

CGP-WPSO Hybrid Algorithm for Gene Regulatory Network Construction

Ming Zheng^{1, a} and Mugui Zhuo^{1, b*}

¹Guangxi Colleges and Universities Key Laboratory of Professional Software Technology, Wuzhou University, Wuzhou, China

^a370505375@qq.com, ^b756456050@qq.com

*The corresponding author

Keywords: Gene Regulatory Network; Hybrid Algorithm; Particle Swarm Optimization; Inference of Gene Network; Optimal Strategy

Abstract. It is of great significance for the global food security to rely on the gene regulatory network to predict the performance of crops. A hybrid algorithm based on Cartesian genetic programming and linear decreasing inertia weight particle swarm optimization is proposed. Furthermore, in order to verify the effectiveness of the algorithm, the algorithm is applied to the problem of model reconstruction of Arabidopsis flowering regulatory system. Finally, the simulation results show that the algorithm can be based on crop genotypes and environmental conditions, the reconstruction can accurately predict the gene regulatory network model crop type. And the proposed algorithm can be suitable for gene regulatory network.

Introduction

It is an important trend in the field of bioinformatics to predict the phenotype of crops using gene regulatory networks[1]. In recent years, with the increasing demand of the world's population for food, it is more and more urgent to model the performance of crops. In the traditional method, the modeling of crops mainly depends on the understanding of biologists and the combination with practical experience. With the development of molecular biology has gradually become a hot topic in biological research, the current trend is at the gene level and the gene regulatory network, reveals the interaction between genes, so as to realize the accurate prediction of crop type.

The biological information contained in the genome of an organism is capable of controlling the growth of biomolecules and their response to environmental stimuli. In recent years, gene sequencing and gene chip technology have developed rapidly. As of September 2011, 3842 genome sequencing projects have been completed, and the other 7629 genome sequencing projects are still in progress[2]. How to mine the useful information in these biological data has become a hotspot in the field of bioinformatics is a big question. One of the most important issues is how to reconstruct the gene regulatory network from a large amount of observational data, such as the environment and biological phenotype. This kind of problem is called the mapping relationship between genotype and phenotype in biology[3]. The modeling of gene regulatory network is one of the important methods to solve the mapping relationship.

At present, there are a large number of research results on the modeling of gene regulatory networks. The proposed recognition algorithm of Boolean gene regulatory network based on network literature[4]; the literature[5] on the reconstruction of the dynamic Bayesian network model SEM algorithm is improved by using linear differential equation; literature[6]expression data regression analysis of gene, gene regulatory network is proposed using a heuristic algorithm to build sparse, and artificial data based on the experiments verify the effectiveness of the algorithm. The reconstruction methods are mostly applied in small gene regulatory networks, and most of the actual situation in the gene regulatory network belongs to the high dimension, the above method due to the large amount of calculation and the failure to achieve good results. Different from the above work, this paper proposed a new hybrid algorithm CGP-WPSO, which Cartesian Genetic Programming (CGP)[7] structure for searching model of gene regulatory network, and then through the linear decreasing inertia weight

particle swarm optimization (WPSO)[8] parameters of gene regulatory network are estimated. The PSO algorithm has fast convergence, and the idea of decreasing inertia weight in particle swarm makes the algorithm not fall into local optimum when the parameter estimation is too fast. Therefore, this new algorithm has faster convergence and global optimization characteristics of PSO and Genetic Programming, can quickly reconstruct the gene regulatory network.

Data and Problem Definition

Acquisition of Environmental and Phenotypic Data. This paper studies the environmental data from major research project of the National Science Foundation The Evolutionary Aspectsof Gene Network Pathway Signal Integration. Environmental data contains the optical cycle and temperature, which were obtained from Portugal's Coimbra and 8 locations in Finland, Jokioinen. The temperature is selected from 1971 to 1998, the local average daily temperature from March 1st to June 30th. The same period of time is obtained from the U.S. naval observatory.

Constructing an artificial gene regulatory network with parameters. The function of the Arabidopsis flowering control system was simulated[9]. The 2.2 section describes the functional characteristics of individual genes. Each gene has a two - type parameter, which represents different mutant alleles. G = 100 different genotypes were constructed by the combination of these different alleles[10].

Problem Definition. The use of artificial data to reconstruct the gene regulatory network model, flowering date gene regulation model after the reconstruction of the forecast should as far as possible and therefore need to be close to the artificial data, the root mean square error minimization E artificial data and model data:

$$E = \sqrt{\frac{\sum_{i=1}^n (D_i^{data} - D_i^{model})^2}{n}} \quad (1)$$

D_i^{data} and D_i^{model} is In the i experiment, the flowering time was predicted by the artificial flowering time and the reconstructed gene regulatory network model. I was selected for different genotypes, planting time and planting date.

CGP-WPSO Hybrid Algorithm

Descartes hybrid algorithm of genetic programming and linear decreasing inertia weight based on particle swarm optimization (CGP-WPSO) is divided into 3 steps (see Fig. 1). The first step is to identify the key genes, that is, from the 100 genes selected in a larger probability of gene regulatory network. These genes are selected as candidate genes, as the input of the second step CGP.

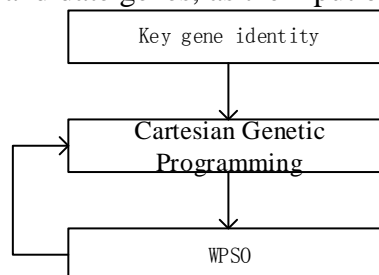


Figure 1. Flow chart CGP-WPSO algorithm

CGP initializes the population with N solutions (gene regulatory network structure), and generates a new network structure population by mutation operation. For each network structure, the WPSO algorithm is used to optimize the parameters to minimize the root mean square error of the bloom date, as shown in Eq. (1). The second step CGP and the third step WPSO will be repeated until the algorithm converges.

Identification of Key Genes. As mentioned in the preceding paragraph, each genotype contains 100 genes, only some of which are part of the flowering regulatory network. The basic idea of gene identification is to exclude those genes that cannot change the flowering time. The specific process is as follows: the identification of genes for each gene loci, genotype according to the gene variant allele was divided into two groups.

Organization Structure of CGP Network. Improved Cartesian Genetic Programming (CGP), so that it can be more suitable for the network structure model. In CGP, each solution (network structure) is represented by a string with M fields. The main goal of mutation operation is to search the search space effectively. Each item in the network structure is adjusted according to the predefined probability (called mutation probability). In addition, mutation operation is also affected by following constraints: (1) as mentioned earlier, the network topology, most of gene regulatory network only with feedforward so we set mutation operation not allowed to produce feedback regulation with network structure. (2) for the same network structure, the gene index must be unique.

WPSO Algorithm. In order to optimize the parameters of CGP algorithm, the particle swarm optimization (WPSO) algorithm with linear decreasing inertia weight is used to optimize the parameters. The parameter C_1 to C_M is represented as a vector C in the parameter domain space. At the beginning of the WPSO algorithm, a particle swarm $C_{(j)}$ is randomly generated, $j = 1, \dots, P$, where P is the number of particles. Each vector $C_{(j)}$ corresponds to the position of the j th particle.

So the function C , V and W can be obtained as below:

$$c_{t+1}(i) = c_t(i) + v_t(i) \quad (2)$$

$$v_{t+1}(j) = w \times v(j) + C_1 \times U[0,1] \times (c_{ib}(j) - c_t(j)) + C_2 \times U[0,1] \times (c_{gb,t} - c_t(j)) \quad (3)$$

$$w = (w_1 - w_2) \times \frac{(MAXT - t)}{MAXT} + w_2 \quad (4)$$

The position of the particle must be controlled in a predefined search space. When the particle exceeds the space, the particle will return to the space. In addition, its speed will be multiplied by (-1), which flew to the opposite direction. The fitness of each gene regulatory network model, which is composed of the network structure and the associated parameters, is defined as the root mean square error (E) (see Fig. 1). At the same time, in order to better simulate the real experiment, we add noise to the date of flowering (STD = 2 days).

Experiment

To test our algorithm, Artificial gene regulatory network structure for artificial flowering date was used. And the results was shown as below:

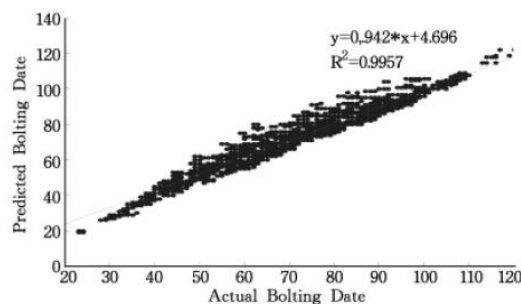


Figure 2. Artificial flowering date data vs reconstruction gene regulation network
Fig. 2 shows that CCG-WPSO is efficient for the gene expression data.

Summary

The automatic generation of gene network structure from the experimental data is a very important research direction in the field of bioinformatics. In this paper, we propose an algorithm to derive the network structure and the related parameters directly from the experimental data. Descartes genetic programming and linear decreasing inertia weight particle swarm optimization algorithm. In this paper, we apply it to the problem of model reconstruction of Arabidopsis flowering regulatory system. Experimental results show the effectiveness of the proposed algorithm.

Acknowledgements

This work was supported by grants from The National Natural Science Foundation of Chi-na (No. 61502343, No. 61373051, and No. 61402423), China Postdoctoral Science Foun-dation funded(No. 2016M590260), the Guangxi Natural Science Foundation (No. 2015GXNSFBA139262), the Science Research Funds for the Guangxi Universities (No. KY2015ZD122), Guangxi Colleges and Universities Key Laboratory of Professional Software Technology, Wuzhou University.

References

- [1] Chai, L.E., Loh, S.K., Low, S.T., Mohamad, M.S., Denis, S., and Zakaria, Z.: ‘A review on the computational approaches for gene regulatory network construction’, *Computers in Biology and Medicine*, 2014, 48, pp. 55-65
- [2] Gaudet, P., Chisholm, R., Berardini, T., Dimmer, E., Engel, S.R., Fey, P., Hill, D.P., Howe, D., Hu, J.C., Huntley, R., Khodiyar, V.K., Kishore, R., Li, D., Lovering, R.C., McCarthy, F., Ni, L., Petri, V., Siegele, D.A., Tweedie, S., Van Auken, K., Wood, V., Basu, S., Carbon, S., Dolan, M., Mungall, C.J., Dolinski, K., Thomas, P., Ashburner, M., Blake, J.A., Cherry, J.M., Lewis, S.E., and Gene Ontology, C.: ‘The Gene Ontology's Reference Genome Project: A Unified Framework for Functional Annotation across Species’, *Plos Computational Biology*, 2009, 5, (7)
- [3] Xu, Z.Y., Zhu, W.G., Wang, C., Huang, L., Zhou, Q.Q., Hu, J.Z., Cheng, X.S., and Hong, K.: ‘Genotype-phenotype relationship in patients with arrhythmogenic right ventricular cardiomyopathy caused by desmosomal gene mutations: A systematic review and meta-analysis’, *Sci Rep*, 2017, 7, pp. 8
- [4] He, Q.B., Xia, Z.L., and Lin, B.: ‘An efficient approach of attractor calculation for large-scale Boolean gene regulatory networks’, *Journal of Theoretical Biology*, 2016, 408, pp. 137-144
- [5] Liu, F., Zhang, S.W., Guo, W.F., Wei, Z.G., and Chen, L.N.: ‘Inference of Gene Regulatory Network Based on Local Bayesian Networks’, *Plos Computational Biology*, 2016, 12, (8), pp. 17
- [6] Lu, T., and Wang, M.: ‘Investigate Data Dependency for Dynamic Gene Regulatory Network Identification through High-dimensional Differential Equation Approach’, *Communications in Statistics-Simulation and Computation*, 2016, 45, (7), pp. 2377-2391
- [7] Goldman, B.W., and Punch, W.F.: ‘Analysis of Cartesian Genetic Programming's Evolutionary Mechanisms’, *Ieee Transactions on Evolutionary Computation*, 2015, 19, (3), pp. 359-373
- [8] Arasomwan, M.A., and Adewumi, A.O.: ‘On the Performance of Linear Decreasing Inertia Weight Particle Swarm Optimization for Global Optimization’, *Scientific World Journal*, 2013
- [9] Wils, C.R., and Kaufmann, K.: ‘Gene-regulatory networks controlling inflorescence and flower development in Arabidopsis thaliana’, *Biochimica et Biophysica Acta-Gen Regulatory Mechanisms*, 2017, 1860, (1), pp. 95-105
- [10] van Veen, H., Vashisht, D., Akman, M., Girke, T., Mustroph, A., Reinen, E., Hartman, S., Kooiker, M., van Tienderen, P., Schranz, M.E., Bailey-Serres, J., Voeselek, L., and Sasidharan,

R.: 'Transcriptomes of Eight *Arabidopsis thaliana* Accessions Reveal Core Conserved, Genotype- and Organ-Specific Responses to Flooding Stress', *Plant Physiology*, 2016, 172, (2), pp. 668-689