Discovering Structures in Gene Regulatory Networks Using Genetic Programming and Particle Swarms

Xinye Cai
Electrical and Computer Engineering Department
Kansas State University
xinye@ksu.edu

Stephen M. Welch
Agronomy Department
Kansas State University
welchsm@ksu.edu

Praveen Koduru
Electrical and Computer Engineering Department
Kansas State University
praveen@ksu.edu

Sanjoy Das
Electrical and Computer Engineering Department
Kansas State University
sdas@ksu.edu

ABSTRACT
In this paper, we describe a genetic programming and particle swarm hybrid algorithm for gene network discovery.

Categories and Subject Descriptors
I.2.8 [**Artificial Intelligence**]: Problem Solving, Control Methods and Search – *graph and tree search strategies, heuristic methods.*

General Terms
Algorithms, Theory.

Keywords
Genetic programming, gene regulatory networks, clonal selection, particle swarm optimization, bioinformatics.

1. PROBLEM DEFINTION
Synthetic data were generated for this problem based on temperatures and day lengths from 18 European sites [1]. The data represented the day of the year that the first inflorescence bud would become visible for *Arabidopsis thaliana* plants shown on each of three different dates *ca.* one month apart.

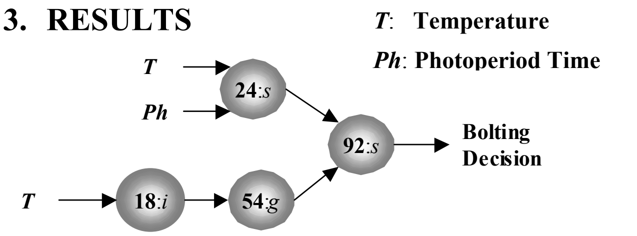
The goal is to obtain a simplified genetic network whose model bolting dates would be as close to the synthetic data as possible. Synthetic bolting dates are predicted by a particular model structure for the i^{th} combination of genotype, geographic site, and planting date.

The genes in the model genetic network are allowed to implement any of the following four functions: (*i*) *gain*: $o = c_g \cdot i_1$, (*ii*) *summer*: $o = c_s \cdot i_1 + i_2$, (*iii*) *multiplier*: $o = c_m \cdot i_1 \cdot i_2$, and (*iv*) *integrator*: c_i, i.e. $o(t) = o(t-1) + c_i \cdot i_1(t)$. In each case, i_1 and i_2 are the inputs and *o* the output. Each gene has a single parameter associated with it ($c =$ c_g, c_s, c_m or c_i). These parameters are assigned two separate numerical values, one for each allele. The inputs to each gene can be either the outputs from other genes in the network, or environmental, i.e. photoperiod or temperature.

2. APPROACH
Genetic Programming is used for network representation. Each solution is represented in the form of a string containing fields equal in number to that of putative network genes. Each field contains four entries - the two gene inputs, an entry representing gene function (*g*, *s*, *m*, or *i*), and an index identifying the gene in the data that this particular field in the string represents.

The optimization algorithm to discover the gene network's structure is based on clonal selection [2]. To estimate the parameters (c_g, c_s, c_m or c_i) of the network's genes, particle swarm optimization (PSO) is applied [3]. In each iteration, the finesses of each network in the population is computed by estimating its parameters through PSO. The solutions obtained are cloned based on their ranks. Each clone is then mutated and its fitness computed based on PSO. The best solutions from all the original networks and the clones are retained and the rest discarded.

3. RESULTS



$$Bolting(t) = C_{92} * (C_{24} * T(t) + Ph(t)) + C_{54} * (C_{18} * \sum_{n=0}^{t} T(n))$$

Figure 1. One of gene structures we found and its mathematical expression

Figure 1 shows a gene network obtained by our algorithm. It has correct genes and two-pathway. The real *Arabidopsis thaliana* network, after which this system is modeled, consists of four pathways, two of which mediate responses to features not considered here.

4. REFERENCES
[1] Welch, S.M., Roe, J.L., Das, S., Dong, Z., He, R., Kirkham, M.B. Merging genomic control networks and Soil-Plant-Atmosphere-Contiuum (SPAC) models. *Agricultural Systems* 86:243-74.

[2] de Castro, L.N., and Von Zuben, F.J. Learning and optimization using the clonal selection principle*, IEEE Transactions on Evolutionary Computation,* 6, 3 (June 2002) 239-251.

[3] M. Clerc and J. Kennedy, "The particle swarm-explosion, stability, and convergence in a multidimensional complex space," *IEEE Transactions on Evolutionary Computation*, vol. 6, no. 1, pp. 58-73, 2002.

GECCO'07, July 7–11, 2007, London, England, United Kingdom
ACM 978-1-59593-697-4/07/0007.