a single data point. Matrix form can be used to describe the equation with multiple data points and multiple chemical species:

$$[\dot{\boldsymbol{X}}] = \boldsymbol{RV} \qquad \text{(Equation 3-6)}$$

Where

$$[\dot{\boldsymbol{X}}] = \begin{bmatrix} [\dot{x_1}]_{t=0} & \cdots & [\dot{x_{N_c}}]_{t=0} \\ \vdots & \ddots & \vdots \\ [\dot{x_1}]_{t=N_t} & \cdots & [\dot{x_{N_c}}]_{t=N_t} \end{bmatrix} \quad \boldsymbol{R} = \begin{bmatrix} r_{1,t=0} & \cdots & r_{N_r,t=0} \\ \vdots & \ddots & \vdots \\ r_{1,t=N_t} & \cdots & r_{N_r,t=N_t} \end{bmatrix} \quad \boldsymbol{V} = \begin{bmatrix} v_{1,1} & \cdots & v_{N_c,1} \\ \vdots & \ddots & \vdots \\ v_{1,N_r} & \cdots & v_{N_c,N_r} \end{bmatrix}$$

t is the data point

N_t is the last data point

N_r is the total number of reactions

With each of the reaction rate, r_j defined by:

$$r_j = k_j \prod_{i=1}^{n} [x_i]^{-v_{i,j}}, \forall\, v_{i,j} < 0 \qquad \text{(Equation 3-7)}$$

Applying natural logarithm on both sides of this equation will yield

$$\ln(r_j) = \ln(k_j) + \sum_{i=1}^{n} v_{i,j} \ln([x_i]), \forall\, v_{i,j} < 0 \qquad \text{(Equation 3-8)}$$

Substituting a_j to simplify the summation part:

$$\ln(a_j) = \sum_{i=1}^{n} v_{i,j} \ln([x_i]), \forall\, v_{i,j} < 0 \qquad \text{(Equation 3-9)}$$

The equation becomes:

$$\ln(r_j) = \ln(k_j) + \ln(a_j) \qquad \text{(Equation 3-10)}$$

Applying exponential on both sides of the equation gives:

$$r_j = k_j a_j \qquad \text{(Equation 3-11)}$$

Expanding it for all the reactions within the chemical reaction network and including the data points, it can be written in matrix form as:

$$\boldsymbol{R} = \boldsymbol{AK} \qquad \text{(Equation 3-12)}$$

where

$$\widetilde{\boldsymbol{V}} = -\boldsymbol{V}^T, \forall\, v_{i,j} < 0 \qquad [\boldsymbol{A}] = \begin{bmatrix} [a_1]_{t=0} & \cdots & [a_{N_R}]_{t=0} \\ \vdots & \ddots & \vdots \\ [a_1]_{t=N_t} & \cdots & [a_{N_R}]_{t=N_t} \end{bmatrix} \quad \boldsymbol{K} = \begin{bmatrix} k_1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & k_{N_R} \end{bmatrix}$$

$\boldsymbol{K}$ is a diagonal matrix with the reaction rate constants values on the diagonal position of the matrix.

Substituting $\boldsymbol{R}$ into the differential equations will give:

$$[\dot{\boldsymbol{X}}] = \boldsymbol{AKV}$$

$$[\dot{\boldsymbol{X}}]\boldsymbol{V}^{-1} = \boldsymbol{AK} \qquad \text{(Equation 3-13)}$$

At this point, it becomes a multiple linear regression problem and using ordinary least squares estimation:

$$\boldsymbol{K} = (\boldsymbol{A}^T\boldsymbol{A})^{-\boldsymbol{1}}\boldsymbol{A}^{\boldsymbol{T}}[\dot{\boldsymbol{X}}]\boldsymbol{V}^{-\boldsymbol{1}} \qquad \text{(Equation 3-14)}$$

3.5.5 *Fitness function*

The fitness function chosen for this step is variance weighted sums of squared error (VMSSE) which will be used to determine how closely related is the predicted concentration data with the experimental concentration data for each of the chemical species. The choice is mainly based on experience of using different type of fitness function and VMSSE managed to produce a more consistent result. This is because VMSSE gives equal priority to all the variables it is evaluating and therefore will not favour variables that have a larger value and ignore variables that are relatively small. In the context of the elucidation of the CRN which can consist of chemical species of large and small concentration, it will not only favour the chemical species with larger concentrations.

Variance weighted sums of squared error (VMSSE) can be described by the equation 3.15.

$$VMSSE = \sum_{i=1}^{N_c} \frac{\sum_{t=0}^{N_t}([x_i]_t - [\hat{x}_i]_t)^2}{\sum_{t=0}^{N_t}\left([x_i]_t - \mu_{x_i}\right)^2}, \forall\, u_{x_\text{i}} = 1 \qquad \text{(Equation 3-15)}$$

$[x_i]_t$ is the concentration data of chemical species x_i at time t

$[\hat{x}_i]_t$ is the predicted concentration data of chemical species x_i at time t

N_c = total number of participating chemical species

N_t = the time when the last data point is being evaluated

u_{x_i} is the measured/unmeasured identifier for $[x_i]$. Measured chemical species will be given the value of 1 and unmeasured will be given the value 0.

μ_{x_i} is the standard deviation of the all the concentration data of $[x_i]$ that is being evaluated

The predicted concentration data, $[\hat{x}_i]$ can be calculated by using the reaction rate constants, $\boldsymbol{K}$ that is obtained in the Equation 3-14 by solving ordinary differential equations for each corresponding data point, t. VMSSE for each of the chemical species is calculated and then their values added. As VMSSE is a measure of errors between the predicted values and the input value, the lower it is, the fitter the individual is. The fitter the individual is, the higher the chance of getting its genes passed down to subsequent generations.

3.5.6 *Reproduction process*

The creation of individuals for the next generation or child individuals is done in the next step. The process started off by determining whether the new child individual will be created through mutation or crossover of parent individuals. This is done by using random number generator and the probability of the choice is set beforehand within the GA.

If mutation is chosen, a single parent individual will be obtained from the selection process and a random gene within the individual will be 'mutated'. 'The new reaction is created the same way reactions are created during the initialisation step. If crossover is chosen, two parent individuals will be obtained through the selection process and one gene from each of the parent individuals will be exchanged to create two new child individuals.

The tournament selection process is used in this work. A set number of individuals are chosen from the present generation and the individual with the best fitness function (highest PPMCC) will be chosen as parent individual. This is to create a larger pool of genes to be passed on to the next generation and not restricted to only the very best individual while at the same time able to reject totally unfit individuals. The step ends once all the required number of child individuals has been created.

3.5.7 *Terminate GA*

Once the maximum number of generation has been reached, the GA will terminate itself. The extraction and analysis of results is done and if the GA is successful, the actual chemical reaction network will be presented as the best individual at the final generation.

3.6 Summary

In this chapter, the genetic algorithm is introduced and the function of each of the components within it are explained. This leads to the design of the automated system for the specific reason for elucidation of chemical reaction network. The system can be summarised as the flowchart show in the figure below,

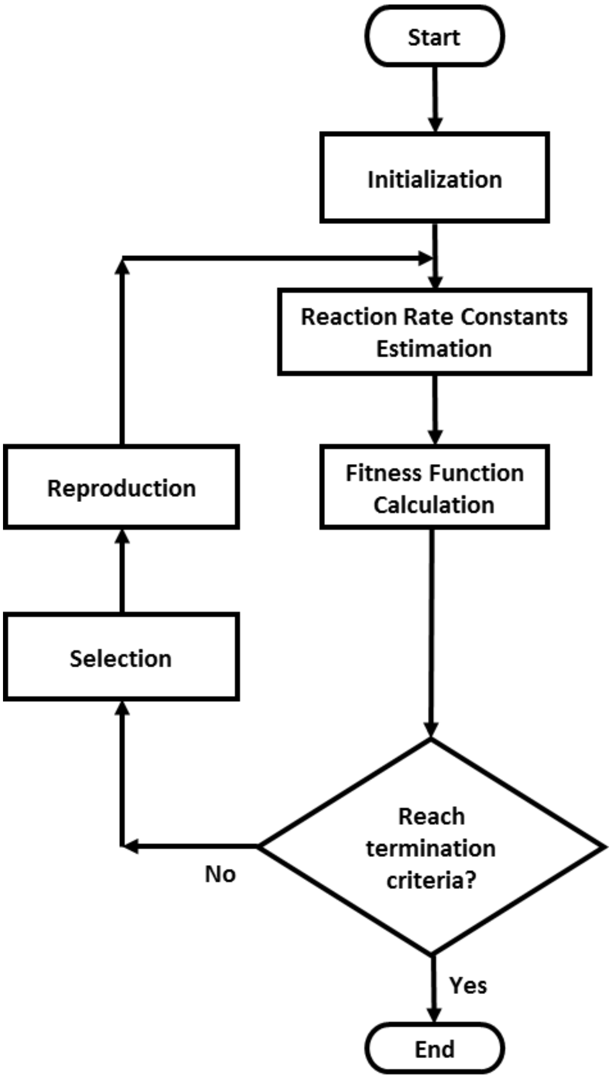

Start
Initialization
Reaction Rate Constants Estimation
Fitness Function Calculation
Reach termination criteria?
No
Selection
Reproduction
Yes
End

Figure 3.6-1 Flow chart for the automated system for chemical reaction network elucidation

The next chapter will test out the capability of this automated system and discuss its weaknesses.

Chapter 4. Application of the Automated System for Chemical Reaction Network Elucidation

4.1 Overview

This chapter introduces two fictitious Chemical Reaction Networks (CRN) to be used to test the automated system for chemical reaction network identification based on Genetic Algorithm (GA) as discussed in the previous chapter. Simulated data are generated for the CRNs and more datasets are created with added noise to increase the challenge of the test. The results of the automated system are presented and discussion on the performance of the automated system are made. Weaknesses of the system are identified and discussed and this follows on to a summary of the chapter.

4.2 Introduction

Two chemical reaction networks (CRN) are used to demonstrate the capability of the automated system in obtaining the actual CRN from the concentration of involved chemical species in an isothermal chemical batch reactor. The CRNs is based on those presented by Searson et al. (2012) and they will be referred as Reaction Network 1 (RN1) and Reaction Network 2 (RN2) accordingly from here onwards. Both of the CRNs are basic enough to test and develop the automated system but also sufficiently complex with the presence of serial reaction, parallel reaction and in RN2, a reversible reaction to test the robustness of the automated system.

The concentration data is generated through the solution of the ordinary differential equations using the Runge-Kutta 4th order method using the stoichiometric matrix, reaction rate constants and the initial concentration for each of the chemical species,

$$[\dot{\boldsymbol{X}}] = \boldsymbol{RV} \qquad \text{(Equation 4-1)}$$

The details of the RN1 and RN2 is as follows:

4.2.1 *Reaction Network 1 (RN1)*

Reaction 1: $2x_1 \overset{k_1}{\rightarrow} x_2$ $k_1 = 0.10\ dm^3mol^{-1}s^{-1}$

Reaction 2: $x_1 \overset{k_2}{\rightarrow} x_3$ $k_2 = 0.20\ s^{-1}$

Reaction 3: $x_3 \overset{k_3}{\rightarrow} x_4$ $k_3 = 0.13\ s^{-1}$

Reaction 4: $x_2 + x_4 \overset{k_4}{\rightarrow} x_5$ $k_4 = 0.30\ dm^3mol^{-1}s^{-1}$

In stoichiometric matrix,

$$\boldsymbol{V_{RN1}} = \begin{bmatrix} -2 & 1 & 0 & 0 & 0 \\ -1 & 0 & 1 & 0 & 0 \\ 0 & 0 & -1 & 1 & 0 \\ 0 & -1 & 0 & -1 & 1 \end{bmatrix} \qquad \text{(Equation 4-2)}$$

This CRN has 5 different chemical species and 4 reactions within it. All the reactions are forward reactions and there exists a single parallel reactions based on the reactant, x_1. 4 different batches of the CRN are simulated and the initial conditions and parameters used are shown in the tables below.

Run time = 0s to 24.0s

Sampling interval = 1.0s

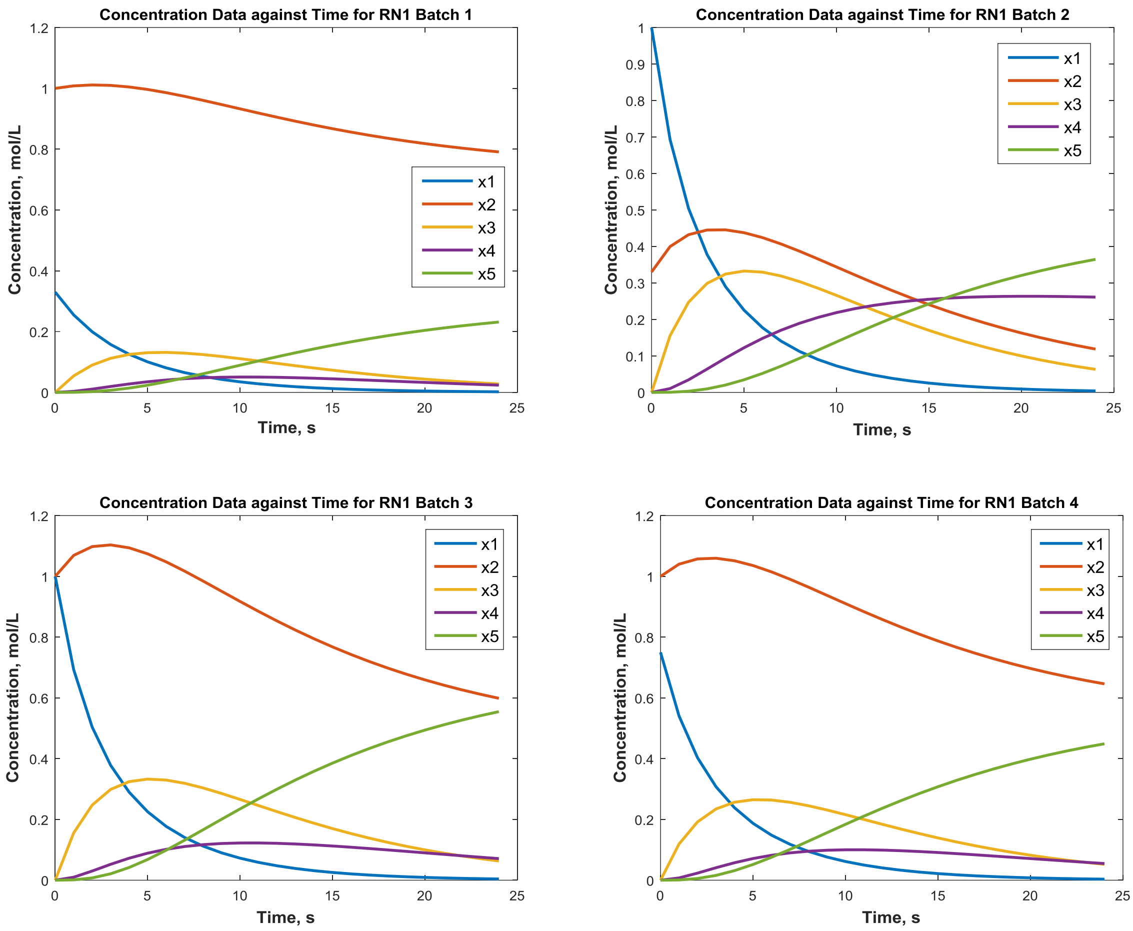
Batch	Initial Concentration, mol/dm^3				
	$[\boldsymbol{x_1}]_{\boldsymbol{t=0}}$	$[\boldsymbol{x_2}]_{\boldsymbol{t=0}}$	$[\boldsymbol{x_3}]_{\boldsymbol{t=0}}$	$[\boldsymbol{x_4}]_{\boldsymbol{t=0}}$	$[\boldsymbol{x_5}]_{\boldsymbol{t=0}}$
1	0.33	1.00	0	0	0
2	1.00	0.33	0	0	0
3	1.00	1.00	0	0	0
4	0.75	1.00	0	0	0

Table 4.2-1: Initial concentration data for Reaction Network 1.

Chemical Species	x_1	x_2	x_3	x_4	x_5
Molecular Weight	1	2	1	1	3

Table 4.2-2 Molecular weight of chemical species in Reaction Network 1

Figure 4.2-1 shows the concentration data of chemical species against time for four of the batches.



Concentration Data against Time for RN1 Batch 1

Concentration Data against Time for RN1 Batch 2

Concentration Data against Time for RN1 Batch 3

Concentration Data against Time for RN1 Batch 4

Figure 4.2-1 Concentration data against time for the four batches in Reaction Network 1

4.2.2 *Reaction Network 2 (RN2)*

Reaction 1: $x_1 + x_2 \xrightarrow{k_1} x_3 + x_4$ $\quad k_1 = 0.20\ dm^3 mol^{-1} s^{-1}$

Reaction 2: $x_2 + x_3 \xrightarrow{k_2} x_5$ $\quad k_2 = 0.10\ dm^3 mol^{-1} s^{-1}$

Reaction 3: $x_1 + x_4 \xrightarrow{k_3} x_6$ $\quad k_3 = 0.15\ dm^3 mol^{-1} s^{-1}$

Reaction 4: $x_6 \xrightarrow{k_4} x_1 + x_4$ $\quad k_4 = 0.05\ \ s^{-1}$

In stoichiometric matrix,

$$\boldsymbol{V_{RN2}} = \begin{bmatrix} -1 & -1 & 1 & 1 & 0 & 0 \\ 0 & -1 & -1 & 0 & 1 & 0 \\ -1 & 0 & 0 & -1 & 0 & 1 \\ 1 & 0 & 0 & 1 & 0 & -1 \end{bmatrix} \qquad \text{(Equation 4-3)}$$

RN2 has 6 chemical species and 4 reactions. This CRN will be used to understand the automated system’s behaviour on reversible reactions as can be seen in the 3rd and

4th reaction. A by-product x_5 is also produced in RN2. 4 different batches of the CRN are simulated and the initial conditions and parameters used are shown in the tables below.

Run time = 0s to 15.0s

Sampling interval = 0.5s

Batch	**Initial Concentration, mol/dm^3**					
	$[x_1]_{t=0}$	$[x_2]_{t=0}$	$[x_3]_{t=0}$	$[x_4]_{t=0}$	$[x_5]_{t=0}$	$[x_6]_{t=0}$
1	2.50	2.50	0	0	0	0
2	2.50	7.50	0	0	0	0
3	7.50	2.50	0	0	0	0
4	10.00	5.00	0	0	0	0

Table 4.2-3: Initial concentration for Reaction Network 2

Chemical Species	x_1	x_2	x_3	x_4	x_5	x_6
Molecular Weight	3	2	1	4	3	7

Table 4.2-4 Molecular weight of the chemical species in Reaction Network 2

Figure 4.2-2 shows the concentration data of chemical species against time for four of the batches.

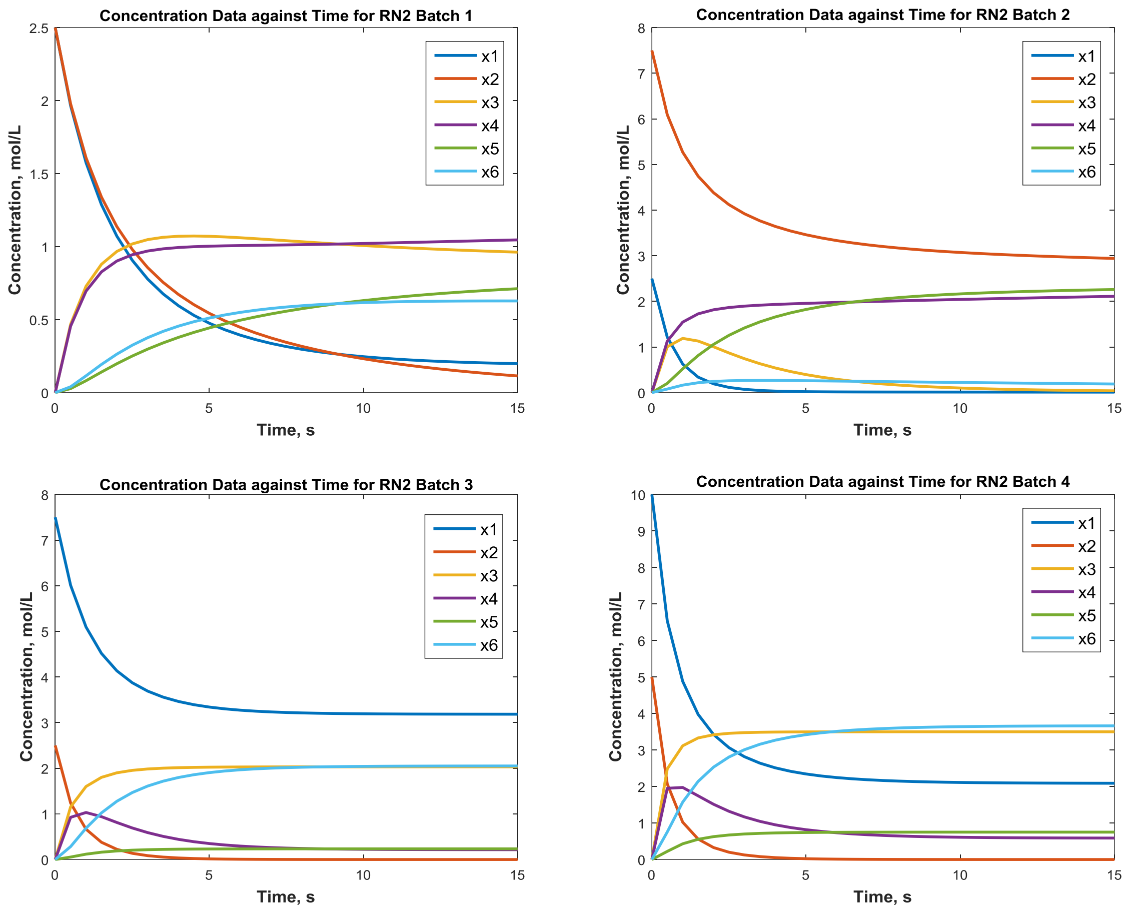


Figure 4.2-2 Plots of concentration data against time for the four batches in Reaction Network 2

4.3 Data processing

In order to use the automated system, the rate of change in concentration data, $[\dot{x_\iota}]$ of each of the chemical species will need to be obtained. To achieve this, the concentration is fitted to a rational polynomial with the suited order in the numerator and denominator through minimisation of the error between the modelled concentration profile from the rational polynomial and the simulated concentration data. An example of a rational polynomial with 2nd order numerator and 2nd order denominator:

$$[x_i]_t = \frac{\alpha_1 t^2 + \alpha_2 t + \alpha_3}{t^2 + \alpha_4 t + \alpha_5} \qquad \text{(Equation 4-4)}$$

Where $[x_i]_t$ is the concentration data of chemical species i at time t

α_1 to α_5 refer to parameters of the rational polynomial.

The rational polynomial can then be differentiated to obtain the rate of change of concentration data, $[\dot{x_i}]_t$ at time t.

4.4 Automated system parameters

The standard parameters used to run the automated system for the elucidation of RN1 and RN2 for the datasets generated above is shown in Table 4.4-1:

Number of individuals per generation	100
Maximum number of generations	50
Tournament size	10
Mutation probability	80%
Crossover probability	20%
Elitism	5% of total individual per generation

Table 4.4-1 Run parameters for the automated system for chemical reaction network elucidation

The selection of the parameters are based on the experience using the automated system in elucidating the CRNs. It is discovered that using a high level of crossover probability inhibits the performance of the automated system and causes the system to converge with higher number of generations and sometimes converge to a local minima. This is shown in Table 4.4-2 when the automated system is tested to elucidate RN1 with 100 population and 100 maximum generations. This can be hypothesised based on the nature of the problem which is highly non-linear. The presence and absence of a single reaction has a significant impact to the fitness function of the CRN. The effect of the presence of each reaction can be very different as well. Combination of certain reactions will also impact the final fitness value of the CRN.

For example in RN1, if the 3rd reaction is missing, even with the 4th reaction present, it will still be impossible to predict the concentration of x_5. This is the calculation of the concentration x_5 is highly dependent on the presence of x_4 which is generated in the 3rd reaction. For reference, the stoichiometric matrix of RN1 is shown below:

$$\boldsymbol{V_{RN1}} = \begin{bmatrix} -2 & 1 & 0 & 0 & 0 \\ -1 & 0 & 1 & 0 & 0 \\ 0 & 0 & -1 & 1 & 0 \\ 0 & -1 & 0 & -1 & 1 \end{bmatrix} \qquad \text{(Equation 4-5)}$$

This will caused the hoarding of poorly performing CRNs because a huge amount variability is needed such as introduction of two or more new reactions into a CRNs before the impact of the reactions can be felt. Crossing over poor performing CRNs will undoubtedly cause delay in convergence and may even missed out on crucial reaction which will ultimately cause the automated system to converge to a local minima.

Crossover probability	**Mutation probability**	**Converged at generation**
90%	10%	Did not reach global minima
70%	30%	37
50%	50%	26
20%	80%	18

Table 4.4-2 Impact of different crossover and mutation probability on convergence of GA.

Therefore, to help explore the highly non-linear search space of the problem, a high mutation rate is used to help produce reactions that may not be useful as a standalone but can help tremendously when in a group of correct reactions, such as the 3rd reaction as above. However, high mutation rate comes with an increased number of incorrect reactions being generated which may cause the next generation produced to perform worse than its predecessors. To control this effect, a small amount of elitism of 5% of the population is introduced into the system to retain the top few performing CRNs to bring forward to the next generation. This gives the system the freedom to choose large number of unique combinations of reactions while retaining the combinations that had shown potential to the next generation.

4.5 Practical implementation of Multiple Linear Regression

In the previous chapter, in the reaction rate constant calculation step, the final equation to obtain the reaction rate constant is as follows,

$$\boldsymbol{K} = (\boldsymbol{A}^T\boldsymbol{A})^{-\mathbf{1}}\boldsymbol{A}^{\boldsymbol{T}}[\dot{\boldsymbol{X}}]\boldsymbol{V}^{-\mathbf{1}} \qquad \text{(Equation 4-6)}$$

What can be noticed is that, it requires the inverse matrix of the stoichiometric matrix of the individual being evaluated, $\boldsymbol{V}$. If $\boldsymbol{V}$ cannot be inverted, the equation cannot be solved. This is likely to occur in RN1 and RN2 given that the number of reactions is less than the number of chemical species present in both RN1 and RN2.

For example the stoichiometric matrix of RN1, $\boldsymbol{V_{RN1}}$,

$$\boldsymbol{V_{RN1}} = \begin{bmatrix} -2 & 1 & 0 & 0 & 0 \\ -1 & 0 & 1 & 0 & 0 \\ 0 & 0 & -1 & 1 & 0 \\ 0 & -1 & 0 & -1 & 1 \end{bmatrix} \qquad \text{(Equation 4-7)}$$

cannot be inverted because it is not a square matrix. To overcome this, the automated system evaluate such matrix as

$$\boldsymbol{V_{RN1}} = \begin{bmatrix} -2 & 1 & 0 & 0 & 0 \\ -1 & 0 & 1 & 0 & 0 \\ 0 & 0 & -1 & 1 & 0 \\ 0 & -1 & 0 & -1 & 1 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix} \qquad \text{(Equation 4-8)}$$

With a redundant reaction as the 5th reaction. With it, $\boldsymbol{V_{RN1}}$ becomes a square matrix. However, the matrix rank is only 4 as there are only 4 linearly independent matrix which subsequently made the matrix singular. Singular matrix cannot be inversed. To overcome this, the inversion of the matrix in the reaction rate calculation is changed to pseudo-inverse.

$$\boldsymbol{K} = (\boldsymbol{A}^{T}\boldsymbol{A})^{-\boldsymbol{1}}\boldsymbol{A}^{\boldsymbol{T}}[\dot{\boldsymbol{X}}]\boldsymbol{V}^{\#} \qquad \text{(Equation 4-9)}$$

Where # refers to the Moore-Penrose pseudo-inverse

Using Equation 4-9, the reaction rate constants calculation becomes possible for stoichiometric matrix that are singular in nature.

4.6 Results and Discussion

All in all, the 4 reaction batches of RN1 and 4 reaction batches of RN2 are simulated and supplied to the automated system. The system is then run to elucidate the CRN of each of the reaction batch and the results of the each of the run is presented in the next section. The Table 4.6-1 shows the detail on the datasets used for each of the automated system’s runs.

Run	Chemical Reaction Network	Batch
4-1	RN1	1
4-2	RN1	2
4-3	RN1	3
4-4	RN1	4
4-5	RN2	1
4-6	RN2	2
4-7	RN2	3
4-8	RN2	4

Table 4.6-1 Run details for Run 4-1 to Run 4-8

4.6.1 *Reaction Network 1 (RN1)*

Table 4.6-2 shows the elucidated CRN from the automated system for each of the reaction batch for RN1. The fitness function used is variance weighted sums of squared error.

Run	Best Performing Individual	Reaction Rate Constant	Fitness Function
4-1	$\begin{bmatrix} -2 & 1 & 0 & 0 & 0 \\ -1 & 0 & 1 & 0 & 0 \\ 0 & 0 & -1 & 1 & 0 \\ 0 & -1 & 0 & -1 & 1 \end{bmatrix}$	$k_1 = 0.0964\ dm^3mol^{-1}s^{-1}$ $k_2 = 0.2016\ s^{-1}$ $k_3 = 0.1300\ s^{-1}$ $k_4 = 0.3005\ dm^3mol^{-1}s^{-1}$	0.0326
4-2	$\begin{bmatrix} -2 & 1 & 0 & 0 & 0 \\ -1 & 0 & 1 & 0 & 0 \\ 0 & 0 & -1 & 1 & 0 \\ 0 & -1 & 0 & -1 & 1 \end{bmatrix}$	$k_1 = 0.0987\ dm^3mol^{-1}s^{-1}$ $k_2 = 0.2016\ s^{-1}$ $k_3 = 0.1297\ s^{-1}$ $k_4 = 0.3002\ dm^3mol^{-1}s^{-1}$	0.0391
4-3	$\begin{bmatrix} -2 & 1 & 0 & 0 & 0 \\ -1 & 0 & 1 & 0 & 0 \\ 0 & 0 & -1 & 1 & 0 \\ 0 & -1 & 0 & -1 & 1 \end{bmatrix}$	$k_1 = 0.1000\ dm^3mol^{-1}s^{-1}$ $k_2 = 0.2010\ s^{-1}$ $k_3 = 0.1299\ s^{-1}$ $k_4 = 0.3013\ dm^3mol^{-1}s^{-1}$	0.0390
4-4	$\begin{bmatrix} -2 & 1 & 0 & 0 & 0 \\ -1 & 0 & 1 & 0 & 0 \\ 0 & 0 & -1 & 1 & 0 \\ 0 & -1 & 0 & -1 & 1 \end{bmatrix}$	$k_1 = 0.1001\ dm^3mol^{-1}s^{-1}$ $k_2 = 0.2003\ s^{-1}$ $k_3 = 0.1296\ s^{-1}$ $k_4 = 0.3009\ dm^3mol^{-1}s^{-1}$	0.0320

Table 4.6-2 Results for Run 4-1 to Run 4-4

From the result, it can easily be seen that the automated system successfully identify all of the reactions within RN1 for each of the batch. The reaction rate constants calculated are also very close to the actual values in RN1 with the biggest percentage error approximately 3.60% (Batch 1's k_1 $0.0964\, dm^3 mol^{-1} s^{-1}$ against RN1's k_1 $0.1000\ dm^3 mol^{-1} s^{-1}$).

$$\text{Percentage error} = \frac{0.1000 - 0.0964}{0.1000} \times 100\% = 3.60\%$$

The fitness function is also low as can be seen in the graph comparing the simulated and predicted concentration data for Run 4-2 which is the batch with the worst fitness, 0.0391.

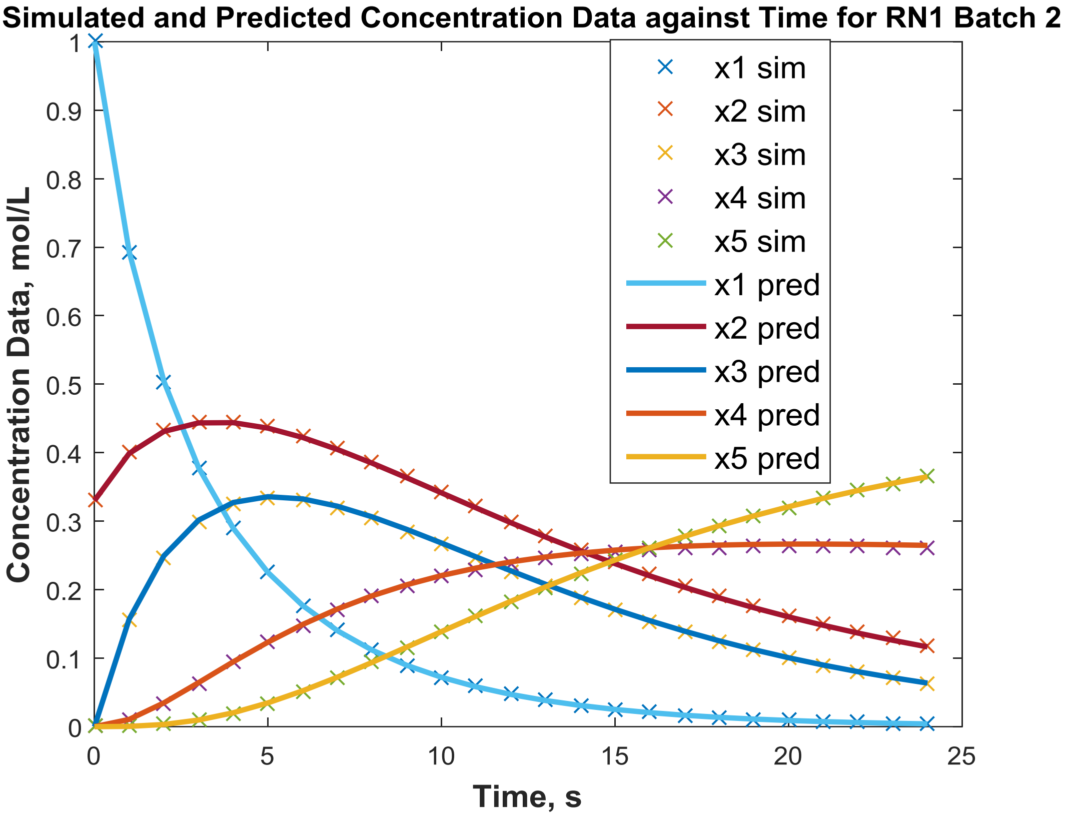


Figure 4.6-1 Simulated and predicted concentration data against for Run 4-2 (sim = simulated, pred = predicted)

From the figure, it can be surmised that the predicted concentration data matched the simulated concentration with no significant error.

However, given that the origin data has no error, it should be expected that the automated system should produce a CRN that match the origin data exactly with no error and the reaction rate constants should be exactly the same. This small error stems from the fact that the rate of concentration change of simulated data are approximated through differentiation of rational polynomials. The approximations introduced a slight error in the calculation for the reaction rate constants and thus resulted in the above results.

Overall, the method had shown effective for RN1.

4.6.2 *Reaction Network 2 (RN2)*

Table 4.6-3 shows the elucidated CRN from the automated system for each of the reaction batch for RN2. The fitness function used is variance weighted sums of squared error.

Run	Best Performing Individual	Reaction Rate Constant	Fitness Function
4-5	$\begin{bmatrix} -1 & -1 & 1 & 1 & 0 & 0 \\ 0 & -1 & -1 & 0 & 1 & 0 \\ -1 & 0 & 0 & -1 & 0 & 1 \\ -1 & 2 & 2 & 0 & -1 & 0 \\ 1 & 0 & 0 & 2 & 1 & -2 \\ 0 & 0 & 1 & -2 & 0 & 1 \end{bmatrix}$	$k_1 = 0.2018\ dm^3mol^{-1}s^{-1}$ $k_2 = 0.1043\ dm^3mol^{-1}s^{-1}$ $k_3 = 0.1179\ dm^3mol^{-1}s^{-1}$ $k_4 = -0.0035\ dm^3mol^{-1}s^{-1}$ $k_5 = 0.0077\ dm^3mol^{-1}s^{-1}$ $k_6 = -0.0005\ dm^3mol^{-1}s^{-1}$	6.1877
4-6	$\begin{bmatrix} -1 & -1 & 1 & 1 & 0 & 0 \\ 0 & -1 & -1 & 0 & 1 & 0 \\ -1 & 0 & 0 & -1 & 0 & 1 \\ 0 & 0 & 0 & -1 & -1 & 1 \\ 1 & 0 & 0 & 0 & -1 & 0 \\ 0 & 1 & -2 & 0 & 0 & 0 \end{bmatrix}$	$k_1 = 0.2051\ dm^3mol^{-1}s^{-1}$ $k_2 = 0.0990\ dm^3mol^{-1}s^{-1}$ $k_3 = 0.1130\ dm^3mol^{-1}s^{-1}$ $k_4 = -0.0002\ dm^3mol^{-1}s^{-1}$ $k_5 = -0.00004\ s^{-1}$ $k_6 = 0.0050\ dm^3mol^{-1}s^{-1}$	18.5476
4-7	$\begin{bmatrix} -1 & -1 & 1 & 1 & 0 & 0 \\ 0 & -1 & -1 & 0 & 1 & 0 \\ -1 & 0 & 0 & -1 & 0 & 1 \\ 0 & 1 & 1 & 1 & 0 & -1 \\ 1 & -1 & 2 & 0 & -1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}$	$k_1 = 0.1960\ dm^3mol^{-1}s^{-1}$ $k_2 = 0.1177\ dm^3mol^{-1}s^{-1}$ $k_3 = 0.1268\ dm^3mol^{-1}s^{-1}$ $k_4 = -0.0001\ dm^3mol^{-1}s^{-1}$ $k_5 = 0.1783\ dm^3mol^{-1}s^{-1}$ $k_6 = 0$	13.8507
4-8	$\begin{bmatrix} -1 & -1 & 1 & 1 & 0 & 0 \\ 0 & -1 & -1 & 0 & 1 & 0 \\ -1 & 0 & 0 & -1 & 0 & 1 \\ 1 & 1 & 1 & 0 & -2 & 0 \\ 0 & 2 & -1 & 1 & 0 & -1 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}$	$k_1 = 0.2018\ dm^3mol^{-1}s^{-1}$ $k_2 = 0.1043\ dm^3mol^{-1}s^{-1}$ $k_3 = 0.1179\ dm^3mol^{-1}s^{-1}$ $k_4 = 0.0004\ dm^3mol^{-1}s^{-1}$ $k_5 = -0.00008\ dm^3mol^{-1}s^{-1}$ $k_6 = 0$	25.6179

Table 4.6-3 Results for Run 4-5 to Run 4-8

Different from the automated system performance in identification of reactions in RN1, it does not seem to be as effective at identification of reactions for RN2. The fitness functions are significantly larger when compared against those in RN1 which meant they have more errors when comparing simulated and predicted concentration data.

It can be seen however that the first three reactions of RN2 are identified correctly. The 4th reaction is not identified in any of them. The reaction rate constants approximated for Reaction 1 and Reaction 2 are good with the worst percentage error of 17.7% (Run 4-7's k_2).

Apart from not identifying the 4th reaction of RN2, the results also shows the automated system producing reactions that are not part of RN2. These reactions sometimes come with a negative reaction rate constants which has no physical meaning as it will suggest reactants are produced in a reaction that is supposed to expend them, for example, Run 4-5's $k_4 = -0.0035\ dm^3 mol^{-1} s^{-1}$.

The prediction accuracy is poorer than RN1 as can be seen in Figure 4.6-2 which depicts the performance of the automated system for Run4-8:

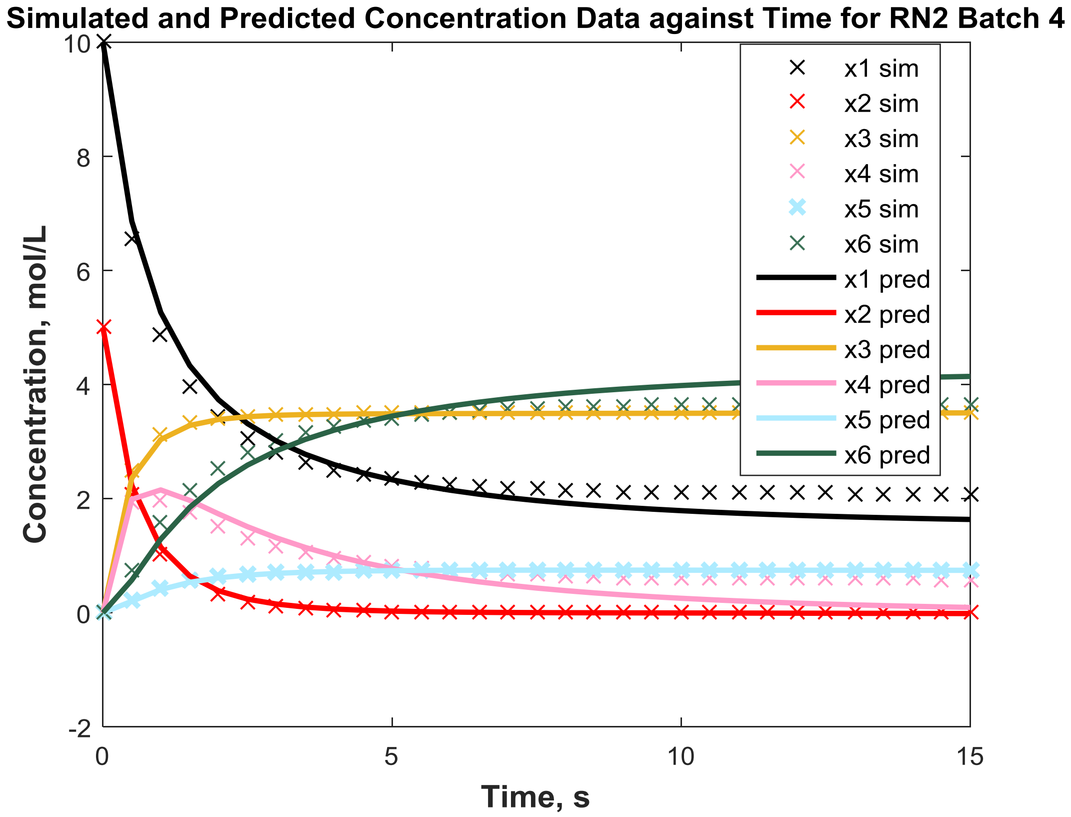


Figure 4.6-2 Simulated and predicted concentration data against time for Run 4-8 (sim = simulated, pred = predicted)

From the figure, it can be seen that x_2, x_3 and x_5 simulated values are fitted well to the predicted concentration data but the rest of the chemical species have poorer fit.

Although the system has partial success in the identification of the CRN for RN2, its performance is much poorer as compared to when it is used for RN1. The source of the automated system's weakness can be traced to the stoichiometric matrix of RN2.

$$\boldsymbol{V_{RN2}} = \begin{bmatrix} -1 & -1 & 1 & 1 & 0 & 0 \\ 0 & -1 & -1 & 0 & 1 & 0 \\ -1 & 0 & 0 & -1 & 0 & 1 \\ 1 & 0 & 0 & 1 & 0 & -1 \end{bmatrix} \qquad \text{(Equation 4-10)}$$

The 3rd reaction and 4th reaction are reverse reactions of each other and this made the 3rd and 4th row of $\boldsymbol{V_{RN2}}$ not linearly independent to each other. Which in turn made the solution for 4 reaction rate constants based on 3 linearly independent equations impossible through multiple linear regression as used in the reaction rate constant calculation.

$$\boldsymbol{K} = (\boldsymbol{A}^T\boldsymbol{A})^{-\mathbf{1}}\boldsymbol{A}^{\boldsymbol{T}}[\dot{\boldsymbol{X}}]\boldsymbol{V}^{\#} \qquad \text{(Equation 4-11)}$$

Negative reaction rate constants are also the result of using this particular reaction rate constant calculation method. The formula does not limit itself to positive values only and it is possible for it to produce negative values for the reaction rate constants it calculates. This occurs more often than not in CRN such as RN2 where the formula is unable to obtain accurate reaction rate constants from the actual CRN.

Even with such inherent weaknesses, the automated system is still able to produce predicted concentration data that although not entirely accurate but at least not too far off from the simulated concentration data. It also still able to identify 3 of the 4 reactions from RN2 and 2 of the reaction has good approximation on the reaction rate constants. However, the limitation of the automated system in evaluating reversible reactions made it unreliable for RN2 and at such will not be used to evaluate RN2 in the next section where the effect of noise is explored on the system's robustness.

4.7 Noise

To test on the robustness of the automated system, the concentration data for the 4 batches for RN1 are perturbed with Gaussian noise with the mean of 0. Two levels of noise with standard deviation equal to the maximum of 4% and 8% of the maximum value of the chemical species' concentration are introduced to the concentration data. For example, a chemical species in a batch run has a maximum value of 1.0 will have Gaussian noise with mean of 0 and standard deviation of 0.04 and 0.08 added accordingly for both of the noise level. RN2 is not used for this section because it has been shown the automated system is not effective in elucidating its CRN.

Table 4.7-1shows the datasets used in this section:

Run	Chemical Reaction Network	Batch	Gaussian Noise Standard Deviation
4-9	RN1	1	4% of max range
4-10	RN1	1	8% of max range
4-11	RN1	2	4% of max range
4-12	RN1	2	8% of max range
4-13	RN1	3	4% of max range
4-14	RN1	3	8% of max range
4-15	RN1	4	4% of max range
4-16	RN1	4	8% of max range

Table 4.7-1 Run details for Run 4-9 to Run 4-16

Similarly to the previous section, each of chemical species in the runs are approximated to rational polynomial using the suitable order for numerator and denominator. The rational polynomials are then differentiated to obtain the rate of concentration change of the chemical species.

In this section, the rational polynomials are also used to smoothen the noisy data to mitigate the effect of noise in the system. Figure 4.7-1 is an example of smoothen concentration data of Run 4-9.

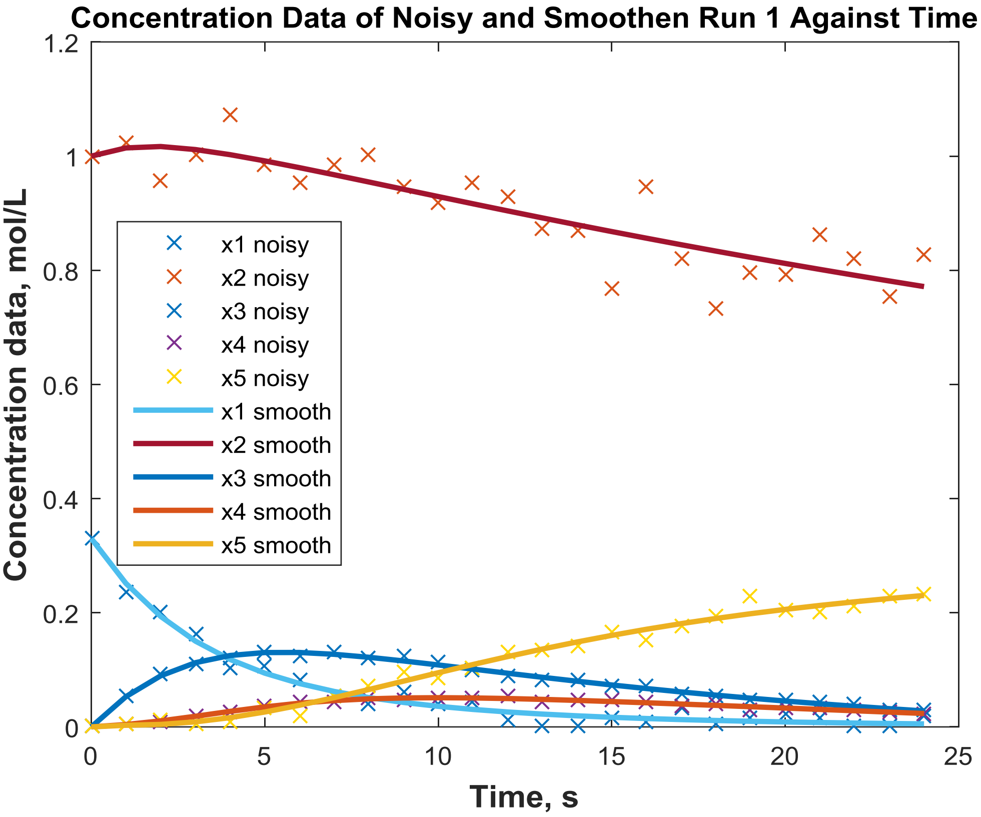


Figure 4.7-1 Noisy and smoothened concentration data against time for Run 4-9

The results from the 8 runs are shown in Table 4.7-2:

Run	Best Performing Individual	Reaction Rate Constant	Fitness Function
4-9	$\begin{bmatrix} -2 & 1 & 0 & 0 & 0 \\ -1 & 0 & 1 & 0 & 0 \\ 0 & 0 & -1 & 1 & 0 \\ 0 & -1 & 0 & -1 & 1 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix}$	$k_1 = 0.1116\ dm^3mol^{-1}s^{-1}$ $k_2 = 0.2088\ s^{-1}$ $k_3 = 0.1335\ s^{-1}$ $k_4 = 0.3025\ dm^3mol^{-1}s^{-1}$ $k_5 = 0$	0.6160
4-10	$\begin{bmatrix} -1 & 0 & 1 & 0 & 0 \\ 0 & 0 & -1 & 1 & 0 \\ 0 & -1 & 0 & -1 & 1 \\ -1 & 0 & 2 & -1 & 0 \\ -1 & 1 & -1 & 0 & 0 \end{bmatrix}$	$k_1 = 0.1741\ s^{-1}$ $k_2 = 0.1468\ s^{-1}$ $k_3 = 0.3371\ dm^3mol^{-1}s^{-1}$ $k_4 = 0.0805\ dm^3mol^{-1}s^{-1}$ $k_5 = 0.0632\ dm^3mol^{-1}s^{-1}$	6.9803
4-11	$\begin{bmatrix} -2 & 1 & 0 & 0 & 0 \\ -1 & 0 & 1 & 0 & 0 \\ 0 & 0 & -1 & 1 & 0 \\ 0 & -1 & 0 & -1 & 1 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix}$	$k_1 = 0.1330\ dm^3mol^{-1}s^{-1}$ $k_2 = 0.2323\ s^{-1}$ $k_3 = 0.1231\ s^{-1}$ $k_4 = 0.2895\ dm^3mol^{-1}s^{-1}$ $k_5 = 0$	3.1846
4-12	$\begin{bmatrix} -2 & 1 & 0 & 0 & 0 \\ -1 & 0 & 1 & 0 & 0 \\ 0 & 0 & -1 & 1 & 0 \\ 0 & -1 & 0 & -1 & 1 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix}$	$k_1 = 0.1107\ dm^3mol^{-1}s^{-1}$ $k_2 = 0.1977\ s^{-1}$ $k_3 = 0.1275\ s^{-1}$ $k_4 = 0.2809\ dm^3mol^{-1}s^{-1}$ $k_5 = 0$	1.3956

Run	Best Performing Individual	Reaction Rate Constant	Fitness Function
4-13	$\begin{bmatrix} -2 & 1 & 0 & 0 & 0 \\ -1 & 0 & 1 & 0 & 0 \\ 0 & 0 & -1 & 1 & 0 \\ 0 & -1 & 0 & -1 & 1 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix}$	$k_1 = 0.0879\ dm^3mol^{-1}s^{-1}$ $k_2 = 0.1923\ s^{-1}$ $k_3 = 0.1273\ s^{-1}$ $k_4 = 0.2896\ dm^3mol^{-1}s^{-1}$ $k_5 = 0$	2.1179
4-14	$\begin{bmatrix} -2 & 1 & 0 & 0 & 0 \\ -1 & 0 & 1 & 0 & 0 \\ 0 & 0 & -1 & 1 & 0 \\ 0 & -1 & 0 & -1 & 1 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix}$	$k_1 = 0.1053\ dm^3mol^{-1}s^{-1}$ $k_2 = 0.2399\ s^{-1}$ $k_3 = 0.1343\ s^{-1}$ $k_4 = 0.2984\ dm^3mol^{-1}s^{-1}$ $k_5 = 0$	12.8375
4-15	$\begin{bmatrix} -2 & 1 & 0 & 0 & 0 \\ -1 & 0 & 1 & 0 & 0 \\ 0 & 0 & -1 & 1 & 0 \\ 0 & -1 & 0 & -1 & 1 \\ -1 & 0 & 2 & -1 & 0 \end{bmatrix}$	$k_1 = 0.0958\ dm^3mol^{-1}s^{-1}$ $k_2 = 0.1832\ s^{-1}$ $k_3 = 0.1391\ s^{-1}$ $k_4 = 0.3126\ dm^3mol^{-1}s^{-1}$ $k_5 = 0.2697\ dm^3mol^{-1}s^{-1}$	1.7917
4-16	$\begin{bmatrix} -2 & 1 & 0 & 0 & 0 \\ -1 & 0 & 1 & 0 & 0 \\ 0 & 0 & -1 & 1 & 0 \\ 0 & -1 & 0 & -1 & 1 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix}$	$k_1 = 0.1085\ dm^3mol^{-1}s^{-1}$ $k_2 = 0.1952\ s^{-1}$ $k_3 = 0.1257\ s^{-1}$ $k_4 = 0.2736\ dm^3mol^{-1}s^{-1}$ $k_5 = 0$	5.4517

Table 4.7-2 Results from the Run 4-9 to Run 4-16

As expected the results from the concentration data with noise included have a higher level of error when compared against when the automated system is run with concentration data without noise included. The automated system managed to obtain the accurate CRN in 6 out of the 8 runs. Run 4-15 contains the correct reactions but also includes an additional misidentified reaction. Run 4-10 did not managed to elucidate all of the RN1 reactions. The impact of noise in the performance of the automated system can be felt here as its effectiveness is now diminished by the presence of noise in the system.

For the runs that correctly identified RN1 CRN, Run 4-14 has the worst performance in terms of fitness function and will used for further analysis. A plot of the predicted concentration data given from the best performing individual in Run 4-14 against the noisy concentration data that has been smoothened and against the original noiseless concentration data is shown in Table 4.7-2.

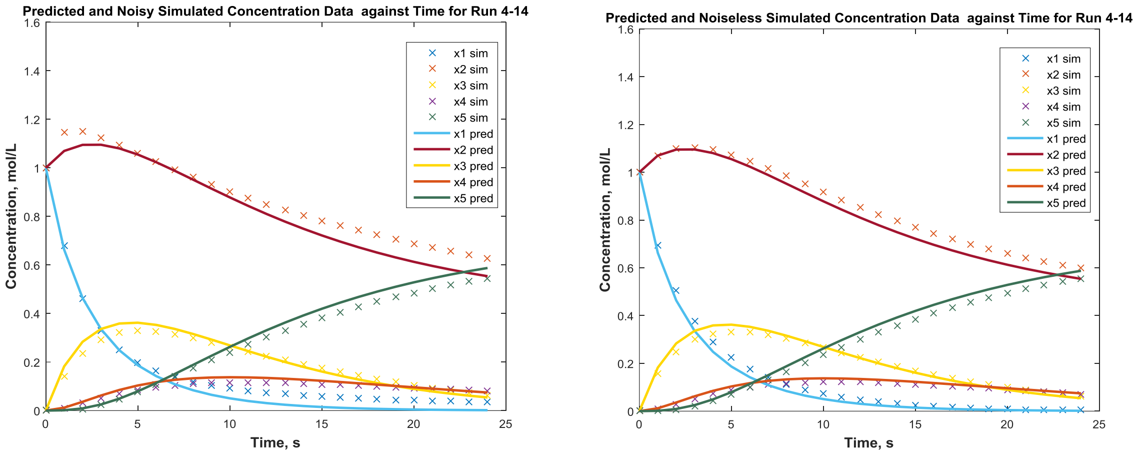


Figure 4.7-2 Comparison between performance of predicted concentration data on noisy and noiseless concentration data

From the figures, it can be seen that the predicted concentration data has a better fit to the original noiseless concentration data than the concentration with noise. This is especially true at the beginning of the run for chemical species x_2 and at the end of the run for chemical species x_1. Using the same reaction rate constants and CRN structure of the best performing individual of Run 4-14, the fitness function is recorded to be 4.9865 when it is compared against the noiseless concentration data. It shows that it performs better against the noiseless concentration data as compared to the noisy data where it only achieve the fitness function of 12.8375 even when the automated system used the noisy concentration data to elucidate the CRN.

The poor fitness function of the run can explained by the poor quality of the concentration data it was provided with. With noisier data, the errors between the data and the actual data is larger and any smoothing process to the data will be affected. The smoothing process of the concentration data is not subjected to mass balance limitations as it is only fitted onto rational polynomials. The automated system on the other hand is based on the law of mass action and the predicted concentration data are reconstructed by solving the ordinary differential equations:

$$[\dot{\boldsymbol{X}}] = \boldsymbol{RV} \qquad \text{(Equation 4-12)}$$

which made the predicted concentration data mass balanced. It is impossible to fit a mass balanced predicted concentration data exactly onto concentration data that do not adhered to mass balance.

However, it can be seen from the Figure that the automated system tried to give the best possible fit given its mass balanced limitations to the smoothing noisy concentration data and still managed to elucidate the correct CRN. It is also shown that when fitted onto the original noiseless concentration data, the predicted concentration data has a better fit and shape to the noiseless concentration data. This is mainly because the original noiseless concentration data is also subjected to the same mass balance constraint when it was simulated through the solution of the ordinary differential Equation 4-12. This is also applicable to the other 5 runs (Run 4-9, Run 4-11, Run 4-12, Run 4-13 and Run 4-16) that correctly identified the CRN but to smaller extent.

The issue with using noisy concentration data is also extended to Run 4-10 and Run 4-15 where it has manifested differently. In Run 4-15, additional reaction that is not part of RN1 is created by the automated system in order for it supplement the predicted concentration data so that it can match the noisy concentration data better. This can be considered as a case of ‘overfitting’ of the CRN because it has gone beyond fitting predicted data to the underlying concentration data in noisy data and has tried to fit the predicted data to the noise. This occurs because when the correct CRN is used, it only achieve the fitness function of 4.1098 as compared to the best performing individual in Run 4-15 which is 1.7917. The automated system will then choose the better performing CRN as its best individual even when it is not the correct one. Figure 4.7-3 shows the plot between the predicted and simulated concentration data against time of the best performing individual in Run 4-15.

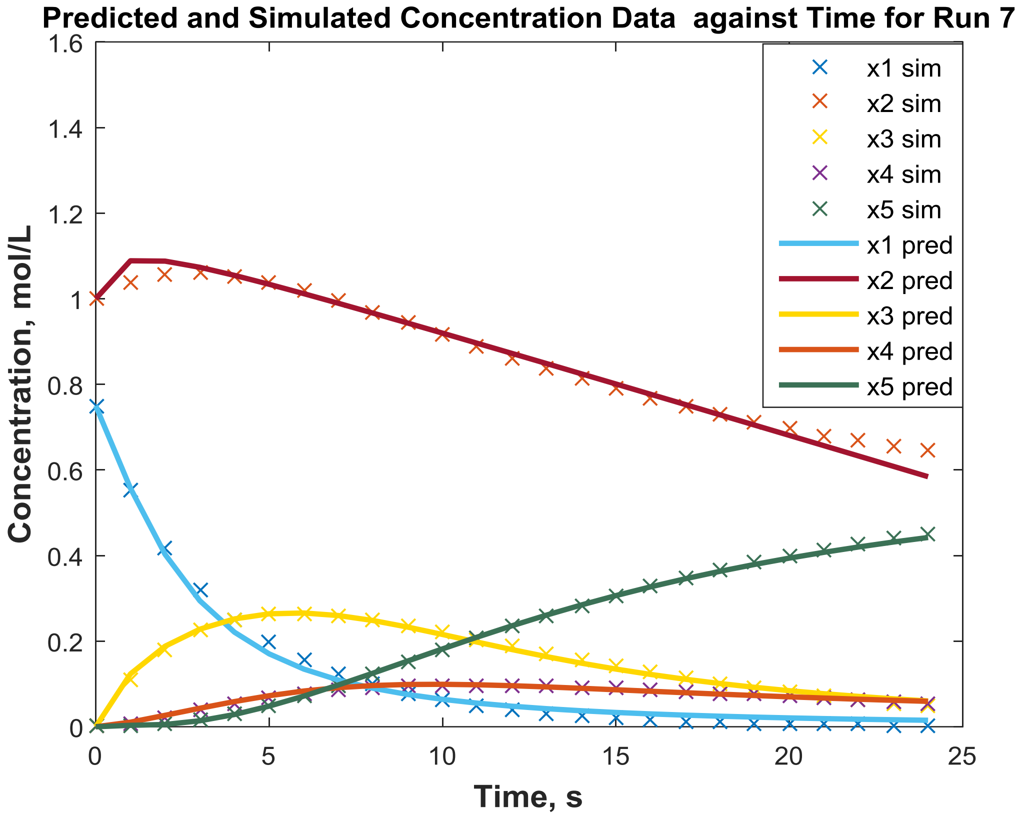


Figure 4.7-3 Predicted and simulated concentration data against time for Run 4-15.

The figure shows that how closely fitted the predicted data is to the simulated noisy data. As far the automated system is concerned, it has achieved its goal in fitting the best possible CRN it can to the simulated data it is provided with but as a user of the system, he will need to further analyse the results to see if there are any reaction that can be discounted. Running the batch further could provide more insight into the performance of the CRN or comparing the results against other reaction batch running with different conditions would bring light to which reaction is not the correct one.

Run 4-10 can be considered as a failure of the automated system. The best performing CRN was unable to identify all the reactions within RN1 and this is a more severe case of ‘overfitting’. The fitness function of the correct CRN only gives 17.9585 while the automated system manage to achieve 6.9803 with the best performing individual. It introduced two other reactions that are not found in RN1 in the CRN in order to match the noise in the simulated data better while omitting one actual reaction from RN1. Figure 4.7-4 shows how effective it is at matching the noisy simulated data even if the CRN it predicts is wrong.

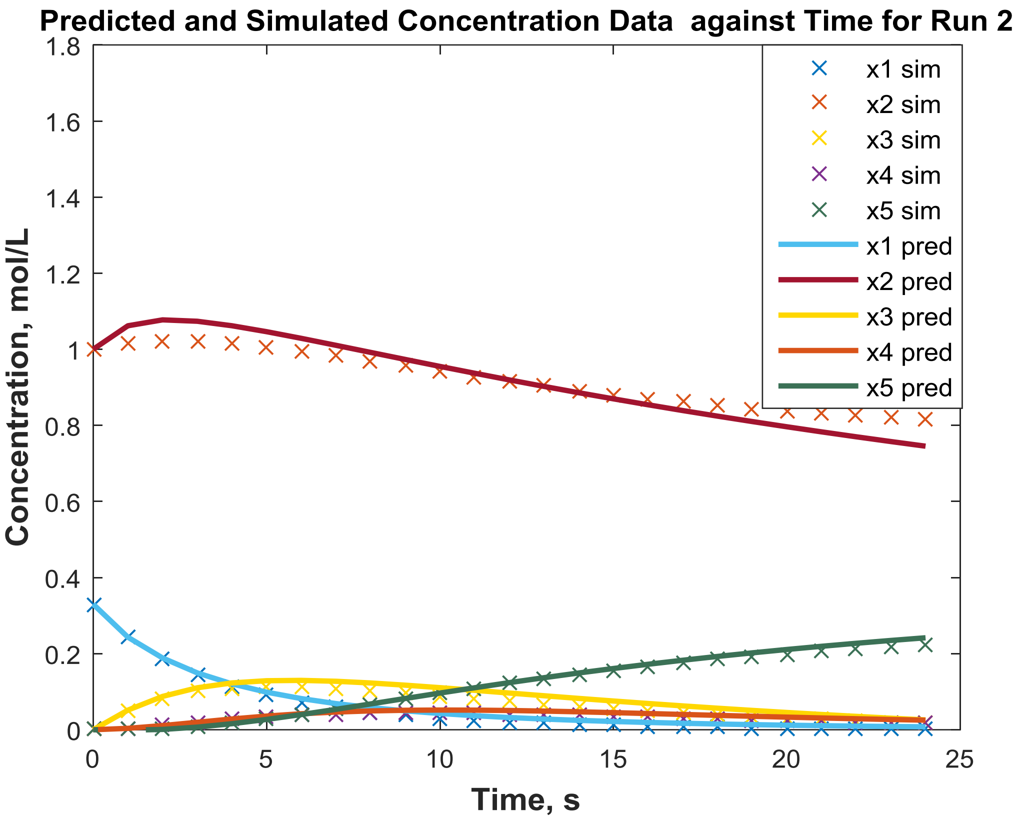


Figure 4.7-4 Predicted and simulated concentration data against time for Run 4-10

This case is a case of GIGO (garbage in, garbage out) where if you provide a poor input data, the computer program in this case, the automated system can only output a poor result.

In general, concentration data with higher amount noise level (Run 4-10, Run 4-12, Run 4-14 and Run 4-14) does gives a poorer fitness function when compared against those with the lower amount of noise level, except for Run 4-12. It may be that the rational polynomial estimated for it is much better than that was done for Run 4-11. It can be concluded that the automated system's is robust enough to run even in the presence of errors or mass imbalance in the concentration data but the results can be better, especially in the case of Run 4-10 where the complete CRN was not deduced.

4.8 Unmeasured chemical species

RN1 and RN2 that are tested in the previous sections did not address the case where there are absence of concentration data in some of the participating chemical species.

This version of the automated system for the elucidation of CRN cannot handle such case. The limitation is caused by its inability to solve for the reaction rate constants through the equation

$$\boldsymbol{K} = (\boldsymbol{A}^{T}\boldsymbol{A})^{-\mathbf{1}}\boldsymbol{A}^{\boldsymbol{T}}[\dot{\boldsymbol{X}}]\boldsymbol{V}^{\#} \qquad \text{(Equation 4-13)}$$

if it does not have all the information in rate of concentration change, $[\dot{\boldsymbol{X}}]$ and $\boldsymbol{A}$ which is dependent on the availability concentration, $[\boldsymbol{X}]$.

4.9 Summary

This chapter presents the application of the automated system for the elucidation of Chemical Reaction Network (CRN) developed in the previous chapter through the use of two fictitious CRN, Reaction Network 1 (RN1) and Reaction Network 2 (RN2). 4 different reaction batches are simulated with different initial conditions for each of the two CRNs. The automated system is tested further for robustness with concentration data that have Gaussian noise added. The automated system had shown to be efficient in elucidating the CRN in RN1 but failed when used to elucidate RN2. The further tests with noise shows that the automated system is able to handle noisy data to a certain extent, identifying all the reactions in the CRN in 5 out of the 8 runs for RN1 without errors and achieve good fitness for all the runs.

Discussions on the weaknesses of this version of the automated system are made and analysed on poor performing runs. It shows inability to handle reversible reactions due to the fact that the system use multiple linear regression which requires the equations it is solving to be linearly independent and reversible reactions produce equations that linearly dependent to each other.

The automated system is also susceptible to noisy data and will cause ‘overfitting’ to occur where the CRN produced by the automated system will try to fit its predicted concentration data to the noise in the simulated concentration data by introducing reactions that are not present in RN1. A sufficiently poor data may cause the automated system to fail to identify some of the reactions and supplement it with other reactions that are not part of RN1 in order to match itself to the simulated concentration data.

Finally, this version of the automated system is unable to operate in the presence of any unmeasured concentration data of participating chemical species because it will unable to solve for the reaction rate constants.

In the next chapter, these weaknesses are addressed through a major upgrade to the automated system.

Chapter 5. Implementation of Two Tiers Optimisation to the Automated System for Chemical Reaction Elucidation

5.1 Overview

In Chapter 3, an automated system using Genetic Algorithm (GA) for the elucidation of chemical reaction network (CRN) are introduced. Using multiple linear regression in line with the artificial intelligence in the form of GA, a suitable CRN can be predicted from the given chemical species' concentration data. In Chapter 4, this automated system has been tested against two fictitious CRNs which unearthed a number of weaknesses with the system. The weaknesses are

1. Unable to process reversible reactions.
2. Susceptible to 'overfitting' when poor noisy quality data are used.
3. Unable to function at all in the presence of any unmeasured concentration data of participating chemical species.

This chapter will address the inability of the automated system in dealing with unmeasured chemical species as this is a major concern especially when faced with reactions which has difficult to measure chemical species or short lived reaction intermediates. The source of the weakness is in the calculation of the reaction rate constants, $\boldsymbol{K}$ and therefore the modifications of the automated system will focus mainly on it.

5.2 Introduction

Unmeasured chemical species in the CRN will have a high chance of causing reaction rate constants calculation to fail when the method suggested in Chapter 3 is used. A simple example of this problem can be seen by observing the following reaction:

$$R_1: \qquad x_1 + x_2 \rightarrow x_3 \qquad \text{(Equation 5-1)}$$

Based on the law of mass action, this reaction will have the reaction rate of

$$r_{R_1} = k_{R_1}[x_1][x_2] \qquad \text{(Equation 5-2)}$$

And the rate of change of $[x_1]$ can be describe as

$$\frac{d[x_1]}{dt} = -r_{R_1} \qquad \text{(Equation 5-3)}$$

The kinetic rate constant, k_{R_1} can be calculated if the data for $[x_1]$ and $[x_2]$ are available using the method as discussed in Chapter 3 which reduces it to a multiple linear regression calculation.

$$\boldsymbol{K} = (\boldsymbol{A}^T\boldsymbol{A})^{-\mathbf{1}}\boldsymbol{A}^{\boldsymbol{T}}[\dot{\boldsymbol{X}}]\boldsymbol{V}^{-\boldsymbol{pinv}} \qquad \text{(Equation 5-4)}$$

However, when one of the chemical species data is unavailable, especially if it is a reactant, the formula cannot be used. For example in the above reaction, R_1, if $[x_1]$ is unavailable, r_{R_1} cannot be obtained which is required for calculation of $\frac{d[x_1]}{dt}$. The rate of concentration change, $\frac{d[x_1]}{dt}$ also cannot be approximated from $[x_1]$ if there is no concentration data for it. Without the ability to determine $\frac{d[x_1]}{dt}$ and r_{R_1} makes the correct determination of the reaction rate constant k_{R_1} through the equation highly improbable.

The automated system will be unable to function without a good reaction rate constant that it can use to reconstruct the predicted concentration data that it uses to compare against the input concentration data. Therefore, a new method of reaction rate constants determination is introduced.

5.3 Dealing with unmeasured chemical species

In order to calculate the reaction rate constants in the presence of unmeasured chemical species, a second tier recursive method based on nonlinear least squares optimiser is introduced into the automated system for elucidation of CRN. This new modified automated system will be referred to as automated system (version 2) for easier reference.

The Figure below shows the flowchart of both tiers in automated system (version 2) and their relationship:

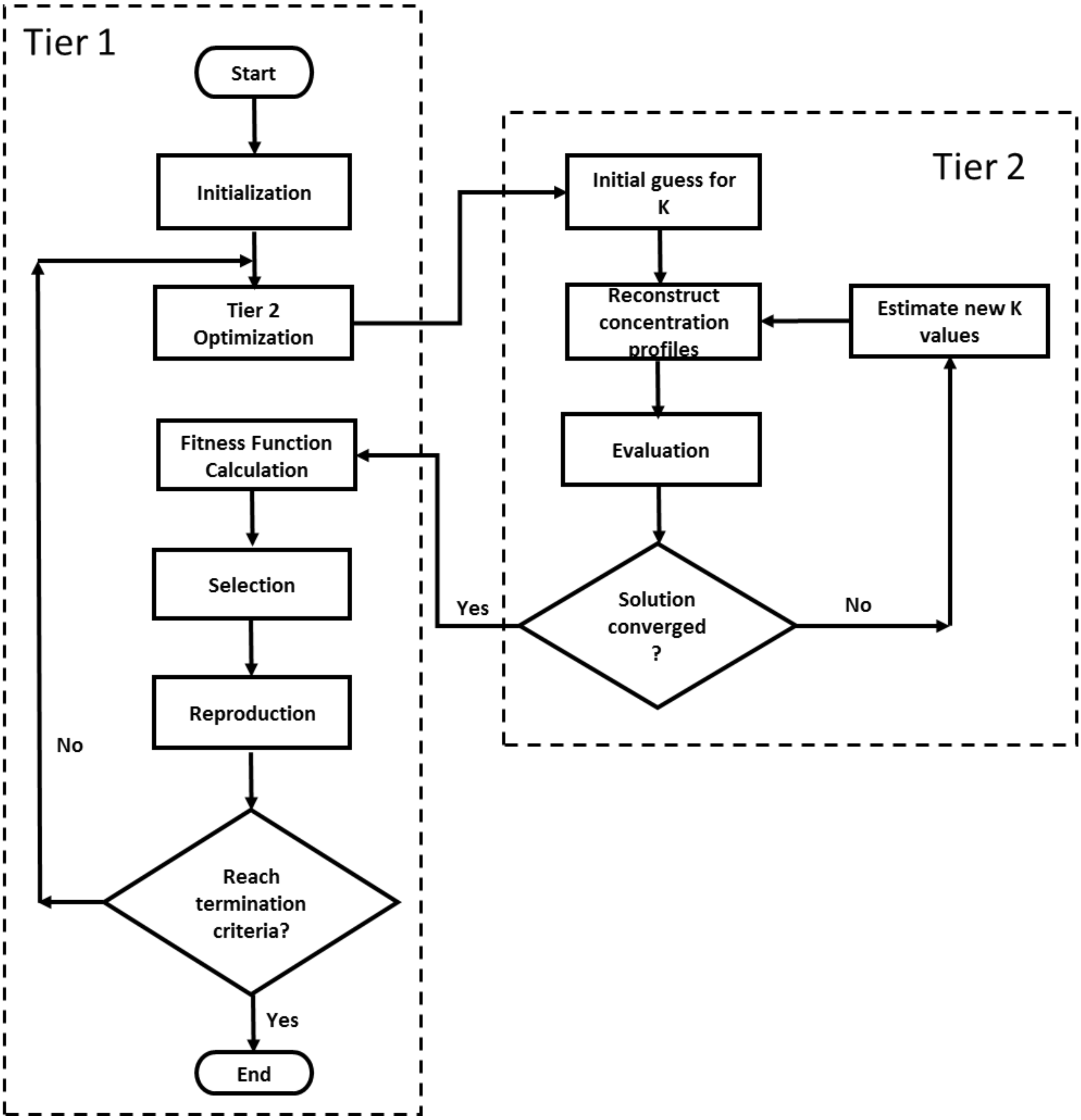


Figure 5.3-1 Flowchart for automated system with two tiers optimisation

Tier 1 of automated system (version 2) follows the basic Genetic Algorithm (GA) flow which also used in the automated system described Chapter 3. In this version of the automated system, the reaction rate constant estimation step is replaced with the Tier 2 optimisation step which is used to estimate the reaction rate constant. The Tier 1 of automated system (version 2) is responsible to produce candidate CRNs through evolution using the GA in the system. The candidate CRNs information will be passed to Tier 2 to estimate each of the candidate CRN's reaction rate constants. The estimation of automated system (version 2) is based on nonlinear least squares optimisation algorithm which is employed using the command 'lsqnonlin' in MATLAB.

The first step in Tier 2 is to have an initial guess for the reaction rate constants, $\boldsymbol{K}$. This value is user defined and will help with the speed of convergence if the user have additional information on the range of expected reaction rate constants. Else, it can be set at 0.1 for all the reaction rate constants to kick-start the loop.

Using the reaction rate constants, $\boldsymbol{K}$ determine in the previous step, initial concentration data of the chemical species at the beginning of the batch, candidate CRN stoichiometric matrix and the time when concentration data points are taken, the predicted concentration profiles of chemical species, $[\widetilde{\boldsymbol{X}}]$ can be reconstructed. This includes the concentration profiles of the unmeasured chemical species as all the required information is present even if the reaction rate constants, $\boldsymbol{K}$ is not accurate in the beginning of the loop.

The predicted concentration profiles, $[\widetilde{\boldsymbol{X}}]$ is then compared against the input concentration data, $[\boldsymbol{X}]$ which can be from a simulated data of a CRN or from experimental data. The comparison used is the same as the one used in the Tier 1 fitness function calculation which is variance weighted sums of squared error, *VMSSE*.

$$VMSSE = \sum_{i=1}^{N_c} \frac{\sum_{t=0}^{N_t}([x_i]_t - [\hat{x}_i]_t)^2}{\sum_{t=0}^{N_t}\left([x_i]_t - \mu_{x_i}\right)^2}, \forall\, u_{x_\mathrm{i}} = 1 \qquad \text{(Equation 5-5)}$$

$[x_i]_t$ is the concentration data of chemical species x_i at time t

$[\hat{x}_i]_t$ is the predicted concentration data of chemical species x_i at time t

N_c = total number of participating chemical species

N_t = the time when the last data point is being evaluated

u_{x_i} is the measured/unmeasured identifier for $[x_i]$. Measured chemical species will be given the value of 1 and unmeasured will be given the value 0.

μ_{x_i} is the standard deviation of the all the concentration data of $[x_i]$ that is being evaluated

With the use of the measured/unmeasured identifier, u_{x_i}, the unmeasured chemical species' concentration will be excluded from the calculation of *VMSSE*. In other words, the fitness function of the current guessed/estimated reaction rate constants, $\boldsymbol{K}$ will not take into account the unmeasured chemical species. It is impossible to grade how good is the predicted concentration data if there is no input concentration data for it to compare to.

Based on the *VMSSE*, the trust-region-reflective algorithm in 'lsqnonlin' function will determine whether it has met the best possible solution for the reaction rate constants, $\boldsymbol{K}$ by comparing it against estimated $\boldsymbol{K}$ values in previous loops. On the first loop, this step is skipped as there are no previous estimations to compare. If the algorithm determines further optimisation is possible, a new $\boldsymbol{K}$ values will be estimated using the trust-region-reflective algorithm. The process will be repeated until it has achieved the best possible *VMSSE* and the final set of $\boldsymbol{K}$ values can be passed back to Tier 1 to continue on the evolution of the CRNs.

With this method, the reaction rate constants, $\boldsymbol{K}$ can be determined even if there is unmeasured concentration data. This method also do away with the need to obtain the rate of concentration change, $[\dot{\boldsymbol{X}}]$ which would reduce the amount of errors that is introduce into the system from using the approximated values through the differentiation of rational polynomials as discussed in the previous chapter.

5.4 Testing against measured data with no noise

Automated system (version 2) is first tested against the datasets that are generated in the previous chapters with all the chemical species marked as measured chemical species. This will show the capability of automated system (version 2) when all the concentration data are known and it will be compared against the automated system developed in Chapter 3.

Table 5.4-1 details the datasets used for this test.

Run	Chemical Reaction Network	Batch	Gaussian Noise Standard Deviation
5-1	RN1	1	0% of max range
5-2	RN1	2	0% of max range
5-3	RN1	3	0% of max range
5-4	RN1	4	0% of max range
5-5	RN2	1	0% of max range
5-6	RN2	2	0% of max range
5-7	RN2	3	0% of max range
5-8	RN2	4	0% of max range

Table 5.4-1 Run details for Run 5-1 to 5-8

The first 8 runs to test out automated system (version 2) will be tested on concentration data without any noise introduced, the same as the one done in Chapter 4 (Run 4-1 to Run 4-8). The parameters used in automated system (version 2) is the same as the one used in the automated system for the runs in Chapter 4. The results of the runs are published in Table 5.4-2 and Table 5.4-3.

Run	Best Performing Individual	Reaction Rate Constant	Fitness Function
5-1	$\begin{bmatrix} -2 & 1 & 0 & 0 & 0 \\ -1 & 0 & 1 & 0 & 0 \\ 0 & 0 & -1 & 1 & 0 \\ 0 & -1 & 0 & -1 & 1 \end{bmatrix}$	$k_1 = 0.1000\ dm^3mol^{-1}s^{-1}$ $k_2 = 0.2000\ s^{-1}$ $k_3 = 0.1300\ s^{-1}$ $k_4 = 0.3000\ dm^3mol^{-1}s^{-1}$	0
5-2	$\begin{bmatrix} -2 & 1 & 0 & 0 & 0 \\ -1 & 0 & 1 & 0 & 0 \\ 0 & 0 & -1 & 1 & 0 \\ 0 & -1 & 0 & -1 & 1 \end{bmatrix}$	$k_1 = 0.1000\ dm^3mol^{-1}s^{-1}$ $k_2 = 0.2000\ s^{-1}$ $k_3 = 0.1300\ s^{-1}$ $k_4 = 0.3000\ dm^3mol^{-1}s^{-1}$	0
5-3	$\begin{bmatrix} -2 & 1 & 0 & 0 & 0 \\ -1 & 0 & 1 & 0 & 0 \\ 0 & 0 & -1 & 1 & 0 \\ 0 & -1 & 0 & -1 & 1 \end{bmatrix}$	$k_1 = 0.1000\ dm^3mol^{-1}s^{-1}$ $k_2 = 0.2000\ s^{-1}$ $k_3 = 0.1300\ s^{-1}$ $k_4 = 0.3000\ dm^3mol^{-1}s^{-1}$	0
5-4	$\begin{bmatrix} -2 & 1 & 0 & 0 & 0 \\ -1 & 0 & 1 & 0 & 0 \\ 0 & 0 & -1 & 1 & 0 \\ 0 & -1 & 0 & -1 & 1 \end{bmatrix}$	$k_1 = 0.1000\ dm^3mol^{-1}s^{-1}$ $k_2 = 0.2000\ s^{-1}$ $k_3 = 0.1300\ s^{-1}$ $k_4 = 0.3000\ dm^3mol^{-1}s^{-1}$	0

Table 5.4-2 Results for Run 5-1 to Run 5-4

Run	Best Performing Individual	Reaction Rate Constant	Fitness Function
5-5	$\begin{bmatrix} -1 & -1 & 1 & 1 & 0 & 0 \\ 0 & -1 & -1 & 0 & 1 & 0 \\ -1 & 0 & 0 & -1 & 0 & 1 \\ 1 & 0 & 0 & 1 & 0 & -1 \end{bmatrix}$	$k_1 = 0.2000\ dm^3mol^{-1}s^{-1}$ $k_2 = 0.1000\ dm^3mol^{-1}s^{-1}$ $k_3 = 0.1500\ dm^3mol^{-1}s^{-1}$ $k_4 = 0.0500\ dm^3mol^{-1}s^{-1}$	0
5-6	$\begin{bmatrix} -1 & -1 & 1 & 1 & 0 & 0 \\ 0 & -1 & -1 & 0 & 1 & 0 \\ -1 & 0 & 0 & -1 & 0 & 1 \\ 1 & 0 & 0 & 1 & 0 & -1 \end{bmatrix}$	$k_1 = 0.1979\ dm^3mol^{-1}s^{-1}$ $k_2 = 0.1453\ dm^3mol^{-1}s^{-1}$ $k_3 = 0.0992\ dm^3mol^{-1}s^{-1}$ $k_4 = 0.0521\ dm^3mol^{-1}s^{-1}$	0.2077
5-7	$\begin{bmatrix} -1 & -1 & 1 & 1 & 0 & 0 \\ 0 & -1 & -1 & 0 & 1 & 0 \\ -1 & 0 & 0 & -1 & 0 & 1 \\ 1 & 0 & 0 & 1 & 0 & -1 \end{bmatrix}$	$k_1 = 0.1979\ dm^3mol^{-1}s^{-1}$ $k_2 = 0.1453\ dm^3mol^{-1}s^{-1}$ $k_3 = 0.0992\ dm^3mol^{-1}s^{-1}$ $k_4 = 0.0521\ dm^3mol^{-1}s^{-1}$	0.2077
5-8	$\begin{bmatrix} -1 & -1 & 1 & 1 & 0 & 0 \\ 0 & -1 & -1 & 0 & 1 & 0 \\ -1 & 0 & 0 & -1 & 0 & 1 \\ 1 & 0 & 0 & 1 & 0 & -1 \end{bmatrix}$	$k_1 = 0.2000\ dm^3mol^{-1}s^{-1}$ $k_2 = 0.1000\ dm^3mol^{-1}s^{-1}$ $k_3 = 0.1500\ dm^3mol^{-1}s^{-1}$ $k_4 = 0.0500\ dm^3mol^{-1}s^{-1}$	0

Table 5.4-3 Results for Run 5-5 to Run 5-8

From the two tables, it is clear that automated system (version 2) was able to deduce the CRN perfectly without any additional incorrectly identified reaction or missing any reaction from both RN1 and RN2. This is especially true for RN1 where it also calculates the reaction rate constants perfectly and its predicted concentration data matches the simulated data exactly.

For RN2, Run 5-5 and Run 5-8 also achieve a perfect match with fitness function of 0 and exact estimation of the reaction rate constants of RN2. Run 5-6 and Run 5-7 manages to elucidate the CRN that is the same as RN2 but suffers a slight error in the calculation of the reaction rate constants and achieve a slightly weaker fitness of 0.2077 as compared to Run 5-5 and Run 5-8. The slight error in the reaction rate calculation is caused by the non-linear least squares optimisation routine used converging before the actual result. The choice of the initial guess for the reaction rate constants plays a part in the convergence of the optimisation routine and this could be rectified by choosing a different initial guess. Even if the exact values are not obtained, it still manages to elucidate the CRN structure and the reaction rate constants are very close to the actual values in RN2.

Comparison of fitness function of Run 4-1 to Run 4-8 against Run 5-1 to Run 5-8 is shown in Table 5.4-4:

Run	Fitness Function (Run 4-1 to Run 4-8)	Run	Fitness Function (Run 5-1 to Run 5-8)
4-1	0.0326	5-1	0
4-2	0.0391	5-2	0
4-3	0.0390	5-3	0
4-4	0.0320	5-4	0
4-5	6.1877	5-5	0
4-6	18.5476	5-6	0.2077
4-7	13.8507	5-7	0.2077
4-8	25.6179	5-8	0

Table 5.4-4 Comparison of fitness functions between Run 4-1 to Run 4-8 and Run 5-1 to Run 5-8

The comparison easily shows that Run 5-1 to Run 5-8 outperforms their counterpart in Run 4-1 to Run 4-8. For the datasets created from RN1, automated system (version 2) manage to achieve 0 fitness function or perfect fit to the original CRN as shown from the results of Run 5-1 to Run 5-4. Run 4-1 to Run 4-4 results are just slightly inferior and this as discussed in Chapter 4, affected by inaccuracy when the concentration data are modelled to fit rational polynomials.

For RN2, the significance of automated system (version 2)'s ability can be observed. For Run 4-5 to Run 4-8, the fitness function are relatively much higher as compared to those in Run 5-5 and Run 5-8 and this stems from the fact that the original automated system was unable to elucidate the correct CRN. Automated system (version 2) is able to overcome this shortfall and manage to discover the CRN that matches with RN2. This is mainly because this version of the automated system does not depend on the use of multiple linear regression in order to solve for the reaction rate constants and thus is not faced with the complication of having more variables to solve for than linearly independent equations.

It is clear based on the performance comparison between Run 4-1 to Run 4-8 against Run 5-1 to Run 5-8, automated system (version 2) is a better elucidation system for CRN as compared to the original automated system. Without the need to fit the concentration data to rational polynomials reduces the amount of errors that are introduced into the concentration data. The new automated system is able to handle reversible reactions as shown by the run results from Run 5-5 to Run 5-8 as compared the original automated system which was unable to handle such reactions.

5.5 Noise

Automated system (version 2) robustness is also tested against datasets that are perturbed with noise and subsequently smoothened. These datasets are the same as those used in Run 4-9 to Run 4-16 for RN1. Comparisons between the performance of automated system (version 2) and the original automated system are then made. The details of the run is displayed in Table 5.5-1:

Run	Chemical Reaction Network	Batch	Gaussian Noise Standard Deviation
5-9	RN1	1	4% of max range
5-10	RN1	1	8% of max range
5-11	RN1	2	4% of max range
5-12	RN1	2	8% of max range
5-13	RN1	3	4% of max range
5-14	RN1	3	8% of max range
5-15	RN1	4	4% of max range
5-16	RN1	4	8% of max range

Table 5.5-1 Run results for Run 5-9 to Run 5-16

The results of the runs are displayed in Table 5.5-2:

Run	Best Performing Individual	Reaction Rate Constant	Fitness Function
5-9	$\begin{bmatrix} -2 & 1 & 0 & 0 & 0 \\ -1 & 0 & 1 & 0 & 0 \\ 0 & 0 & -1 & 1 & 0 \\ 0 & -1 & 0 & -1 & 1 \\ 0 & -1 & 2 & 0 & 0 \end{bmatrix}$	$k_1 = 0.1127\ dm^3mol^{-1}s^{-1}$ $k_2 = 0.1999\ s^{-1}$ $k_3 = 0.1340\ s^{-1}$ $k_4 = 0.3064\ dm^3mol^{-1}s^{-1}$ $k_5 = 0.0003\ \ s^{-1}$	0.2539
5-10	$\begin{bmatrix} -2 & 1 & 0 & 0 & 0 \\ -1 & 0 & 1 & 0 & 0 \\ 0 & 0 & -1 & 1 & 0 \\ 0 & -1 & 0 & -1 & 1 \\ 1 & -1 & 1 & 0 & 0 \end{bmatrix}$	$k_1 = 0.1487\ dm^3mol^{-1}s^{-1}$ $k_2 = 0.2083\ s^{-1}$ $k_3 = 0.1370\ s^{-1}$ $k_4 = 0.3081\ dm^3mol^{-1}s^{-1}$ $k_5 = 0.0007\ \ s^{-1}$	2.6685
5-11	$\begin{bmatrix} -2 & 1 & 0 & 0 & 0 \\ -1 & 0 & 1 & 0 & 0 \\ 0 & 0 & -1 & 1 & 0 \\ 0 & -1 & 0 & -1 & 1 \\ 2 & 1 & 0 & 2 & -2 \end{bmatrix}$	$k_1 = 0.0998\ dm^3mol^{-1}s^{-1}$ $k_2 = 0.2040\ s^{-1}$ $k_3 = 0.1260\ s^{-1}$ $k_4 = 0.3132\ dm^3mol^{-1}s^{-1}$ $k_5 = 0.0074\ dm^3mol^{-1}s^{-1}$	0.8643

Run	Best Performing Individual	Reaction Rate Constant	Fitness Function
5-12	$\begin{bmatrix} -2 & 1 & 0 & 0 & 0 \\ -1 & 0 & 1 & 0 & 0 \\ 0 & 0 & -1 & 1 & 0 \\ 0 & -1 & 0 & -1 & 1 \\ 1 & 1 & -1 & 1 & -1 \end{bmatrix}$	$k_1 = 0.1097\, dm^3mol^{-1}s^{-1}$ $k_2 = 0.2101\, s^{-1}$ $k_3 = 0.1283\, s^{-1}$ $k_4 = 0.3399\, dm^3mol^{-1}s^{-1}$ $k_5 = 0.0989\, dm^3mol^{-1}s^{-1}$	0.6973
5-13	$\begin{bmatrix} -2 & 1 & 0 & 0 & 0 \\ -1 & 0 & 1 & 0 & 0 \\ 0 & 0 & -1 & 1 & 0 \\ 0 & -1 & 0 & -1 & 1 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix}$	$k_1 = 0.1056\, dm^3mol^{-1}s^{-1}$ $k_2 = 0.2043\, s^{-1}$ $k_3 = 0.1291\, s^{-1}$ $k_4 = 0.2953\, dm^3mol^{-1}s^{-1}$ $k_5 = 0$	0.5840
5-14	$\begin{bmatrix} -1 & 0 & 1 & 0 & 0 \\ 0 & 0 & -1 & 1 & 0 \\ 0 & -1 & 0 & -1 & 1 \\ -1 & 1 & -1 & 0 & 0 \\ 2 & -1 & 1 & 2 & -1 \end{bmatrix}$	$k_1 = 0.2124\, s^{-1}$ $k_2 = 0.1224\, s^{-1}$ $k_3 = 0.2866\, dm^3mol^{-1}s^{-1}$ $k_4 = 0.2279\ \ dm^3mol^{-1}s^{-1}$ $k_5 = 0.0046\, dm^3mol^{-1}s^{-1}$	6.2602
5-15	$\begin{bmatrix} -2 & 1 & 0 & 0 & 0 \\ -1 & 0 & 1 & 0 & 0 \\ 0 & 0 & -1 & 1 & 0 \\ 0 & -1 & 0 & -1 & 1 \\ -1 & 0 & 2 & -1 & 0 \end{bmatrix}$	$k_1 = 0.1133\, dm^3mol^{-1}s^{-1}$ $k_2 = 0.2101\, s^{-1}$ $k_3 = 0.1293\, s^{-1}$ $k_4 = 0.2989\, dm^3mol^{-1}s^{-1}$ $k_5 = 0.0605\, dm^3mol^{-1}s^{-1}$	1.3177
5-16	$\begin{bmatrix} -2 & 1 & 0 & 0 & 0 \\ -1 & 0 & 1 & 0 & 0 \\ 0 & 0 & -1 & 1 & 0 \\ 0 & -1 & 0 & -1 & 1 \\ -1 & 1 & -1 & 0 & 0 \end{bmatrix}$	$k_1 = 0.0460\, dm^3mol^{-1}s^{-1}$ $k_2 = 0.\ 2168s^{-1}$ $k_3 = 0.1317\, s^{-1}$ $k_4 = 0.2894\, dm^3mol^{-1}s^{-1}$ $k_5 = 0.1164\, dm^3mol^{-1}s^{-1}$	4.4113

Table 5.5-2 Results for Run 5-9 to Run 5-16

Of the 8 runs that are being tested, 7 of the runs managed to elucidate all the reactions in RN1 with Run 5-14 missed one of the four reactions in RN1. The plots below shows the result of Run 5-14 when plot against the noisy simulated data that is used for the run and against the same data but without the noise.

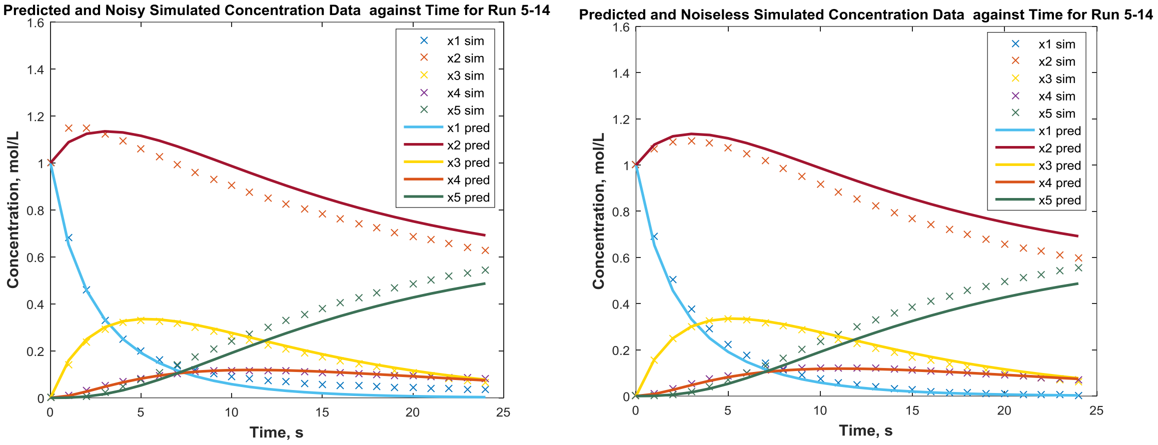


Figure 5.5-1 Concentration data of predicted concentration compared against noiseless and noisy concentration against time for 5-14

The plots are compared against the results that are obtained through the use of the old automated system when run with the same concentration data (Run 4-14). For easier reference the plots are presented in Figure 5.5-2:

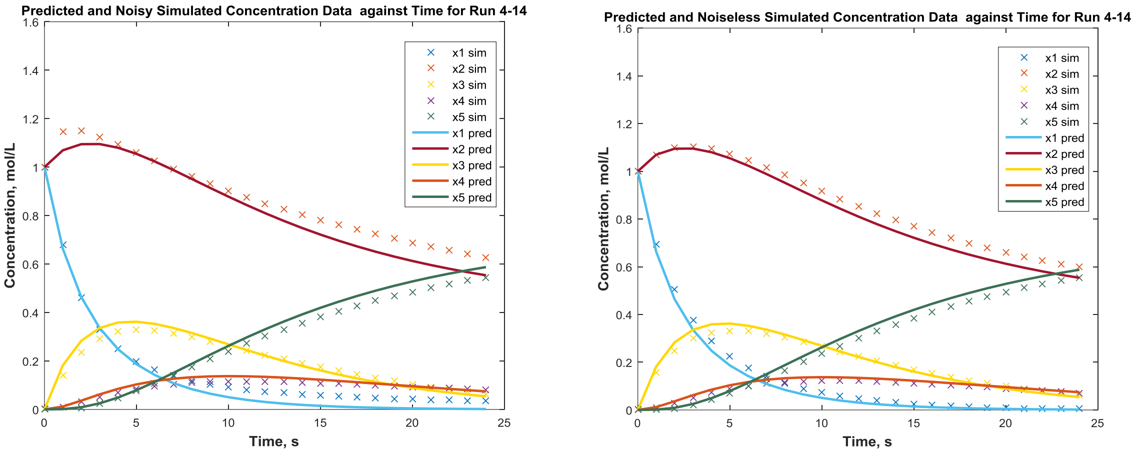


Figure 5.5-2 Predicted concentration data against noisy and noiseless concentration data against time for Run 4-14

Graphically, comparing plot of Run 4-14 and Run 5-14 on noisy data, it can be seen that both perform just as poorly in order to fit their predicted concentration data to the input concentration data. However, when the predicted concentration data are plotted against the noiseless concentration data, Run 4-14 was able to fit it better while Run 5-14 fit is much poorer especially at nearer the end of the batch run. From the results, it seems that the original automated system is able to reject some of the noise within the concentration data and try to simulate the actual CRN while automated system (version 2) is much more aggressive in obtaining a better fit to the input data causing ‘overfitting’. This is further shown by the fact that Run 4-14 managed to elucidate the entire CRN for RN1 but Run 5-14 missed a single reaction and substituting it with two reactions that are not present in RN1. The comparison does highlight the fact that the quality of the noisy data is poor and which is the main reason which causes the poor fit of the predicted concentration data to input simulation data as discussed in Chapter 4.

The 7 runs that managed to detect all the reactions also included reaction that is not part of RN1 in their CRNs apart from Run 5-13 which managed to obtain the exact RN1's CRN. These 6 runs that contain unknown reactions in their CRNs exhibit the behaviour of 'overfitting' as automated system (version 2) try to even include the random errors in the system into its model. To discuss this further a table of comparison of the fitness function between Run 4-9 to Run 4-16 and between Run 5-9 to Run 5-16 is presented.

Run	Fitness Function (Run 4-9 to Run 4-16)	Run	Fitness Function (Run 5-9 to Run 5-16)
4-9	0.6160	5-9	0.2539
4-10	6.9803	5-10	2.6685
4-11	3.1846	5-11	0.8643
4-12	1.3956	5-12	0.6973
4-13	2.1179	5-13	0.5840
4-14	12.8375	5-14	6.2602
4-15	1.7917	5-15	1.3177
4-16	5.4517	5-16	4.4113

Table 5.5-3 Comparison fitness results between Run 4-9 to Run 4-16 and Run 5-9 to Run 5-16

Table 5.5-3 shows that automated system (version 2) consistently record a better fitness function as compared against the original automated system when comparing the runs with their counterpart. This shows that the reaction rate estimation method done in automated system (version 2) is better as it helps to reduce the errors between the predicted data and simulated data more than that of the original automated system. Unfortunately, it also meant that if there is any noise in the system, the automated system (version 2) is more sensitive to it and may cause 'overfitting' to occur.

Noise is also added to the concentration of the 4 batches produced from RN1 in the similar fashion as RN1. Gaussian noise with mean of 0 and standard deviation of 4% and 8% of the maximum range of the chemical species concentration data are introduced to the concentration data. This will be used to evaluate the robustness of automated system (version 2) on RN2 since it has been shown that the system can handle reversible reactions as opposed to the original automated system. Details of the run shown in Table 5.5-4:

Run	Chemical Reaction Network	Batch	Gaussian Noise Standard Deviation
5-17	RN2	1	4% of max range
5-18	RN2	1	8% of max range
5-19	RN2	2	4% of max range
5-20	RN2	2	8% of max range
5-21	RN2	3	4% of max range
5-22	RN2	3	8% of max range
5-23	RN2	4	4% of max range
5-24	RN2	4	8% of max range

Table 5.5-4 Run details for Run 5-17 to Run 5-24

The results of the runs are published in Table 5.5-5:

Run	Best Performing Individual	Reaction Rate Constant	Fitness Function
5-17	$\begin{bmatrix} -1 & -1 & 1 & 1 & 0 & 0 \\ 0 & -1 & -1 & 0 & 1 & 0 \\ -1 & 0 & 0 & -1 & 0 & 1 \\ 1 & 0 & 0 & 2 & 1 & -2 \\ 1 & 0 & 0 & 0 & -1 & 0 \\ 0 & 0 & 0 & -1 & -1 & 1 \end{bmatrix}$	$k_1 = 0.1990\ dm^3mol^{-1}s^{-1}$ $k_2 = 0.1005\ dm^3mol^{-1}s^{-1}$ $k_3 = 0.1349\ dm^3mol^{-1}s^{-1}$ $k_4 = 0.0415\ dm^3mol^{-1}s^{-1}$ $k_5 = 0.0074\ s^{-1}$ $k_6 = 0.0146\ dm^3mol^{-1}s^{-1}$	0.7452
5-18	$\begin{bmatrix} -1 & -1 & 1 & 1 & 0 & 0 \\ 0 & -1 & -1 & 0 & 1 & 0 \\ -1 & 0 & 0 & -1 & 0 & 1 \\ 1 & 0 & 0 & 1 & 0 & -1 \\ -1 & 1 & 0 & 2 & 0 & -1 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}$	$k_1 = 0.2116\ dm^3mol^{-1}s^{-1}$ $k_2 = 0.1003\ dm^3mol^{-1}s^{-1}$ $k_3 = 0.1439\ dm^3mol^{-1}s^{-1}$ $k_4 = 0.0271\ s^{-1}$ $k_5 = 0.0057\ dm^3mol^{-1}s^{-1}$ $k_6 = 0$	2.4376

Run	Best Performing Individual	Reaction Rate Constant	Fitness Function
5-19	$\begin{bmatrix} -1 & -1 & 1 & 1 & 0 & 0 \\ 0 & -1 & -1 & 0 & 1 & 0 \\ -1 & 0 & 0 & -1 & 0 & 1 \\ 1 & 0 & 1 & 0 & 1 & -1 \\ 0 & -1 & 1 & 2 & 0 & -1 \\ 0 & 0 & -1 & 2 & 0 & -1 \end{bmatrix}$	$k_1 = 0.1820\ dm^3mol^{-1}s^{-1}$ $k_2 = 0.1070\ dm^3mol^{-1}s^{-1}$ $k_3 = 0.1566\ dm^3mol^{-1}s^{-1}$ $k_4 = 0.0120\ s^{-1}$ $k_5 = 0.0078\ dm^3mol^{-1}s^{-1}$ $k_6 = 0.0394\ dm^3mol^{-1}s^{-1}$	3.1385
5-20	$\begin{bmatrix} -1 & -1 & 1 & 1 & 0 & 0 \\ 0 & -1 & -1 & 0 & 1 & 0 \\ -1 & 0 & 0 & -1 & 0 & 1 \\ 0 & 2 & 0 & -1 & 0 & 0 \\ -1 & 0 & 0 & 0 & 1 & 0 \\ 2 & 0 & 0 & 1 & -1 & -1 \end{bmatrix}$	$k_1 = 0.1925\ dm^3mol^{-1}s^{-1}$ $k_2 = 0.0901\ dm^3mol^{-1}s^{-1}$ $k_3 = 0.1516\ dm^3mol^{-1}s^{-1}$ $k_4 = 0.0039\ s^{-1}$ $k_5 = 0.0412\ s^{-1}$ $k_6 = 0.0564\ dm^3mol^{-1}s^{-1}$	13.7669
5-21	$\begin{bmatrix} -1 & -1 & 1 & 1 & 0 & 0 \\ 0 & -1 & -1 & 0 & 1 & 0 \\ -1 & 0 & 0 & -1 & 0 & 1 \\ 1 & 0 & 0 & 1 & 0 & -1 \\ 2 & -1 & 0 & -1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}$	$k_1 = 0.1959\ dm^3mol^{-1}s^{-1}$ $k_2 = 0.0986\ dm^3mol^{-1}s^{-1}$ $k_3 = 0.1520\ dm^3mol^{-1}s^{-1}$ $k_4 = 0.0547\ dm^3mol^{-1}s^{-1}$ $k_5 = 0.0034\ dm^3mol^{-1}s^{-1}$ $k_6 = 0\ dm^3mol^{-1}s^{-1}$	2.1002
5-22	$\begin{bmatrix} -1 & -1 & 1 & 1 & 0 & 0 \\ 0 & -1 & -1 & 0 & 1 & 0 \\ -1 & 0 & 0 & -1 & 0 & 1 \\ 1 & 0 & 0 & 1 & 0 & -1 \\ 1 & 1 & -1 & -1 & 0 & 0 \\ 0 & 1 & 1 & 0 & -1 & 0 \end{bmatrix}$	$k_1 = 0.2099\ dm^3mol^{-1}s^{-1}$ $k_2 = 0.0988\ dm^3mol^{-1}s^{-1}$ $k_3 = 0.1229\ dm^3mol^{-1}s^{-1}$ $k_4 = 0.0293\ s^{-1}$ $k_5 = 0.0368\ dm^3mol^{-1}s^{-1}$ $k_6 = 0.0180\ s^{-1}$	13.9896
5-23	$\begin{bmatrix} -1 & -1 & 1 & 1 & 0 & 0 \\ 0 & -1 & -1 & 0 & 1 & 0 \\ -1 & 0 & 0 & -1 & 0 & 1 \\ 1 & 0 & 0 & 1 & 0 & -1 \\ 0 & 0 & 0 & -1 & -1 & 1 \\ 0 & 2 & 0 & -1 & 0 & 0 \end{bmatrix}$	$k_1 = 0.1938\ dm^3mol^{-1}s^{-1}$ $k_2 = 0.1005\ dm^3mol^{-1}s^{-1}$ $k_3 = 0.1382\ dm^3mol^{-1}s^{-1}$ $k_4 = 0.0436\ s^{-1}$ $k_5 = 0.0322\ dm^3mol^{-1}s^{-1}$ $k_6 = 0.0226\ s^{-1}$	1.9532
5-24	$\begin{bmatrix} -1 & -1 & 1 & 1 & 0 & 0 \\ 0 & -1 & -1 & 0 & 1 & 0 \\ -1 & 0 & 0 & -1 & 0 & 1 \\ 1 & 0 & 0 & 1 & 0 & -1 \\ 2 & 1 & 2 & 0 & -1 & -1 \\ 0 & 0 & -1 & 2 & 0 & -1 \end{bmatrix}$	$k_1 = 0.1871\ dm^3mol^{-1}s^{-1}$ $k_2 = 0.0977\ dm^3mol^{-1}s^{-1}$ $k_3 = 0.1545\ dm^3mol^{-1}s^{-1}$ $k_4 = 0.0516\ s^{-1}$ $k_5 = 0.0019\ dm^3mol^{-1}s^{-1}$ $k_6 = 0.0006\ dm^3mol^{-1}s^{-1}$	12.7394

Table 5.5-5 Results of Run 5-17 to Run 5-24

Of the 8 runs between Run 5-17 to Run 5-24, 6 of them managed to obtain all the reactions that are present in RN2. Two of them, namely Run 5-17 and Run 5-20 only managed to discover 3 of the 4 reactions in in RN2. Both of them are missing the 4th reaction of RN2 and further analysis into the CRN structure that the runs uncovered shows that the missing reaction manifested differently in the two runs. The 4th reaction of RN2 and the one missing from both Run 5-17 and Run 5-20 is:

$$x_6 \rightarrow x_1 + x_4 \qquad \text{(Equation 5-6)}$$

In Run 5-17, there are three reactions that was included in the final CRN that are not part of RN2 and two of them are of particular interest, the 4th and 5th reaction of the CRN

$$\begin{gathered} 2x_6 \rightarrow x_1 + 2x_4 + x_5 \\ x_5 \rightarrow x_1 \end{gathered} \qquad \text{(Equation 5-7)}$$

These two reactions suggest that it is possible for the chemical species, x_5 to serve as a reaction intermediate and will react further to produce x_1. Combining both the reactions will give:

$$2x_6 \rightarrow 2x_1 + 2x_4 \qquad \text{(Equation 5-8)}$$

which is the RN2's 4th reaction but at a higher reaction order.

Similarly in Run 5-20, combining two of the additional reactions it suggested that are not part of RN2,

$$\begin{gathered} x_5 + x_6 \rightarrow 2x_1 + x_4 \\ x_1 \rightarrow x_5 \end{gathered} \qquad \text{(Equation 5-9)}$$

produces

$$x_6 \rightarrow x_1 + x_4 \qquad \text{(Equation 5-10)}$$

which is the 4th reaction of RN2.

The reaction pathway also consisted of using x_5 as reaction intermediate where it is first expended then the reproduce again in the following reaction through the dissociation of x_1.

Although both of the runs did not elucidate the 4th reaction of RN2, the other reactions their CRNs consist that are not part RN2 suggest the existence of the RN2's 4th reaction occurring in more than one reaction step. Thus, the two runs managed to discover the reaction indirectly and although it may not be correct or at the right order, it gives the user automated system (version 2) information that the reverse reaction of RN2's 3rd reaction is possible within the CRN.

Run 5-22 shows the worst performance in terms of fitness function among the runs from Run 5-17 to Run 5-24. The following Figure shows how its predicted concentration data fit the simulated concentration data for Run 5-22.

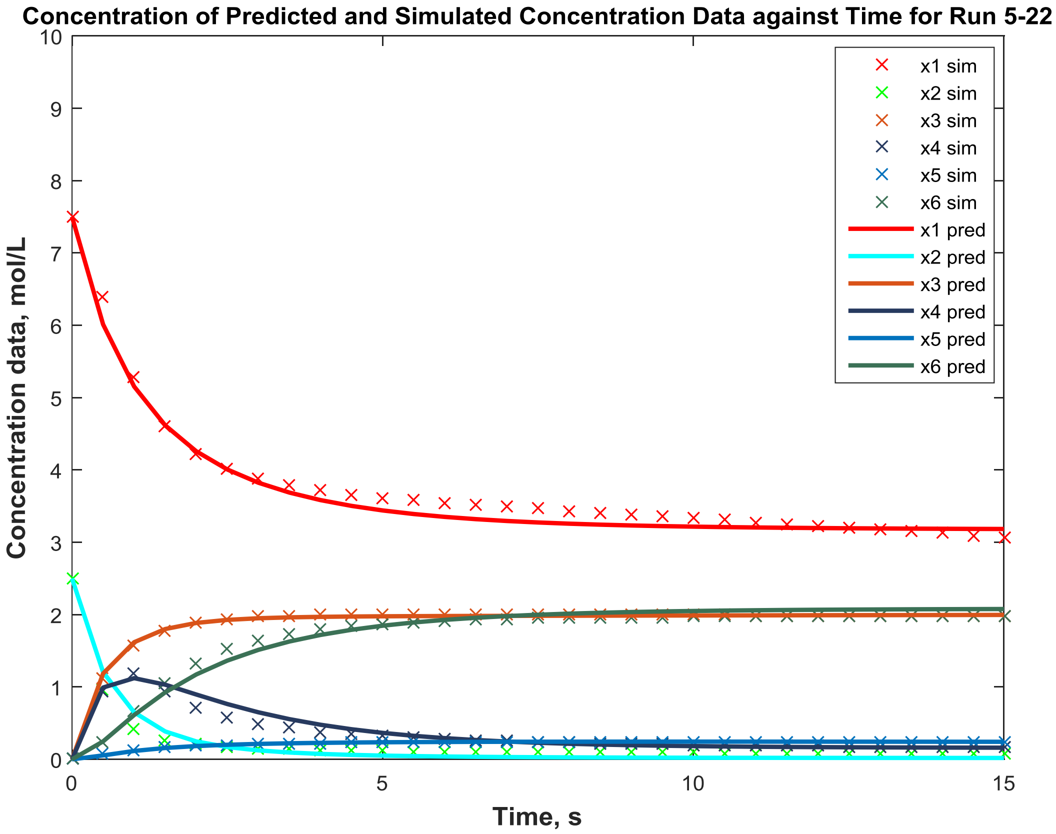


Figure 5.5-3 Plot of concentration of predicted and simulated concentration data against time for Run 5-22

It shows graphically in Figure 5.5-3 that the predicted concentration is fitted quite well to the simulated noisy data for even the worst performing run. This shows the capability of automated system (version 2) ability to elucidate the CRN correctly with an additional reaction not found in RN2 while providing a good fit to the concentration data.

Similar to Run 5-9 to 5-16, Run 5-17 to Run 5-24 are also subjected to the problem of 'overfitting' due to the noise in the data which causes the rise of the additional artificial reactions that are not part of RN2. Some of these reactions as discussed for Run 5-17 and Run 5-20 are the results of them modelling a reaction in RN2 as more than a single step reaction.

The runs from Run 5-9 to Run 5-24 shows that automated system (version 2) is able to elucidate the CRNs even when noise is introduced into the input concentration data. It is shown here again that even with reversible reaction in Run 5-17 to Run 5-24, it can still discover the required CRNs in 6 of the 8 runs. The 2 runs that did not elucidate all the reactions within RN2 are able to portray the missing reaction in the form of combination of two reactions. Automated system (version 2) is also shown to be more aggressive than the original automated system when fitting the data onto the provided concentration data, to the point the noise within the concentration data are being modelled into the final CRN structure. This caused 'overfitting' in the final CRN structures which led to the structures containing reactions that do not belong to CRN that describes the input concentration data.

5.6 Unmeasured chemical species

The previous sections had shown that automated system (version 2) had surpassed the original automated system in elucidating CRN from concentration data of participating chemical species with and without noise perturbation. It has consistently shown a better fitness function compared to the previous version of the automated system and is shown to be able to handle reversible reaction which is one of the major weakness of the original automated system.

The next part is to prove the capability of automated system (version 2)'s main design goal, which is for it to work even in the absence of some of the concentration data of the participating chemical species. The datasets used will be the same as the one used in the previous sections but with concentration data of some of the chemical species removed. Two chemical species from RN1 and two from RN2 will be hidden from automated system (version 2) when the run is performed. These chemical species act as reaction intermediates in their CRNs and chosen because it will provide a better challenge to the system's capability. If it is unable to elucidate the presence of the reaction intermediates correctly, automated system (version 2) will highly likely fail in predicting the actual CRN because it will not be able to model the concentration profiles of the CRN's final product. The system will first be tested against datasets that have no noise and only hidden concentration data of certain chemical species. Table 5.6-1 shows the simulation parameters used.

Run	Chemical Reaction Network	Batch	Gaussian Noise Standard Deviation	Unmeasured Chemical Species
5-25	RN1	1	0% of max range	x_3 and x_4
5-26	RN1	2	0% of max range	x_3 and x_4
5-27	RN1	3	0% of max range	x_3 and x_4
5-28	RN1	4	0% of max range	x_3 and x_4
5-29	RN2	1	0% of max range	x_3 and x_4
5-30	RN2	2	0% of max range	x_3 and x_4
5-31	RN2	3	0% of max range	x_3 and x_4
5-32	RN2	4	0% of max range	x_3 and x_4

Table 5.6-1 Run details for Run 5-25 to Run 5-32

The run parameters for automated system (version 2) is the same as those of Run 5-1 to Run 5-24. The fitness function used on runs with unmeasured chemical species will exclude the calculation of the fitness of the unmeasured chemical species.

The results of the run are compiled in Table 5.6-2 and Table 5.6-3:

Run	Best Performing Individual	Reaction Rate Constant	Fitness Function
5-25	$\begin{bmatrix} -2 & 1 & 0 & 0 & 0 \\ -1 & 0 & 1 & 0 & 0 \\ 0 & -1 & -1 & 0 & 1 \\ 0 & 0 & 1 & -1 & 0 \\ 1 & 1 & 0 & 0 & -1 \end{bmatrix}$	$k_1 = 0.1006\, dm^3mol^{-1}s^{-1}$ $k_2 = 0.2004\, s^{-1}$ $k_3 = 0.2034\, dm^3mol^{-1}s^{-1}$ $k_4 = 0.1762\, s^{-1}$ $k_5 = 0.0011\ \ s^{-1}$	0.0034
5-26	$\begin{bmatrix} -2 & 1 & 0 & 0 & 0 \\ -1 & 0 & 1 & 0 & 0 \\ 0 & 0 & -1 & 1 & 0 \\ 0 & -1 & 0 & -1 & 1 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix}$	$k_1 = 0.1001\, dm^3mol^{-1}s^{-1}$ $k_2 = 0.1994\, s^{-1}$ $k_3 = 0.1481\, s^{-1}$ $k_4 = 0.2701\, dm^3mol^{-1}s^{-1}$ $k_5 = 0$	0.0031
5-27	$\begin{bmatrix} -2 & 1 & 0 & 0 & 0 \\ -1 & 0 & 1 & 0 & 0 \\ 0 & 0 & -1 & 1 & 0 \\ 0 & -1 & 0 & -1 & 1 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix}$	$k_1 = 0.1002\, dm^3mol^{-1}s^{-1}$ $k_2 = 0.1998\, s^{-1}$ $k_3 = 0.1414\, s^{-1}$ $k_4 = 0.2638\, dm^3mol^{-1}s^{-1}$ $k_5 = 0$	0.0008

Run	Best Performing Individual	Reaction Rate Constant	Fitness Function
5-28	$\begin{bmatrix} -2 & 1 & 0 & 0 & 0 \\ -1 & 0 & 1 & 0 & 0 \\ 0 & 0 & -1 & 1 & 0 \\ 0 & -1 & 0 & -1 & 1 \\ 0 & 1 & 2 & -1 & -1 \end{bmatrix}$	$k_1 = 0.0993\ dm^3mol^{-1}s^{-1}$ $k_2 = 0.2004\ s^{-1}$ $k_3 = 0.1904\ s^{-1}$ $k_4 = 0.1875\ dm^3mol^{-1}s^{-1}$ $k_5 = 0.0057\ dm^3mol^{-1}s^{-1}$	0.0016

Table 5.6-2 Results for Run 5-25 to 5-28

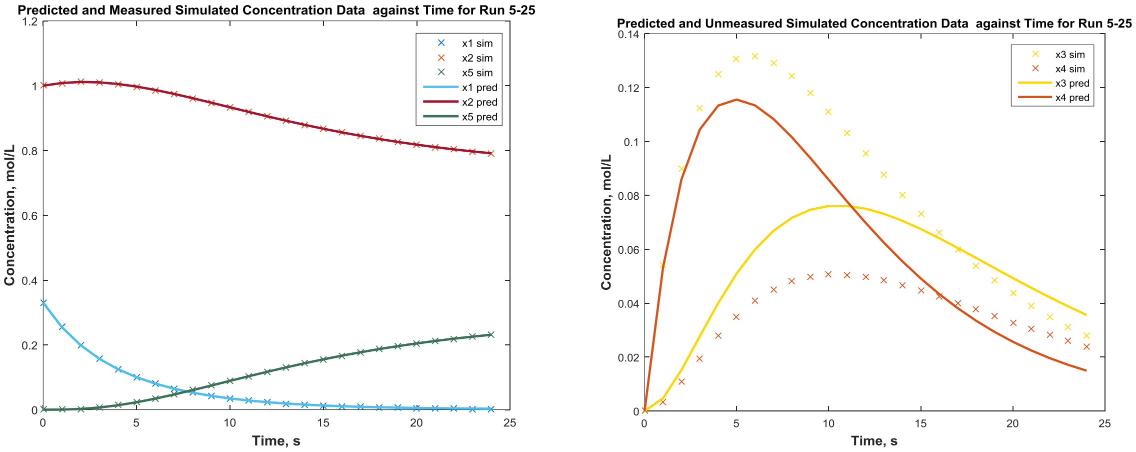
Run	Best Performing Individual	Reaction Rate Constant	Fitness Function
5-29	$\begin{bmatrix} -1 & -1 & 1 & 1 & 0 & 0 \\ -1 & 0 & 0 & -1 & 0 & 1 \\ 0 & 0 & 0 & -1 & -1 & 1 \\ 0 & -1 & 2 & 0 & 0 & 0 \\ 0 & 0 & 0 & -1 & -1 & 1 \\ -1 & 1 & 1 & 0 & 0 & 0 \end{bmatrix}$	$k_1 = 0.1737\ dm^3mol^{-1}s^{-1}$ $k_2 = 2033\ dm^3mol^{-1}s^{-1}$ $k_3 = 0.0821\ dm^3mol^{-1}s^{-1}$ $k_4 = 0.1508\ s^{-1}$ $k_5 = 0.2319\ dm^3mol^{-1}s^{-1}$ $k_6 = 0.0432\ mol^{-1}s^{-1}$	0.2511
5-30	$\begin{bmatrix} -1 & -1 & 1 & 1 & 0 & 0 \\ 0 & -1 & -1 & 0 & 1 & 0 \\ -1 & 0 & 0 & -1 & 0 & 1 \\ 0 & 0 & 0 & 1 & 1 & -1 \\ 0 & 2 & 0 & 1 & 2 & -2 \\ 1 & -1 & 2 & 1 & 0 & -1 \end{bmatrix}$	$k_1 = 0.1908\ dm^3mol^{-1}s^{-1}$ $k_2 = 0.0947\ dm^3mol^{-1}s^{-1}$ $k_3 = 0.1460\ dm^3mol^{-1}s^{-1}$ $k_4 = 0.0200\ s^{-1}$ $k_5 = 0.0304\ dm^3mol^{-1}s^{-1}$ $k_6 = 0\ .0014\ dm^3mol^{-1}s^{-1}$	1.0986
5-31	$\begin{bmatrix} -1 & -1 & 1 & 1 & 0 & 0 \\ -1 & 0 & 0 & -1 & 0 & 1 \\ 1 & 0 & 2 & -2 & 1 & 0 \\ 0 & 1 & 2 & 0 & 1 & -1 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}$	$k_1 = 0.2159\ dm^3mol^{-1}s^{-1}$ $k_2 = 0.1739\ dm^3mol^{-1}s^{-1}$ $k_3 = 0.1606\ dm^3mol^{-1}s^{-1}$ $k_4 = 0.0002\ s^{-1}$ $k_5 = 0$ $k_6 = 0$	0.8913
5-32	$\begin{bmatrix} -1 & -1 & 1 & 1 & 0 & 0 \\ -1 & 0 & 0 & -1 & 0 & 1 \\ 0 & -1 & 2 & 1 & 1 & -1 \\ 1 & -1 & 0 & -1 & 1 & 0 \\ 1 & 0 & 1 & -1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}$	$k_1 = 0.1936\ dm^3mol^{-1}s^{-1}$ $k_2 = 0.1925\ dm^3mol^{-1}s^{-1}$ $k_3 = 0.0872\ dm^3mol^{-1}s^{-1}$ $k_4 = 0.1044\ s^{-1}$ $k_5 = 0.0175\ s^{-1}$ $k_6 = 0$	0.7571

Table 5.6-3 Results for Run 5-29 to Run 5-32

From the first Table, Run 5-25 had been shown not be able to elucidate RN1 and its final CRN structure is missing two reactions from RN1. Run 5-26 to Run 5-28 managed to elucidate RN1 structure with Run 5-28 including a reaction that is not part of RN1 in its CRN structure.

The lack of information on concentration data x_3 and x_4 caused automated system (version 2) to ignore any errors between the predicted and simulated concentration data on x_3 and x_4. Without the need to match the concentration data on x_3 and x_4

caused the Tier 2 algorithm to focus mainly on minimising the errors between the predicted and simulated concentration data of measured chemical species. Without information on x_3 and x_4, also caused the system have no reference on the accuracy of its estimation for reaction rate constants of any reactions that consume or produces x_3 and x_4. It can only refers to concentration data of known measured chemical species as its guide which could be in the prior or next reaction step. Doing this increases the amount of inaccuracy in its estimation of the reaction rate constants which ultimately affects the CRN deduced in the run even when the datasets have no random error. Figure 5.6-1 shows the plot of the final CRN model predicted concentration data against measured and unmeasured chemical species for Run 5-25 as an example:


Predicted and Measured Simulated Concentration Data against Time for Run 5-25

Predicted and Unmeasured Simulated Concentration Data against Time for Run 5-25

Figure 5.6-1 Comparison between predicted concentration data against measured and unmeasured concentration data

Figure 5.6-1 shows clearly how automated system (version 2) fits the measured concentration data nearly perfectly but ignored entirely on how bad the fit is for the unmeasured concentration data. Automated system (version 2) basically ‘overfits’ the model to the measured concentration data given that it is not provided information on the unmeasured one. This weakness affects all the other runs, some more than others.

Run 5-29 to Run 5-32 did not manage to elucidate the true structure of RN2 with the worst run missing 2 RN2’s reactions. They are all affected by the weakness where automated system (version 2) overly prioritised the fitness of measured chemical species.

The Table below shows the comparison between the runs’ fitness against Run 5-1 to Run 5-8 where all the chemical species are measured. For the comparison to be meaningful, the fitness function is divided against the number of measured chemical species because for the fitness of unmeasured chemical species is not taken into account for the runs with unmeasured chemical species.

Run	Fitness Function (Run 5-1 to Run 5-8) per measured chemical species	Run	Fitness Function (Run 5-25 to Run 5-32) per measured chemical species
5-1	0	5-25	0.0011
5-2	0	5-26	0.0010
5-3	0	5-27	0.0003
5-4	0	5-28	0.0005
5-5	0	5-29	0.0837
5-6	0.0415	5-30	0.3662
5-7	0. 0415	5-31	0.2971
5-8	0	5-32	0.2524

Table 5.6-4 Comparison between Run 5-1 to Run 5-8 and Run 5-25 to Run 5-32

It can be concluded that the presence of unmeasured chemical species in the CRN affects the discovery of the CRN which caused automated system (version 2) unable to perform better estimation on the rate reaction constants. This leads to a poorer fitness function from the runs when compared to runs with all the chemical species measured. However, these runs prove that automated system (version 2) can be used to elucidate the CRN, albeit with a lower accuracy as compared to the original automated system.

5.7 Unmeasured chemical species with noise

Similar to the previous section, the performance of automated system (version 2) is tested against concentration data that is perturbed with noise. The datasets used is the same as for Run 5-9 to 5-16 but with the concentration data for x_3 and x_4 missing. Table 5.7-1 shows the details of the runs:

Run	Chemical Reaction Network	Batch	Gaussian Noise Standard Deviation	Unmeasured Chemical Species
5-33	RN1	1	4% of max range	x_3 and x_4
5-34	RN1	1	8% of max range	x_3 and x_4
5-35	RN1	2	4% of max range	x_3 and x_4
5-36	RN1	2	8% of max range	x_3 and x_4
5-37	RN1	3	4% of max range	x_3 and x_4
5-38	RN1	3	8% of max range	x_3 and x_4
5-39	RN1	4	4% of max range	x_3 and x_4
5-40	RN1	4	8% of max range	x_3 and x_4

Table 5.7-1 Run details for Run 5-33 to Run 5-40

Similarly, the run parameters for automated system (version 2) remains unchanged and is the same the previous section. The results from the runs are presented Table 5.7-2:

Run	Best Performing Individual	Reaction Rate Constant	Fitness Function
5-33	$\begin{bmatrix} -2 & 1 & 0 & 0 & 0 \\ -1 & 0 & 1 & 0 & 0 \\ 0 & 0 & -1 & 1 & 0 \\ 0 & -1 & 0 & -1 & 1 \\ 2 & 0 & 0 & 1 & -1 \end{bmatrix}$	$k_1 = 0.1238\ dm^3 mol^{-1} s^{-1}$ $k_2 = 0.2023\ s^{-1}$ $k_3 = 0.1852\ s^{-1}$ $k_4 = 0.2201\ dm^3 mol^{-1} s^{-1}$ $k_5 = 0.0040\ s^{-1}$	0.1161
5-34	$\begin{bmatrix} -2 & 1 & 0 & 0 & 0 \\ 0 & -1 & 0 & -1 & 1 \\ -1 & 0 & 0 & 1 & 0 \\ 0 & -1 & 0 & 2 & 0 \\ -1 & 1 & 0 & -1 & 0 \end{bmatrix}$	$k_1 = 0.4581\ dm^3 mol^{-1} s^{-1}$ $k_2 = 0.1422\ dm^3 mol^{-1} s^{-1}$ $k_3 = 0.0561\ s^{-1}$ $k_4 = 0.0072\ s^{-1}$ $k_5 = 0.5303\ s^{-1}$	0.5762
5-35	$\begin{bmatrix} -2 & 1 & 0 & 0 & 0 \\ -1 & 0 & 0 & 1 & 0 \\ 0 & -1 & -1 & 0 & 1 \\ 1 & 0 & -1 & 0 & 0 \\ 0 & 0 & 1 & -1 & 0 \end{bmatrix}$	$k_1 = 0.1049\ dm^3 mol^{-1} s^{-1}$ $k_2 = 0.2378\ s^{-1}$ $k_3 = 0.2860\ dm^3 mol^{-1} s^{-1}$ $k_4 = 0.0337\ s^{-1}$ $k_5 = 0.1430\ s^{-1}$	0.1698
5-36	$\begin{bmatrix} -2 & 1 & 0 & 0 & 0 \\ 2 & 0 & -1 & -1 & 0 \\ 0 & -1 & -1 & 0 & 1 \\ -1 & 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & -1 & 0 \end{bmatrix}$	$k_1 = 0.1111\ dm^3 mol^{-1} s^{-1}$ $k_2 = 0.0682\ dm^3 mol^{-1} s^{-1}$ $k_3 = 0.2333\ dm^3 mol^{-1} s^{-1}$ $k_4 = 0.2035\ s^{-1}$ $k_5 = 0.2109\ s^{-1}$	0.1794

Run	Best Performing Individual	Reaction Rate Constant	Fitness Function
5-37	$\begin{bmatrix} -2 & 1 & 0 & 0 & 0 \\ 0 & -1 & -1 & 0 & 1 \\ 0 & 0 & 1 & -1 & 0 \\ -1 & 0 & 0 & 1 & 0 \\ -1 & 1 & 0 & -1 & 0 \end{bmatrix}$	$k_1 = 0.0836\, dm^3mol^{-1}s^{-1}$ $k_2 = 0.2280\ dm^3mol^{-1}s^{-1}$ $k_3 = 0.1460 s^{-1}$ $k_4 = 0.2017\, s^{-1}$ $k_5 = 0.0298\ dm^3mol^{-1}s^{-1}$	0.1991
5-38	$\begin{bmatrix} -2 & 1 & 0 & 0 & 0 \\ -1 & 0 & 1 & 0 & 0 \\ 1 & 0 & -1 & 0 & 0 \\ 0 & -1 & -1 & 0 & 1 \\ 1 & -1 & 1 & 0 & 0 \end{bmatrix}$	$k_1 = 0.1766\, dm^3mol^{-1}s^{-1}$ $k_2 = 0.1693\, s^{-1}$ $k_3 = 0.0051\, s^{-1}$ $k_4 = 0.1012\, dm^3mol^{-1}s^{-1}$ $k_5 = 0.0056\, s^{-1}$	0.7206
5-39	$\begin{bmatrix} -2 & 1 & 0 & 0 & 0 \\ -1 & 0 & 1 & 0 & 0 \\ 0 & 0 & -1 & 1 & 0 \\ 0 & -1 & 0 & -1 & 1 \\ 1 & 0 & 0 & 2 & -1 \end{bmatrix}$	$k_1 = 0.1189\, dm^3mol^{-1}s^{-1}$ $k_2 = 0.1967\, s^{-1}$ $k_3 = 0.1918\, s^{-1}$ $k_4 = 0.1964\, dm^3mol^{-1}s^{-1}$ $k_5 = 0.0107\, s^{-1}$	0.2629
5-40	$\begin{bmatrix} 0 & -1 & 2 & 0 & 0 \\ -1 & 0 & 0 & 1 & 0 \\ 0 & -1 & -1 & 0 & 1 \\ 0 & 1 & 0 & -2 & 0 \\ 2 & 0 & 2 & -1 & -1 \end{bmatrix}$	$k_1 = 0.0120\, s^{-1}$ $k_2 = 0.2798\, s^{-1}$ $k_3 = 0.3420\, dm^3mol^{-1}s^{-1}$ $k_4 = 0.5178\, dm^3mol^{-1}s^{-1}$ $k_5 = 0.0777\, dm^3mol^{-1}s^{-1}$	1.2473

Table 5.7-2 Results for Run 5-33 to Run 5-40

Only 2 of the runs managed to elucidate the actual CRN structure of RN1 (Run 5-33 and Run 5-39) with additional reaction that are not part of RN1. The other runs had various level of failure in elucidating the reactions in RN1, ranging from missing all four of the reactions, Run 5-40 to missing two reactions, Run 5-34 and Run 5-38. The performance of the worst run, Run 5-40 by fitness function and missing reactions can be seen in Figure 5.7-1:

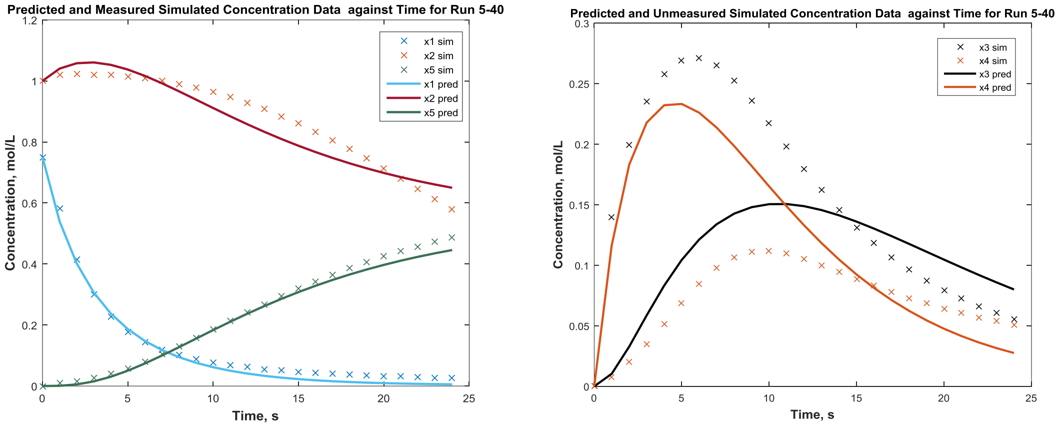


Figure 5.7-1 Comparison between predicted concentration data against measured and unmeasured concentration data

Similar to the figures presented for Run 5-25, Run 5-40 also shows automated system (version 2) ‘overfitting’ attempt on the measured concentration data. This is exacerbated by the noise that is present in the system as can be seen by the poor fit of chemical species x_1.

In Run 5-34, chemical species x_3 is not even used as reactant or product. As the existence or accuracy of the concentration profile of x_3 is not even considered by automated system (version 2) because it is unmeasured, the system can neglect it altogether if it can fit the predicted data of the measured chemical species to the input noisy concentration data. This again shows the effect ‘overfitting’ on the system.

Comparison between the fitness function per measured chemical species for the runs using all measured chemical species concentration data (Run 5-9 to Run 5-16) and those with the presence of unmeasured chemical species (Run 5-33 to Run 5-40) is shown in Table 5.7-3:

Run	Fitness Function (Run 5-9 to Run 5-16) per measured chemical species	Run	Fitness Function (Run 5-33 to Run 5-40) per measured chemical species
5-9	0.0508	5-33	0.0387
5-10	0.5337	5-34	0.1921
5-11	0.1729	5-35	0.0566
5-12	0.1395	5-36	0.0598
5-13	0.1168	5-37	0.0664
5-14	1.2520	5-38	0.2402
5-15	0.2635	5-39	0.0876
5-16	0.8823	5-40	0.4158

Table 5.7-3 Comparison between Run 5-9 to Run 5-16 and Run 5-33 to Run 5-40

From the comparison, it can be seen that when automated system (version 2) is used on unmeasured chemical species it can achieved better fitness. This is as expected as the number of chemical species that automated system (version 2) has to fit to the input concentration data is less for the runs with unmeasured chemical species. Although the runs (Run 5-33 to Run 5-40) achieve a better fitness compared to their counterpart consistently, they are much worse when comparing the final elucidated CRN results with the Run 5-9 to Run 5-16. This again shows how the Run 5-33 to Run 5-40 ‘overfitting’ their predicted concentration data to the measured chemical species.

Run 5-41 to Run 5-48 is run to test automated system (version 2) against noisy concentration data from RN2 with two unmeasured chemical species, x_3 and x_4. The Table 5.7-4 shows the parameters for the datasets used for the runs. The concentration data used the same as those used for Run 5-17 to Run 5-24.

Run	Chemical Reaction Network	Batch	Gaussian Noise Standard Deviation	Unmeasured Chemical Species
5-41	RN2	1	4% of max range	x_3 and x_4
5-42	RN2	1	8% of max range	x_3 and x_4
5-43	RN2	2	4% of max range	x_3 and x_4
5-44	RN2	2	8% of max range	x_3 and x_4
5-45	RN2	3	4% of max range	x_3 and x_4
5-46	RN2	3	8% of max range	x_3 and x_4
5-47	RN2	4	4% of max range	x_3 and x_4
5-48	RN2	4	8% of max range	x_3 and x_4

Table 5.7-4 Run parameters for Run 5-41 to Run 5-48

Run parameters for automated system (version 2) are the same as the previous sections for fair comparison. Table 5.7-5 shows the performance of the runs.

Run	Best Performing Individual	Reaction Rate Constant	Fitness Function
5-41	$\begin{bmatrix} -1 & -1 & 1 & 1 & 0 & 0 \\ 0 & -1 & -1 & 0 & 1 & 0 \\ -1 & 0 & 0 & -1 & 0 & 1 \\ 0 & 0 & 2 & 2 & -1 & -1 \\ 1 & 0 & 2 & -2 & 1 & 0 \\ 2 & -1 & 0 & -1 & 0 & 0 \end{bmatrix}$	$k_1 = 0.2038\, dm^3mol^{-1}s^{-1}$ $k_2 = 0.0789\, dm^3mol^{-1}s^{-1}$ $k_3 = 0.1387\, dm^3mol^{-1}s^{-1}$ $k_4 = 0.0504\, dm^3mol^{-1}s^{-1}$ $k_5 = 0.0217\, dm^3mol^{-1}s^{-1}$ $k_6 = 0.0012\, dm^3mol^{-1}s^{-1}$	0.1192
5-42	$\begin{bmatrix} -2 & 1 & 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & -2 & 0 & 1 \\ 0 & -1 & 2 & 0 & 0 & 0 \\ -1 & 1 & 1 & 0 & 0 & 0 \\ 0 & -2 & 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & -1 & 1 & 0 \end{bmatrix}$	$k_1 = 0.1258\, dm^3mol^{-1}s^{-1}$ $k_2 = 0.1784\, dm^3mol^{-1}s^{-1}$ $k_3 = 0.1769\, s^{-1}$ $k_4 = 0.1304\, s^{-1}$ $k_5 = 0.1752\, dm^3mol^{-1}s^{-1}$ $k_6 = 0.1174\, s^{-1}$	0.5620

Run	Best Performing Individual	Reaction Rate Constant	Fitness Function
5-43	$\begin{bmatrix} -1 & -1 & 1 & 1 & 0 & 0 \\ 0 & -1 & -1 & 0 & 1 & 0 \\ -1 & 0 & 0 & -1 & 0 & 1 \\ 0 & 0 & -1 & 2 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 & -1 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}$	$k_1 = 0.1856\ dm^3mol^{-1}s^{-1}$ $k_2 = 0.1276\ dm^3mol^{-1}s^{-1}$ $k_3 = 0.1614\ dm^3mol^{-1}s^{-1}$ $k_4 = 0.0671\ s^{-1}$ $k_5 = 0.0340\ dm^3mol^{-1}s^{-1}$ $k_6 = 0$	2.7019
5-44	$\begin{bmatrix} -1 & -1 & 1 & 1 & 0 & 0 \\ 0 & -1 & -1 & 0 & 1 & 0 \\ -1 & 0 & 0 & -1 & 0 & 1 \\ 2 & 1 & 2 & 0 & -1 & -1 \\ -1 & 2 & -1 & 0 & 0 & 0 \\ 1 & 1 & -1 & -1 & 0 & 0 \end{bmatrix}$	$k_1 = 0.1862\ dm^3mol^{-1}s^{-1}$ $k_2 = 0.0861\ dm^3mol^{-1}s^{-1}$ $k_3 = 0.1427\ dm^3mol^{-1}s^{-1}$ $k_4 = 0.0779\ dm^3mol^{-1}s^{-1}$ $k_5 = 0.0458\ dm^3mol^{-1}s^{-1}$ $k_6 = 0.0348\ dm^3mol^{-1}s^{-1}$	10.9617
5-45	$\begin{bmatrix} -1 & -1 & 1 & 1 & 0 & 0 \\ -1 & 0 & 0 & -1 & 0 & 1 \\ 0 & 0 & 0 & 1 & 1 & -1 \\ 0 & 1 & 2 & -1 & 0 & 0 \\ 1 & 1 & -1 & -1 & 0 & 0 \\ 1 & 0 & 2 & -2 & 1 & 0 \end{bmatrix}$	$k_1 = 0.2273\ dm^3mol^{-1}s^{-1}$ $k_2 = 0.1710\ dm^3mol^{-1}s^{-1}$ $k_3 = 0.0002\ s^{-1}$ $k_4 = 0.0119\ s^{-1}$ $k_5 = 0.0615\ dm^3mol^{-1}s^{-1}$ $k_6 = 0.1662\ dm^3mol^{-1}s^{-1}$	2.0820
5-46	$\begin{bmatrix} -1 & -1 & 1 & 1 & 0 & 0 \\ -1 & 0 & 0 & -1 & 0 & 1 \\ 1 & -1 & 0 & -1 & 1 & 0 \\ 2 & 1 & -1 & 0 & 0 & -1 \\ 1 & 1 & 0 & -2 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}$	$k_1 = 0.2643\ dm^3mol^{-1}s^{-1}$ $k_2 = 0.1349\ dm^3mol^{-1}s^{-1}$ $k_3 = 0.0835\ dm^3mol^{-1}s^{-1}$ $k_4 = 0.0115\ dm^3mol^{-1}s^{-1}$ $k_5 = 0.0469\ dm^3mol^{-1}s^{-1}$ $k_6 = 0$	4.8747
5-47	$\begin{bmatrix} -1 & -1 & 1 & 1 & 0 & 0 \\ -1 & 0 & 0 & -1 & 0 & 1 \\ 0 & 1 & -2 & 0 & 0 & 0 \\ 1 & 0 & 2 & -2 & 1 & 0 \\ 1 & 0 & 1 & 0 & 1 & -1 \\ 0 & 2 & 0 & -1 & 0 & 0 \end{bmatrix}$	$k_1 = 0.2416\ dm^3mol^{-1}s^{-1}$ $k_2 = 0.1955\ dm^3mol^{-1}s^{-1}$ $k_3 = 0.0002\ dm^3mol^{-1}s^{-1}$ $k_4 = 0.1798\ dm^3mol^{-1}s^{-1}$ $k_5 = 0.0002\ dm^3mol^{-1}s^{-1}$ $k_6 = 0.0004\ s^{-1}$	2.7871
5-48	$\begin{bmatrix} -1 & -1 & 1 & 1 & 0 & 0 \\ -1 & 0 & 0 & -1 & 0 & 1 \\ 1 & -1 & 0 & -1 & 1 & 0 \\ 1 & 1 & -1 & -1 & 0 & 0 \\ 0 & 1 & 0 & -2 & 2 & 0 \\ 0 & 2 & 0 & -1 & 0 & 0 \end{bmatrix}$	$k_1 = 0.2176\ dm^3mol^{-1}s^{-1}$ $k_2 = 0.1576\ dm^3mol^{-1}s^{-1}$ $k_3 = 0.0973\ dm^3mol^{-1}s^{-1}$ $k_4 = 0.0596\ dm^3mol^{-1}s^{-1}$ $k_5 = 0.0235\ dm^3mol^{-1}s^{-1}$ $k_6 = 0.0082\ s^{-1}$	4.8446

Table 5.7-5 Results for Run 5-41 to Run 5-48

The results of Run 5-41 to Run 5-48 are very similar to those in Run 5-33 to Run 5-40 where none of the runs' CRNs match that of RN2. Comparing it against Run 5-17 to Run 5-24, these runs perform worse off because among the Run 5-17 to Run 5-24, 6 out of 8 managed to obtain all the reactions in RN2. No further discussion is required as the reason for the failure is the same as it is with Run 5-33 to Run 5-40 which are discussed previously, which is mainly 'overfitting' to measured concentration data.

Comparisons of the fitness function per measured chemical species between Run 5-17 to Run 5-24 (no unmeasured chemical species) and Run 5-41 to Run 5-48 (two unmeasured chemical species) are shown in Table 5.7-6:

Run	Fitness Function (Run 5-17 to Run 5-24) per measured chemical species	Run	Fitness Function (Run 5-41 to Run 5-48) per measured chemical species
5-17	0.1242	5-41	0.0298
5-18	0.4063	5-42	0.1405
5-19	0.5231	5-43	0.6755
5-20	2.2945	5-44	2.7404
5-21	0.3500	5-45	0.5205
5-22	2.3316	5-46	1.2187
5-23	0.3255	5-47	0.6968
5-24	2.1232	5-48	1.2112

Table 5.7-6 Comparison between Run 5-17 to Run 5-24 and Run 5-41 to Run 5-48

Here it can be seen that the fitness functions between the runs with unmeasured chemical species and those without are similar, with some better and some worse off than their counterpart. This could mean that automated system (version 2) that the missing concentration data can affect the reaction rate approximation to the point it suffers even when it does not need to consider how accurate it predicts the unmeasured chemical species.

5.8 Summary

In his chapter, the weaknesses of the original automated system for discovery of CRN from concentration data of chemical species are addressed through the introduction of the second tier optimisation routine. The new automated system, referred as automated system (version 2) is tested against noiseless and noisy data and has shown it can handle reversible reactions that is present in Reaction Network 2 (RN2). Further testing on data with unmeasured chemical species had shown the system can be used to elucidate the correct CRNs even when it has missing data. As expected, automated system (version 2)'s performance with unmeasured chemical species is worse off when compared against its performance with measured chemical species in terms of elucidating the correct CRNs.

Automated system (version 2) has repeatedly shown that it tends to fit to data more aggressively as compared to the previous version causing more 'overfitting' to occur which in turn generates false reactions that do not occur in the actual CRN. When concentration data of unmeasured chemical species are used, the automated system will only fit its predicted concentration data to the input concentration data for the measured chemical species, ignoring the unmeasured chemical species. At the extreme case, the automated system can even exclude existence of the unmeasured chemical species in its final deduced CRN.

Although the developed automated system for CRN elucidation can now be used for reversible reactions and unmeasured chemical species, it still suffers from ‘overfitting’ and the effect is more severe when dealing with unmeasured chemical species. The next chapter will modify the system further to take into account the ‘overfitting’ issue.

Chapter 6. Automated System with Multi Objective Optimisation for Elucidation of Chemical Reaction Network

6.1 Overview

This chapter presents another addition to the automated system of chemical reaction network (CRN) elucidation. A multi-objective optimisation routine is included into the system so it can now push its search directions towards diversity instead of just focussing on fitness. The top most occurring reactions from the population are then gathered and sorted. The top 10 most occurring reactions will be extracted and considered. The system is tested with the same datasets presented in the previous chapters and the results are discussed.

6.2 Introduction

‘Overfitting’ is a problem that is caused by modeller, in this case the automated system for identification of CRN models modelling in the noise from the input concentration data when building the CRN. In order to model in the noise, the automated system will sometimes need more reactions than the actual CRN or even substitute it with some other reactions that can model the noise better. The effect is a high level of fit for the predicted concentration data to the input concentration data but at the cost of not elucidating the actual CRN or including additional reactions that are not part of the actual CRN. When unmeasured chemical species is present in the input concentration data, the issue becomes more severe because the automated system will only focus on fitting the measured chemical species while ignoring the unmeasured chemical species because it has no information on the concentration profiles of the unmeasured chemical species. To counter the ‘overfitting’ issue, the version 2 of the automated system will be modified further to implement multi-objective algorithm into the genetic algorithm (GA) used in the system.

6.3 Multi-objective optimisation

Multi-objective optimisation is as the name suggests, optimise a function or problem with more than one objective (Deb, 2001). In the case of single objective optimisation, there is only one objective to achieve and the optimiser will only focus on achieving the optimal solution in relation to that objective. However, when more than one objective is desired, the optimal solution becomes more difficult to be determined. These optimisation objectives usually are in competition with each other for example, consider the statement below,

The more powerful a computer is, the more expensive it is to build it.

Based on the statement, a person wanting to buy a cheap and powerful computer will have the following optimisation problem:

Optimisation problem	*: Buy cheap and powerful computer*
First objective	*: Low computer price*
Second objective	*: High computer power*

There is no one optimal solution for this as it is not possible to have the most powerful computer with the lowest cost based on the statement that more powerful computer costs more. So it becomes a multi-objective optimisation problem where the optimal results is a range of results lying on the Pareto front. Results on the Pareto front are results that are not better off or worse off than each other and cannot be improved further without sacrificing any of the objectives. Continue on from the computer example, a plot of cost against power of the computer can be plotted like the one shown in Figure 6.3-1.

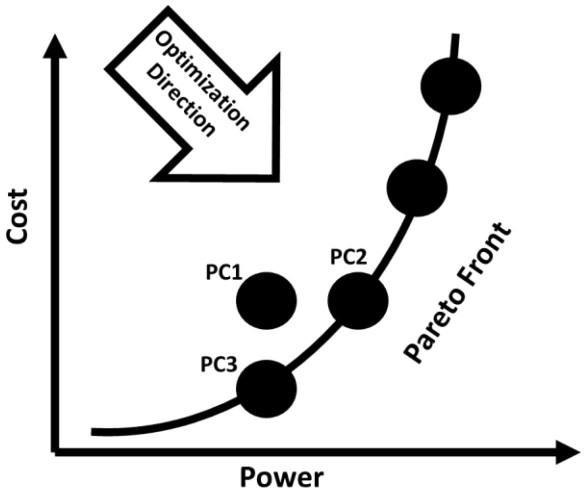


Figure 6.3-1 Plot showing Pareto Front and optimisation direction

The goal of getting high powered computer at a low cost will force the optimisation to proceed to the bottom of the plot for better price and to the right for higher power which when combined causes the direction of the optimisation towards lower right of the plot. Once it can increase the power no further at the same cost or when it can reduce the price no further at the same level of power, the particular computer model can be considered to be lying on the Pareto Front. Comparing PC2 and PC3 from the Figure, it can be seen none dominates the other in term of the objective. PC3 is cheaper than PC2 but PC2 is more powerful than PC3 and choosing between the two will involve trade-off between power and price. However, PC1 when compared against PC2 and PC3 is inferior and can still be optimised. This is because at the same level of power, PC3 is cheaper than PC1 and at the same price, PC2 is more powerful than PC1. Those computer models that lies on the Pareto Front can be considered as Pareto optimal results. Therefore, the final result of a multi-objective optimisation is not a single most optimal result but a group of Pareto optimal results which are not any way inferior from one and another.

In the field of optimisation through evolutionary algorithm which GA is part of, there exists a selection of multi-objective evolutionary algorithm (MOEA) to choose from. Examples of MOEA are such as Multi Objective Genetic Algorithm (Murata and Ishibuchi, 1995), Differential Evolution Multi Objective (Mlakar et. al., 2015), Multi Objective Particle Swarm Optimisation (Reyes-Sierra and Coello, 2006) and Elitist Non-dominated Sorting Genetic Algorithm (NSGA-II) by Deb et al. (2002).

6.4 Elitist Non-dominated Sorting Genetic Algorithm (NSGA-II)

NSGA-II, an upgraded version of the NSGA (Srinivas and Deb, 1995) is one of the most popular MOEA and its framework is used most from the available MOEA (Zhou et. al., 2011). The paper describing the design of NSGA-II (Deb et al., 2002) has been cited in more than 4000[1] publications and patents as of March 2016. It had been used in numerous applications such as antenna array design (Pandura et al., 2006), DNA sequence design (Shin et al., 2005), robot grippers design (Saravanan et al., 2009), dynamic controller design (Wozniak, 2011) and optimisation of petroleum processing units (Ivanov and Ray, 2014).

The NSGA-II algorithm is employed after the reproduction stage of the evolutionary algorithm. The first step of NSGA-II is combining both parent and the newly reproduced population and assigning Pareto fronts to them. The Pareto optimal individuals are

[1] Based on the “cited by” data at where the paper is located in the IEEE website address: http://ieeexplore.ieee.org/xpl/abstractCitations.jsp?tp=&arnumber=996017&url=http%3A%2F%2Fieeexplore.ieee.org%2Fxpls%2Fabs_all.jsp%3Farnumber%3D996017

considered to be on Pareto Front 1 and they are not inferior or dominated by each other. Then, the individuals from Pareto Front 1 are removed from considerations and the current Pareto optimal of the leftover individuals will be assigned Pareto Front 2. Those on Pareto Front 2 are then removed and the process repeats until no individual is left. Those on Pareto Front 2 is dominated by those on Pareto Front 1 and so on. Figure 6.4-1 shows how the results of the ranking by Pareto Fronts will looked like graphically for the minimisation of Objective 1 and maximisation of Objective 2.

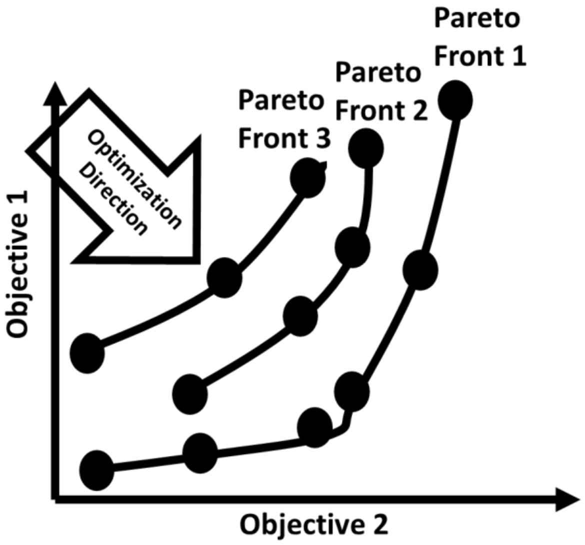


Figure 6.4-1 Plot showing multi-level Pareto Front

Once the individuals’ Pareto Fronts are determined, each of the individuals’ crowding distance are calculated. Crowding distance is defined as the distance between individuals’ two immediate neighbours of the same Pareto Front on the scatter plot of the objective functions. The individuals at the edge will be assigned crowding distance of infinity.

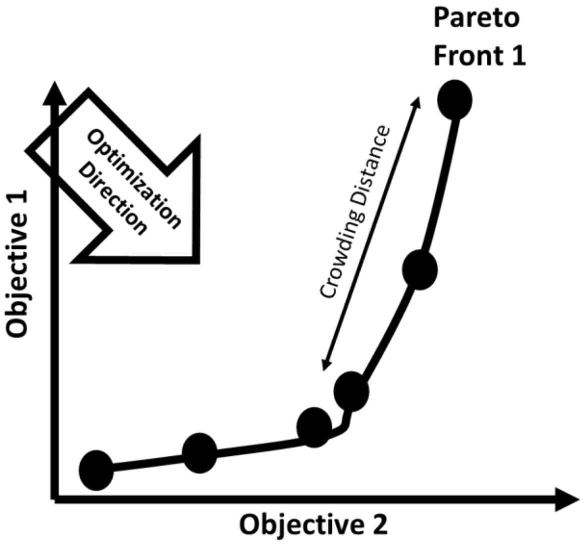


Figure 6.4-2 Plot showing how crowding distance is determined

The individuals are then sorted according to their Pareto Front with those with a smaller front index are more desirable than those with a larger front index. Individuals on the same Pareto Front are then sorted against each other according to the crowding distance of the individual with larger crowding distance more desirable. In short, the sorted list of individuals will have individual nearer to the first Pareto front and larger crowding factor nearer to the top of the list (more desirable). The top half of the list of individuals is then passed on to the next generation while the bottom half is discarded. With this the best of the parent generation and the newly created population are passed on to the next generation. The best here refers to low Pareto Front index and those at the same Front, higher crowding factor. Figure 6.4-3 gives a graphical view on how child generation is decided through the NSGA-II algorithm.

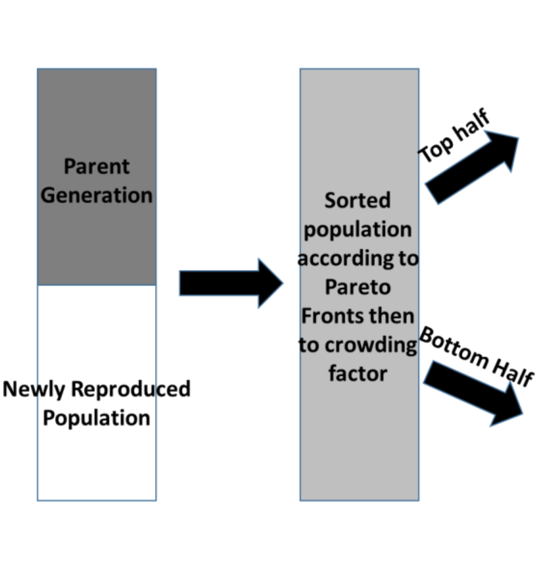

Parent Generation
Newly Reproduced Population
Sorted population according to Pareto Fronts then to crowding factor
Top half
Child Generation
Bottom Half
Discarded Population

Figure 6.4-3 NSGA-II sorting and elimination of population

By keeping individuals nearer to the Pareto optimal solutions in the GA ensured that subsequent generations will push the Pareto Front further and further providing better Pareto optimal solutions. This is because individuals that are nearer to the Pareto optimal solutions are fitter than those further away and when reproduction between individuals that have good genes, chances are individual with better genes will appear. Those with better genes or fitter individuals in subsequent generations may even surpassed the performance of their predecessors and may be placed beyond the Pareto Front of their predecessors, pushing the Pareto Front further. The large crowding factor is used to increase diversity in the population with the idea that individuals with larger crowding distance are more isolated and thus more unique compared to the rest. Unique individual may contain gene that are not present in the other individual and retaining them helps increase diversity in the population.

The push for diversity through the use of crowding distance is what NSGA-II useful to overcome the problem of ‘overfitting’ in the automated system designed in this work.

6.5 Modification done on the automated system

Instead of hunting for the best CRN possible based on the fitness function only, the goal of the automated system is changed to obtaining a group of CRNs from the results that are of equal value (Pareto optimal). To do so, it is to the best interest of the automated system to be designed to gain as much diversity in the reaction as possible. The objective functions for the automated system thus become

Objective 1 : Minimise fitness function

Objective 2 : Maximise diversity in the population

Increment of diversity in the NSGA-II is done through crowding distance. This particular measure is not constant and changes one generation to another depending on the individual immediate neighbours and therefore cannot be used as an objective function. To describe the uniqueness of the CRNs, average relative reactants' molecular weight (*ARRMW*) is proposed. Relative reactants' molecular weight (*RRMW*) is calculated by dividing the molecular weight of two reactants in a reaction. *ARRMW* is the average of all the *RRMW* of all the reactions in the CRN. *RRMW* will be designated as 0 if there is only one reactant in the reaction.

$$if\ m_{reactant\ 1} \neq m_{reactant\ 2}, \quad RRMW = \frac{m_{reactant\ 1}}{m_{reactant\ 2}} \quad where\ m_{reactant\ 1} > m_{reactant\ 2}$$

$$if\ no\ m_{reactant\ 2}, \quad RRMW = 0$$

$$ARRMW = \frac{\sum_1^{N_r} RRMW}{N_r}$$ Equation 6-1

where

$M_{reactant\ 1}$ refers to molecular weight of the heavier reactant

$M_{reactant\ 2}$ refers to molecular weight of the lighter reactant

ARRMW does not serve as a conflicting objective function for the fitness function used for the automated system as *ARRMW* relationship to fitness function is not straight forward. It is however a good measure on how unique a CRN is as the *ARRMW* of will be different from one CRN to another if the reactions within the CRNs are not the same. Having more reactions or less reactions within the CRN will affect the *ARRMW* as well. Therefore, to increase diversity, the goal is populate the GA with as many CRNs with different *ARRMW* as possible. Due to the structure of the goal, Pareto optimal solutions for the system is defined as lowest possible fitness functions for any particular *ARRMW*. Note that this is not how normally Pareto optimal solutions is described as but the same terminology as original NSGA-II is kept to reduce the need to introduce new terms.

NSGA-II algorithm will need to modify slightly in light of the second objective function that does not follow the typical maximisation or minimisation problem. The method to assign the individuals into Pareto Fronts is changed to assigning Pareto Front 1 to any individuals who has the lowest fitness function for a particular *ARRMW*. Once determined, these individuals are removed and the Pareto optimal individuals of the leftover individuals will be assigned Pareto Front 2 and so on. The only difference in the individuals within the same Pareto Front in this case is only that they have different set of *ARRMW* and it does not matter if their fitness function is higher or lower than each other. This is different from the original NSGA-II algorithm where each of the individuals in a Pareto Front is not inferior to each other when comparing both of the objective functions. Apart from the assignment of Pareto Fronts, the other part of the algorithm is the same as the original NSGA-II, including calculation of crowding distance which would help in further increase the diversification of the results. The Figure 6.5-1 gives a better picture on how the shape of the Pareto Front had changed and the direction of the optimisation of the automated system.

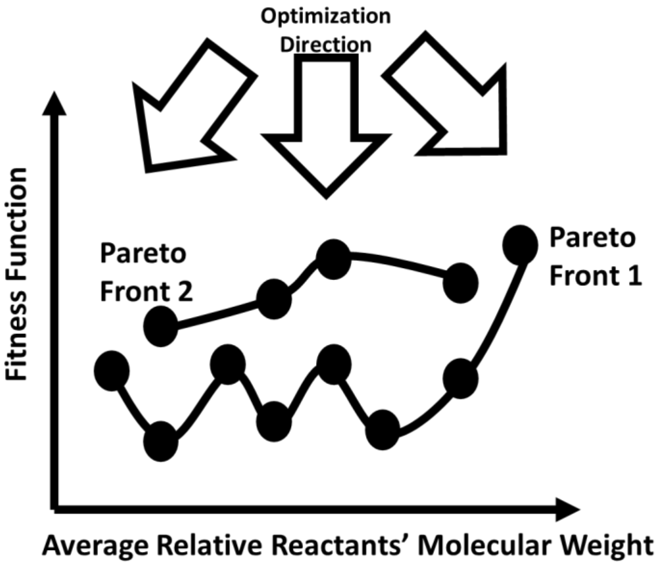


Figure 6.5-1 Pareto Front for the designed automated system

In this case, the shape of the Pareto front can be extremely irregular and the optimisation direction is in general downwards for lower fitness function and outwards for more diversity.

Figure 6.5-2 shows the automated system flow chart with the NSGA-II implemented.

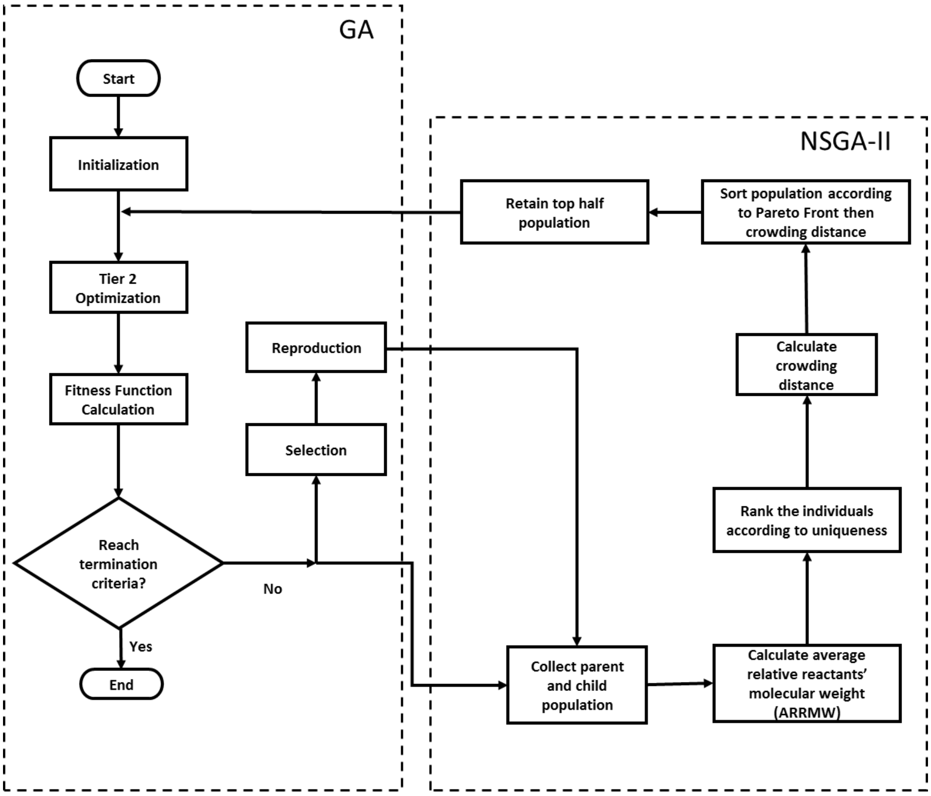

GA
Start
Initialization
Tier 2 Optimization
Fitness Function Calculation
Reach termination criteria?
Yes
End
No
Selection
Reproduction
NSGA-II
Retain top half population
Sort population according to Pareto Front then crowding distance
Calculate crowding distance
Rank the individuals according to uniqueness
Collect parent and child population
Calculate average relative reactants' molecular weight (ARRMW)

Figure 6.5-2 Flowchart for automated system (NSGA-II)

This version of the automated system will be referred to as automated system (NSGA-II) for here on.

6.6 Test against runs with unmeasured chemical species

To prove the feasibility of the system, it is tested against the datasets with unmeasured chemical species used in the previous chapter. The goal of the test is check on how diverse the population of CRN is at the final generation as compared to the runs conducted by automated system (version 2). The second goal is to peruse through the individuals at the Pareto Front of the final generation and obtain a list of reactions and the number of times they appear.

The datasets used for this run is the same as the one used in the previous chapter for unmeasured chemical species with Gaussian noise with mean of 0 and standard deviation equal to 8% of the maximum range of the concentration data.

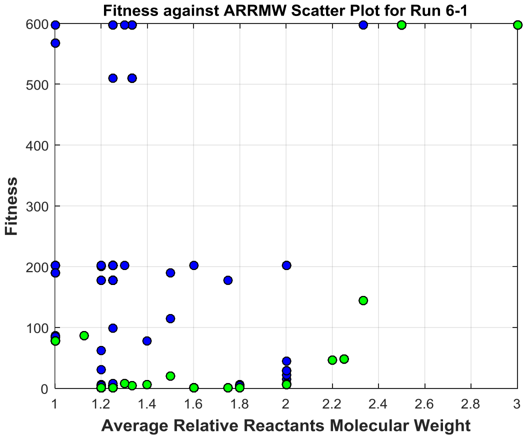
Run	Chemical Reaction Network	Batch	Gaussian Noise Standard Deviation	Unmeasured Chemical Species
6-1	RN1	1	8% of max range	x_3 and x_4
6-2	RN1	2	8% of max range	x_3 and x_4
6-3	RN1	3	8% of max range	x_3 and x_4
6-4	RN1	4	8% of max range	x_3 and x_4
6-5	RN2	1	8% of max range	x_3 and x_4
6-6	RN2	2	8% of max range	x_3 and x_4
6-7	RN2	3	8% of max range	x_3 and x_4
6-8	RN2	4	8% of max range	x_3 and x_4

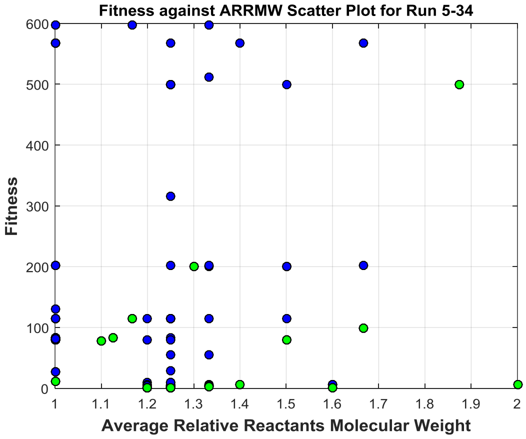
Table 6.6-1 Run parameters for Run 6-1 to Run 6-8

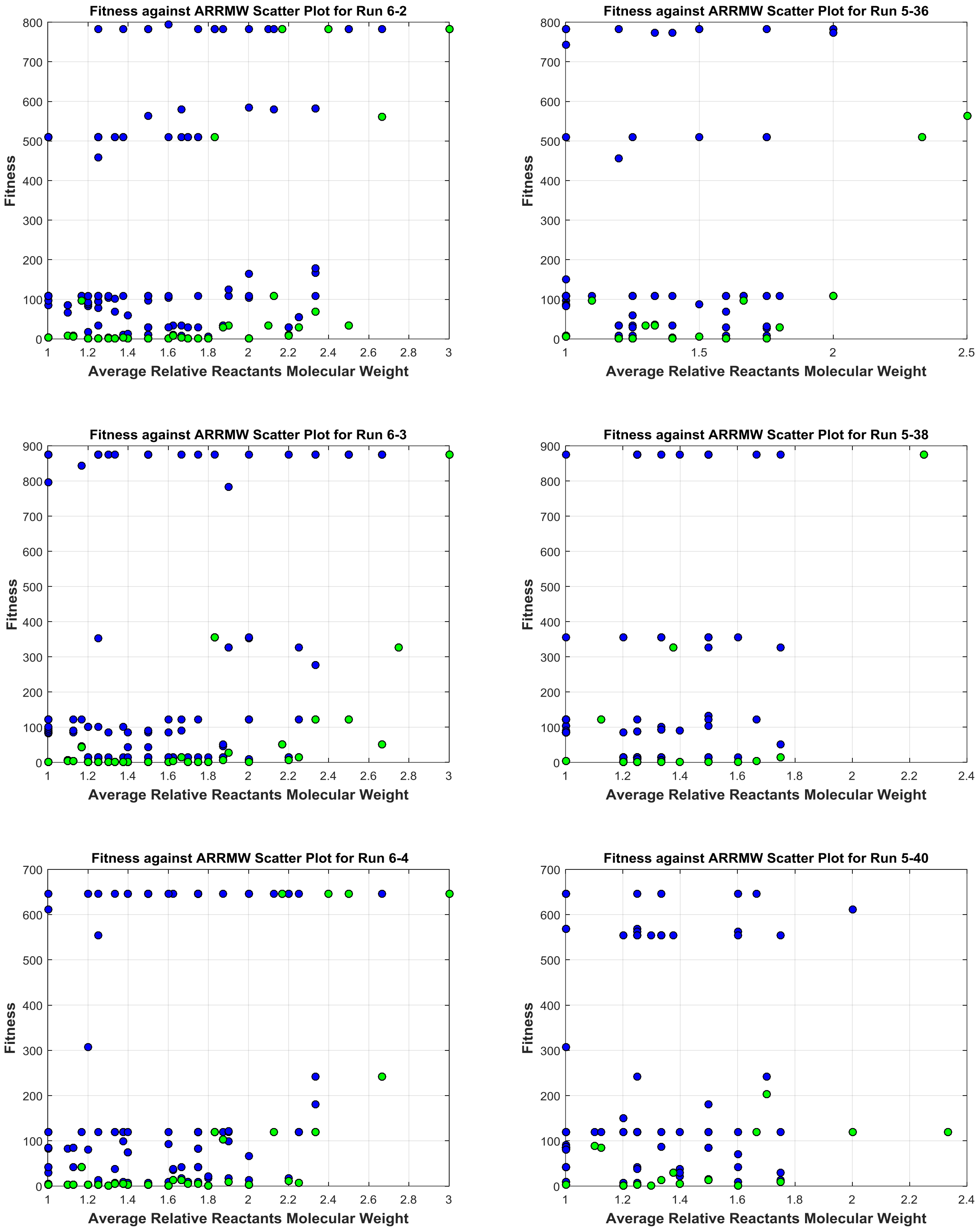
The run parameters is exactly the same as the one used by automated system (version 2) in the previous chapter but with elitism removed because NSGA-II algorithm incorporated it by nature.

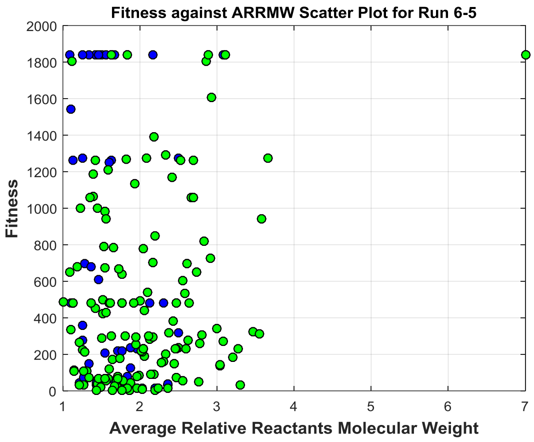
6.7 Results

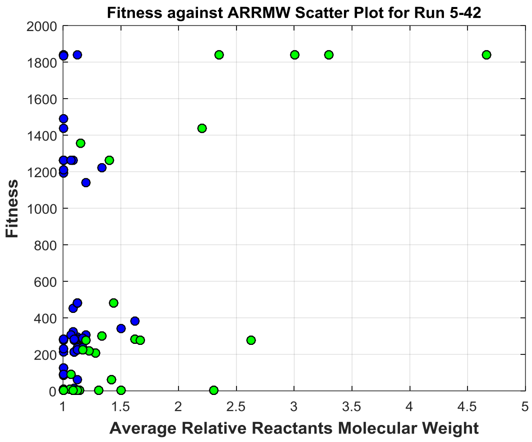
The results' scatter plot is compared against results generated by automated system (version 2) through runs using the same datasets (Run 5-34, Run 5-36, Run 5-38, Run 5-40, Run 5-42, Run 5-44, Run 5-46 and Run 5-48). Figure 6.7-1 shows the scatter plot of fitness function against *ARRMW*.

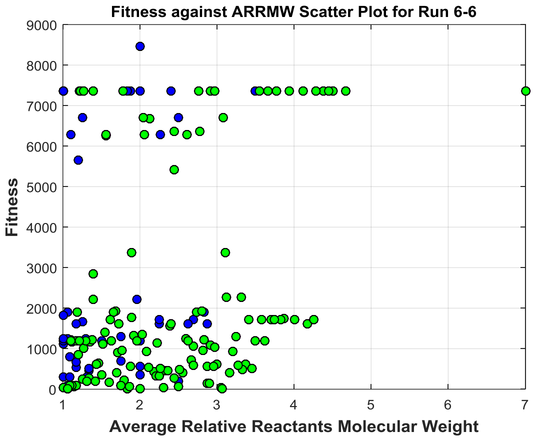


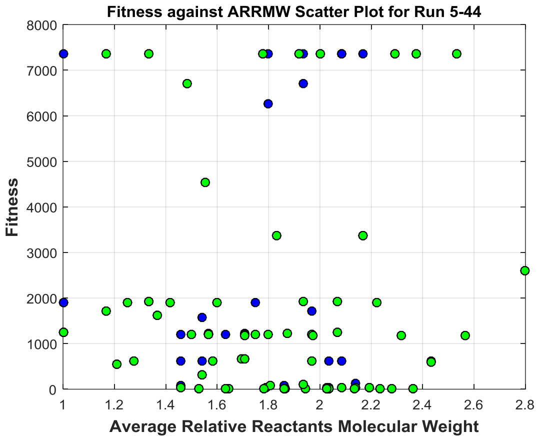


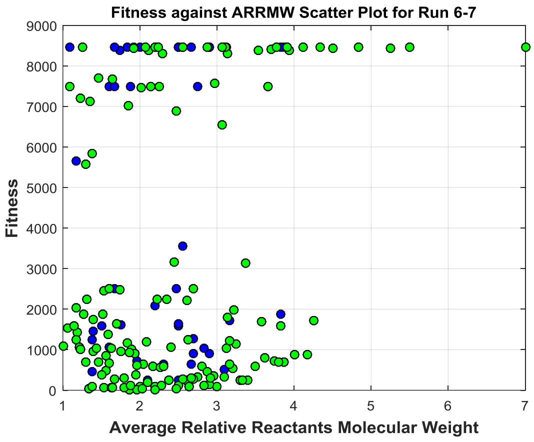

Fitness against ARRMW Scatter Plot for Run 6-2
Fitness against ARRMW Scatter Plot for Run 5-36
Fitness against ARRMW Scatter Plot for Run 6-3
Fitness against ARRMW Scatter Plot for Run 5-38
Fitness against ARRMW Scatter Plot for Run 6-4
Fitness against ARRMW Scatter Plot for Run 5-40
Fitness
Average Relative Reactants Molecular Weight
0
100
200
300
400
500
600
700
800
900
1
1.2
1.4
1.6
1.8
2
2.2
2.4
2.6
2.8
3
1.5
2.5

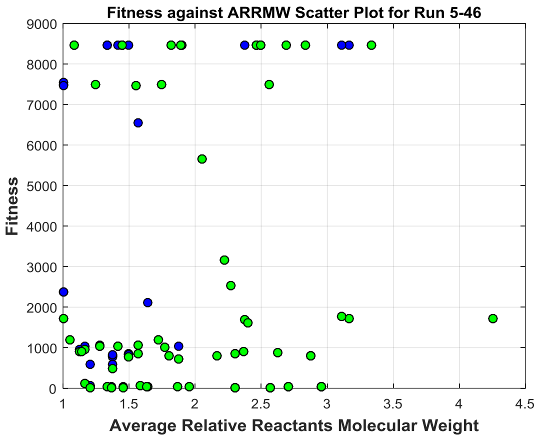

Fitness against ARRMW Scatter Plot for Run 6-5


Fitness against ARRMW Scatter Plot for Run 5-42


Fitness against ARRMW Scatter Plot for Run 6-6


Fitness against ARRMW Scatter Plot for Run 5-44


Fitness against ARRMW Scatter Plot for Run 6-7


Fitness against ARRMW Scatter Plot for Run 5-46

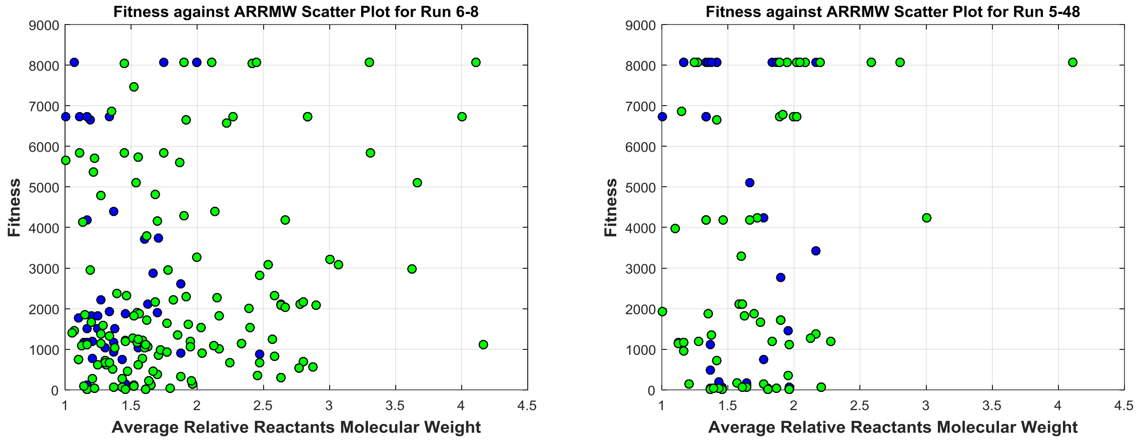


Figure 6.7-1 Comparison between scatter plots of Run 6-1 to Run 6-8 and Run 5-34, Run 5-36, Run 5-38, Run 5-40, Run 5-42, Run 5-44, Run 5-46 and Run 5-48. Green data points refer to Pareto Optimal CRNs.

From the scatter plots, a few observations can be made. Automated system (NSGA-II) produces individuals with larger ARRMW than automated system (version 2). For example, in Run 6-7 the individual with the largest ARRMW is 7 while its counterpart, Run 5-46 largest individual have less than 4.5 ARRMW. This shows automated system (NSGA-II) push outwards to increase diversity in its system.

Automated system (NSGA-II) also yields more densely at the Pareto Front (green colour data point) as compared to those produced by automated system (version 2). It also shows more densely packed individuals at the lower end of the fitness function as compared to automated system (version 2) runs which are more sparsely populated. Run 6-5 against Run 5-42 is a good example of this with significantly large amount of individuals in Run 5-42 having the same ARRMW. This shows that automated system (NSGA-II) not only push outwards, it also push inwards to the area between individuals to increase the diversity.

There are also more individuals at Pareto Front for runs using RN2 as compared to RN1 and this is because there are 6 chemical species in RN2 and only 5 chemical species in RN1. More chemical species meant that there are more combinations of reactions that automated system (NSGA-II) can create which will gives different ARRMW. The difference in ARRMW among the CRNs is what drives NSGA-II in the automated system and this therefore caused more points at the Pareto Front to appear.

The range for the fitness function axis used here is extremely high, for example Run 6-6 with maximum value of 9000 in order to include all the individuals. Figure 6.7-2 shows the same scatter plot for Run 6-6 but cap the range for the fitness function to 150.

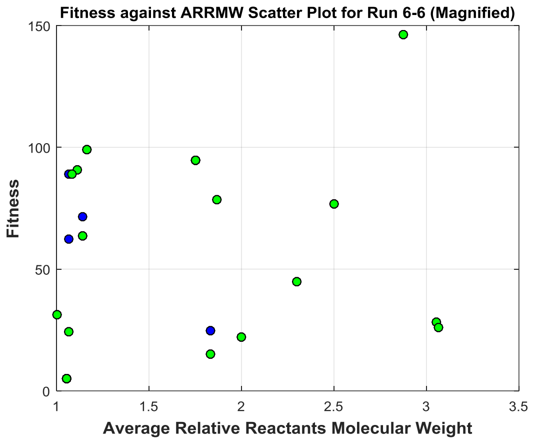


Figure 6.7-2 Magnified view of the scatter plots for the results for Run 6-6. Green data points refer to Pareto Optimal CRNs.

Figure 6.7-2 shows that the behaviour of the results from automated system (NSGA-II) remains the same at low fitness values.

It shows that automated system (NSGA-II) does increase the diversity in the individual in its final generation when compared again automated system (version 2). Whether this will help with the 'overfitness' problem is investigated in the next section.

6.8 CRN elucidation based on number of occurrences

Rather than trying to single out one best performing CRN based on fitness function, automated system (NSGA-II) will choose a group of CRNs among its final generation. This is set as the top 25% of the population when the individuals are sorted according to Pareto Front and then fitness function. This meant to be part the final group of CRNs, the CRN should have low Pareto Front index and low fitness value. The top 10 most occurring reactions for each of the runs are presented in Table 6.8-1:

Run	Top 10 Most Occurring Reactions	No. of Occurrence	Part of Actual CRN
6-1	$[0\ \ -1\ \ -1\ \ 0\ \ 1]$	30	No
	$[\mathbf{-2\ \ 1\ \ 0\ \ 0\ \ 0}]$	**25**	**Yes**
	$[-1\ \ 0\ \ 2\ \ -1\ \ 0]$	22	No
	$[0\ \ -1\ \ 1\ \ 1\ \ 0]$	20	No
	$[1\ \ 0\ \ 0\ \ -1\ \ 0]$	10	No
	$[\mathbf{0\ \ -1\ \ 0\ \ -1\ \ 1}]$	**9**	**Yes**
	$[0\ \ -1\ \ 2\ \ 0\ \ 0]$	6	No
	$[-1\ \ 0\ \ 0\ \ 1\ \ 0]$	6	No
	$[-1\ \ 2\ \ 0\ \ 0\ \ -1]$	5	No
	$[-1\ \ -1\ \ 0\ \ 0\ \ 1]$	5	No
6-2	$[\mathbf{-2\ \ 1\ \ 0\ \ 0\ \ 0}]$	**34**	**Yes**
	$[\mathbf{-1\ \ 0\ \ 1\ \ 0\ \ 0}]$	**31**	**Yes**
	$[\mathbf{0\ \ -1\ \ 0\ \ -1\ \ 1}]$	**22**	**Yes**
	$[0\ \ -1\ \ -1\ \ 0\ \ 1]$	21	No
	$[\mathbf{0\ \ 0\ \ -1\ \ 1\ \ 0}]$	**14**	**Yes**
	$[-1\ \ 1\ \ 2\ \ 0\ \ -1]$	11	No
	$[1\ \ -1\ \ 2\ \ 2\ \ -1]$	9	No
	$[-1\ \ -1\ \ 2\ \ 1\ \ 0]$	8	No
	$[2\ \ -1\ \ 1\ \ 2\ \ -1]$	8	No
	$[0\ \ 2\ \ 0\ \ -1\ \ -1]$	8	No
6-3	$[-1\ \ 0\ \ 0\ \ 1\ \ 0]$	32	No
	$[\mathbf{-2\ \ 1\ \ 0\ \ 0\ \ 0}]$	**27**	**Yes**
	$[\mathbf{0\ \ -1\ \ 0\ \ -1\ \ 1}]$	**24**	**Yes**
	$[0\ \ -1\ \ -1\ \ 0\ \ 1]$	15	No
	$[2\ \ -1\ \ 1\ \ 2\ \ -1]$	12	No
	$[\mathbf{-1\ \ 0\ \ 1\ \ 0\ \ 0}]$	**9**	**Yes**
	$[1\ \ -1\ \ 0\ \ 1\ \ 0]$	8	No
	$[1\ \ -2\ \ 0\ \ 0\ \ 1]$	7	No
	$[1\ \ -1\ \ 2\ \ 2\ \ -1]$	7	No
	$[2\ \ 0\ \ -1\ \ 2\ \ -1]$	7	No
6-4	$[-1\ \ 0\ \ 0\ \ 1\ \ 0]$	39	No
	$[-1\ \ 1\ \ 0\ \ -1\ \ 0]$	28	No
	$[0\ \ -1\ \ -1\ \ 0\ \ 1]$	24	No
	$[1\ \ -1\ \ 2\ \ 2\ \ -1]$	20	No
	$[0\ \ 1\ \ 2\ \ -1\ \ -1]$	15	No
	$[1\ \ -2\ \ 0\ \ 0\ \ 1]$	15	No
	$[\mathbf{0\ \ -1\ \ 0\ \ -1\ \ 1}]$	**15**	**Yes**
	$[0\ \ 0\ \ 1\ \ -1\ \ 0]$	8	No
	$[2\ \ -1\ \ 1\ \ -1\ \ 0]$	6	No
	$[-1\ \ -1\ \ 1\ \ 2\ \ 0]$	5	No

Run	Top 10 Most Occurring Reactions	No. of Occurrence	Part of Actual CRN
6-5	$\boldsymbol{[-1\ \ 0\ \ 0\ \ -1\ \ 0\ \ 1]}$	**45**	**Yes**
	$\boldsymbol{[-1\ \ -1\ \ 1\ \ 1\ \ 0\ \ 0]}$	**38**	**Yes**
	$[1\ \ 0\ \ 1\ \ 0\ \ 1\ \ -1]$	19	No
	$[0\ \ 1\ \ 0\ \ 2\ \ -1\ \ -1]$	14	No
	$[0\ \ 0\ \ -1\ \ 1\ \ -1\ \ 0]$	11	No
	$[-1\ \ 0\ \ -1\ \ 1\ \ 0\ \ 0]$	11	No
	$[-1\ \ 0\ \ 0\ \ 0\ \ 1\ \ 0]$	9	No
	$[0\ \ -1\ \ 1\ \ 1\ \ -1\ \ 0]$	9	No
	$[0\ \ -1\ \ 2\ \ 0\ \ 0\ \ 0]$	8	No
	$[1\ \ 1\ \ -1\ \ -1\ \ 0\ \ 0]$	7	No
6-6	$\boldsymbol{[-1\ \ -1\ \ 1\ \ 1\ \ 0\ \ 0]}$	**42**	**Yes**
	$[0\ \ 0\ \ 1\ \ -2\ \ 0\ \ 1]$	27	No
	$\boldsymbol{[-1\ \ 0\ \ 0\ \ -1\ \ 0\ \ 1]}$	**23**	**Yes**
	$[2\ \ -1\ \ 0\ \ 0\ \ 1\ \ -1]$	13	No
	$[0\ \ 1\ \ -1\ \ 0\ \ 2\ \ -1]$	12	No
	$[-2\ \ 0\ \ 2\ \ 1\ \ 0\ \ 0]$	11	No
	$[0\ \ -2\ \ 1\ \ 0\ \ 1\ \ 0]$	11	No
	$\boldsymbol{[0\ \ -1\ \ -1\ \ 0\ \ 1\ \ 0]}$	**10**	**Yes**
	$[0\ \ 0\ \ -1\ \ 2\ \ 0\ \ -1]$	9	No
	$[1\ \ 0\ \ 0\ \ 0\ \ -1\ \ 0]$	9	No
6-7	$\boldsymbol{[-1\ \ -1\ \ 1\ \ 1\ \ 0\ \ 0]}$	**45**	**Yes**
	$[1\ \ -1\ \ 0\ \ -1\ \ 1\ \ 0]$	36	No
	$[0\ \ 0\ \ 1\ \ -2\ \ 0\ \ 1]$	30	No
	$[-1\ \ 0\ \ -1\ \ 1\ \ 0\ \ 0]$	26	No
	$[0\ \ 1\ \ -1\ \ -1\ \ 1\ \ 0]$	17	No
	$[0\ \ 0\ \ -1\ \ 2\ \ 0\ \ -1]$	17	No
	$\boldsymbol{[-1\ \ 0\ \ 0\ \ -1\ \ 0\ \ 1]}$	**15**	**Yes**
	$[1\ \ -1\ \ 2\ \ 1\ \ 0\ \ -1]$	13	No
	$[0\ \ 0\ \ 0\ \ -1\ \ -1\ \ 1]$	8	No
	$[0\ \ 1\ \ -2\ \ 0\ \ 0\ \ 0]$	8	No
6-8	$\boldsymbol{[-1\ \ -1\ \ 1\ \ 1\ \ 0\ \ 0]}$	**32**	**Yes**
	$[-1\ \ 0\ \ 0\ \ 0\ \ 1\ \ 0]$	27	No
	$[0\ \ 0\ \ 1\ \ -2\ \ 0\ \ 1]$	24	No
	$\boldsymbol{[-1\ \ 0\ \ 0\ \ -1\ \ 0\ \ 1]}$	**15**	**Yes**
	$[-1\ \ 2\ \ 2\ \ 1\ \ 0\ \ -1]$	12	No
	$[0\ \ -1\ \ 0\ \ -1\ \ 2\ \ 0]$	9	No
	$[2\ \ -1\ \ 0\ \ -1\ \ 0\ \ 0]$	8	No
	$[0\ \ 1\ \ -2\ \ 0\ \ 0\ \ 0]$	7	No
	$[1\ \ 0\ \ 0\ \ 0\ \ -1\ \ 0]$	7	No
	$\boldsymbol{[1\ \ 0\ \ 0\ \ 1\ \ 0\ \ -1]}$	**7**	**Yes**

Table 6.8-1 Results showing most occurring reactions for Run 6-1 to Run 6-8

For Run 6-1 to Run 6-4, only Run 6-2 managed to discover all the reactions within RN1 when using the top 10 most occurring reactions. Of the four reactions in RN1, the 3rd reaction is the rarest within the CRNs obtained from the runs. The reaction is

$$x_3 \rightarrow x_4 \qquad \text{(Equation 6-2)}$$

which contains both the unmeasured chemical species making it difficult to be detected by automated system (NSGA-II). However from the results, it is noticeable that the reaction $[-1 \quad 0 \quad 0 \quad 1 \quad 0]$ and $[0 \quad -1 \quad -1 \quad 0 \quad 1]$ occurs at high frequency within the runs. The two reactions translates to

$$x_1 \rightarrow x_4 \qquad \text{(Equation 6-3)}$$

$$x_2 + x_3 \rightarrow x_5$$

The first reaction upon further analysis is a combination of the 2nd and 3rd reaction of RN1,

$$x_1 \rightarrow x_3 \qquad \text{Equation (6-4)}$$

$$x_3 \rightarrow x_4$$

with x_3 as the reaction intermediate. Meanwhile, the second reaction is a combination of the 3rd and 4th reaction of RN1

$$x_3 \rightarrow x_4 \qquad \text{(Equation 6-5)}$$

$$x_2 + x_4 \rightarrow x_5$$

with x_4 as the reaction intermediate.

Both of these cases happen because they use an unmeasured chemical species as reaction intermediates and automated system having no information about it attempt to model it without the additional reaction step. Although the two reactions are not the part of RN1, the user of automated system (NSGA-II) will have a better idea on how the reaction progressed with the information as compared to having a single CRN which the user have no idea how much trust he can give it.

For Run 6-5 to Run 6-8, none of the runs managed to elucidate all of the RN2's reactions. The results are much poorer compared to that of Run 6-1 to Run 6-4 and it may be because there are more chemical species in this CRN and this leads to more freedom for automated system (NSGA-II) to design reactions that are not part of the RN2 but able to be used to reduce the gap between the predicted and simulated noisy concentration data of measured chemical species.

If the results of the runs are not evaluated separately but together for Run 6-1 to Run 6-4 and Run 6-5 to Run 6-8, it may be possible to obtain the actual CRN. This is especially true for Run 6-5 to Run 6-8 because although not one of the runs managed to discover all the reactions within the RN2, each of the reactions in RN2 does appear in one or more of the runs. This shows that some of the datasets may be more suitable for unearthing certain reactions, depending on the initial condition of the chemical species. The effect of noise will also lessen if all the datasets are evaluated together.

6.9 Summary

In summary, this chapter introduced a multi objective evolutionary algorithm (MOEA) to be included into the automated system in order to deal with the ‘overfitting’ problem in the automated system. Elitist Non-Dominated Sorting Genetic Algorithm (NSGA-II) is used to upgrade the automated system and the algorithm is modified to use Average Relative Reactants’ Molecular Weight (ARRMW) to increase diversity in the population so more possible combination of reactions in the CRN can be tested.

The goal of the runs has been changed from focusing on a single best performing CRN in respect to the fitness function to obtaining a group of CRNs and evaluate the reactions in them. Reactions with higher frequency of occurrence in high performing individuals (low fitness function and low Pareto Front index) can be extracted and considered.

Eight different runs are done with four datasets from RN1 and four from RN2 with Gaussian noise with mean of 0 and standard deviation of 8% of the maximum range of the concentration data are added. Two of chemical species in RN1 and two from RN2 are designated as unmeasured.

The results show that NSGA-II in the system do increase the diversity of the final result when compared against the runs done using automated system (version 2). However, of the eight runs, only one manage to elucidate all of the reactions of the actual CRN. Reactions that are not part of CRN are constructed in the runs as well but some of those reactions are combination of reactions from the actual CRN. The range of reactions will give the user more information on the possible reactions within the CRN.

It is noted that different datasets face different difficulty in discovering the different reactions. Running all the datasets in one go may actually help in discovering all the reactions in the actual CRN while having more data helps the automated system discount any errors or noise in the concentration data.

Chapter 7. Application on Experimental Data

7.1 Overview

In this chapter, the automated system for elucidation of chemical reaction network (CRN) capability is demonstrated on experimental data. The data comes from the reaction of trimethyl orthoacetate (TMOA) and allyl alcohol (AA) the expected reaction is introduced. Automated system (version 2) and automated system (NSGA-II) will be used for the runs. Further test on the robustness of the system is done through addition of chemical species that is not part of the reaction. The results of the runs are discussed and conclusions are made.

7.2 Introduction

For the purpose of demonstrating the capability of the automated system for CRN elucidation that has been developed in this work, three sets of experimental data from the reaction of TMOA and AA is obtained from experimental work done by the Department of Chemistry, Durham University. The experiments are conducted isothermally at three different temperature, 80 ^{o}C, 90 ^{o}C and 100 ^{o}C in a 2 litres glass lined batch reactor. The initial concentration of the starting reactants of TMOA, x_{TMOA} and AA, x_{AA} are the same for the three experiments. Samples are taken at different time intervals and each samples is analysed using Gas Chromatography Mass Spectrometry to obtain the concentration of the chemical species present in the reactor. Each of the experiments are run for different length of time, 1000 minutes for 80 ^{o}C experiment, 1440 minutes for 90 ^{o}C experiment and 300 minutes for 100 ^{o}C experiment. Chemical species allyl dimethyl orthoacetate (ADMOA), x_{ADMOA} , diallyl methyl orthoacetate (DMOA), x_{DMOA} and triallyl orthoacetate (TOA), x_{TOA} are detected to be present in the experiments and produced through the reaction between TMOA and AA. To make reference easier, the 80 ^{o}C experiment will be named Exp 1, 90 ^{o}C named as Exp 2 and 100 ^{o}C named as Exp 3 from here onwards.

7.3 Chemical reaction network details

The paper presented by Bollyn and Wright (1998) reveals the reaction chemistry of the reaction between triethyl orthoacetate (TEOA) and allyl alcohol (AA) which is similar to the reaction between TMOA and AA as part of the Claisen condensation reaction. The structure of the CRN is taken and used to determine the performance of the automated systems. According to Bollyn and Wright (1998), the TEOA and AA reacts to produce allyl diethyl orthoacetate, ADEOA and ethanol, EA. ADEOA then reacts with AA again to produce diallyl ethyl orthoacetate (DEOA) and EA. DEOA reacts with AA further to produce triallyl orthoaceteate (TOA) and EA. These reactions are all reversible reactions.

ADEOA, DEOA and TOA can then be reacted further to produce pentenoic acid ethyl ester and pentenoic acid allyl ester but the two chemical species are not present in the Exp 1, Exp 2 and Exp 3, so it will not be discussed further.

Following the reaction mechanisms of the reaction between TEOA and AA, the reaction between TMOA and AA can be summed up as

$$\begin{gathered}\mathrm{TMOA} + \mathrm{AA} \rightleftarrows \mathrm{ADMOA} + \mathrm{MA} \\ \mathrm{ADMOA} + \mathrm{AA} \rightleftarrows \mathrm{DMOA} + \mathrm{MA} \\ \mathrm{DMOA} + \mathrm{AA} \rightleftarrows \mathrm{TOA} + \mathrm{MA}\end{gathered} \qquad \text{(Equation 7-1)}$$

where MA is methanol.

Figure 7.3-1 shows the chemical structures of the chemical species in Equation 7-1.

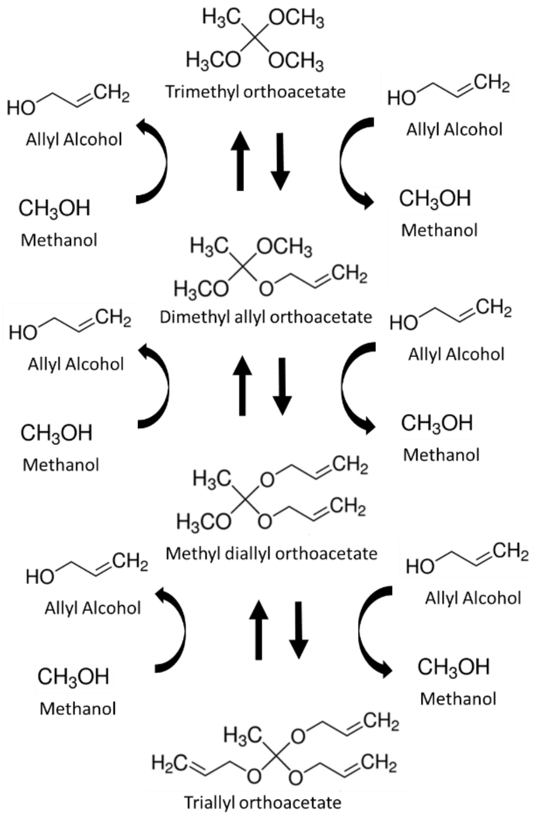


Figure 7.3-1 Chemical structures of reactants, intermediates and products of the reaction between TMOA and AA.

In stoichiometric form,

$$\boldsymbol{V_{Exp1}} = \begin{bmatrix} -1 & 1 & 0 & 0 & -1 & 1 \\ 1 & -1 & 0 & 0 & 1 & -1 \\ 0 & -1 & 1 & 0 & -1 & 1 \\ 0 & 1 & -1 & 0 & 1 & -1 \\ 0 & 0 & -1 & 1 & -1 & 1 \\ 0 & 0 & 1 & -1 & 1 & -1 \end{bmatrix} \qquad \text{(Equation 7-2)}$$

where the columns of the matrix refers to the chemical species as per below matrix

$$[x_{TMOA} \quad x_{ADMOA} \quad x_{DMOA} \quad x_{TOA} \quad x_{AA} \quad x_{MA}] \qquad \text{(Equation 7-3)}$$

The molecular weight of the chemical species is shown in table below:

Chemical Species	TMOA	ADMOA	DMOA	TOA	AA	MA
Molecular Weight	120	146	172	198	58	32

Table 7.3-1 Molecular weight of the chemical species in the experimental data

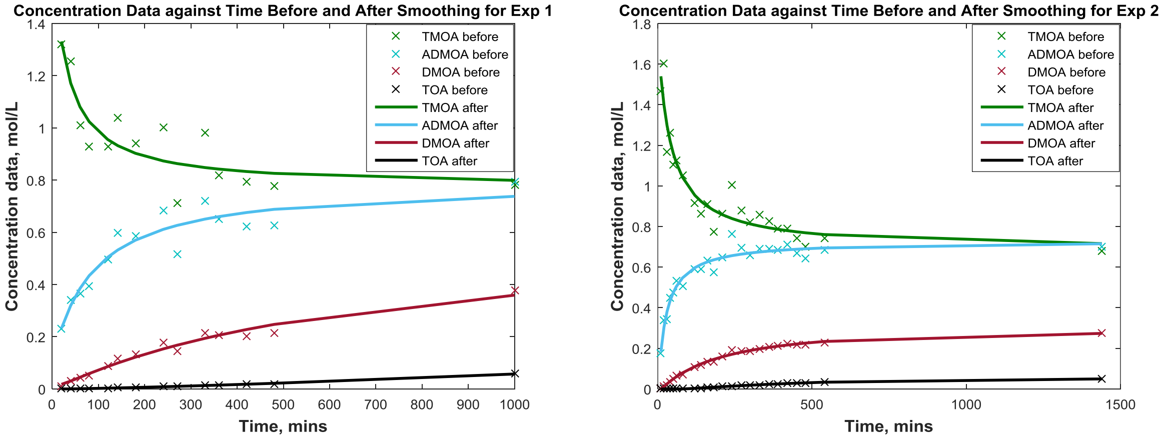
Both the alcohols in the experimental data are unmeasured chemical species. AA and MA is not measured for any part of the experiment.

With this details, the automated system can be employed to elucidate the CRN from the experimental data.

7.4 Data preprocessing

The concentration data from the experimental data is processed by fitting the data to a rational polynomial to smoothen the data. This is done because of the existence of substantial noise in the data and this can be down to human error when conducting the experiments, when taking measurements, variability in the process temperature and pressure and sensitivity of the measurement device.

Figure 7.4-1 that show before and after smoothing are employed:



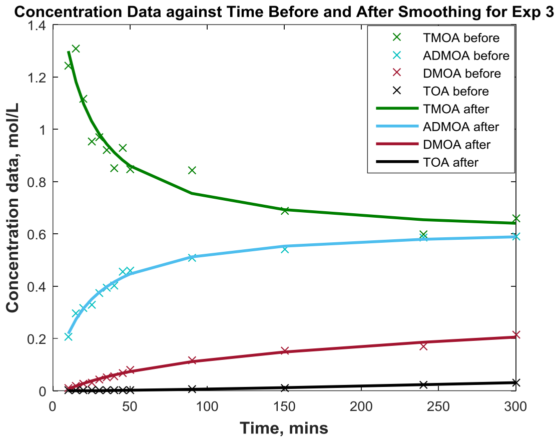


Figure 7.4-1 Plots showing concentration data smoothing

Smoothened data will help easier convergence of the automated system as the automated system reconstructs the concentration profiles based on the assumption of no noise. It will also assume that the data provided are mass balanced.

7.5 Testing the experimental data using automated system (version 2)

For the first part of the test, automated system (version 2) is used first. The run parameters remains the same as they are in the previous chapters.

Details of the run are as shown in Table 7.5-1:

Run	Experimental data source
7-1	Smoothened Exp 1
7-2	Smoothened Exp 2
7-3	Smoothened Exp 3

Table 7.5-1 Run details for Run 7-1, Run 7-2 and Run 7-3

The results of the run are shown in Table 7.5-2:

Run	Best Performing Individual	Reaction Rate Constant	Fitness Function
7-1	$\begin{bmatrix} -1 & 1 & 0 & 0 & -1 & 1 \\ 1 & -1 & 0 & 0 & 1 & -1 \\ 0 & -1 & 1 & 0 & -1 & 1 \\ 0 & 1 & -1 & 0 & 1 & -1 \\ 0 & 1 & -2 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}$	$k_1 = 0.0046\, dm^3 mol^{-1} min^{-1}$ $k_2 = 0.0028\, dm^3 mol^{-1} min^{-1}$ $k_3 = 0.0030\, dm^3 mol^{-1} min^{-1}$ $k_4 = 0.0039\, dm^3 mol^{-1} min^{-1}$ $k_5 = 0.0017\, dm^3 mol^{-1} min^{-1}$ $k_6 = 0$	8.8647
7-2	$\begin{bmatrix} -1 & 1 & 0 & 0 & -1 & 1 \\ 0 & -1 & 1 & 0 & -1 & 1 \\ 0 & 0 & -1 & 1 & -1 & 1 \\ -1 & 2 & -1 & 0 & 0 & 0 \\ 1 & -2 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}$	$k_1 = 0.0035\, dm^3 mol^{-1} min^{-1}$ $k_2 = 0.0018\, dm^3 mol^{-1} min^{-1}$ $k_3 = 0.0009\, dm^3 mol^{-1} min^{-1}$ $k_4 = 0.0028\, dm^3 mol^{-1} min^{-1}$ $k_5 = 0.0007\, dm^3 mol^{-1} min^{-1}$ $k_6 = 0$	26.4821
7-3	$\begin{bmatrix} -1 & 1 & 0 & 0 & -1 & 1 \\ 0 & 1 & -1 & 0 & 1 & -1 \\ 1 & -1 & -1 & 1 & 0 & 0 \\ -1 & 2 & -1 & 0 & 0 & 0 \\ 1 & -2 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}$	$k_1 = 0.0104\, dm^3 mol^{-1} min^{-1}$ $k_2 = 0.0150\, dm^3 mol^{-1} min^{-1}$ $k_3 = 0.0014\, dm^3 mol^{-1} min^{-1}$ $k_4 = 0.0052\, dm^3 mol^{-1} min^{-1}$ $k_5 = 0.0119\, dm^3 mol^{-1} min^{-1}$ $k_6 = 0$	2.3152

Table 7.5-2 Results for Run 7-1, Run 7-2 and Run 7-3

Run 7-1 managed to elucidate 4 of the 6 expected reactions in the CRN and contain an extra reaction that is not part of the CRN. Upon further investigations, the reaction is actually a combination of two reactions from the actual CRN. The extra reaction is as below:

$$2x_{DMOA} \rightarrow x_{ADMOA} + x_{TOA} \qquad \text{(Equation 7-7)}$$

and is actually a combination of two reactions from the expected CRN.

$$\begin{gathered} x_{DMOA} + x_{MA} \rightarrow x_{ADMOA} + x_{AA} \\ x_{DMOA} + x_{AA} \rightarrow x_{TOA} + x_{MA} \end{gathered} \qquad \text{(Equation 7-8)}$$

As expected, because MA and AA are not measured, automated system (version 2) disregards their role as reaction intermediates and fit the predicted concentration data to the experimental data. Even with the deconstruction of the extra reaction, the run is still missing the last reaction from the expected CRN.

The plot for the run between the predicted concentration data and the experimental data is shown in Figure 7.5-2 Predicted and experimental concentration data against time for Run 7-2:

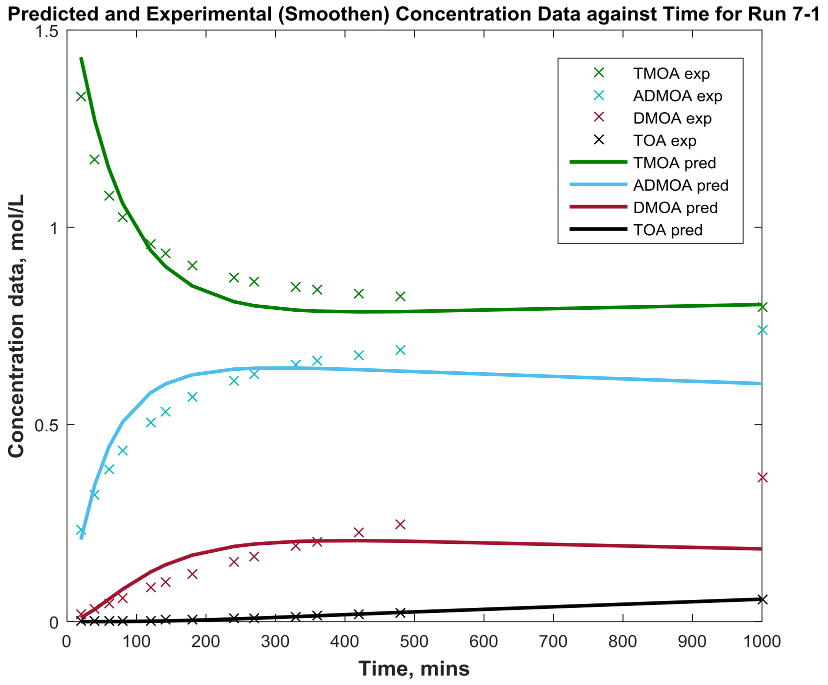


Figure 7.5-1 Concentration of predicted and experimental data against time for Run 7-1

Although automated system (version 2) managed to obtain 4 of the 6 reactions of the expected CRN, the predicted concentration accuracy is poor as can be seen from the Figure. Delving further into the experimental data, it is revealed that measurement error must had occur during data collection because of the material imbalance that exist in the data. Table 7.5-3shows concentration data of the Exp 1 before the smoothing process.

t	Concentration data (mol/L)				
	$[x_{TMOA}]$	$[x_{TMOA}]$	$[x_{DMOA}]$	$[x_{TOA}]$	$[x_{TMOA}]+[x_{TMOA}]+[x_{DMOA}]+[x_{TOA}]$
20	1.3214	0.2295	0.0108	0	1.5617
40	1.2548	0.3425	0.029	0	1.6263
60	1.0097	0.3635	0.0417	0	1.4149
80	0.9268	0.3958	0.0527	0.0009	1.3762
120	0.9278	0.4974	0.0893	0.0028	1.5173
142	1.0378	0.5982	0.118	0.0043	1.7583
180	0.9416	0.5877	0.1322	0.006	1.6675
240	1.0004	0.6843	0.1767	0.0099	1.8713
270	0.7127	0.5156	0.1456	0.0092	1.3831
330	0.9801	0.7222	0.2129	0.0149	1.9301

t	Concentration data (mol/L)				
	$[x_{TMOA}]$	$[x_{TMOA}]$	$[x_{DMOA}]$	$[x_{TOA}]$	$[x_{TMOA}]+[x_{TMOA}]$ $+[x_{DMOA}]+[x_{TOA}]$
360	0.8189	0.6501	0.208	0.0161	1.6931
420	0.7946	0.6211	0.2026	0.0164	1.6347
480	0.7761	0.6275	0.2163	0.0196	1.6395
1000	0.7818	0.7928	0.3773	0.0576	2.0095

Table 7.5-3 Concentration data for Exp 1

Based on expected CRN of

$$\boldsymbol{V_{Exp1}} = \begin{bmatrix} -1 & 1 & 0 & 0 & -1 & 1 \\ 1 & -1 & 0 & 0 & 1 & -1 \\ 0 & -1 & 1 & 0 & -1 & 1 \\ 0 & 1 & -1 & 0 & 1 & -1 \\ 0 & 0 & -1 & 1 & -1 & 1 \\ 0 & 0 & 1 & -1 & 1 & -1 \end{bmatrix} \qquad \text{(Equation 7-9)}$$

The molar balance of

$$[x_{TMOA}] + [x_{TMOA}] + [x_{DMOA}] + [x_{TOA}] \qquad \text{(Equation 7-10)}$$

should always remain the same as conversion of the chemical species is all 1:1 ratio. No reaction in the CRN is capable in increasing or reducing the amount of mols in the reactor.

Refer back to the Table above, the total summation of the concentration data of the four chemical species, ranges from 1.3831 to 2.0095 when it should have remained constant. This type of error in the concentration data will definitely cause any predicted concentration from automated system (version 2) to go off because the predicted concentration data still adhered to mass and material balance. Even then, the system still managed to elucidate the CRN with only two missing reactions from the expected CRN.

For Run 7-2, only 3 of the 6 reactions of expected CRN are identified at the final CRN structure from automated system (version 2). Similar to Run 7-1, it also consists of reactions that are not part of the expected CRN and they are combination of reactions of the expected CRN. The reactions are as follows:

$$\begin{gathered} x_{TMOA} + x_{DMOA} \rightarrow 2x_{ADMOA} \\ 2x_{ADMOA} \rightarrow x_{TMOA} + x_{DMOA} \end{gathered} \qquad \text{(Equation 7-11)}$$

which are combinations of

$$\begin{aligned} x_{TMOA} + x_{AA} &\rightarrow x_{ADMOA} + x_{MA} \\ x_{DMOA} + x_{MA} &\rightarrow x_{ADMOA} + x_{AA} \\ x_{ADMOA} + x_{MA} &\rightarrow x_{TMOA} + x_{AA} \\ x_{ADMOA} + x_{AA} &\rightarrow x_{DMOA} + x_{MA} \end{aligned} \qquad \text{(Equation 7-12)}$$

The predicted data from the CRN is plotted against the experimental data and is shown in Figure 7.5-2:

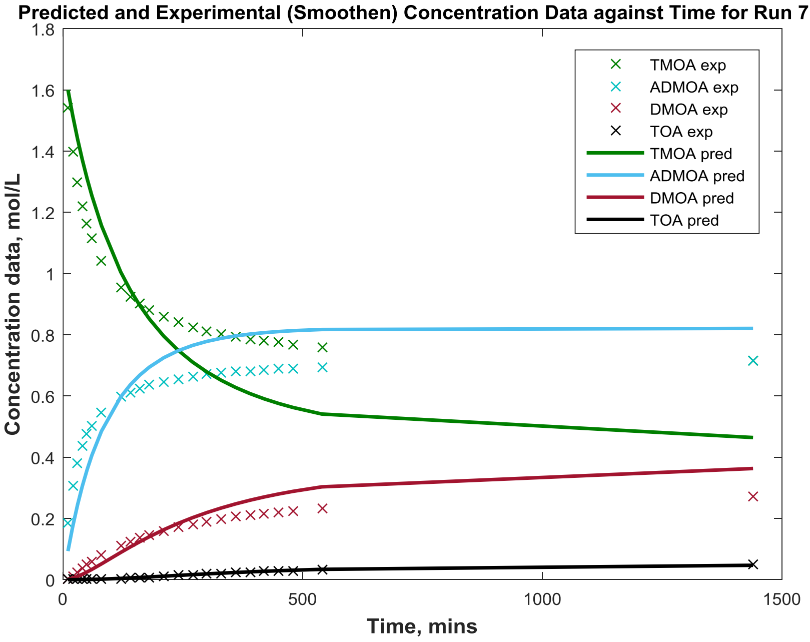


Figure 7.5-2 Predicted and experimental concentration data against time for Run 7-2

The poor fit of the predicted concentration data to the experimental data is apparent from the Figure. Similar to Run 7-1, the experimental data is contaminated by noise or inaccurate measurements. Table 7.5-4 tabled the concentration data of Exp 2.

t	Concentration data (mol/L)				
	$[x_{TMOA}]$	$[x_{TMOA}]$	$[x_{DMOA}]$	$[x_{TOA}]$	$[x_{TMOA}] + [x_{TMOA}] + [x_{DMOA}] + [x_{TOA}]$
10	1.4643	0.1743	0.0020	0.0000	1.6406
20	1.6058	0.3384	0.0130	0.0000	1.9572
30	1.1700	0.3421	0.0207	0.0000	1.5328
40	1.2613	0.4506	0.0363	0.0000	1.7482
50	1.1031	0.4755	0.0486	0.0000	1.6272
60	1.1253	0.5339	0.0632	0.0000	1.7223

t	Concentration data (mol/L)				
	$[x_{TMOA}]$	$[x_{TMOA}]$	$[x_{DMOA}]$	$[x_{TOA}]$	$[x_{TMOA}] + [x_{TMOA}] + [x_{DMOA}] + [x_{TOA}]$
80	1.0522	0.5064	0.0729	0.0000	1.6315
120	0.9132	0.5881	0.1091	0.0030	1.6134
140	0.8623	0.5896	0.1198	0.0048	1.5766
160	0.9123	0.6324	0.1345	0.0063	1.6856
180	0.7734	0.5768	0.1361	0.0079	1.4942
210	0.8659	0.6476	0.1591	0.0105	1.6831
240	1.0046	0.7628	0.1938	0.0147	1.9759
270	0.8816	0.6944	0.1865	0.0163	1.7788
300	0.8234	0.6596	0.1858	0.0180	1.6868
330	0.8604	0.6927	0.1955	0.0202	1.7688
360	0.8275	0.6923	0.2055	0.0231	1.7484
390	0.7904	0.6852	0.2137	0.0262	1.7155
420	0.7910	0.7108	0.2253	0.0288	1.7559
450	0.7434	0.6666	0.2171	0.0282	1.6553
480	0.7008	0.6443	0.2156	0.0299	1.5905
540	0.7408	0.6857	0.2284	0.0317	1.6865
1440	0.6788	0.6989	0.2752	0.0495	1.7024

Table 7.5-4 Concentration data for Exp 2

The molar balance ranged from 1.4942 to 1.9759 and such errors will cause any predicted concentration data that is bound by material balance to not be able to fit to the experimental data.

Again similar to Run 7-1 and Run 7-2, Run 7-3 does not obtain all the reactions in the expected CRN. It only managed to elucidate 3 of the 6 reactions and include 2 additional unexpected reactions. The two reactions is the same as those in Run 7-3 and as explained previously, they are combination reactions of the reaction in the expected CRN. The plot of the predicted concentration of the best individual from the run and experimental data is done in Figure 7.5-3.

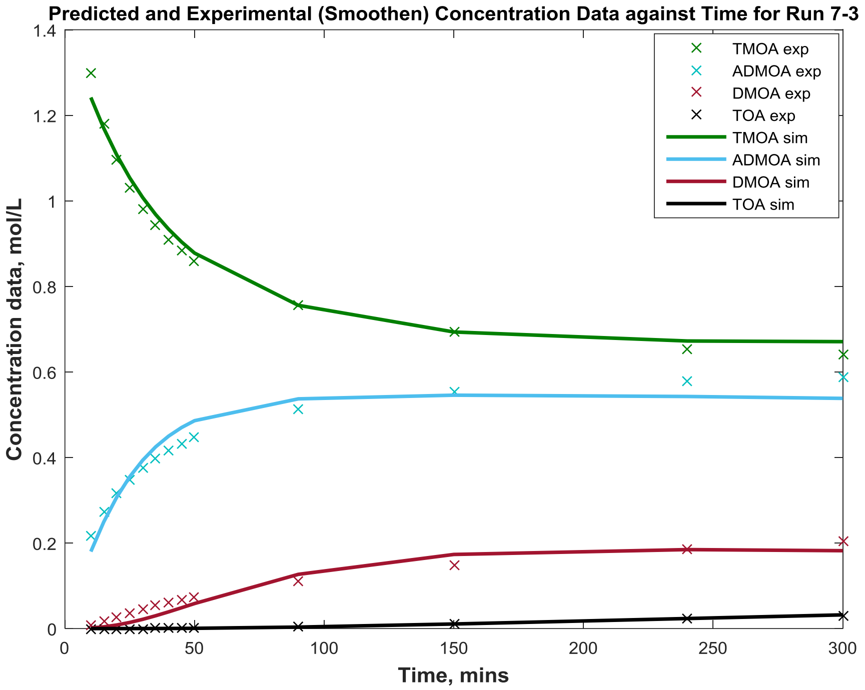


Figure 7.5-3 Predicted and experimental concentration data against time for Run 7-3

The fit of the predicted concentration to the experimental data for Run 7-3 is much better as compared to those in Run 7-1 and Run 7-2 as can be observed from the Figures. The experimental concentration data is shown in Table 7.5-5 Concentration data for Exp 3.

t	Concentration data (mol/L)				
	$[x_{TMOA}]$	$[x_{TMOA}]$	$[x_{DMOA}]$	$[x_{TOA}]$	$[x_{TMOA}]+[x_{TMOA}]$ $+[x_{DMOA}]+[x_{TOA}]$
10	1.2418	0.2079	0.0100	0.0000	1.4596
15	1.3080	0.2978	0.0202	0.0000	1.6260
20	1.1162	0.3177	0.0270	0.0000	1.4609
25	0.9521	0.3291	0.0322	0.0000	1.3134
30	0.9710	0.3721	0.0425	0.0000	1.3857
35	0.9187	0.3947	0.0523	0.0006	1.3662
40	0.8506	0.4008	0.0557	0.0010	1.3081
45	0.9285	0.4565	0.0692	0.0016	1.4558
50	0.8449	0.4572	0.0789	0.0022	1.3832
90	0.8445	0.5085	0.1144	0.0057	1.4731
150	0.6882	0.5422	0.1532	0.0111	1.3947
240	0.5998	0.5857	0.1701	0.0233	1.3788

$\boldsymbol{t}$	**Concentration data (mol/L)**				
	$[\boldsymbol{x_{TMOA}}]$	$[\boldsymbol{x_{TMOA}}]$	$[\boldsymbol{x_{DMOA}}]$	$[\boldsymbol{x_{TOA}}]$	$[\boldsymbol{x_{TMOA}}] + [\boldsymbol{x_{TMOA}}] + [\boldsymbol{x_{DMOA}}] + [\boldsymbol{x_{TOA}}]$
300	0.6600	0.5889	0.2137	0.0312	1.4939

Table 7.5-5 Concentration data for Exp 3

The molar balance for Exp 3 is much more balanced than when compared to Exp 1 and Exp 2. Although it has a spike at one single data point to 1.6260, the data pre-processing would have smoothened this data point out.

7.6 Experimental data adjustment

For further more meaningful investigations into the capability of the automated system, the concentration data has to be adjusted as at the current level, the errors in the data is too overwhelming to produce a good fit for the predicted concentration data. This adjustment is done with the knowledge that it will make the CRN deduced not applicable for the original experimental data as the data has essentially been modified. However, the adjusted data can still serve as a good test for the automated system. The adjustment is done by first classifying the addition of all the concentration data in each data point as

$$h_t = [x_{TMOA}]_t + [x_{TMOA}]_t + [x_{DMOA}]_t + [x_{TOA}]_t \qquad \text{(Equation 7-13)}$$

Taking the mean of h_t

$$\bar{h} = \frac{\sum_0^{N_t} h_t}{N_t} \qquad \text{(Equation 7-14)}$$

The difference of each data points h_t to $\bar{h}$

$$d_t = \bar{h} - h_t \qquad \text{(Equation 7-15)}$$

Taking the weight of each of the concentration data, $W_{[x_i]_t}$ as compared to the other concentration data of the same data point

$$W_{[x_i]_t} = \frac{[x_i]_t}{h_t} \qquad \text{(Equation 7-16)}$$

The adjusted concentration data, $[\widetilde{x_i}]_t$ can then be made by the following formula

$$[\widetilde{x_i}]_t = W_{[x_i]_t} d_t + [x_i]_t \qquad \text{(Equation 7-17)}$$

This will ensure all the adjusted concentration data point will be balanced materially. Each of the adjusted concentration data is adjusted according to how large their value is compare to other concentration data of the same data point. The larger their weight, the bigger adjustments they will experienced compared to the other concentration data of the same data point. Table 7.6-1 shows the adjusted concentration data of Exp 1.

t	Concentration data (mol/L)				
	$[x_{TMOA}]$	$[x_{TMOA}]$	$[x_{DMOA}]$	$[x_{TOA}]$	$[x_{TMOA}]+[x_{TMOA}]+[x_{DMOA}]+[x_{TOA}]$
20	1.3892	0.2413	0.0182	0.0000	1.64878
40	1.2667	0.3469	0.0352	0.0000	1.64878
60	1.1779	0.4198	0.0511	0.0000	1.64878
80	1.1106	0.4710	0.0657	0.0015	1.64878
120	1.0175	0.5363	0.0917	0.0033	1.64878
142	0.9806	0.5594	0.1046	0.0043	1.64878
180	0.9312	0.5870	0.1246	0.0059	1.64878
240	0.8760	0.6122	0.1519	0.0086	1.64878
270	0.8550	0.6199	0.1640	0.0099	1.64878
330	0.8211	0.6296	0.1855	0.0126	1.64878
360	0.8071	0.6325	0.1951	0.0140	1.64878
420	0.7835	0.6360	0.2124	0.0168	1.64878
480	0.7641	0.6373	0.2277	0.0197	1.64878
1000	0.6716	0.6219	0.3070	0.0483	1.64878

Table 7.6-1 Adjusted concentration data of Exp 1

The most significant change is that all the datasets now add up to the same molar balance. The same is done Exp 2 and Exp 3 and using these datasets, the automated system is tested further.

7.7 Testing the experimental data using automated system (version 2) using adjusted data

The adjusted data obtained in Section 7.6 is then tested using the automated system (version 2) again. The details of the run is as Table 7.7-1:

Run	Experimental data source
7-4	Adjusted Exp 1
7-5	Adjusted Exp 2
7-6	Adjusted Exp 3

Table 7.7-1 Run details for Run 7-4 to Run 7-6

The automated system run parameters remain the same as previous.

The results is presented in Table 7.7-2:

Run	Best Performing Individual	Reaction Rate Constant	Fitness Function
7-4	$\begin{bmatrix} -1 & 1 & 0 & 0 & -1 & 1 \\ 1 & -1 & 0 & 0 & 1 & -1 \\ 0 & 0 & 1 & -1 & 1 & -1 \\ -1 & 2 & -1 & 0 & 0 & 0 \\ 1 & -1 & -1 & 1 & 0 & 0 \\ 1 & -2 & 1 & 0 & 0 & 0 \end{bmatrix}$	$k_1 = 0.0040\ dm^3mol^{-1}min^{-1}$ $k_2 = 0.0013\ dm^3mol^{-1}min^{-1}$ $k_3 = 0.0013\ dm^3mol^{-1}min^{-1}$ $k_4 = 0.0079\ dm^3mol^{-1}min^{-1}$ $k_5 = 0.0007\ dm^3mol^{-1}min^{-1}$ $k_6 = 0.0042\ dm^3mol^{-1}min^{-1}$	2.5927
7-5	$\begin{bmatrix} -1 & 1 & 0 & 0 & -1 & 1 \\ 1 & -1 & 0 & 0 & 1 & -1 \\ 0 & 0 & -1 & 1 & -1 & 1 \\ -1 & 2 & -1 & 0 & 0 & 0 \\ 1 & -2 & 1 & 0 & 0 & 0 \\ 0 & -1 & 2 & -1 & 0 & 0 \end{bmatrix}$	$k_1 = 0.0047\ dm^3mol^{-1}min^{-1}$ $k_2 = 0.0015\ dm^3mol^{-1}min^{-1}$ $k_3 = 0.0011\ dm^3mol^{-1}min^{-1}$ $k_4 = 0.0082\ dm^3mol^{-1}min^{-1}$ $k_5 = 0.0039\ dm^3mol^{-1}min^{-1}$ $k_6 = 0.0035\ dm^3mol^{-1}min^{-1}$	4.4383
7-6	$\begin{bmatrix} -1 & 1 & 0 & 0 & -1 & 1 \\ 1 & -1 & 0 & 0 & 1 & -1 \\ 0 & -1 & 1 & 0 & -1 & 1 \\ 0 & 1 & -1 & 0 & 1 & -1 \\ 0 & 0 & -1 & 1 & -1 & 1 \\ -1 & 2 & -1 & 0 & 0 & 0 \end{bmatrix}$	$k_1 = 0.0101\ dm^3mol^{-1}min^{-1}$ $k_2 = 0.0073\ dm^3mol^{-1}min^{-1}$ $k_3 = 0.0055\ dm^3mol^{-1}min^{-1}$ $k_4 = 0.0014\ dm^3mol^{-1}min^{-1}$ $k_5 = 0.0119\ dm^3mol^{-1}min^{-1}$ $k_6 = 0.0056\ dm^3mol^{-1}min^{-1}$	1.0166

Table 7.7-2 Results for Run 7-4 to Run 7-6

It can be concluded from the results presented in Table 7.7-2 shows similar behaviour to those in Table 7.5-2. Run 7-4, Run 7-5 and Run 7-6 all did not manage to elucidate the expected CRN and had as well introduced reactions that are in essence, combination of reactions from the expected CRN. This is similar to Run 7-1, Run 7-2 and Run 7-3 which also have the similar type of results but using data before adjustment. It can be observed however that the fitness function for Run 7-4, Run 7-5 and Run 7-6 is significantly lower than those of Run 7-1, Run 7-2 and Run 7-3 and this is down to using data that are actually molar balanced. This can be seen in the comparison of the concentration profile of the Run 7-4 of the experimental and predicted data.

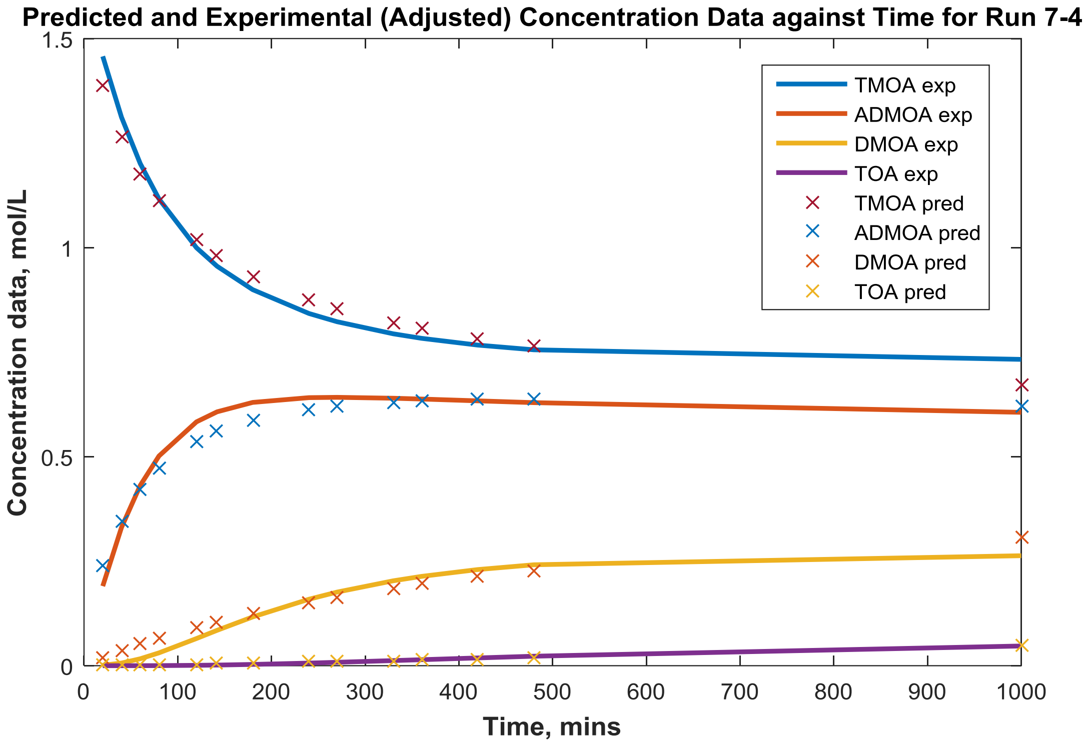


Figure 7.7-1 Predicted and experimental concentration data against time for Run 7-4

Figure 7.7-1 shows how closely the predicted concentration data follows the concentration data of the adjusted experimental data. This shows that the automated system (version 2) can produce CRNs that can closely follow the concentration profile provided the data used is balanced or does not deviate from the balance significantly.

The runs in Table 7.7-2 also present the limitation of automated system (version 2) in handling this type of data even when it has been molar balanced. However, it has to be noted that the method to molar balanced the data may not be accurate as it is based on averages and in reality the molar balance could have been skewed from chemical species to the other based on the noise that is causing the inaccuracies in the data.

7.8 Including ‘fake’ chemical species into the concentration data

Using the adjusted concentration data, the capability of the automated system (version 2) to operate with wrong information is tested. As it is needed to provide the molecular weight of each of the participating chemical species at the start of the run, this test will examine what happened when automated system (version 2) is provided with information of chemical species that are unrelated to the CRN at all. The ‘fake’ chemical species molecular weight is a combination of molecular weights of participating chemical species.

Two ‘fake’ chemical species is added into the system and will be called as chemical A and chemical B. The molecular weight these two are

Chemical Species	A	B
Molecular Weight	178	204

Table 7.8-1 Molecular weight for 'fake' chemical species

Naturally, both of these chemical species are designated as unmeasured chemical species. Details of the runs is as Table 7.8-2,

Run	Experimental data source	‘Fake’ chemical species
7-7	Adjusted Exp 1	A and B
7-8	Adjusted Exp 2	A and B
7-9	Adjusted Exp 3	A and B

Table 7.8-2 Run details for Run 7-7 to Run 7-9

The automated system run parameters remain the same as previous.

The results of the run is shown in the table below. The 7th and 8th column in the stoichiometric matrix refers to chemical species A and B accordingly.

Run	Best Performing Individual	Reaction Rate Constant	Fitness Function
7-7	$\begin{bmatrix} -1 & 1 & 0 & 0 & -1 & 1 & 0 & 0 \\ 0 & 1 & -2 & 1 & 0 & 0 & 0 & 0 \\ 0 & -1 & 2 & -1 & 0 & 0 & 0 & 0 \\ 1 & -2 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & -1 & 0 & 0 & -1 & 0 & 0 & 1 \\ -1 & 1 & 1 & -1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 & -1 & 0 \\ 0 & 0 & 0 & 0 & 1 & -1 & 1 & -1 \end{bmatrix}$	$k_1 = 0.0035\ dm^3mol^{-1}min^{-1}$ $k_2 = 0.0241\ dm^3mol^{-1}min^{-1}$ $k_3 = 0.0456\ dm^3mol^{-1}min^{-1}$ $k_4 = 0.0034\ dm^3mol^{-1}min^{-1}$ $k_5 = 0.0010\ dm^3mol^{-1}min^{-1}$ $k_6 = 0.0464\ dm^3mol^{-1}min^{-1}$ $k_7 = 0.0096\ dm^3mol^{-1}min^{-1}$ $k_8 = 0.0106\ dm^3mol^{-1}min^{-1}$	4.595
7-8	$\begin{bmatrix} -1 & 1 & 0 & 0 & -1 & 1 & 0 & 0 \\ 0 & 1 & -2 & 1 & 0 & 0 & 0 & 0 \\ -1 & 1 & 1 & -1 & 0 & 0 & 0 & 0 \\ 1 & -1 & -1 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & -1 & 1 & -1 & 1 \\ 0 & 1 & 0 & 0 & 0 & 1 & -1 & 0 \\ 0 & -1 & 2 & -1 & 0 & 0 & 0 & 0 \\ 1 & -2 & 1 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}$	$k_1 = 0.0040\ dm^3mol^{-1}min^{-1}$ $k_2 = 0.1013\ dm^3mol^{-1}min^{-1}$ $k_3 = 0.8758\ dm^3mol^{-1}min^{-1}$ $k_4 = 0.0611\ dm^3mol^{-1}min^{-1}$ $k_5 = 0.8706\ dm^3mol^{-1}min^{-1}$ $k_6 = 0.0002\ dm^3mol^{-1}min^{-1}$ $k_7 = 0.2136\ dm^3mol^{-1}min^{-1}$ $k_8 = 0.0063\ dm^3mol^{-1}min^{-1}$	14.7809

Run	Best Performing Individual	Reaction Rate Constant	Fitness Function
7-9	$\begin{bmatrix} -1 & 1 & 0 & 0 & -1 & 1 & 0 & 0 \\ -1 & 2 & -1 & 0 & 0 & 0 & 0 & 0 \\ -1 & 1 & 1 & -1 & 0 & 0 & 0 & 0 \\ 0 & 1 & -2 & 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 2 & -1 & 0 & -1 \\ 0 & 0 & 1 & 0 & 0 & 1 & 0 & -1 \\ 1 & -2 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & -1 & 1 & -1 & 0 & 0 \end{bmatrix}$	$k_1 = 0.0085\ dm^3mol^{-1}min^{-1}$ $k_2 = 0.0239\ dm^3mol^{-1}min^{-1}$ $k_3 = 0.0536\ dm^3mol^{-1}min^{-1}$ $k_4 = 0.0571\ dm^3mol^{-1}min^{-1}$ $k_5 = 0.0081\ dm^3mol^{-1}min^{-1}$ $k_6 = 0.9876\ dm^3mol^{-1}min^{-1}$ $k_7 = 0.0143\ dm^3mol^{-1}min^{-1}$ $k_8 = 0.0514\ dm^3mol^{-1}min^{-1}$	3.9881

Table 7.8-3 Results for Run 7-7 to Run 7-9

From Table 7.8-3, it can be observed that all three of runs did not elucidate all of the reactions from expected CRN. The ‘fake’ chemical species is also used by the automated system to model the final CRN structure. These results show how aggressive the automated system cannot differentiate between involved and uninvolved chemical species among those that are provided to them. This is within expectation, after without any data to compare to, automated system (version 2) is not restricted by anything and will use whatever it can use to get a good fit. The following figure shows what happens when the predicted concentration data is plotted onto the adjusted experimental concentration data.

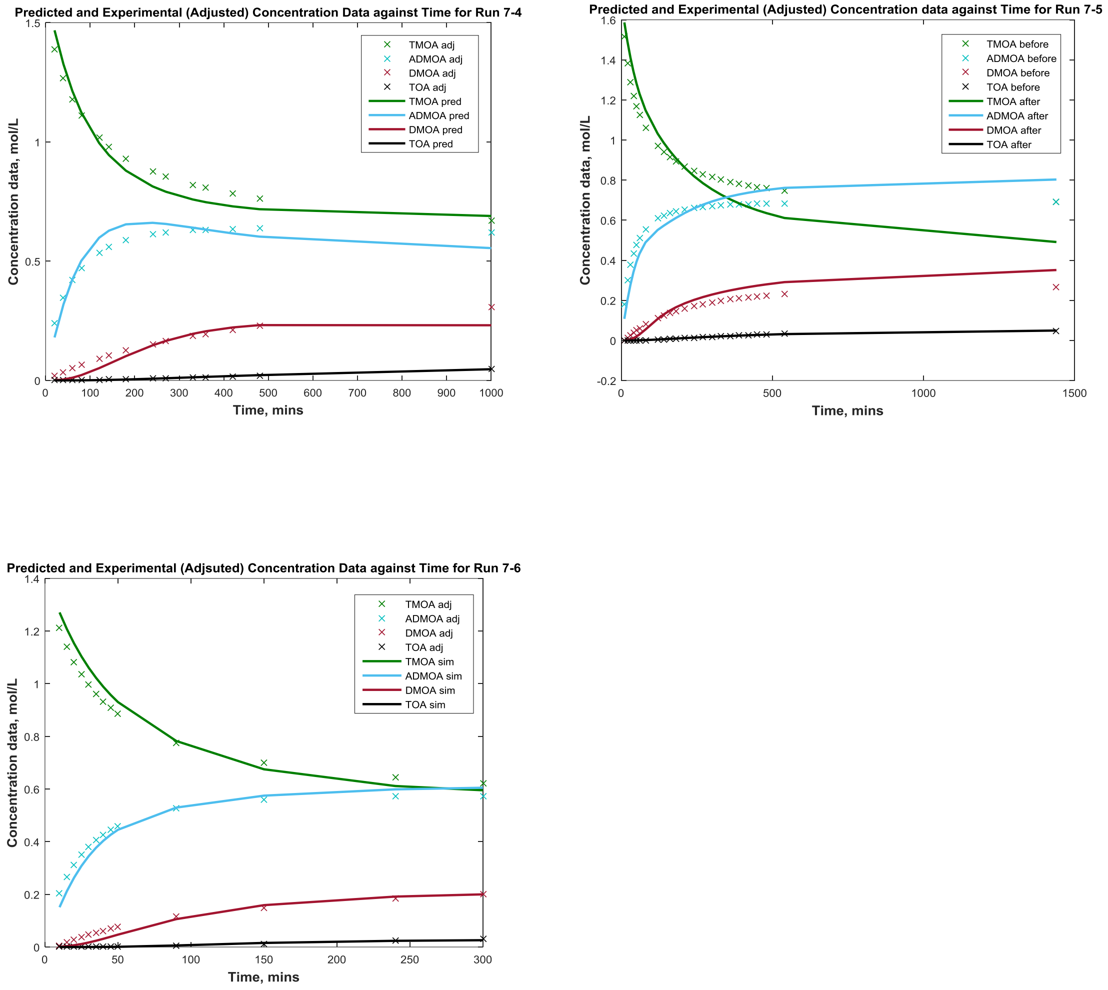


Figure 7.8-1 Plot of predicted concentration and experimental data against time for Run 7-7 to Run 7-9

Looking at Figure 7.8-1, the fit is poor even when the concentration data has been adjusted especially for Run 7-8 and this can only be attributed to the presence of the ‘fake’ chemical species. In comparison, Figure 7.7-1 shows that without the effect of the ‘fake’ chemical species, automated system (version 2) will be able to elucidate a CRN with a good fit to the experimental data.

When exposed to ‘fake’ chemical species, automated system (version 2) is not able to elucidate most of the reactions from the expected CRN. It even included some reactions where the ‘fake’ chemical species plays a part in the final CRN of the run. The same datasets are used again but automated system (NSGA-II) is used to elucidate the CRN. The details of the run as follows:

Run	Experimental data source	'Fake' chemical species
7-7	Adjusted Exp 1	A and B
7-8	Adjusted Exp 2	A and B
7-9	Adjusted Exp 3	A and B

Table 7.8-4 Run details for Run 7-7 to Run 7-9

The run parameters for automated system (NSGA-II) remain the same as the one used in the Chapter 6. The results of the run are presented in the Table 7.8-5.

Run	Top 10 Most Occurring Reactions	No. of Occurrence	Part of Actual CRN
7-7	$[\mathbf{-1 \quad 1 \quad 0 \quad 0 \quad -1 \quad 1 \quad 0 \quad 0}]$	**130**	**Yes**
	$[1 \quad -2 \quad 1 \quad 0 \quad 0 \quad 0 \quad 0 \quad 0]$	103	No
	$[0 \quad -1 \quad 0 \quad 0 \quad -1 \quad 0 \quad 0 \quad 1]$	75	No
	$[0 \quad 1 \quad -2 \quad 1 \quad 0 \quad 0 \quad 0 \quad 0]$	72	No
	$[1 \quad 0 \quad 0 \quad 0 \quad 1 \quad 0 \quad -1 \quad 0]$	59	No
	$[0 \quad -1 \quad 2 \quad -1 \quad 0 \quad 0 \quad 0 \quad 0]$	58	No
	$[-1 \quad 1 \quad 1 \quad -1 \quad 0 \quad 0 \quad 0 \quad 0]$	46	No
	$[0 \quad 0 \quad 0 \quad 0 \quad 1 \quad -1 \quad 1 \quad -1]$	35	No
	$[\mathbf{0 \quad 0 \quad -1 \quad 1 \quad -1 \quad 1 \quad 0 \quad 0}]$	**33**	**Yes**
	$[\mathbf{0 \quad 0 \quad 1 \quad -1 \quad 1 \quad -1 \quad 0 \quad 0}]$	**29**	**Yes**
7-8	$[\mathbf{-1 \quad 1 \quad 0 \quad 0 \quad -1 \quad 1 \quad 0 \quad 0}]$	**133**	**Yes**
	$[1 \quad -2 \quad 1 \quad 0 \quad 0 \quad 0 \quad 0 \quad 0]$	120	No
	$[0 \quad 1 \quad -2 \quad 1 \quad 0 \quad 0 \quad 0 \quad 0]$	103	No
	$[0 \quad 1 \quad 0 \quad 0 \quad 0 \quad 1 \quad -1 \quad 0]$	91	No
	$[-1 \quad 1 \quad 1 \quad -1 \quad 0 \quad 0 \quad 0 \quad 0]$	72	No
	$[\mathbf{0 \quad 0 \quad -1 \quad 1 \quad -1 \quad 1 \quad 0 \quad 0}]$	60	**Yes**
	$[\mathbf{0 \quad 0 \quad 1 \quad -1 \quad 1 \quad -1 \quad 0 \quad 0}]$	59	**Yes**
	$[\mathbf{0 \quad 1 \quad -1 \quad 0 \quad 1 \quad -1 \quad 0 \quad 0}]$	29	**Yes**
	$[0 \quad 0 \quad 0 \quad 0 \quad -1 \quad 1 \quad -1 \quad 1]$	29	No
	$[1 \quad 1 \quad 0 \quad 0 \quad 2 \quad 0 \quad -1 \quad -1]$	25	No
7-9	$[-1 \quad 0 \quad 0 \quad 0 \quad -1 \quad 0 \quad 1 \quad 0]$	125	No
	$[0 \quad 1 \quad 0 \quad 0 \quad 0 \quad 1 \quad -1 \quad 0]$	107	No
	$[1 \quad -2 \quad 1 \quad 0 \quad 0 \quad 0 \quad 0 \quad 0]$	100	No
	$[0 \quad 0 \quad 1 \quad 0 \quad -1 \quad 2 \quad -1 \quad 0]$	75	No
	$[-1 \quad 2 \quad -1 \quad 0 \quad 0 \quad 0 \quad 0 \quad 0]$	68	No
	$[1 \quad -1 \quad 0 \quad 0 \quad 0 \quad 0 \quad -1 \quad 1]$	62	No
	$[1 \quad 0 \quad 0 \quad 1 \quad 0 \quad 2 \quad -1 \quad -1]$	61	No
	$[0 \quad 1 \quad -2 \quad 1 \quad 0 \quad 0 \quad 0 \quad 0]$	60	No
	$[\mathbf{-1 \quad 1 \quad 0 \quad 0 \quad -1 \quad 1 \quad 0 \quad 0}]$	**35**	**Yes**
	$[-1 \quad 1 \quad 1 \quad -1 \quad 0 \quad 0 \quad 0 \quad 0]$	32	No

Table 7.8-5 Results for the Run 7-7 to Run 7-9

From Table 7.8-5, it can be seen that Run 7-7, Run 7-8 and Run 7-9, do not obtain all the reactions from the expected CRN. There are only 3 of the 6 expected reactions at the top 10 most occurring reactions for Run 7-7. As for Run 7-8, it does manage to obtain 4 out of 6 of the reactions from the expected CRN and finally Run 7-9 only manages to have one of the expected reactions in its top 10 most occurring reaction. All the three runs also have reactions that the ‘fake’ chemical species play a part in which showed that automated system (NSGA-II) cannot distinguish between involved or uninvolved chemical species

7.9 Summary

The reaction of trimethyl orthoacetate and allyl alcohol is discussed and the expected chemical reaction network for the reaction is presented. The experimental data is then run using automated system (version 2) and it was discovered that there is material imbalance in the concentration data which could be due to contamination or any errors that occur during the experimental stage. Adjustments are done on the experimental data sets so that further testing on it can be conducted. With the adjustment done, it is shown that the automated system (version 2) can elucidate the CRN with a good fit to the concentration profile of the experimental data so long it does not deviate too significantly from the molar balance. The next test is done with the presence of two ‘fake’ chemical species. The inclusion of the ‘fake’ chemical species severely affects the performance of both automated system (version 2) and automated system (NSGA-II). The next course of action is to try to run all the batches together and see whether with more data, it will improves the performance of the automated systems.

Chapter 8. Implement Batch Running into the Automated System for Chemical Reaction Elucidation

8.1 Overview

This chapter detail the work to mitigate the effect of ‘overfitting’ of the generated CRNs to the single dataset used to elucidate the CRN by using multiple datasets with different initial conditions and with multiple process temperature. Modifications are done to the algorithm in the automated system to accommodate for the changes so that it can evaluate datasets with different initial conditions and temperature. The automated system is then tested against the datasets for RN1 and RN2 that had different initial conditions and against the datasets for the experimental data for TMOA and AA that had different process temperatures. The results of the modified automated system are then presented at the end of the chapter.

8.2 Introduction

In the previous chapters, through the different iterations of the automated system that is develop have all suffer from high ‘overfitting’ problems. One of the reasons is the lack of the variability in the data that was used to elucidate the CRN which helps to propagate the effect of noise in the system. Introducing datasets that incorporate a larger range of concentration profiles changes into the system will help reduce the effect of the noise on a single dataset. This is because the noise from one dataset will not be compatible with another dataset and vice versa, causing their impact to be reduced in the final elucidated CRN. To implement these, the automated system’s algorithm need to be augmented to include running datasets with different initial conditions and different temperatures. The modifications to include different initial conditions datasets will be done first and tested on the datasets from RN1 and RN2 as discussed in Chapter 5 with 8% Gaussian noise and 2 unmeasured chemical species for each of the reaction networks. The final modification will be done on the automated system so that it will be able to run on datasets that belonged to different process temperatures such as the datasets from the adjusted experimental data for the reaction of TMOA and AA as presented in Chapter 7.

8.3 Modifications on automated system to include multiple batches with different initial conditions

In order for the evaluation of concentration data from multiple batches with different initial batch concentrations, the automated system will need to determine the reaction rate constants for the reactions in the elucidated CRN that can be used across all the batches with different initial conditions. Figure 7.8-1 shows how the modifications is implemented to the automated system algorithm.

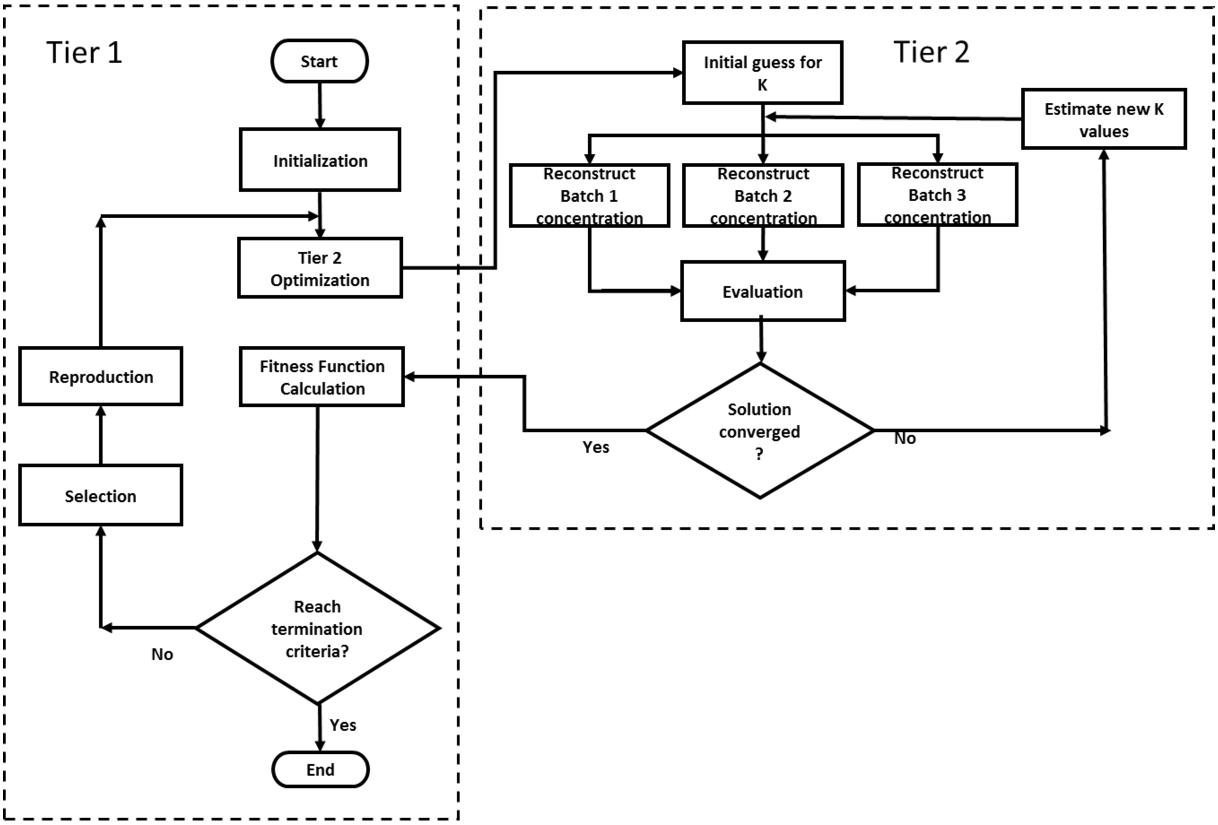
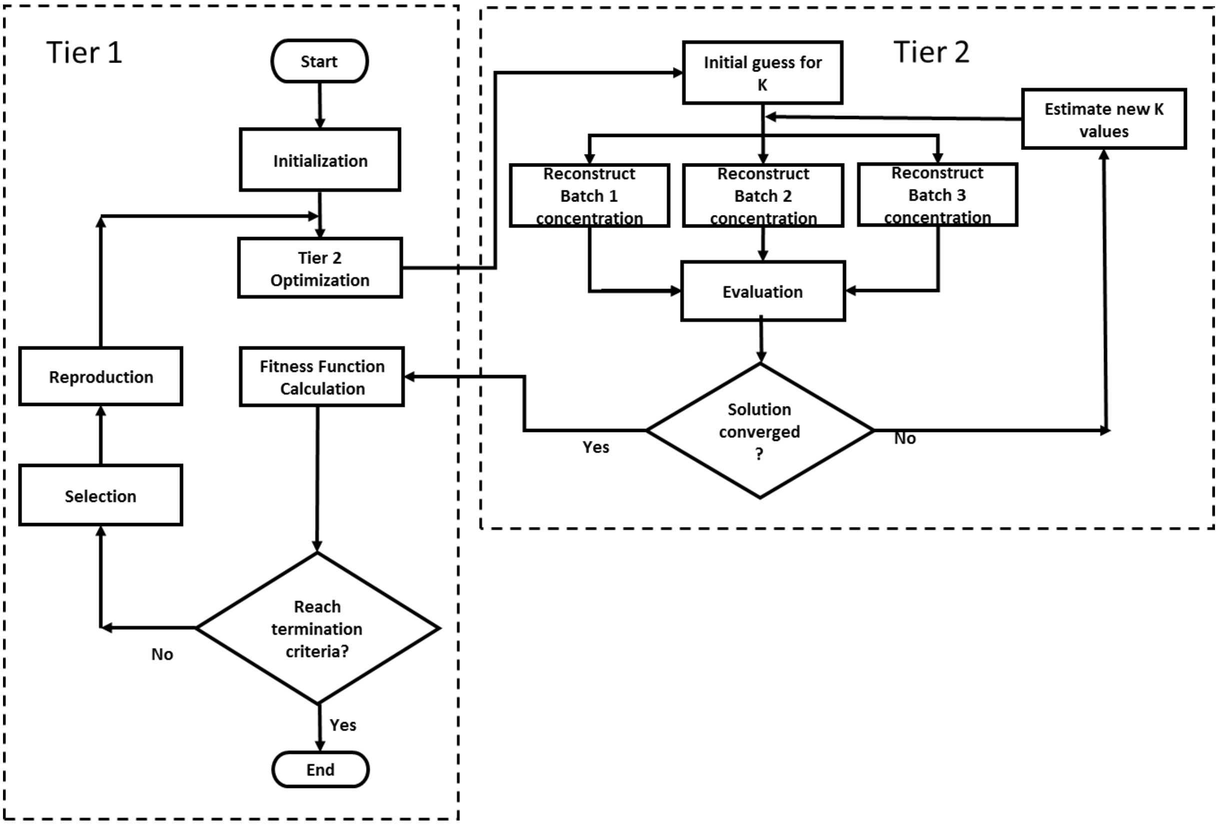

Tier 1
Start
Initialization
Tier 2 Optimization
Reproduction
Fitness Function Calculation
Selection
Reach termination criteria?
No
Yes
End
Tier 2
Initial guess for K
Estimate new K values
Reconstruct Batch 1 concentration
Reconstruct Batch 2 concentration
Reconstruct Batch 3 concentration
Evaluation
Solution converged ?
Yes
No

Figure 8.3-1 Flowchart for automated system to handle multiple datasets with different initial conditions

The modifications are done on the Tier 2 optimisation loop introduced in Chapter 5 (Figure 5.3-1) during the estimation of the reaction rate constants. Every reaction rate constants that are to be considered in the Tier 2 optimisation loop will need to be tested against all of the concentration data from different batches compared to only one before the modifications. In the example in Figure 8.3-1


Tier 1
Start
Initialization
Tier 2 Optimization
Reproduction
Fitness Function Calculation
Selection
Reach termination criteria?
No
Yes
End
Tier 2
Initial guess for K
Estimate new K values
Reconstruct Batch 1 concentration
Reconstruct Batch 2 concentration
Reconstruct Batch 3 concentration
Evaluation
Solution converged ?
Yes
No

, 3 different batches with different initial conditions are considered. If the reaction rate constants that are calculated are affected by the noise in Batch 1, it will unsuitable for the use for Batch 2 and so on. Based on this reasoning, this may help to reduce the impact of the noise in the system. The modifications can be done for automated system (version 2) and automated system (NSGA-II) but for the purpose of the testing out the modification alone, automated system (version 2) is modified.

8.4 Results from modification to include multiple batches with different initial conditions

8.4.1 *Run set up and results*

RN1 and RN2 are used to test the modification.

Run	CRN	Batch	Unmeasured	Gaussian noise
8-1	RN1	1,2,3,4	x_3 and x_4	8% of max value
8-2	RN2	1,2,3,4	x_3 and x_4	8% of max value

Table 8.4-1 Run details for Run 8-1 and 8-2

The automated system's parameters remains unchanged. Below is the result for the two runs.

Run	Best Performing Individual	Reaction Rate Constant
8-1	$\begin{bmatrix} -2 & 1 & 0 & 0 & 0 \\ 0 & -1 & 0 & -1 & 1 \\ 2 & 0 & 0 & 1 & -1 \\ 0 & 0 & -1 & 1 & 0 \\ -1 & 0 & 1 & 0 & 0 \end{bmatrix}$	$k_1 = 0.1208\, dm^3mol^{-1}s^{-1}$ $k_2 = 0.2707\, dm^3mol^{-1}s^{-1}$ $k_3 = 0.0057\, s^{-1}$ $k_4 = 0.1738\, s^{-1}$ $k_5 = 0.1886\, s^{-1}$
8-2	$\begin{bmatrix} -1 & -1 & 1 & 1 & 0 & 0 \\ -1 & 0 & 0 & -1 & 0 & 1 \\ 1 & 0 & 0 & 1 & 0 & -1 \\ 0 & -1 & -1 & 0 & 1 & 0 \\ 1 & 1 & -1 & 0 & 1 & -1 \\ -1 & 0 & 0 & 1 & 2 & -1 \end{bmatrix}$	$k_1 = 0.2074\, dm^3mol^{-1}s^{-1}$ $k_2 = 0.1504\, dm^3mol^{-1}s^{-1}$ $k_3 = 0.0485\, s^{-1}$ $k_4 = 0.1000\, dm^3mol^{-1}s^{-1}$ $k_5 = 0.0002\, dm^3mol^{-1}s^{-1}$ $k_6 = 0.0001\, dm^3mol^{-1}s^{-1}$

Table 8.4-2 Results for Run 8-1 and Run 8-2

Fitness function is not useful for comparison against Run 8-1 and Run 8-2 because they are of different CRN. It is also not useful to use to compare against previous runs using previous versions of the automated system because of the nature of the this automated system which run all 3 datasets at the same time.

8.4.2 *Discussion for Run 8-1*

Table 8.4-2 shows that for Run 8-1, all of the reactions from RN1 is elucidated by the automated system and it also contains a reaction that is not part of RN1. The unrelated reaction, the 3rd reaction has only a reaction rate of $0.0057\ s^{-1}$ which is relatively small as compared to the other reactions.

Table 8.4-3 shows a reduced view of Table 5.7-2 which are runs that are done on single batch basis using automated system (version 2) with the same level of noise for comparison purposes.

Run	Best Performing Individual	Reaction Rate Constant
5-34	$\begin{bmatrix} -2 & 1 & 0 & 0 & 0 \\ 0 & -1 & 0 & -1 & 1 \\ -1 & 0 & 0 & 1 & 0 \\ 0 & -1 & 0 & 2 & 0 \\ -1 & 1 & 0 & -1 & 0 \end{bmatrix}$	$k_1 = 0.4581\, dm^3mol^{-1}s^{-1}$ $k_2 = 0.1422\, dm^3mol^{-1}s^{-1}$ $k_3 = 0.0561\, s^{-1}$ $k_4 = 0.0072\, s^{-1}$ $k_5 = 0.5303\, s^{-1}$
5-36	$\begin{bmatrix} -2 & 1 & 0 & 0 & 0 \\ 2 & 0 & -1 & -1 & 0 \\ 0 & -1 & -1 & 0 & 1 \\ -1 & 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & -1 & 0 \end{bmatrix}$	$k_1 = 0.1111\, dm^3mol^{-1}s^{-1}$ $k_2 = 0.0682\, dm^3mol^{-1}s^{-1}$ $k_3 = 0.2333\, dm^3mol^{-1}s^{-1}$ $k_4 = 0.2035\, s^{-1}$ $k_5 = 0.2109\, s^{-1}$

Run	Best Performing Individual	Reaction Rate Constant
5-38	$\begin{bmatrix} -2 & 1 & 0 & 0 & 0 \\ -1 & 0 & 1 & 0 & 0 \\ 1 & 0 & -1 & 0 & 0 \\ 0 & -1 & -1 & 0 & 1 \\ 1 & -1 & 1 & 0 & 0 \end{bmatrix}$	$k_1 = 0.1766\ dm^3mol^{-1}s^{-1}$ $k_2 = 0.1693\ s^{-1}$ $k_3 = 0.0051\ s^{-1}$ $k_4 = 0.1012\ dm^3mol^{-1}s^{-1}$ $k_5 = 0.0056\ s^{-1}$
5-40	$\begin{bmatrix} 0 & -1 & 2 & 0 & 0 \\ -1 & 0 & 0 & 1 & 0 \\ 0 & -1 & -1 & 0 & 1 \\ 0 & 1 & 0 & -2 & 0 \\ 2 & 0 & 2 & -1 & -1 \end{bmatrix}$	$k_1 = 0.0120\ s^{-1}$ $k_2 = 0.2798\ s^{-1}$ $k_3 = 0.3420\ dm^3mol^{-1}s^{-1}$ $k_4 = 0.5178\ dm^3mol^{-1}s^{-1}$ $k_5 = 0.0777\ dm^3mol^{-1}s^{-1}$

Table 8.4-3 Summarised results for Run 5-34, 5-36, 5-38 and 5-40

Previously discussed in Section 5.7, the performance of the automated system (version 2) is highly affected by the presence of noise in the system and presence of unmeasured chemical species, causing it unable to elucidate the entire CRN effectively. From the 4 runs presented in Table 8.4-3, none of them managed to elucidate the CRN that successfully identified all the involved reactions in RN1.

In comparison, Run 8-1 is much more successful in identifying all the reactions with just an additional reaction that are not part of the original RN1 which only have a marginally small effect on the performance of the entire elucidated CRN.

8.4.3 *Discussions for Run 8-2*

The results for Run 8-2 is also presented in Table 8.4-2 which again managed to elucidate all the reactions within RN2. It does however contain 2 additional reactions that do not belong to RN2 which similar to Run 8-1 have a very small value and have insignificant effect on the overall CRN.

Table 8.4-4 shows a reduced view of Table 5.7-5 which are runs that are done on single batch basis using automated system (version 2) with the same level of noise for comparison purposes.

Run	Best Performing Individual	Reaction Rate Constant
5-42	$\begin{bmatrix} -2 & 1 & 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & -2 & 0 & 1 \\ 0 & -1 & 2 & 0 & 0 & 0 \\ -1 & 1 & 1 & 0 & 0 & 0 \\ 0 & -2 & 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & -1 & 1 & 0 \end{bmatrix}$	$k_1 = 0.1258\ dm^3mol^{-1}s^{-1}$ $k_2 = 0.1784\ dm^3mol^{-1}s^{-1}$ $k_3 = 0.1769\ s^{-1}$ $k_4 = 0.1304\ s^{-1}$ $k_5 = 0.1752\ dm^3mol^{-1}s^{-1}$ $k_6 = 0.1174\ s^{-1}$

Run	Best Performing Individual	Reaction Rate Constant
5-44	$\begin{bmatrix} -1 & -1 & 1 & 1 & 0 & 0 \\ 0 & -1 & -1 & 0 & 1 & 0 \\ -1 & 0 & 0 & -1 & 0 & 1 \\ 2 & 1 & 2 & 0 & -1 & -1 \\ -1 & 2 & -1 & 0 & 0 & 0 \\ 1 & 1 & -1 & -1 & 0 & 0 \end{bmatrix}$	$k_1 = 0.1862\, dm^3 mol^{-1} s^{-1}$ $k_2 = 0.0861\, dm^3 mol^{-1} s^{-1}$ $k_3 = 0.1427\, dm^3 mol^{-1} s^{-1}$ $k_4 = 0.0779\, dm^3 mol^{-1} s^{-1}$ $k_5 = 0.0458\, dm^3 mol^{-1} s^{-1}$ $k_6 = 0.0348\, dm^3 mol^{-1} s^{-1}$
5-46	$\begin{bmatrix} -1 & -1 & 1 & 1 & 0 & 0 \\ -1 & 0 & 0 & -1 & 0 & 1 \\ 1 & -1 & 0 & -1 & 1 & 0 \\ 2 & 1 & -1 & 0 & 0 & -1 \\ 1 & 1 & 0 & -2 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}$	$k_1 = 0.2643\, dm^3 mol^{-1} s^{-1}$ $k_2 = 0.1349\, dm^3 mol^{-1} s^{-1}$ $k_3 = 0.0835\, dm^3 mol^{-1} s^{-1}$ $k_4 = 0.0115\, dm^3 mol^{-1} s^{-1}$ $k_5 = 0.0469\, dm^3 mol^{-1} s^{-1}$ $k_6 = 0$
5-48	$\begin{bmatrix} -1 & -1 & 1 & 1 & 0 & 0 \\ -1 & 0 & 0 & -1 & 0 & 1 \\ 1 & -1 & 0 & -1 & 1 & 0 \\ 1 & 1 & -1 & -1 & 0 & 0 \\ 0 & 1 & 0 & -2 & 2 & 0 \\ 0 & 2 & 0 & -1 & 0 & 0 \end{bmatrix}$	$k_1 = 0.2176\, dm^3 mol^{-1} s^{-1}$ $k_2 = 0.1576\, dm^3 mol^{-1} s^{-1}$ $k_3 = 0.0973\, dm^3 mol^{-1} s^{-1}$ $k_4 = 0.0596\, dm^3 mol^{-1} s^{-1}$ $k_5 = 0.0235\, dm^3 mol^{-1} s^{-1}$ $k_6 = 0.0082\, s^{-1}$

Table 8.4-4 Summarised results for Run5-42, 5-44, 5-46 and 5-48.

From Table 8.4-4 and as previously discussed in Section 5.7, the previous algorithm did not manage to unearth the complete set of reactions for RN2. In comparison, Run 8-2 shows that the newly modified algorithm for the automated system can elucidate all the reactions for RN2. Again, similar to Run 8-1, there are unrelated reactions to RN2 but these reactions have relatively small value in reaction rate constants and thus have insignificant impact to the overall performance of the RN2.

8.4.4 *Summary of results for Run 8-1 and Run 8-2*

In summary, the adjustment to the automated system that enables it to evaluate datasets with different initial conditions improve its ability in elucidating all the reactions in RN1 and RN2. This is a much better performance in comparison to automated system (version 2) with the results discussed in Section 5.7, which did not manage to elucidate the complete CRN of RN1 and RN2. However, there are still unrelated reaction to RN1 and RN2 that are misidentified in the Run 8-1 and Run 8-2 but they do have a relatively small reaction rate constants which give them very small impact to the overall performance of the CRN. The user of the automated system can also make further studies to investigate on these unrelated reactions to see the viability of the reactions in order to rule them out.

8.5 Modifications on automated system to include multiple batches with different process temperatures

Another method to increase variability in the input data to the automated system is to use datasets that belonged to batches that are run at different temperatures. The increase in variability in the datasets will suppress the effect of noise and increase the significance of the underlying reactions.

Due to the fact that each of the datasets comes from batches operated at different temperatures, the reaction rate constants for each of the reactions will be different from one batch to another. The reaction rate constants are after all dependent on the process temperature as shown in the Arrhenius equation in Equation 2-19.

Figure 8.5-1 shows the process flow of the modified automated system that incorporates the evaluation of datasets from batches operated at different temperatures.

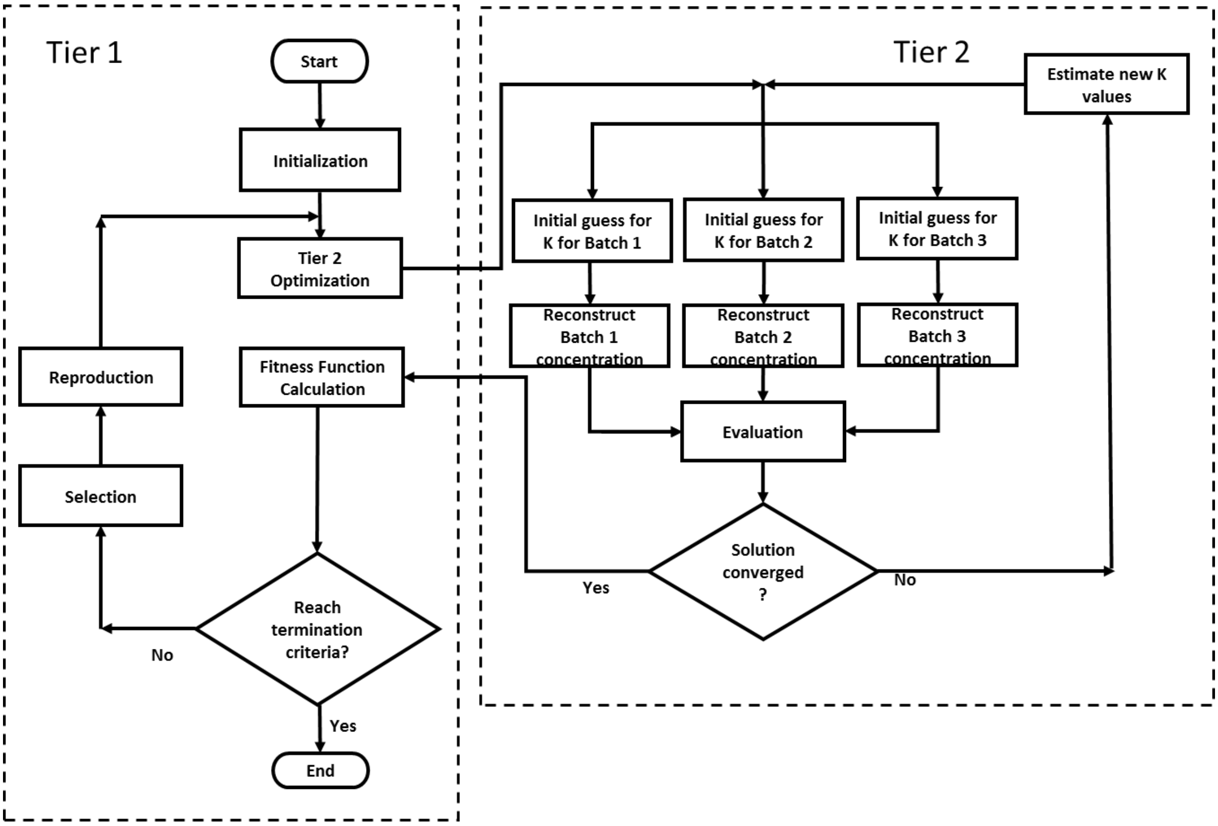

Tier 1
Start
Initialization
Tier 2 Optimization
Reproduction
Fitness Function Calculation
Selection
Reach termination criteria?
No
Yes
End
Tier 2
Estimate new K values
Initial guess for K for Batch 1
Initial guess for K for Batch 2
Initial guess for K for Batch 3
Reconstruct Batch 1 concentration
Reconstruct Batch 2 concentration
Reconstruct Batch 3 concentration
Evaluation
Solution converged ?
Yes
No

Figure 8.5-1 Flowchart for automated system to handle multiple datasets with different operating temperatures

For this modification, each of the batches with different operating temperature are evaluated separately. For the example shown in Figure 8.5-1, there are 3 batches with different operating temperature evaluated at the same time. Each CRN that is evaluated will be optimised for its reaction rate constants for each of the batch with different operating temperature. So, in the example in Figure 8.5-1, Batch 1 will have its own set of reaction rate constants, Batch 2 and Batch 3 likewise. The batches data will be reconstructed based on the determined reaction rate constants and their fitness will be used to evaluate the performance of the candidate CRN.

8.6 Results from modification to include multiple batches with different operating temperatures

8.6.1 *Run set up and results*

The adjusted TMOA and AA experimental data prepared in Section 7.6 are used to test out the modified automated system. All 3 datasets from batches operating at 80^{o}C, 90^{o}C and 100^{o}C are used to elucidate the CRN for the reactions between TMOA and AA. Similar to Section 7.7, 2 additional ‘fake’ chemical species A and B are introduced into the system to test its capability.

Run	Reaction	Batch	Unmeasured	‘Fake’ chemical species
8-3	TMOA and AA	Exp 1, Exp 2 and Exp 3	AA and MA	Chemical species A and B as introduced in Section 7.7

Table 8.6-1 Run details for Run 8-3

The automated system’s parameters remains unchanged. Table 8.6-2 shows the results from the run.

Run	Best Performing Individual
8-3	$\begin{bmatrix} -1 & 1 & 0 & 0 & -1 & 1 & 0 & 0 \\ 1 & -1 & 0 & 0 & 1 & -1 & 0 & 0 \\ 0 & -1 & 1 & 0 & -1 & 1 & 0 & 0 \\ -1 & 2 & -1 & 0 & 0 & 0 & 0 & 0 \\ 1 & -2 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & -2 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}$

Table 8.6-2 Results for Run 8-3

Similar to Run 8-1 and Run 8-2, fitness function data is not presented because it does not serve as a good comparison for previous runs because it consist of 3 different datasets in one. The reaction rate constants are also dissimilar from one batch to another and no comparison can be derived from it.

8.6.2 *Discussion and summary for Run 8-3*

The results presented in Table 8.6-2 shows that the modified algorithm had managed to elucidate 3 of the 6 reactions from the expected reaction between TMOA and AA as presented in Equation 7-2 which is shown again below for better reference.

$$\boldsymbol{V_{Exp1}} = \begin{bmatrix} -1 & 1 & 0 & 0 & -1 & 1 \\ 1 & -1 & 0 & 0 & 1 & -1 \\ 0 & -1 & 1 & 0 & -1 & 1 \\ 0 & 1 & -1 & 0 & 1 & -1 \\ 0 & 0 & -1 & 1 & -1 & 1 \\ 0 & 0 & 1 & -1 & 1 & -1 \end{bmatrix} \qquad \text{(Equation 7-2)}$$

The elucidated CRN shown in Table 8.6-2 contained 3 reactions that are combination of reactions in Equation 7-2. The 4th reaction in the elucidated CRN is combination of 1st and 4th reaction in Equation 7-2, the 5th reaction in the elucidated CRN is combination of the 2nd and 3rd reaction in Equation 7-2 and the 6th reaction in the elucidated CRN is combination of 4th and 5th reaction. Through this combination, it can be seen that the elucidated CRN still did not manage to capture all the reactions but did manage to capture the dynamics of at least 5 reactions of the 6 reactions in Equation 7-2.

For comparison purposes, the elucidated CRN in Table 8.6-2 is compared against the CRNs elucidated for Run 7-7, Run 7-8 and Run 7-9 which were presented in Section 7.7 and the summary of the runs are presented again below in Table 8.6-3.

Run	Best Performing Individual
7-7	$\begin{bmatrix} -1 & 1 & 0 & 0 & -1 & 1 & 0 & 0 \\ 0 & 1 & -2 & 1 & 0 & 0 & 0 & 0 \\ 0 & -1 & 2 & -1 & 0 & 0 & 0 & 0 \\ 1 & -2 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & -1 & 0 & 0 & -1 & 0 & 0 & 1 \\ -1 & 1 & 1 & -1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 & -1 & 0 \\ 0 & 0 & 0 & 0 & 1 & -1 & 1 & -1 \end{bmatrix}$
7-8	$\begin{bmatrix} -1 & 1 & 0 & 0 & -1 & 1 & 0 & 0 \\ 0 & 1 & -2 & 1 & 0 & 0 & 0 & 0 \\ -1 & 1 & 1 & -1 & 0 & 0 & 0 & 0 \\ 1 & -1 & -1 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & -1 & 1 & -1 & 1 \\ 0 & 1 & 0 & 0 & 0 & 1 & -1 & 0 \\ 0 & -1 & 2 & -1 & 0 & 0 & 0 & 0 \\ 1 & -2 & 1 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}$

Run	Best Performing Individual
7-9	$\begin{bmatrix} -1 & 1 & 0 & 0 & -1 & 1 & 0 & 0 \\ -1 & 2 & -1 & 0 & 0 & 0 & 0 & 0 \\ -1 & 1 & 1 & -1 & 0 & 0 & 0 & 0 \\ 0 & 1 & -2 & 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 2 & -1 & 0 & -1 \\ 0 & 0 & 1 & 0 & 0 & 1 & 0 & -1 \\ 1 & -2 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & -1 & 1 & -1 & 0 & 0 \end{bmatrix}$

Table 8.6-3 Summarised results for Run 7-7 to Run 7-9

The main difference that can be captured is that the 'fake' chemical species is not present at all in the elucidated CRN with the modified automated system which is a vast improvement against the previous runs with automated system (version 2). Run 7-7, Run 7-8 and Run 7-9 are not able to distinguish between involved chemical species and those that are not involved in the system but when given more variable datasets, Run 8-3 managed to exclude the 'fake' chemical species.

8.7 Modifications on automated system to include multiple batches with different initial conditions and different operating temperatures

From the results presented in Section 8.4 and Section 8.6, it can be seen the benefits of using more datasets with a larger range of variations through having different initial conditions and operating temperatures. It is therefore safe to assume that combining the both of the methods will have a massive improvement in the performance of the automated system in the discovery of the CRN. The flowchart shown in Figure 8.7-1.

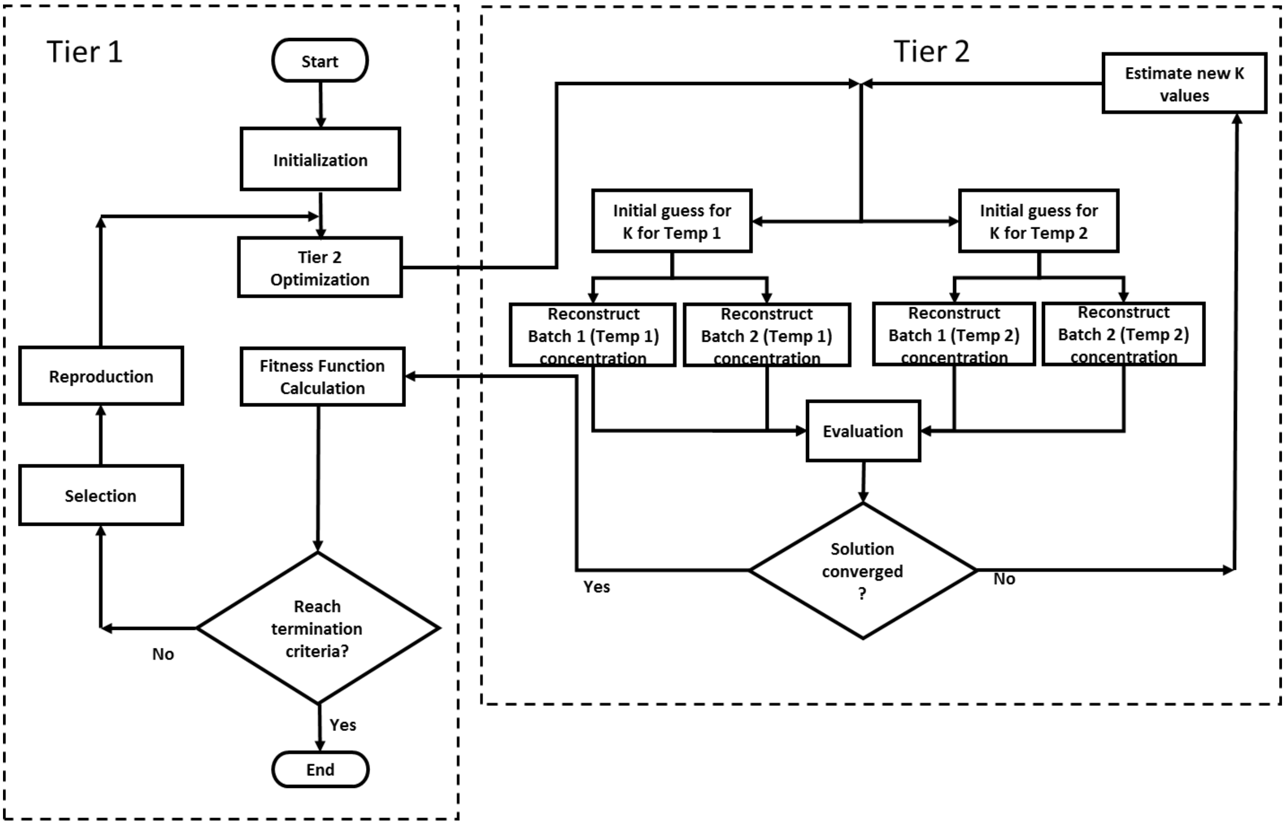

Tier 1
Start
Initialization
Tier 2 Optimization
Reproduction
Fitness Function Calculation
Selection
Reach termination criteria?
No
Yes
End
Tier 2
Estimate new K values
Initial guess for K for Temp 1
Initial guess for K for Temp 2
Reconstruct Batch 1 (Temp 1) concentration
Reconstruct Batch 2 (Temp 1) concentration
Reconstruct Batch 1 (Temp 2) concentration
Reconstruct Batch 2 (Temp 2) concentration
Evaluation
Solution converged ?
Yes
No

Figure 8.7-1 Flowchart for automated system that can analyse datasets from batches with different initial conditions and operating temperatures

The flowchart in Figure 8.7-1 represents the final form of the automated system that is develop in this work. The NSGA-II algorithm presented in Chapter 6 can still be included into this final form by including the algorithm in the reproduction stage in Tier 1 of the automated system. NSGA-II will help to expand the possible number of reactions in the system for the user to determine their viability.

Due to the lack of suitable datasets to test out the automated system's final form, no investigation is conducted on the effectiveness of the automated system. Based on the positive results obtained from Section 8.4 and Section 8.6, it can be assumed that this system will perform better than those obtained in Section 8.4 and Section 8.6. Unfortunately, the work will be have to be done in future work when there is availability of suitable datasets.

8.8 Summary

It is clear that the modifications done to the automated system for it to process datasets that belonged to batches with different initial conditions simultaneously increases its performance in elucidating the CRN accurately. It had been shown to work with RN1 and RN2 and able to elucidate all the reactions in both of the CRNs. There are some unrelated reactions that are generated but these come with very small reaction rate constants and can be investigated further by the user to discount them from the CRNs. This is positive result in comparison to the results in Section 5.7 which did not manage to obtain all the reactions in the CRN.

When the automated system is modified to take in process data from batches with different operating temperatures, it shows positive result of able to elucidate most of the reactions in the expected reaction and not being misdirected to include any ‘fake’ chemical species in the elucidated CRN. Similar to the previous modification, this is a positive direction as in comparison to the results obtained in Section 7.7 which is not able to differentiate between involved and uninvolved chemical species.

Lastly, with both of the conclusion drawn on the modifications, it leads to the combination of both modifications to the automated system to form the final form of the automated system. It can be assumed that this automated system will be superior to the previous iterations of the automated system. Unfortunately, the system is not tested with any datasets due to the lack of a suitable one in this work. The work to prove the viability of the automated system final form will be left for future work.

Chapter 9. Discussion and Conclusion

9.1 Discussions

This thesis proposes and designs an automated system to facilitate in the discovery of chemical reaction network from the concentration data of chemical species in an isothermal chemical batch reactor. The automated system is designed based on genetic algorithm, an approach that has never been done before. It is noticed early on that the biggest challenge to the problem is calculating for the reaction rate constants.

Linearisation of the solution for the reaction rate constants is done and made it possible to use multiple linear regression to solve for it. Small successes are achieved from it as the automated system is able to elucidate one of the two fictitious chemical reaction network tested in this work. It was later on proven not feasible because of multiple limitations such as unable to work with reversible reactions and inability to handle unmeasured chemical species in its elucidation work.

Further modifications are done to the system by incorporating a second tier optimisation loop to the automated system using a non-linear optimiser to estimate and calculate the values of the reaction rate constants. The automated system is shown to be successful when dealing with datasets that has no noise present in the input concentration data. When noise is introduced, the automated system still managed to cope but started to miss reaction in its discovered network and including reaction that are not part of the chemical reaction network. This is down to ‘overfitting’ problem where the automated system will try to force the predicted concentration to fit the input concentration even if it is fitting to pure noise.

‘Overfitting’ is a major issue when unmeasured chemical species is used. The automated system will fit the predicted data to the measured chemical species while ignoring the unmeasured chemical species because it lacks the data to do so. This results in highly fitted measured chemical species but terrible fit unmeasured chemical species. The final CRN structure suffers as a result of such behaviour.

Multi objective optimisation is included into the automated system in order to curb its behaviour where it will only looked to increase fitness to the input concentration data. Average relative reactants’ molecular weight is introduced so the system is able to look for more diversity in its end result. This is important especially in the case where there are unmeasured chemical species causing reactions from the actual chemical reaction network unable to perform well because it lacks the information of the unmeasured chemical species. With this, there is still a chance they get included in the final result of the runs. Tests run using the automated system with multi objective optimisation capability comes to the conclusion it may require more data in order to perform well.

An experimental result from the reaction of trimethyl orthoacetate and allyl alcohol is used to test the automated system. It has been noted that the experimental data is poor in quality due to the large material imbalances detected within the data. However, the automated system still managed to elucidate most of the reaction from the expected chemical reaction network. Further tests were conducted by including ‘fake’ chemical species which does not participate in the reaction into the automated system. It proves that their presence will reduce the system’s performance and cause the automated system to even include reactions that contain those ‘fake’ chemical species in the end result.

In Chapter 8, the automated system is further developed so that it will be able to handle datasets that consists of runs from multiple batches with different initial conditions. RN1 and RN2 are used to test the automated system and the results had shown to be superior to the previous iterations of the automated system. All the reactions from the actual CRNs are elucidated and although the final CRN structure contains reactions that are not part of RN1 and RN2, it is noted that the values are relatively small enough that they have insignificant impact on the chemical reaction networks. The automated system is then further modified to handle data from multiple batches from different operating temperatures and adjusted experimental data for the reaction between trimethyl orthoacetate and allyl alcohol is used to test the system. It is discovered that the automated system is able to discern between involved and uninvolved chemical species as it does not elucidate the ‘fake’ chemical species that are introduced in the system. The final CRN structure that the automated system deduced is also comparable to the expected reaction between trimethyl orthoacetate and allyl alcohol with only one reaction missing. This again is a more positive results from work done in previous iterations of the automated system. A final form of the automated system is introduced at the end of the chapter which combines both of the modifications which is expected to perform better than any of the previous iterations of the automated system. However, the test for the performance of the final form is beyond the scope of this work as there is a lack of suitable datasets to test the system on. This work will be left for future work.

9.2 Conclusion

The conclusions that can be drawn from this thesis are

- Linear solution to reaction rate constants is possible but is highly sensitive to the content of the reactions and will require them to be linearly independent.
- Automated system with two tier optimisation loop is capable of elucidating CRNs but suffers when noise is introduced into the system.
- Presence of unmeasured chemical species will cause automated system to skew its fitting target towards the measured chemical species causing the fitness function on unmeasured chemical species to suffer.
- Multi-objective optimisation is shown to be successful in create more diversity in the final result but is still unable to overcome the ‘overfitting’ issue.
- Providing the automated system with larger variation in data by modifying the automated system to run the datasets from batches made with different initial conditions or different operating temperature simultaneously.

Final conclusion of the thesis is that an automated system that can identify the involved reactions and elucidate the chemical reaction network through the use of genetic algorithm has been developed. The final form of the automated system although untested, is expected to be able to perform better than the previous iterations. This shows the viability of using evolutionary algorithm in progressing the work for the development of automated identification of complex chemical reaction networks and further work should be invested into development of a more efficient algorithm that can take in larger number of variables than those that had been considered in this thesis.

9.3 Future work

This work had shown the viability of genetic algorithm, one of the many possible options for evolutionary algorithms to assist in the elucidation of chemical reaction networks through data mining of the concentration data of chemical species.

Two possible trajectories for future work can be considered, the first is to continue development of the work that had been presented in this thesis starting from the final form of the automated system shown in Chapter 8. The final form of the automated system performance remains untested and should be the initial focus of the future work. It should be tested against more complex chemical reaction networks using datasets from batches with different initial conditions and different operating temperatures. This will undoubtedly unearthed further complications in the system and will need to be resolved. Further considerations needs to be made to increase the speed of the algorithm and the current system will take up substantial computational resource to converge to a result. Options may involve designing the codes to take into account parallel processing of several machines to complete the work, introduce additional rule sets to discount unviable chemical reaction networks through consideration of Gibb's free energy of the reactions and use linear approximations of the reaction rate constants whenever possible. The linear approximations of the reaction rate constants can be done by isolating the reactions that consisted of only measured chemical species that are not involved in any other reactions. These isolated reactions' reaction rate constants can then be approximated separately by using the multiple linear regression as discussed in Chapter 4. The identifiability of the reaction rate constants can be considered and those CRNs with non-identifiable reaction rate constants will not be considered to reduce the load of the system. Finally, it will be imperative to collect more datasets of a single chemical reaction network so that a sufficiently large training, validation and testing datasets can be set up to test if it helps to increase accuracy of the automated system.

The second focus for future work will be to involve the use of different applications of the evolutionary algorithm such as evolutionary strategies, estimation of distribution algorithm and genetic expression programming. The differences in these algorithms to genetic algorithm is not major and the automated system developed in this thesis is transferable to other algorithms. The main aim of this future work is to compare the efficiencies of the algorithms and although genetic algorithm may be considered one of the easiest to understand, it may not necessarily the most efficient in terms of computational power. Further investigations on the impact of each of the algorithms on the final results should be made and if possible, combine the best portion of each algorithms to create a superior automated system for elucidation of chemical reaction networks. The work can further be extended beyond evolutionary algorithm to include swarm intelligence which involved algorithm such as ant colony optimisation and particle swarm optimisation. The end goal will be to develop an effective and efficient automated system through the investigations of all the available algorithms.

Finally, the two trajectories can converge to form a wholesome program that will identify the chemical reaction networks effectively and efficiently. The program can determine what input data it requires to improve its performance. This will be in line with the main goal of this thesis and further it by developing an entity that can not only analyse data to elucidate chemical reaction network but self-introspect to determine what weaknesses it possessed and request for further input to increase its own performance.

References

Aytug, A., & Koehler, G. J. (1996). Stopping Criteria for Finite Length Genetic Algorithms. *ORSA Journal on Computing, 8*(2), 183-191.

Aytug, H., & Koehler, G. J. (2000). New Stopping Criterion for Genetic Algorithms. *European Journal of Operational Research, 126*(2), 662-674.

Bonvin, D., & Rippin, D. W. (1990). Target Factor Analysis for the Identification of Stoichiometric Models. *Chemical Engineering Science, 45*, 3417–3426.

Brendel, M., Bonvin, D., & Marquardt, W. (2006). Incremental Identification of Kinetic Models for Homogeneous Reaction Systems. *Chemical Engineering Science, 61*, 5404-5420.

Budin, L., Golub, M., & Budin, A. (1996). Traditional Techniques of Genetic Algorithms. *KoREMA ‘96. - 41st Annual Conference*, (pp. 93-96). Opatija.

Burnham, S. C. (2007). Towards the Automated Determination of Chemical Reaction Networks. Newcastle Upon Tyne, United Kingdom: Newcastle University.

Burnham, S. C., Searson, D. P., Willis, M. J., & Wright, A. R. (2008). Inference of Chemical Reaction Networks. *Chemical Engineering Science, 63*(4), 862-873.

Cao, H., Yu, J., Kang, L., Chen, Y., & Chen, Y. (1999). The Kinetic Evolutionary Modeling of Complex Systems of Chemical Reactions. *Computers & Chemistry, 23*(2), 143-1551.

Crampin, E. J., Schnell, S., & McSharry, P. E. (2004). Mathematical and Computational Techniques to Deduce Complex Biochemical Reaction Mechnisms. *Progress in Biophysics and Molecular Biology, 86*, 77-112.

Davis, M. E., & Davis, R. J. (2003). *FUNDAMENTALS OF CHEMICAL REACTION ENGINEERING.* New York: McGraw-Hili.

De Jong, K. (1988). Learning with Genetic Algorithms: An Overview. *Machine Learning, 3*, 121-138.

Deb, K. (n.d.). *Multi-Objective Optimization Using Evolutionary Algorithms.* John Wiley & Sons.

Deb, K., Pratap, A., Agarwal, S., & Meyarivan, T. (2002). A fast and elitist multiobjective genetic algorithm: NSGA-II. *IEEE Transactions on Evolutionary Computation 6 (2)*, 182–197.

Elliott, L., Ingham, D. B., Kyne, A. G., Mera, N. S., Pourkashanian, M., & Wilson, C. W. (2004). Genetic Algorithms for Optimisation of Chemical Kinetics Reaction Mechanisms. *Progress in Energy and Combustion Science, 30*(3), 297-328.

Filippi, C., L., G. J., Bordet, J., & J., V. (1986). Tendency Modeling of Semibatch Reactors for Optimization and Control. *Chemical Engineering Science, 41*(4), 913-920.

Fotopoulos, J., Georgakis, C., & Stenger Jr., H. G. (1994a). Uncertainty Issues in the Modeling and Optimization of Batch Reactors with Tendency Models. *Chemical Engineering Science, 49*, 5533-5547.

Fotopoulos, J., Georgakis, C., & Stenger Jr., H. G. (1994b). Structured Target Factor Analysis for the Stoichiometric Modeling of Batch Reactors. *American Control Conference*, *1*, pp. 495-499. doi:10.1109/ACC.1994.751786

Gen, M., Cheng, R., & Wang, D. (1997). Genetic Algorithms for Solving Shortest Path Problems. *IEEE Transactions on Evolutionary Computation*, (pp. 401-406).

Goldberg, D. E., & Holland, J. H. (1998). Genetic Algorithms and Machine Learning. *Machine Learning, 3*, 95-99.

Goldberg, D. E., Korb, B., & Deb, K. (1989). Messy Genetic Algorithms: Motivations, Analysis, and First Results. *Complex Systems, 3*, 493-530.

Guldberg, C. M., & Waage, P. (1879). Concerning Chemical Affinity. *Erdmann's Journal fur Practische Chemie, 127*, 69-114.

Harris, S. D., Elliott, L., Ingham, D. B., Pourkashanian, M., & Wilson, C. W. (2000). The Optimisation of Reaction Rate Parameters for Chemical Kinetic Modelling of Combustion using Genetic Algorithms. *Computer Methods in Applied Mechanics and Engineering, 190*(8-10), 1065-1090.

Haupt, R. L., & Haupt, S. E. (2004). *Practical Genetic Algorithms, Second Edition.* John Wiley & Sons, Inc.

Hedar, A. R., Ong, B. T., & Fukushima, M. (2007). *Genetic Algorithms with Automatic Accelerated Termination.* Technical Report, Dept. of Applied Mathematics and Physics, Kyoto University.

Hernandez, J. J., Ballesteros, R., & Sanz-Argent, J. (2010). Reduction of Kinetic Mechanisms for Fuel Oxidation through Genetic Algorithms. *Mathematical and Computer Modelling, 52*(7-8), 1185-1193.

Holland, J. (1975). *Adaptation in Natural and Artificial Systems.* Ann Arbor: University of Michigan Press.

Jackson, R. A. (2004). *Mechanisms in Organic Reactions.* Cambridge: Royal Society of Chemistry.

Kaelo, P., & Ali, M. M. (2007). Integrated Crossover Rules in Real Coded Genetic Algorithms. *European Journal of Operational Research, 176*, 60-76.

Keyvanloo, K., Mehdi, S., & Towfighi, J. (2012). Genetic Algorithm Model Development for Prediction of Main Products in Thermal Cracking of Naphtha: Comparison with Kinetic Modeling. *Chemical Engineering Journal, 209*, 255-262.

Kikuchi, S., Tominaga, D., Arita, M., Takahashi, K., & Tomita, M. (2003). Dynamic Modeling of Genetic Networks using Genetic Algorithm and S-system. *Bioinformatics, 19*(5), 643-650.

Koza, J. R., Bennet, F. H., Angre, D., & Keane, M. A. (1999). *Genetic Programming III: Darwinian Invention and Problem Solving.* San Francisco, CA: Morgan Kaufmann.

Koza, J. R., Mydlowec, W., Lanza, G., Yu, J., & Keane, M. A. (2007). Automatic Computational Discovery of Chemical Reaction Networks Using Genetic Programming. In D. Sašo, & T. Ljupco, *Computational Discovery of Scientific Knowledge* (pp. 205-227). Springer-Verlag Berlin Heidelberg.

Kudryavtsev, A., Jameson, R., & Linert, W. (2001). *The Law of Mass Action.* Springer-Verlag Berlin and Heidelberg GmbH & Co. KG.

Le Lann, M. V., Cabassud, M., & Casamatta, G. (1999). Modeling, Optimization and Control of Batch Chemical Reactors in Fine Chemical Production. *Annual Reviews in Control, 23*, 25-34.

Levenspiel, O. (1999). *Chemical Reaction Engineering Third Edition.* John Wiley and Sons.

Lin, K. H. (2004). Reaction Kinetics. In R. C. Dorf, *The Engineering Handbook, Second Edition* (p. Chapter 78). CRC Press.

Maeder, M., & Neuhold, Y. P. (2004). Application of a Genetic Algorithm: Near Optimal Estimation of the Rate and Equilibrium Constants of Complex Reaction Mechanisms. *Chemometrics and Intelligent Laboratory Systems, 70*(2), 193-203.

Maria, G. (2004). A Review of Algorithms and Trends in Kinetic Model Identification for Chemical and Biochemical Systems. *Chem. Biochem. Eng., 18*(3), 195–222.

McNaught, A. D., & Wilkinson, A. (1997). *IUPAC. Compendium of Chemical Terminology, 2nd ed. (the "Gold Book").* Oxford: Blackwell Scientific Publications. doi:10.1351/goldbook

Perini, F., Brakora, J. L., Reitz, R. D., & Cantore, G. (2012). Development of Reduced and Optimized Reaction Mechanisms based on Genetic Algorithms and Element Flux Analysis. *Combustion and Flare, 159*(1), 103-119.

Raidl, G. R. (1999). Weight Codings in a Genetic Algorithm for the Multiconstraint Knapsack Problem. *IEEE Congress on Evolutionary Computation*, (pp. 596-603). Washington DC.

Rangarajan, S., Bhan, A., & Daoutidis, P. (2012). Language-oriented Rule-based Reaction Network Generation and Analysis: Description of RING. *Computers & Chemical Engineering*, In press.

Rastogi, A., Fotopoulos, J., Georgakis, C., & Stenger, H. G. (1992). The Identification of Kinetic Expressions and the Evolutionary Optimization of Specialty Chemical Batch Reactors using Tendency Models. *Chemical Engineering Science, 47*(9-11), 2487-2492.

Rastogi, A., Vega, A., Gerogakis, C., & G., S. H. (1990). Optimization of Catalyzed Epoxidation of Unsaturated Fatty Acids by Using Tendency Models. *Chemical Engineering Science, 45*(8), 2067-2074.

Reyes-Sierra, M., & Coello, C. A. (2006). Multi-Objective particle swarm optimizers: A survey of the state-of-the-art. *Int. J. Comput. Intell. Res. 2*, 287-308.

Rostrup-Nielsen, J. (2000). Reaction Kinetics and Scale-up of Catalytic Processes. *Journal of Molecular Catalysis A: Chemical, 163*, 157–162.

Savageau, M. A. (1976). *Biochemical System Analysis: a Study of Function and Design in Molecular Biology.* Reading, MA.: Addison-Wesley.

Searson, D. P., Willis, M. J., & Wright, A. R. (2012). Reverse Enginering Chemical Reaction Networks from Time Series Data. In M. Dehmer, K. Varmuza, & D. Bonchev, *Statistical Modelling of Molecular Descriptors in QSAR/QSPR* (pp. 327-348). John Wiley & Sons Incorporated.

Searson, D. P., Willis, M. J., Horne, S. J., & Wright, A. R. (2007). Inference of Chemical Reaction Networks Using Hybrid S-system Models. *Chemical Product and Process Modeling, 2*(1).

Storn, R., & Price, K. (1997). Differential Evolution - A Simple and Efficient Heuristic for Global Optimization over Continuous Spaces. *Journal of Global Optimization, 11*, 341-359.

Tsoulos, I. G. (2008). Modifications of Real Code Genetic Algorithm for Global Optimization. *Applied Mathematics and Computation, 203*(2), 598-607.

Voit, E. O., & Almeida, J. (2004). Decoupling Dynamical Systems for Pathway Identification from Metabolic Profiles. *Bioinformatics, 20*(11), 1670-1681.

Waago, P., & Guldherg, C. M. (1864). Forhandlinger: Videnskabs-Selskabet i Christiana. *35*.

Wang, Z., Yang, B., Chen, C., Yuan, J., & Wang, L. (2007). Modeling and Optimization for the Secondary Reaction of FCC Gasoline based on the Fuzzy Neural Network and Genetic Algorithm. *Chemical Engineering and Processing: Process Intensification, 46*(3), 175-180.

Whitley, D. (1994). A Genetic Algorithm Tutorial. *Statistics and Computing, 4*, 65-85.

Yin, B., Wei, Z., & Meng, Q. (2005). The Study of Special Encoding in Genetic Algorithms and a Sufficient Convergence Conditions of GAs. In L. Wang, K. Chen, & Y. Ong, *Advances in Natural Computation* (pp. 426-426). Springer-Verlag Berlin Heidelberg.