2014 IEEE Congress on Evolutionary Computation (CEC)
July 6-11, 2014, Beijing, China

An Immune Network Approach to Learning Qualitative Models of Biological Pathways

Wei Pang
School of Natural and Computing Sciences
University of Aberdeen
Aberdeen, UK AB24 3UE
Email: pang.wei@abdn.ac.uk

George M. Coghill
School of Natural and Computing Sciences
University of Aberdeen
Aberdeen, UK AB24 3UE
Email: g.coghill@abdn.ac.uk

***Abstract*—In this paper we continue the research on learning qualitative differential equation (QDE) models of biological pathways building on previous work. In particular, we adapt opt-AiNet, an immune-inspired network approach, to effectively search the qualitative model space. To improve the performance of opt-AiNet on the discrete search space, the hypermutation operator has been modified, and the affinity between two antibodies has been redefined. In addition, to accelerate the model verification process, we developed a more efficient Waltz-like inverse model checking algorithm. Finally, a Bayesian scoring function is incorporated into the fitness evaluation to better guide the search. Experimental results on learning the detoxification pathway of Methylglyoxal with various hypothesised hidden species validate the proposed approach, and indicate that our opt-AiNet based approach outperforms the previous CLONALG based approach on qualitative pathway identification.**

I. Introduction

Qualitative Differential Equation Model Learning (QML) [1] involves inferring qualitative models in the form of qualitative differential equation (QDE) for a dynamic system from available data and background knowledge. QML is particularly suitable for situations where only sparse, noisy data and/or incomplete knowledge about the system are available. In the last three decades, a number of QML systems have been proposed to solve different problems and address various issues of QML. Examples of these systems include GENMODEL [2], MISQ [3], QSI [4], QME [5], ILP-QSI [6] (formerly known as QOPH [7]), and the most recent QML-Morven framework [8], [9], an earlier description of which can be found in [10].

In particular, in our previous work [11] we developed a special-purpose QML system for qualitative system identification of biological pathways. In this QML system we used an immune-inspired algorithm named CLONALG (the CLONal selection ALGorithm) [12] as a search strategy. For ease of description, in this paper this QML system will be named QML$_{PI}$-CLONALG, where "*PI*" means pathway identification. QML$_{PI}$-CLONALG aimed to address two issues of QML: first, how to make better use of domain specific knowledge (biological knowledge); second, how to improve the scalability of QML when dealing with large-sized model spaces. In that research we proposed a CLONALG based algorithm for searching multimodal model spaces (search spaces containing multiple global or local optima), and promising results were obtained. However, due to the expensive computational cost of qualitative simulation, for complicated candidate pathways it was not possible to perform the actual qualitative simulation, and this prevented us from further investigating the performance of immune inspired QML for pathway identification.

In this paper, given the assumption that in a complicated pathway there are many hidden variables (those variables that cannot be measured by biological experiments) and only a few measured variables, which is a very common situation in biology, we first develop a more efficient way for model verification. This allows us to perform in-depth experiments on testing the performances of immune-inspired QML systems. In particular, we focus on exploring the potential of an alternative immune-inspired approach, opt-AiNet [13], [14], on learning QDE models of pathways because of its previously proven performance on multi-model search spaces. More importantly, as reported in our previous research [15] opt-AiNet is an effective search strategy for general-purpose QML systems, and it can achieve comparable performance to CLONALG. This motivates us to explore the potential of opt-AiNet as a search strategy for special-purpose QML systems, in particular, the QML system for pathway identification problems. The resulting QML system is named QML$_{PI}$-AiNet.

The rest of this paper is organised as follows: we first briefly introduce the basics about QDE models in Section II. This is followed by a description of the algorithm for converting pathways to QDE models in Section III. In Section IV we give a formal description of the search space of the problem and define different kinds of pathways. The proposed QML$_{PI}$-AiNet will be presented in Section V, and the experiments to validate QML$_{PI}$-AiNet are detailed in Section VI. Finally Section VII concludes the paper.

II. Qualitative Differential Equations

In this research we use the *Morven* framework [16], [17] to represent QDE models. Formally, a QDE is defined as a tuple <V, Q, C, T> [18], where V represents the set of *qualitative variables*; Q is the set of *quantity spaces*, each of which is associated with a qualitative variable in V; C is a set of *qualitative constraints* that apply to the variables in

978-1-4799-1488-3/14/$31.00 ©2014 IEEE

TABLE I. THE SIGNS QUANTITY SPACE

Quantity	Range	
negative(-)	$(-\infty,$	$0)$
zero(0)	0	
positive(+)	$(0,$	$\infty)$

TABLE II. FUNCTION MAPPINGS UNDER THE SIGNS QUANTITY SPACE

Function(A,B)	negative	zero	positive
negative	1	0	0
zero	0	1	0
positive	0	0	1

V; T is a set of transitions between *qualitative states*. Simply speaking, a QDE is the conjunction of all its qualitative constraints, which link the qualitative variables and express the relations among these variables.

As for the set of quantity spaces Q, different qualitative reasoning engines may have different forms of representation, but all qualitative variables are restricted to only take qualitative values from their associated quantity spaces. The most commonly used and simplest quantity space is the *signs quantity space*, in which there are only three qualitative values: *positive*, *zero*, and *negative*, as shown in Table I.

The set of qualitative constraints C are of two types: *algebraic constraints* and *functional constraints*. The former represent algebraic relations between variables as in quantitative mathematics, for instance, *addition*, *subtraction*, and *multiplication*; the latter describe incomplete knowledge between two variables, for example, the monotonically increasing and decreasing relations, which state that one variable will monotonically increase with the increase/decrease of another.

$function$ constraints in the *Morven* framework are the above-mentioned functional constraints, and they define many-to-many mappings which allow flexible empirical descriptions between two variables without knowing the exact mathematical relation. One example of function mappings in *Morven* is shown in Table II. In this table variables A and B use the signs quantity space as shown in Table I; "1" stands for the existence of a mapping between variables A and B, and "0" otherwise.

Table IV lists some *Morven* constraints and their corresponding mathematical equations. In this table variables in the right column such as *X(t)* are continuous functions of time *t*. f is a function that is continuously differentiable over its domain. In the constraints listed in the left column of the table, the label *dt* means *derivative*, and the integer immediately following it indicates which derivative of the variable (0 means the magnitude). This means each place in a *Morven* constraint can represent not only the magnitude, but also arbitrary derivative of a variable.

We use the single tank system described in Fig. 1 as an example, and show how we represent this system by a QDE model under the *Morven* framework. The quantitative model for a linear version of the single tank system is as follows:

$$q_o = k * V,$$
$$dV/dt = q_i - q_o,$$

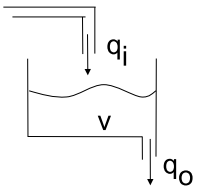

Fig. 1. The Single Tank System

TABLE III. THE *Morven* MODEL FOR THE SINGLE TANK SYSTEM

Differential Plane 0	
C1: Function (dt 0 q_o, dt 0 V)	$(q_o = k * V)$
C2: sub (dt 1 V, dt 0 q_i, dt 0 q_o)	$(V' = q_i - q_o)$
Differential Plane 1	
C3: Function (dt 1 q_o, dt1 V)	$(q_o' = k * V')$
C4: sub (dt 2 V, dt1 q_i, dt1 q_o)	$(V'' = q_i' - q_o')$

TABLE IV. SOME QUALITATIVE CONSTRAINTS IN *Morven* AND THEIR CORRESPONDING MATHEMATICAL EQUATIONS

Morven Constraints	Mathematical Equations
sub (dt 0 Z, dt 0 X, dt 0 Y)	$Z(t) = X(t) - Y(t)$
mul (dt 0 Z, dt 0 Y, dt 0 X)	$Z(t) = Y(t) \times X(t)$
div (dt 0 Z, dt 0 Y, dt 0 X)	$Z(t) = Y(t)/X(t)$
Function (dt 0 Y, dt 0 X)	$Y(t) = f(X(t))$
sub (dt 1 Z, dt 0 X, dt 0 Y)	$dZ(t)/dt = X(t) - Y(t)$
Function (dt 1 Y, dt 0 X)	$dY(t)/dt = f(X(t))$

V=<pos, zer, zer>
q_i=<pos, zer>
q_o=<pos,zer>

Fig. 2. A Qualitative State of the Single Tank in *Morven*

where V is the volume of the liquid in the tank, q_i is the inflow, q_o is the outflow, and k is a positive constant coefficient determined by the cross sectional area of the tank and the density of the liquid.

The corresponding *Morven* model is shown in Table III. This model is composed of four constraints, *C1* to *C4*, which are distributed over two *differential planes* [16]. The meaning of these constraints has been explained in Table IV, and the corresponding quantitative relation for each constraint is shown on the right hand side in the brackets. For variable V, the magnitude, the first and second derivatives are used; for variable q_o and q_i, only the magnitude and the first derivative are used.

If all the qualitative variables (including their magnitudes and derivatives) use the signs quantity space defined in Table I, the mappings of the *Function* in constraint *C1* and *C3* are given in Table II.

After qualitative simulation of a QDE model, the output could be either an *envisionment* containing all possible *qualitative states* and their legal transitions, or a behaviour tree which is part of the envisionment. A qualitative state is a complete assignment of qualitative values to all qualitative variables of the system. One possible qualitative state of the QDE model described in Table III is shown in Fig. 2. In this figure the assignment $V =< pos, zer, zer >$ means that the magnitude of V is *positive*, the first and second derivatives are *zero* (all values are taken from the signs quantity space defined in Table I). It is similar for the assignments of q_i and q_o.

III. From Pathways to QDE Models

As in [11], we consider that a pathway *P* is composed of several biochemical reactions, including the enzymatic and non-enzymatic ones. We also make standard biological assumptions on the pathway, that is, all enzymatic reactions follow the Michaelis-Menten kinetics, and all non-enzymatic reactions obey the law of mass action. For a non-enzymatic reversible reaction, A+B$\longleftrightarrow$C+D, according to the law of mass action the reaction rate is:

$$\begin{aligned} V &= K_1[A][B] - K_2[C][D] \quad (1) \\ &= -\frac{1}{a} \times \frac{d[A]}{dt} = -\frac{1}{b} \times \frac{d[B]}{dt} = \frac{1}{c} \times \frac{d[C]}{dt} = \frac{1}{d} \times \frac{d[D]}{dt}, \end{aligned}$$

where K_1 and K_2 are the rate constants of the forward and backward reaction respectively; a, b, c, and d are stoichiometric coefficients; [A], [B], [C] and [D] stand for concentrations of the corresponding species. For an enzymatic reaction A$\longrightarrow$B, the reaction rate V is defined as follows:

$$V = -\frac{d[A]}{dt} = \frac{d[B]}{dt} = V_{max} \times \frac{[A]}{k_s + [A]}. \quad (2)$$

In the above, V_{max} and k_s are constants. We can see that with the increase of the concentration of A, the reaction rate will increase too. This can be considered as a monotonically increasing relation in the qualitative context.

Based on Equations (1) and (2), a possible pathway can be converted into a QDE model by the converting algorithm, details of which can be found in [11]. In this way we can perform the search in the pathway level, that is, search all reasonable pathways, rather than in the qualitative constraint level as in [15], e.g., search all possible QDE models directly. This will significantly reduce the size of the search space. For instance, considering the following simple pathway which is composed of only three reactions:

r1: A+B $\longleftrightarrow$C+D

r2: B $\longrightarrow$ E

r3: E $\longrightarrow$ C

Using the converting algorithm, we can convert the above pathway into a QDE model (using the *Morven* formalism), as shown in Table V. In this table, Constraints *c1*$\sim$ *c5* and *c8*$\sim$ *c11* are related to Reaction *r1*; Constraints *c6*, *c9*, and *c12* are related to Reaction *r2*; Constraints *c7*, *c10*, and *c12* are related to Reaction *r3*. In this table variables whose names start with "Aux" are called auxiliary variables, which are used to break down long equations so that the qualitative constraints can be used (more details may be found in [11]). All Function constraints in this table represent monotonically increasing relations, and their mappings are as shown in Table II.

From Table V we see that for a simple pathway consisting of three reactions, the corresponding QDE model contains 12 constraints. This means it will be easier to perform the search at the reaction level rather than at the qualitative constraint level, because at the qualitative constraint level the search space is much bigger.

TABLE V. The Qualitative Model for an Example Pathway

Index	Qualitative Constraints	Mathematical Equations
c1	mul (dt 0 Aux1, dt 0 A, dt 0 B)	Aux1=A $\times$ B
c2	Function (dt 0 Aux2, dt 0 Aux1)	Aux2=f(Aux1) ($f' > 0$)
c3	mul (dt 0 Aux3, dt 0 C, dt 0 D)	Aux3=C*D
c4	Function (dt 0 Aux4, dt 0 Aux3)	Aux4=f(Aux3) ($f' > 0$)
c5	sub (dt 0 Aux5, dt 0 Aux2, dt 0 Aux4)	Aux5=Aux2-Aux4
c6	Function (dt 0 Aux6, dt 0 B)	Aux6= f(B) ($f' > 0$)
c7	Function (dt 0 Aux7, dt 0 E)	Aux7=f(E) ($f' > 0$)
c8	sub (dt 1 A, dt 0 Aux2, dt 0 Aux4)	d A/dt=Aux2-Aux4
c9	sub (dt 1 B, dt 0 Aux5, dt 0 Aux6)	d B/dt=Aux5-Aux6
c10	sub (dt 1 C, dt 0 Aux7, dt 0 Aux5)	d C/dt = Aux7-Aux5
c11	sub (dt 1 D, dt 0 Aux4, dt 0 Aux2)	d D/dt= Aux4-Aux2
c12	sub (dt 1 E, dt 0 Aux6, dt 0 Aux7)	d E/dt= Aux6-Aux7

IV. The Search Space, Reasonable and Candidate Pathways

For a pathway *P* to be identified, given all the species involved in this pathway and the standard assumptions about enzymatic and non-enzymatic reactions mentioned in Section III, we can generate all possible reactions by enumerating all combinations of species and reaction types. These reactions are further partitioned into several subgroups, each of which contains all reactions having the same reactants. Suppose SS is the set containing all these subgroups:

$$SS = \{S_1, S_2, ..., S_n\}. \quad (3)$$

In the above S_i ($1 \leq i \leq n$) contains all possible reactions with the same combination of reactants. In addition, for ease of implementation, we add a "dummy" reaction in each S_i, denoted ϕ. If a dummy reaction ϕ is selected in S_i, the pathway will not include any reaction from S_i.

Definition 1: **Possible Pathway.** A possible pathway is constructed by selecting one and only one reaction (including the dummy reaction) from each S_i in SS.

This is because one combination of reactants can only lead to unique products. Accordingly, the size of the search space containing all possible pathways is

$$|SS_P| = \prod_{i=1}^{n} |S_i|. \quad (4)$$

In the above SS_P stands for the search space for the pathway *P*. $|SS_P|$ and $|S_i|$ denote the size of the search space and the number of reactions in the subset S_i, respectively.

Definition 2: **Reasonable Pathway.** A reasonable pathway is a possible pathway that satisfies the following constraints: (1) Completeness: a pathway must include at least all given species. (2) Consistency: there are no conflicting reactions. (3) Connection: there is no disjoint section in the pathway. (4) Domain-specific constraints: a pathway must satisfy additional constraints extracted from domain knowledge.

Definition 3: **Candidate Pathway.** If the QDE model of a reasonable pathway can cover given qualitative data (*GQD*), this pathway is a candidate pathway with respect to *GQD*.

It is noted there are often many candidate pathways for GQD because the model space is often highly multimodal.

For each candidate pathway, we calculate its Bayesian score to indicate the probability of this pathway being the right one.

Definition 4: **Bayesian Score of a Candidate Pathway.** The Bayesian Score of a candidate pathway is the Bayesian score of the QDE model converted from this candidate pathway.

The Bayesian score of a QDE model is calculated according to Muggleton's learning from positive data framework [19], as shown below:

$$Bayes(M) = p \ln \frac{1}{g(M)} - \ln sz(M) \quad (5)$$

In the above *sz(M)* is the size of the given QDE model *M*, *g(M)* is the generality of the model, and p is the number of positive examples. So this Bayesian scoring is the tradeoff between the size and generality of a model. Based on previous work [6], in Equation (5) *sz(M)* is estimated by summing up the sizes of all constraints; *g(M)* is defined as the proportion of qualitative states obtained from simulation to all possible qualitative states generated from given variables and their associated quantity spaces; *p* is the number of given qualitative states.

The bigger the Bayesian score of a candidate pathway, the higher the probability that this pathway is the correct model. In this research, the above described Bayesian score is incorporated into the fitness evaluation to guide the search.

V. QML$_{PI}$-AiNet

In this section the detailed implementation of QML$_{PI}$-AiNet will be presented.

A. Antibody Encoding and Decoding

Similar to QML$_{PI}$-CLONALG, an antibody in QML$_{PI}$-AiNet is composed of several slots, each of which corresponds to a reaction subset S_i in SS described in Equation (3). In contrast to the integer encoding for antibodies in QML$_{PI}$-CLONALG, in QML$_{PI}$-AiNet the real number encoding is used, which is the same encoding strategy as in the original opt-AiNet. An antibody is represented as follows:

$$Ab = \{Sl_1, Sl_2, ..., Sl_n\}. \quad (6)$$

In the above Ab stands for an antibody; Sl_i ($1 \leq i \leq n$) represents the value assigned to the corresponding slot of Ab, satisfying $Sl_i \in \mathbf{R}$ and $1 \leq Sl_i \leq |S_i|$.

As the real number encoding strategy is used, when we decode an antibody, each value Sl_i will be rounded off to its nearest integer, denoted as $[Sl_i]$. If Sl_i is in the middle of two integers, the smaller integer will be taken. Then the newly obtained integer for each slot will be used as an index to retrieve the corresponding biochemical reaction in each S_i. So after the decoding of an antibody represented by Equation (6), the following pathway P will be obtained:

$$P = \{R_{[Sl_1]}, R_{[Sl_2]}, ..., R_{[Sl_n]}\}. \quad (7)$$

In the above $R_{[Sl_i]}$ means the $[Sl_i]$-*th* reaction in S_i.

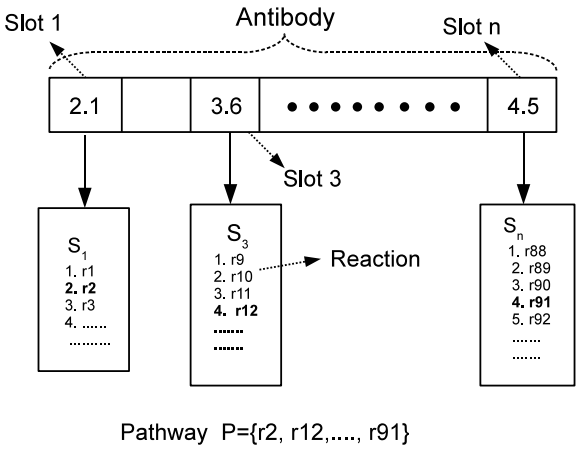

Fig. 3. The Antibody Encoding and Decoding of QML$_{PI}$-AiNet

Figure 3 shows an example of the antibody encoding and decoding in QML$_{PI}$-AiNet. In this figure, the antibody has *n* slots, which correspond to S_1, S_2, ... , S_n in SS described in Equation (3), respectively. In Slot 1 the current value is 2.1. After decoding we get an integer 2, so the second reaction $r2$ in S_1 is selected (indicated in bold font). It is similar for the other slots. After decoding the pathway contains reactions $r2$, $r12$, and $r91$.

B. Fitness Evaluation

We note here that in QML$_{PI}$-CLONALG this process is called the affinity evaluation. In QML$_{PI}$-AiNet the affinity has a different meaning which will be defined later in Section V-D. In the fitness evaluation process of QML$_{PI}$-AiNet, an antibody is first decoded into a pathway, then this pathway is checked against the reasonable pathway constraints, as given in **Definition** 2. The more constraints a model satisfies, the higher fitness value this model will get.

If this pathway is a reasonable one, it will be converted to a QDE model (as described in Section III) and checked against the given data. In previous work [11], we checked the model coverage by qualitative simulation with *Morven*. However the qualitative simulation is very computationally expensive for large-sized models. In this research, considering the situation that in a complex pathway only a few variables can be measured, instead of simulating the model converted from a pathway, we inversely check whether the given qualitative states can make the model consistent, or in other words, make all qualitative constraints of the model consistent. This is done by using a Waltz-like constraint propagation algorithm as described in [20].

The details of the model checking by the Walltz-like algorithm are described as follows: we first extract qualitative values from a qualitative state (described in Section II and one example is given in Figure 2), say *QS*, and substitute these values into variables in the candidate qualitative model, say *M*, to be checked. Because *QS* only contains values of observed variables in *M*, the values of unobserved (hidden) variables of *M* cannot be determined at the very beginning. However, If we consider a QDE model as a constraint

network with each variable being a node and each qualitative constraint being a connection of variables, we can apply the Waltz-like constraint propagation algorithm [20] to check whether the substituted values can lead to inconsistency in the network. In the process of the Waltz algorithm, we propagate the values of the observed variables throughout the constraint network (the model), and this may enable us to calculate the values of some hidden variables, which may lead to the calculation of more hidden variables. We do this in an iterative manner until either an inconsistency occurs in a qualitative constraint, which means *QS* cannot be covered by the model *M*, or the value of any variable in the model cannot be updated further and there is no inconsistency in the model, which means *QS* is consistent with *M*.

In the above inverse checking algorithm, the smaller the proportion of observed variables to all variables, the more effective the algorithm is. This efficient model checking algorithm enables us to perform the learning tasks when there are many hidden variables.

After the data coverage of a pathway has been checked, if this pathway is a candidate pathway according to **Definition** 3, we can further calculated the Bayesian score of this pathway according to **Definition** 4, and the obtained Bayesian score will be added into the total fitness value to guide further search.

C. Mutation

The original mutation operator of opt-AiNet was defined for continuous problems. Considering the discrete qualitative model space, the following mutation operation is proposed for each slot of the antibody:

$$C' = \begin{cases} U(1,n) & \text{if } U(0,1) < \alpha N(0,1) \\ C & \text{otherwise} \end{cases} \tag{8}$$

$$\alpha = \frac{1}{\beta} e^{-f^*} \tag{9}$$

In the above, C' and C are the mutated value and current value for one slot of the antibody, respectively. $U(0,1)$ is a uniformly distributed random number with the range [0,1]. Similarly, $U(1,n)$ stands for a uniformly distributed random number with the range [1, n], where n is the number of constraints in the current slot of the antibody. $N(0,1)$ is a Gaussian random variable which has a mean value of 0 and standard deviation of 1. f^* is the normalised fitness with the range [0,1]. e^{-f^*} is the inverse exponential function. α stands for the amount of mutation, and β is a parameter that adjusts the exponential function. This new mutation operator first determines whether a slot should be mutated. The probability of mutating is proportional to the fitness value of the current antibody. Once the current slot is set to mutate, the mutation will follow the uniform distribution.

D. Affinity

In opt-AiNet the affinity is defined as the Euclidean distance between two antibodies. In QML$_{PI}$-AiNet because we use the integer decoding strategy, and each antibody represents a possible pathway composed of several reactions, we define the affinity between two antibodies as "the degree of dissimilarity" between two pathways represented by the two antibodies. The degree of dissimilarity between two pathways is calculated by simply counting the number of different reactions in these two pathways.

TABLE VI. PARAMETERS IN QML$_{PI}$-AINET

Name	*Meaning*
N_i	Number of initial antibodies in the population
N_c	Number of clones for each antibody
$AvgFitError$	Threshold determines the stability of population
$Supp$	The suppression threshold
d	The percentage of new antibodies to be added into the population
β	control parameter for mutation

E. The Detailed Steps of QML$_{PI}$-AiNet

The steps of QML$_{PI}$-AiNet follow the framework of opt-AiNet. First we list the parameters used by the algorithm in Table VI. The steps of the proposed QML$_{PI}$-AiNet algorithm are given in detail as follows:

Step 1: Randomly generate N_i antibodies.
While (stop criteria are not satisfied) iteratively execute *Step 2* $\sim$ *Step 4*:

Step 2: Clonal Selection

- *Step 2-1*: Antibody fitness evaluation: calculate the fitness values of all antibodies according to the description in Section V-B.
- *Step 2-2*: Clone: Generate N_c clones for each antibody.
- *Step 2-3*: Mutation: Each antibody will be mutated according to the description in Section V-C. In particular, the original and modified mutation operators will both be tested.
- *Step 2-4*: Fitness Evaluation: evaluate all the newly cloned antibodies. Calculate the normalised fitness value for each antibody.
- *Step 2-5*: Selection: Select the antibody which has the biggest fitness value from each parent antibody and its clones. All the selected antibodies construct a new antibody population.
- *Step 2-6*: Average Fitness Error Calculation: Calculate the average fitness of the new population. If the difference between the old average fitness and new average fitness is bigger than the given threshold $AvgFitError$, repeat Step 2; otherwise proceed to Step 3.

Step 3: Network Suppression: Each antibody interacts with others. If the affinity of two antibodies (defined in Section V-D), is less than the suppression threshold $Supp$, the one with the smaller fitness value will be removed.

Step 4: Add d percent of the randomly generated antibodies to the population.

VI. EXPERIMENTS

In this section we will test the performance of QML$_{PI}$-AiNet by a series of experiments, each of which is designed

to learn a pathway with a different number of hypothesised hidden variables. The more hidden variables there are in the pathway, the larger the search space and hence the more challenging the learning task. We will compare the scalability of QML$_{PI}$-AiNet and QML$_{PI}$-CLONALG when applied to learning the pathways with different levels of complexity.

More specifically, we will complete and extend the experiments in [11]. In [11], for learning the MG pathway, we also assumed different numbers of hidden variables and tested how well QML-CLONALG could find candidate pathways compared to other algorithms. However, the use of qualitative simulation to test the data coverage of models restricted us from testing the algorithm on more complicated pathways, and we have to ignore the data coverage tests for such complicated pathways because the corresponding qualitative simulation is very expensive. In this sense some of the previous experiments performed in [11] are not complete.

In this research, we will use the more efficient Waltz-like inverse checking algorithm to verify models as mentioned in Section V-B. We will also use the Bayesian Score to evaluate each candidate pathway as described in Secion IV. This enables us to perform the *full* experiments which include the data coverage test.

A. The MG Detoxification Pathway

According to the current (incomplete) understanding [21], the MG detoxification pathway is composed of one non-enzymatic reaction and two enzymatic reactions, as shown below:

MG+G $\longleftrightarrow$ H
H $\longrightarrow$ S
S $\longrightarrow$ G

In the above, MG stands for Methylglyoxal; G stands for glutathione; H is hemithioacetal; S is S-lactoyl-glutathione. The first reaction is a reversible one and follows the mass action law. The second and third enzymatic reactions are irreversible and catalysed by GlxI (glyoxalase I) and GlxII (glyoxalase II), respectively, and they are assumed to conform to Michaelis-Menton kinetics. As the exact mechanisms of the MG detoxification are still not fully understood, we hypothesise different numbers of hidden variables and try to reconstruct the pathway from qualitative data.

B. Qualitative Data

The qualitative data are obtained by simulating the qualitative model converted from the current understanding of the MG pathway. In the simulation, all variables take the signs quantity space as described in Table I. In all experiments, the same qualitative data are provided. There are a total of 33 qualitative states, which are listed in Table VII.

C. Experimental Settings

Based on the current understanding of the MG pathway, we hypothesise three, five, seven, and nine hidden variables, which gives us four sets of experiments, and these four sets of experiments are called MG-3Hid, MG-5Hid, MG-7Hid, and MG-9Hid, respectively. For all four sets of experiments, we make the following reasonable assumptions: (1) there is one mass action reaction, and one of the reactants of this reaction must contain Methylglyoxal; (2) the number of enzymatic reactions is unknown. Each set of experiments will be learnt by both QML$_{PI}$-CLONALG and QML$_{PI}$-AiNet, and we will also use the completely random algorithm as baselines.

TABLE VII. QUALITATIVE DATA USED FOR LEARNING

MG	G	H	S
<pos,pos>	<pos,pos>	<pos,neg>	<pos,neg>
<pos,zer>	<pos,pos>	<pos,neg>	<pos,neg>
<pos , neg>	<pos , neg>	<pos , neg>	<pos , neg>
<pos , neg>	<pos , zer>	<pos , neg>	<pos , neg>
<pos , neg>	<pos , pos>	<pos , neg>	<pos , neg>
<pos , neg>	<pos , neg>	<pos , zer>	<pos , neg>
<pos , neg>	<pos , zer>	<pos , zer>	<pos , neg>
<pos , neg>	<pos , pos>	<pos , zer>	<pos , neg>
<pos , neg>	<pos , neg>	<pos , pos>	<pos , neg>
<pos , neg>	<pos , zer>	<pos , pos>	<pos , neg>
<pos , neg>	<pos , pos>	<pos , pos>	<pos , neg>
<pos , pos>	<pos , pos>	<pos , neg>	<pos , zer>
<pos , zer>	<pos , pos>	<pos , neg>	<pos , zer>
<pos , neg>	<pos , neg>	<pos , neg>	<pos , zer>
<pos , neg>	<pos , zer>	<pos , neg>	<pos , zer>
<pos , neg>	<pos , pos>	<pos , neg>	<pos , zer>
<pos , neg>	<pos , neg>	<pos , zer>	<pos , zer>
<pos , neg>	<pos , zer>	<pos , zer>	<pos , zer>
<pos , neg>	<pos , pos>	<pos , zer>	<pos , zer>
<pos , neg>	<pos , neg>	<pos , pos>	<pos , zer>
<pos , neg>	<pos , zer>	<pos , pos>	<pos , zer>
<pos , neg>	<pos , pos>	<pos , pos>	<pos , zer>
<pos , pos>	<pos , pos>	<pos , neg>	<pos , pos>
<pos , zer>	<pos , pos>	<pos , neg>	<pos , pos>
<pos , neg>	<pos , neg>	<pos , neg>	<pos , pos>
<pos , neg>	<pos , zer>	<pos , neg>	<pos , pos>
<pos , neg>	<pos , pos>	<pos , neg>	<pos , pos>
<pos , neg>	<pos , neg>	<pos , zer>	<pos , pos>
<pos , neg>	<pos , zer>	<pos , zer>	<pos , pos>
<pos , neg>	<pos , pos>	<pos , zer>	<pos , pos>
<pos , neg>	<pos , neg>	<pos , pos>	<pos , pos>
<pos , neg>	<pos , zer>	<pos , pos>	<pos , pos>
<pos , neg>	<pos , pos>	<pos , pos>	<pos , pos>

TABLE VIII. EXPERIMENTAL SETTINGS

Experiment Set	Species	Search Space
MG-3Hid	M,G, H, S, A, B, C	1.38×10^{13}
MG-5Hid	M,G, H, S, A, B, C, D, E	1.93×10^{20}
MG-7Hid	M,G, H, S, A, B, C, E, E, F, I	1.21×10^{28}
MG-9Hid	M,G, H, S, A, B, C, E, E, F, I, J, K	2.48×10^{36}

The experimental settings are listed in Table VIII. In this table M, G, S, H are the four identified species in the pathway, and A$\sim$K are hyphothesised hidden species. All the experiments were performed on a computer cluster with 43 compute nodes, and each node has two Intel XEON E5520 (2.268GHz) quad-core processors and 12GB RAM). To ensure a fair competition, all algorithms are restricted to use a maximum of 4GB memory for all experiments.

D. Experimental Results

The experimental results are listed in Table IX. In this table we tested the performance of three algorithms on the four experiment sets, and recorded the number of candidate pathways and pathways with highest Bayesian scores found by each algorithm. All algorithms were run for ten trials, and the best and average performance (with standard deviation) were recorded. Each algorithm was run for 2,000 seconds. The parameter settings for QML$_{PI}$-AiNet are as follows:

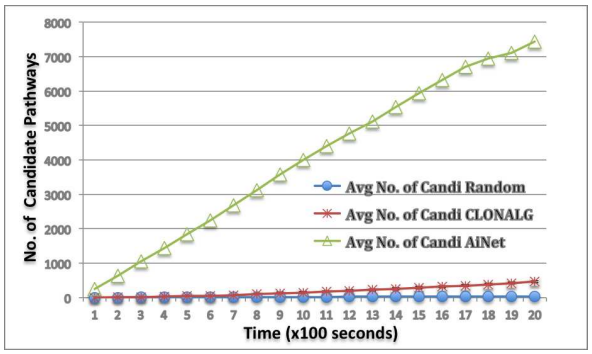

Fig. 4. Experimental Results for MG-7Hid: Average Number of Candidate Models Found Over Ten Trials

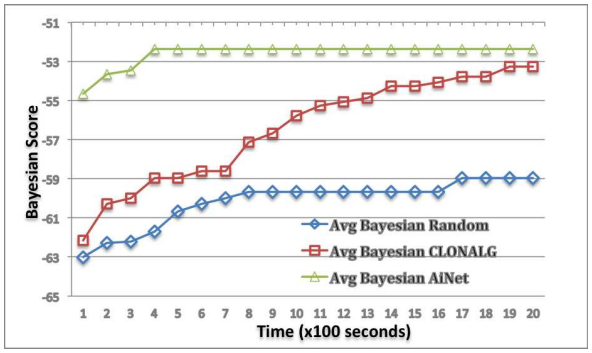

Fig. 5. Experimental Results for MG-7Hid: Average Bayesian Score of the Best Candidate Model Over Ten Trials

N_i=20; N_c=10; $AvgFitError$=0.001; $supp$ is 10 for MG-9Hid, 9 for MG-7Hid, 7 for MG-5Hid, 5 for MG-3Hid; d=0.2; β=1. The parameter settings for QML$_{PI}$-CLONALG are as follows: the clonal size is 10; the hyper-mutation probability is 0.1; the population size is 100 for MG-3Hid, 1000 for other experiment sets. The values of parameters are chosen according to either classical values taken in both algorithms or considering the complexity of the search space and the performance of the search.

From the results shown in Table IX we see that with the increase of the size of the search space, QML$_{PI}$-AiNet performs better and better than QML$_{PI}$-CLONALG in terms of the number of candidate pathways found as well as the quality of the best solutions measured by Bayesian scores. This is also illustrated in Figures 4 and 5, which show the detailed experimental results for MG-7Hid. From Figure 4 we can see that in average QML$_{PI}$-AiNet found one order of magnitude more candidate pathways than QML$_{PI}$-CLONALG, which well demonstrated the ability of QML$_{PI}$-AiNet to deal with multimodal search spaces. From Figure 5 we can also see that QML$_{PI}$-AiNet found pathways with higher Bayesian scores and converged to the highest Bayesian score more quickly compared with QML$_{PI}$-CLONALG. This indicates that QML$_{PI}$-AiNet can better deal with large-scale multimodal qualitative model spaces than QML$_{PI}$-CLONALG.

VII. Conclusions and Future Work

In this paper, we have presented an immune network approach to learning QDE models of biological pathways. The proposed QML$_{PI}$-AiNet employs an opt-AiNet based search strategy to search the qualitative model space, which could be highly multimodal given incomplete knowledge and data.

A comparison of the performance of QML$_{PI}$-AiNet with the previous system QML$_{PI}$-CLONALG indicates that the proposed QML$_{PI}$-AiNet can better deal with highly multimodal qualitative model spaces, and is also more scalable to large search spaces. Given the same computational resources, in all experiments QML$_{PI}$-AiNet outperformed QML$_{PI}$-CLONALG. This indicates that QML$_{PI}$-AiNet is a very suitable special-purpose QML system for qualitative pathway identification.

Finally, it is noted that the proposed special-purpose immune network approach to QML can be generalised to solve other real-world applications, such as identification of economic, logistics [22], mechanical, and electrical systems, provided a method of converting models representing such real-world applications to QDE models is developed. In the future work, we will consider the situations where there are noisy qualitative states or only a few qualitative states are available, which is similar to previous study [23], [6] on general-purpose QML systems. How to make QML$_{PI}$-AiNet adapt to these situations will become a challenging task.

Acknowledgment

GMC is supported by the CRISP project (*Combinatorial Responses In Stress Pathways*) funded by the BBSRC (BB/F00513X/1) under the Systems Approaches to Biological Research (SABR) Initiative. WP and GMC are also supported by the partnership fund from dot.rural, RCUK Digital Economy research.

References

[1] W. Pang and G. M. Coghill, "Learning qualitative differential equation models: a survey of algorithms and applications," *The Knowledge Engineering Review*, vol. 25, no. 1, pp. 69–107, 2010.

[2] D. T. Hau and E. W. Coiera, "Learning qualitative models of dynamic systems," *Machine Learning*, vol. 26, pp. 177–211, 1993.

[3] S. Ramachandran, R. J. Mooney, and B. J. Kuipers, "Learning qualitative models for systems with multiple operating regions," in *the Eighth International Workshop on Qualitative Reasoning about Physical Systems*. Nara, Japan, June 1994, pp. 213–223.

[4] A. C. C. Say and S. Kuru, "Qualitative system identification: deriving structure from behavior," *Artificial Intelligence*, vol. 83, pp. 75–141, 1996.

[5] A. Varšek, "Qualitative model evolution," in *Proceedings of the Twelfth International Joint Conference on Artificial Intelligence*, J. Mylopoulos and R. Reiter, Eds. Sydney, Australia: AAAI Press, 1991, pp. 1311–1316.

[6] G. M. Coghill, A. Srinivasan, and R. D. King, "Qualitative system identification from imperfect data," *Journal of Artificial Intelligence Research*, vol. 32, pp. 825–877, 2008.

[7] G. M. Coghill, S. M. Garrett, and R. D. King, "Learning qualitative models in the presence of noise," in *Proc. 16th International Workshop on Qualitative Reasoning (QR'02)*, Sitges, Spain, June 2002, pp. 27–36.

[8] W. Pang, "QML-Morven: A framework for learning qualitative models," Ph.D. dissertation, University of Aberdeen, 2009.

[9] W. Pang and G. M. Coghill, "QML-Morven: A novel framework for learning qualitative models," Department of Computer Science, University of Aberdeen, Tech. Rep. ABDN-CS-12-03, 2012.

TABLE IX. EXPERIMENTAL RESULTS

Experiment ID	Algorithm	No. of Candidate Pathways		No. of Best Pathways		Highest Bayesian Scores	
		Best	Average (STDEV)	Best	Average (STDEV)	Best	Average (STDEV)
MG-3Hid-1	Random	668	600.3(67.91)	30	4.9(8.88)	-24.37	-24.86(1.58)
MG-3Hid-2	CLONALG	2148	1760(235.38)	20	15.6(4.62)	-24.37	-24.37(0)
MG-3Hid-3	AiNet	5993	5244(292.62)	29	27.3(2.50)	-24.37	-24.37(0)
MG-5Hid-1	Random	170	126.3(26.75)	12	3(3.30)	-41.37	-43.39(1.00)
MG-5Hid-2	CLONALG	1529	1156.1(242.16)	19	9(5.62)	-38.37	-38.37(0)
MG-5Hid-3	AiNet	8677	7770.4(573.38)	153	118.4(19.22)	-38.37	-38.37(0)
MG-7Hid-1	Random	51	39.2(7.33)	22	4.2(6.43)	-52.37	-58.96(2.63)
MG-7Hid-2	CLONALG	652	462.3(143.54)	9	3.5(2.42)	-52.37	-53.57(1.55)
MG-7Hid-3	AiNet	9271	7439(789.27)	217	163.6(45.75)	-52.37	-52.37(0)
MG-9Hid-1	Random	19	12.6(4.300)	4	1.9(1.10)	-73.60	-76.30(0.95)
MG-9Hid-2	CLONALG	265	182.4(56.07)	1	1(0)	-71.37	-71.97(0.52)
MG-9Hid-3	AiNet	11406	7455.5(2827.18)	238	131.7(53.53)	-66.37	-66.37(0)

[10] ——, "Modified clonal selection algorithm for learning qualitative compartmental models of metabolic systems," in *Genetic and Evolutionary Computation Conference (GECCO07)*, D. Thierens, Ed. New York, NY, USA: ACM Press, 2007, pp. 2887–2894.

[11] ——, "An immune-inspired approach to qualitative system identification of biological pathways," *Natural Computing*, vol. 10, no. 1, pp. 189–207, 2011.

[12] de Castro and F. J. V. Zuben, "Learning and optimization using the clonal selection principle," in *IEEE Transactions on Evolutionary Computation, Special Issue on Artificial Immune Systems*, vol. 6, no. 3. IEEE Press, 2002, pp. 239–251.

[13] de Castro and J. Timmis, "An artificial immune network for multimodal function optimization," in *Proceedings of IEEE Congress on Evolutionary Computation (CEC'02)*. IEEE Press, 2002, pp. 674–699.

[14] J. Timmis and C. Edmonds, "A comment on opt-AINet: An immune network algorithm for optimisation," in *Genetic and Evolutionary Computation (GECCO 2004), Lecture Notes in Computer Science*, D. Kalyanmoy, Ed., vol. 3102. Springer, 2004, pp. 308–317.

[15] W. Pang and G. M. Coghill, "QML-AiNet: An immune-inspired network approach to qualitative model learning," in *LNCS 6209, Proceedings of 9th International Conference on Artificial Immune Systems (ICARIS 2010)*, E. H. et al., Ed. Edinburgh, UK: Springer, 2010, pp. 223–236.

[16] G. M. Coghill and M. J. Chantler, "Mycroft: a framework for qualitative reasoning," in *Second International Conference on Intelligent Systems Engineering*, Hamburg, Germany, September 1994, pp. 43 – 48.

[17] A. M. Bruce and G. M. Coghill, "Parallel fuzzy qualitative reasoning," in *Proceedings of the 19th International Workshop on Qualitative Reasoning*. Graz, Austria, 2005, pp. 110–116.

[18] B. Kuipers, *Qualitative Reasoning: Modeling and Simulation with Incomplete Knowledge*. Cambridge, MA: MIT Press, 1994.

[19] S. Muggleton, "Learning from positive data," in *Lecture Notes in AI*, 1996, pp. 358–376.

[20] E. Davis, "Constraint propagation with interval labels," *Artificial Intelligence*, vol. 32, no. 3, pp. 281–331, July 1987.

[21] G. P. Ferguson, S. Totemeyer, M. J. MacLean, and I. R. Booth, "Methylglyoxal production in bacteria: suicide or survival?" *Archives of Microbiology*, vol. 170, no. 4, pp. 209–218, September 1998.

[22] Y. Liu, C. Zhou, D. Guo, K. Wang, W. Pang, and Y. Zhai, "A decision support system using soft computing for modern international container transportation services," *Applied Soft Computing*, vol. 10, no. 4, pp. 1087–1095, 2008.

[23] W. Pang and G. M. Coghill, "Advanced experiments for learning qualitative compartment models," in *The 21st International Workshop on Qualitative Reasoning*, Aberystwyth, UK, June 2007, pp. 109–117.