A Hierarchical Pareto Dominance based Multi-objective Approach for the Optimization of Gene Regulatory Network Models

Xinye Cai  Zhenzhou Hu  Sanjoy Das  Stephen M. Welch
College of Computer Science & Technology, Nanjing University of Aeronautics & Astronautics, Nanjing, Jiangsu, China
Email: xinye@nuaa.edu.cn  hu890612@163.com
Sanjoy Das  Stephen M. Welch
Electrical & Computer Engineering Department, Kansas State University, Manhattan, KS, USA
Email: sdas@ksu.edu
Agronomy Department, Kansas State University, Manhattan, KS, USA
Email: welchsm@ksu.edu

Abstract—In this paper, a hierarchical Pareto dominance based multi-objective evolutionary approach is proposed for the optimization of gene regulatory network models. The approach is presented based on the neglected observations in GRN optimization that (i) structural dependencies exist among objectives; and (ii) some objectives may be more important than others. The hierarchical Pareto dominance is able to reduce the number of objectives during optimization process and increase the selection pressure to relieve the many objective problem. The proposed hierarchical Pareto dominance based multi-objective approach is verified and compared with classical Pareto dominance based algorithm NSGAII on the gene regulatory network optimization problem. The results obtained indicate that the presented approach has great performance when no noise exist. Also it shows superior results compared to NSGAII.

I. INTRODUCTION

There are a small number of model organisms whose genetic networks have been studied in detail including Arabidopsis (a flowering plant), bakers yeast, a nematode, the sea urchin, and the fruit fly, Drosophila, among others. At the same time, the complete DNA sequence has been determined for some 3047 organisms (as of January 2012) with another 7784 in progress (http://www.genomesonline.org/cgi-bin/GOLD/index.cgi). Thus, there is a major and growing gap between available genomic data and a functional knowledge of the networks whose operations the DNA encodes. For this reason, there have been a great deal of work to predict phenotype by directly modeling the gene interactions at the expression level [1][2][3]. In these cases, genes are typically modeled using differential equations, Boolean logic, linear units, oscillators, etc [1][2][3][4][5]. Estimating parameters for the gene regulatory network (GRN) models so that simulated expression levels match experimentally observed data is a key modeling step. This is a cumbersome task that requires an effective, derivative-free approach, such as the evolutionary algorithms [6]. In addition, the existence of multiple data sets, e.g. expression data of various genes, make it by nature a multi-objective optimization problem (MOP) [7]. Accordingly, a number of multi-objective evolutionary algorithms (MOEAs) have been proposed [7][8][9][10][11][12]. Among them, Koduru et. al applied fuzzy dominance in multi-objective evolutionary algorithms hybridized with simplex method [7]. Alternatively, Datta et. al presented a multi-objective differential evolution algorithm for GRN optimization problem [9]. Cai et. al proposed a GP-PSO hybrid algorithm to reconstruct the GRN structure and estimate model parameters simultaneously [8]. Lately, Lee et. al proposed a similar method but used the concept of network decomposition [12]. The survey of multi-objective evolutionary approach for GRN optimization problem can be found in [10].

Fig. 1: Arabidopsis flowering time control model

Nevertheless, all the MOEAs designed for GRN optimization problem, to our best knowledge, are based on the hypothesis that all the objectives are independent to each other and equally important as well [7][8][9][10][11][12]. But the reality of GRN optimization problem is that (i) structural dependencies exist among objectives; and (ii) some objectives may be more important than others. For instance, in the optimization of Arabidopsis flowering time control model, as shown in Fig. 1, the expression data of the switch genes (LFY and AP1) that signal plant commitment to flowering are apparently more important than that of other upstream genes,
such as $TFL1$. In addition, the expression data of gene $LFY$ is likely to correlate with that of gene $AP1$ as gene $LFY$ and $AP1$ regulate each other. In other words, dependencies may exist among objectives (expression of different genes) in such multi-objective optimization problem.

In this paper, we propose a hierarchical multi-objective evolutionary approach for the optimization of GRN. Our proposed approach is able to make use of the structural information among genes and treat objectives (expression data of genes) differently based on their importance in the GRN. The property analysis of the proposed approach show its capability to reduce the dimensions of objectives and relieve many-objective problem. More details of many-objective problems can be referred to in [13][14]. The rest of the paper is organized as follows. Section II introduces the optimization problem of the flowering time control model in Arabidopsis. In Section III, the concept of multi-objective optimization problem is formally defined. Section IV details the hierarchical multi-objective optimization method. Properties of the proposed approach is also analyzed in this section. In section V, results obtained for GRN optimization problem is discussed. We also compare the hierarchical multi-objective method with NSGAI. Finally, Section VI draws the conclusions and the future work.

II. MODELING FLOWERING TIME CONTROL IN ARABIDOPSIS

In the molecular genetic model plant, Arabidopsis thaliana, three genes TERMINAL FLOWERING 1 ($TFL1$), APETALA 1 ($AP1$), and LEAFY ($LFY$) play a special role in flowering [2][7]. OFF to ON state changes in two of them ($AP1$ and $LFY$) signal plant commitment to flowering. Although it is not possible to completely disentangle the linkages based on extant experimental data, one possible model is that of a three element positive feedback loop comprising a bistable switch. A model for this switch is the coupled set of differential equations

$$
\frac{d}{dt}LFY = R_L g(SOC1,TFL1) - \lambda_L LFY \\
\frac{d}{dt}AP1 = R_H h_{up}(LFY) - \lambda_H AP1 \\
\frac{d}{dt}TFL1 = R_T h_{down}(AP1) - \lambda_T TFL1
$$

where $h_{up}$ and $h_{down}$ are respective, promotive $(n = 3)$ and repressive $(n = -3)$ Hill function [15] defined as

$$h_i(x) = \frac{a_i^n}{a_i^n + K_i^n}$$

where $n$ is a cooperativity coefficient. The function $g$ is a repressive Hill function whose input is ($TFL1 - SOC1$). The difference input to $g$ is restricted to positive values; negative biochemical concentrations are impossible.

External switch input is provided by the expression level of the $SUPRESSOR$ $OF$ $OVEREXPRESSION$ $OF$ $CO$ ($SOC1$) gene, which is a linear, ramped sinusoid. The steepness of the ramp and the amplitude of its oscillations relate to the rate of progress toward flowering. An equation for $SOC1$ expression levels is

$$SOC1(t_m) = b_s t_m + \frac{(a_s - b_s)}{2} t_m \sin \left( \frac{2 \pi t_m}{p_s} \right)$$

GRN Optimization Problem Definition: The 9 parameters to be estimated are: $R_L$, $R_H$, $\lambda_L$, $\lambda_H$, $\lambda_T$, $K_L$, $FY$, $K_{AP1}$ and $K_T$. A total of 318 data points were generated covering 5 days, using realistic parameter values. The three objectives to be minimized are the root mean squared error (RMSE) between the synthetic and simulated data for $LFY$, $AP1$ and $TFL1$.

Since multiple gene expression data($LFY$, $AP1$ and $TFL1$) are available for the above problem, the problem is multi-objective by nature. Accordingly, the concept of multi-objective optimization problem will be introduced in the next section.

III. MULTI-OBJECTIVE OPTIMIZATION PROBLEM

A multi-objective optimization problem can be defined as:

Definition 1 (Multi-objective optimization problem): Given a problem involving $N$ decision variables $x_1, x_2, \ldots, x_N$ in a search space $X \subset \mathbb{R}^N$, we assume, without loss of generality, $M$ objectives $f_1(\cdot), \ldots, f_M(\cdot)$ in objective function space $Y \subset \mathbb{R}^M$, are to be minimized.

Minimize

$$f(\vec{x}) = (f_1(x_1, x_2, \ldots, x_N), \ldots, f_M(x_1, x_2, \ldots, x_N))$$

The vector function is a mapping $f : X \rightarrow Y$.

A. Pareto Dominance and Pareto Front

In MOP, it is usually not possible to find a single solution which is optimal for all the objectives. Instead, many good solutions may exist. These solutions are always “trade-offs” or good compromises among the objectives. A concept of Pareto dominance is introduced to address MOP in MOEAs. Definition 2 (Pareto dominance and non-dominated solutions): Let $\vec{x}, \vec{y}$ be two vectors of decision variables in MOP. $\vec{x}$ is considered to dominate $\vec{y}$ (written as $\vec{x} \prec \vec{y}$) iff they satisfy the conditions:

$$\vec{x}, \vec{y} \in X, \forall i \in 1, \ldots, M | f_i(\vec{x}) \leq f_i(\vec{y})$$

$$\vec{x}, \vec{y} \in X, \exists j \in 1, \ldots, M | f_j(\vec{x}) < f_j(\vec{y})$$

(4)

On the contrary, a decision vector $\vec{x}$ is considered to be a non-dominated solution if there is no other solution that satisfies Eq. 4. The set of all non-dominated solutions form a Pareto set.

Definition 3 (Pareto front): The projection of the Pareto set $P$ in the $M$ dimensional objective function space $Y$ is called Pareto front, $F$.

$$F = \{(f_1(\vec{x}), f_2(\vec{x}), \ldots, f_M(\vec{x})) | \vec{x} \in P\}$$

(5)

IV. HIERARCHICAL PARETO BASED MULTI-OBJECTIVE APPROACH

To address MOP, MOEAs usually adopt the concept of Pareto dominance to discriminate among solutions in the multi-objective context, and therefore it has been the basis to develop most of the MOEAs proposed so far, e.g., [7] and [16].
However, in GRN optimization problem, as multiple genes comprise the gene regulatory network structures, the relations among objective (expression of genes) have a hierarchical structure, that is, the expression of certain genes, such as the flowering switch genes AP1 and LFY in Flowering Time Control model in Arabidopsis, may be much more important than other genes, such as TFL1. Under this circumstance, objectives need to be discriminated based on their importance in GRN and the hierarchical Pareto dominance based multi-objective optimization for GRN become necessary. The goal can be achieved by extending the concept of Pareto dominance as follows.

A. Hierarchical Pareto Dominance

Although the method can be extended to apply to any number of objective hierarchies, we only consider two sets of objectives in the GRN optimization problem - a set of primary objectives $P = \{f_1(x)\text{: RMSE for LFY}, f_2(x)\text{: RMSE for AP1}\}$ and a set of secondary objectives $S = \{f_3(x)\text{: RMSE for TFL1}\}$. Given two solutions $\vec{x}$ and $\vec{y}$, if $\vec{x}$ has smaller RMSE than $\vec{y}$ along all the primary objectives $P$, $\vec{x}$ is considered to dominate $\vec{y}$ in the primary objective set $P$, that is $\vec{x} \preceq_1 \vec{y}$. Similarly, $\vec{x}$ dominating $\vec{y}$ along all the secondary objectives $S$ can be noted as $\vec{x} \prec_2 \vec{y}$.

The Pareto dominance can be extended to hierarchical Pareto dominance by applying the following rules:

1. if $\vec{x} \preceq_1 \vec{y}$, then $\vec{x} \prec \vec{y}$

2. if ($\vec{x} \not\preceq_1 \vec{y}$) $\land$ ($\vec{y} \not\preceq_1 \vec{x}$) $\land$ ($\vec{x} \prec_2 \vec{y}$), then $\vec{x} \prec \vec{y}$

The above rules can be summarized as

$$\vec{x} \prec \vec{y} \leftrightarrow (\vec{x} \preceq_1 \vec{y}) \lor ((\vec{x} \not\preceq_1 \vec{y}) \land (\vec{y} \not\preceq_1 \vec{x}) \land (\vec{x} \prec_2 \vec{y})).$$  \(6\)

B. Properties of Hierarchical Pareto Dominance

Multi-objective evolutionary algorithms defined Pareto dominance as criteria to select better non-dominated solutions. However, when the number of objectives to be optimized increases, the performance of MOEAs decreases considerably [17][18]. In this situation, almost all the solutions become non-dominated to each other and the selection pressure based on classical Pareto dominance become very ineffective. This is termed the many-objective problem [17][13]. Existing methods to relieve the many objective problem include dimension reduction in objective domain [19][20][21][22] and design relaxed form of Pareto dominance [17] [23]. Inspired from the concept of hierarchy of objectives [24] and lexicographic optimality [25], a new relaxed form of Pareto dominance, termed hierarchical Pareto dominance (HPD), is proposed for multi-objective evolutionary optimization in this paper. We analyze the properties of HPD as follows.

1. From Eq. 6, we can observe that when solution ($\vec{x} \not\preceq_1 \vec{y}$) $\land$ ($\vec{y} \not\preceq_1 \vec{x}$), which indicates solutions $\vec{x}$ and $\vec{y}$ do not dominate each other in both primary and secondary objective, all three objectives are compared to discriminate between $\vec{x}$ and $\vec{y}$. In this case, HPD is equivalent to three-objectives Pareto dominance.

2. We can also observe that when ($\vec{x} \prec_1 \vec{y}$) $\lor$ ($\vec{y} \prec_1 \vec{x}$), only two objectives in the primary objective set (RMSE for LFY and AP1) are compared to discriminate between $\vec{x}$ and $\vec{y}$. In this case, HPD actually reduce the number of objectives from 3 (RMSE for LFY, AP1 and TFL1) to 2 (RMSE for LFY and AP1).

3. On the other hand, when ($\vec{x} \not\prec_1 \vec{y}$) $\land$ ($\vec{y} \not\prec_1 \vec{x}$), which indicates solution $\vec{x}$ and $\vec{y}$ does not dominate each other in the primary objective set, the comparison of the secondary objective set is to be activated. In this case, HPD can be considered as a relaxed form of Pareto dominance [23]. It is able to relax the dominance relation and have the non-dominated solutions of the primary objective set to continue comparison with each other in the secondary objective set. Thus HPD is able to increase the selection pressure and enhance the performance of MOEAs.

C. Hierarchical Multi-objective Evolutionary Approach

The hierarchical Pareto dominance is very generic and can be incorporated on any MOEAs framework. But in purpose to compare the proposed approach with classical Pareto dominance based multi-objective algorithms, we incorporate HPD into NSGAII for comparison. The general pseudocode for the proposed algorithm is given in Algorithm 1. Initially, a random parent population $P_0$ with size of $N$ is created. Tournament selection, crossover, and mutation operators are applied to create a child population $Q_0$ of size $N$.

### Algorithm 1: A Hierarchical Pareto dominance based Multi-objective Evolutionary Approach (main loop)

1. while \(t \leq \text{max\_generation}\)
2. Let $R_t = P_t \cup Q_t$
3. Let $F = \text{Hierarchical Pareto dominance}(R_t)$
4. Let $P_{t+1} = \emptyset$ and $i = 1$
5. while \(|P_{t+1}| + |F_t| \leq N\)
6. Apply crowding-distance-assignment($F_t$)
7. Let $P_{t+1} = P_{t+1} \cup F_t$
8. Let $i = i + 1$
9. end
10. Sort($F_t$, $\prec$)
11. Let $P_{t+1} = P_{t+1} \cup F_t[1 : (N - |P_{t+1}|)]$
12. Let $Q_{t+1} = \text{make - new - pop}(P_{t+1})$
13. Let $t = t + 1$
14. end

V. Experimental Result and Discussions

A. Experimental Setup

All simulations are conducted with a population size of 100, crossover rate of 0.8, mutation rate of 0.2 and generation number of 5000 for all implementations. In addition, we use SBX crossover and mutation. Tournament selection is adopted in a recombination and replacement scheme. These design parameters are chosen to be consistent with what were used in [16].
B. Results of GRN Optimization Problem

The proposed algorithm is tested on the GRN optimization problem defined in Section II. The three objectives to be minimized are the root mean squared error between the synthetic and simulated data for $LFY$, $AP1$ and $TFL1$. In order to mimic real expression data, the synthetic data obtained in the aforementioned manner was corrupted by adding random noise that followed a two parameter lognormal distribution. The noise was generated at four different levels, with the variance parameter kept at 1%, 5%, 10%, and 20% of actual values. The proposed approach was run 20 times at each noise level.

Fig. 2 shows the Pareto front produced by merging the non-dominated solutions of all 20 runs at different noise level for the proposed algorithm. We can see that the the convergence of the proposed approach decreased with the increasing noise. It is clear that noise plays a very important role in the performance of the proposed algorithm. When no noise exists, the root mean square error between the simulated and synthetic data for $LFY$ and $AP1$ can reach as small as $10^{-7}$. However, the performance of the proposed approach dropped dramatically with the increase of noise, which indicates the necessity of further noise handling techniques in the algorithm design.

C. Comparison of Proposed Approach with NSGAII

The proposed approach is compared with the classical Pareto dominance based NSGAII on the GRN optimization problem. Fig. 3 shows the non-dominated solutions obtained from 30 runs of approaches - 3-objective($LFY$, $AP1$ and $TFL1$) NSGAII, 2-objective($LFY$ and $AP1$) NSGAII and 2-objective($LFY$ and $AP1$) HPD based approach, represented as symbols “*”, big “.” and big “○”, respectively. Due to the objective reduction property of hierarchical Pareto dominance, as specified in Section IV(B), the non-dominated solutions obtained by the proposed HPD based approach is two dimensional while non-dominated solutions obtained by 3-objective NSGAII remain three dimensional. For the convenience of comparison, we only show the 2-dimensional projected non-dominated solutions obtained by 3-objective NSGAII in Fig. 3. As we can see in Fig. 3, the proposed approach outperforms both 2 and 3-objective NSGAII in terms of prediction error. More specifically, the proposed approach is 10 times better than 3-objective NSGAII; and 2 times better than 2-objective NSGAII, respectively, in terms of root mean squared error between the synthetic and simulated data for $LFY$ and $AP1$($10^{-7}$ vs. $10^{-6}$). The observation in Fig. 3 fit the property analysis of Hierarchical Pareto Dominance quite well.

In addition, the proposed approach is compared with the classical Pareto dominance based NSGAII on the GRN optimization problem at different noise levels, as shown in Fig. 4-7. As we can see in the figures, the performance of our proposed approach become close to, although slightly better than that of NSGAII in prediction error under noisy circumstances. We believe this is because inaccurate information (noisy data) has lead to the failure of taking advantage of the secondary objective set($TFL1$) to further increase the selection pressure for the proposed approach.

VI. CONCLUSION

In this paper, a hierarchical Pareto dominance based multi-objective approach is proposed for the optimization of gene regulatory network models. The approach is presented based on the neglected observation that (i) structural dependencies exist among objectives (expression of different genes); and (ii) some objectives may be more important than others in GRN optimization problem. Through the property analysis of hierarchical Pareto dominance, we conclude that the proposed approach is able to reduce the number of objectives and increase the selection pressure to relieve many objective problem.

We conducted experiments of the proposed approach for the GRN optimization problem. The results show great convergence (RMSE is as small as $10^{-7}$) when no noise exist. The performance of the algorithm deteriorated with the increase
of the noise level. We also carried out a comparison of the proposed approach with NSGAII. Our proposed approach shows superior results over both 2 and 3-objective NSGAII in terms of error in prediction of gene expression. Given proper hierarchical strategy, we can conclude that hierarchical Pareto dominance based multi-objective algorithm is a more appropriate alternative for GRN optimization problems than the Pareto dominance based MOEAs, such as NSGAII. We believe that the proposed algorithm is able to incorporate the prior knowledge (the hierarchies of objectives) into the multi-objective optimization process thus it can be applied to any optimization problem that needs to (i) discriminate between objectives and (ii) relieve many objective problem.

The work in this paper is only preliminary. Future work includes the consideration of noise handling techniques in algorithm design and the investigation of the proposed approach on more complex problem where more objectives need to be considered.

REFERENCES


