Apply the Feature of Entropy Convergence of ACO to Short the Runtime of Gene Order

Ben-Qiong Hu
Group of Gene Computation,
Key Lab. of Visual Computing and Virtual Reality of
Sichuan Province, Chengdu, 610068, China;
College of Information Management,
Chengdu University of Technology,
Chengdu, 610059, China

Gang Jiang
Group of Gene Computation,
College of Mathematics and Software Science,
Sichuan Normal University, Chengdu, 610068, China

Shi-Peng Wang
Group of Gene Computation,
College of Mathematics and Software Science,
Sichuan Normal University, Chengdu, 610068, China

Chao-Yang Pang
Group of Gene Computation,
Key Lab. of Visual Computing and Virtual Reality of
Sichuan Province, Chengdu, China;
College of Mathematics and Software Science,
Sichuan Normal University, Chengdu, 610068, China
Corresponding Email: cypang@live.com

Abstract—**Alzheimer's disease (AD) is the most common form of dementia. To find a way of cure it, gene study is necessary. And gene order is a new conception of gene study currently, where gene order refers to a permutation of genes in which similar genes are ordered together one by one, and optimal gene order can be abstracted as shortest TSP route. Currently only two types of tools are reported to calculate gene order, which are Genetic Algorithm (GA) and Ant Colony Optimization (ACO). In these two types, one bottleneck of computation is that their runtime is too long while gene data is too large. To weaken the bottleneck, in this paper, the feature of entropy convergence of ACO is used as the termination criterion of ACO to speed up the computation of AD gene order. Experiment shows that the method proposed in this paper has obvious advantage on runtime and solution quality.**

Keywords-Gene Order; Ant Colony Optimization (ACO); Entropy; Alzheimer's Disease

I. Introduction

A. Introduction of Alzheimer's Disease (AD), DNA Microarray Data, and Gene Order

Alzheimer disease (AD) is the most common form of dementia which damages several parts of brain [1]. Accounting for the majority of dementia cases, AD Type Dementia is about 50-75% of cases. And according to an estimate using the 2000 census, AD, is the most common form of dementia in the elderly [2]. For example, Charles Kuen Kao who is the father of fiber optic communications and was awarded Nobel Prize on 2009 is suffering a hideous torment of AD, and the American president Ronald Wilson Reagan was killed by AD.

Currently the factors which cause AD are not clear, and there is no good way to cure it. To find the factors of AD and cure it, identifying the correlative genes is critical. Currently, there are a few genes found correlative with AD, such as APP, APOE, PSEN [3], and most of finding ways are to do biology experiments, which cost a lot of money and time. To find more genes correlative with AD, DNA microarray technology is a good tool possibly, and Pang, Hu and *et al.* did an attempt [4]. DNA microarray technology is introduced as below.

DNA microarray technology has been applied to many biological domains, such as drug discovery, molecular diagnosis, and toxicological research. And its important application is to generate gene data, which hides a lot of biological information. And one common data structure of microarray data set is a matrix (table) [5]. In this table, one column data represents one test result (experiment), and the number of column is the number of tests. Every data in the table is the expression level of a gene. Each line data is a vector corresponding to a gene, and it represents the results of the experiments. For a given gene, a line vector is often regarded as a basic unit hiding gene information, and is operated by computation.

One important work of biology is to make similar genes cluster together, which is the base for identifying gene correlative with disease. And gene clustering often refers to cluster similar vectors together because line vectors of matrix (table) contain the information of genes. A number of algorithms were proposed to cluster gene expression profiles [6-11]. To get much better quality of clustering, the conception of gene order is put forward recently and it is introduced as below.

A gene is associated with a line vector of matrix. **Gene order** is the permutation of all line vectors in which all vectors are ordered one by one and similar vectors are ordered together. And the **optimal gene order** refers to the permutation which distance is minimal when traveling every vector one by one along the order. Holding view point of mathematics, optimal gene order is a route of traveling salesman problem (TSP) if every vector is abstracted as a virtual city [11-14].

To calculate optimal gene order approximately, Tsai *et al.* apply family competition genetic algorithm (FCGA) [11], Seung-Kyu *et al.* applied a hybrid genetic algorithm

978-0-7695-4281-2/10 $26.00 © 2010 IEEE
DOI 10.1109/ICGEC.2010.77

(NNGA) to solve the problem [13]. Chao-Yang Pang, Gang Jiang and *et al*. did the first attempt to apply another method named ant colony optimization (ACO) to calculate gene order [5].

B. *Introduction of Ant Colony Optimization (ACO) and ACO-Entropy Algorithm and Genetic Algorithm*

The optimal route of the traveling salesman problem (TSP) is the route all cities must be traveled once and only once and the length of route is the shortest.

To solve the problem of TSP, a method named Ant Colony Optimization (ACO) is presented in 1997 [14]. ACO is essentially a system based on agents that simulate the natural behavior of ants, in which real ants are able to find the shortest route from a food source to their nest, without using visual cues by exploiting pheromone information. Pheromone is deposited when ants walking on a route. It provides heuristic information for other ants to choose their routes. The more dense the pheromone trail of a route is, the more possibly the route is selected by ants. At last, nearly all ants select the route that has the densest pheromone trail, and this route is the shortest route potentially.

ACO has been applied to solve optimization problems widely and successfully [15, 21-23], such as TSP, quadratic assignment problem, image processing. So as to the application at biology, includes DNA fragment assembly and protein folding problem. However, its application on gene order is rare currently, and possibly ref. [5] is the first paper. And the reason of rareness may lies on that ACO costs too many running time and the data size of gene is large. For example, for a group of 500 cities and a general personal computer, a few hours will be cost by ACO. On the other hand, the quality of ACO solution is higher than other methods in general, that is attractive for the computation of gene order. Therefore, fast improved ACO algorithm is useful for computation of gene order. A new termination criterion of ACO named entropy convergence is presented recently by Pang [16]. And using this criterion, ACO stops at the approximated minimal iteration steps at which ACO solution becomes stable, and this property results in the running speed of ACO is improved heavily.

The computation tools of gene order are few currently, which are Genetic Algorithms (GA) [11] and ACOs [5]. Genetic algorithm can be understood as an intelligent probabilistic search algorithm which works on the Darwin's principle of natural selection and can be applied to a variety of combinatorial optimization problems [17].

C. *Motivation*

As gene order is a good method for cluster similar genes together which may be useful for the biologists, and according to the authors' understanding of this paper, there are two types of algorithms to calculated gene order currently, which are GA [11] and ACO [5]. And as the size of AD gene data is too large, general algorithm of GA or ACO will cost too long runtime possibly. To fast process large AD gene data, the motivation of this paper is to apply the combination of ACO and its feature of entropy convergence to calculate AD gene order.

II. Methods

A. *Measurement of Similarity of Genes*

A gene associates with a line vector of a matrix, as it is introduced above. And the similarity of two genes can be estimated by the distance between the two associated vectors. There are a few measurement have been proposed to measure gene similarity [13]. Pang, Jiang and *et al*. show that squared Euclidean distance generates best quality of gene order compared with three other distance formulas [5], and it will be used in this paper.

In addition, gene order is a TSP route. Therefore, we can use the length of route to review the quality of gene order. The smaller the length is, the better the quality of the gene order is. And the length is called **fitness function** [5].

B. *Apply ACO and Its Feature of Entropy Convergence to Calculate Optimal Gene Order*

The typical application of ACO is to solve TSP, and the outline of its method is listed as below:

```
Step1. Initialize parameters and pheromone on every
       edge between two cities. Pre-assign maximum
       iteration number $t_{\max}$. And suppose $t$ denotes
       the t-th iteration step.
Step2. While ( $t \le t_{\max}$ )
    {
       Step2.1 Let every ant travel all cities to find a
               TSP route, and the pheromone on the
               edges will conduct it to select its route.
       Step2.2 If an ant passes through an edge, it will
               release pheromone on this edge. Update
               the amount of pheromone depositing on
               every edge.
    }
Step3. Select the route which has minimum length as
       output.
```

As it is showed above, the iteration number of ACO $t_{\max}$ is pre-assigned, which is set to an empirical value in general. To save runtime, we need a criterion to find the approximation of optimal $t_{\max}$, at which ACO is convergent approximately (Note: ACO converges to local optimal solution in general) . To find the criterion, Pang studied the relation between quantity of pheromone, entropy and convergence, and find the approximate optimum $t_{\max}$ can be marked by entropy convergence [16].

Assume that there are m ants a_1, a_2, …, a_m to seek their TSP routes. And assume that the *i-th* ant a_i finds route $r_i^{(t)}$ at *t-th* iteration step. There is pheromone depositing on every edge of route $r_i^{(t)}$ and suppose the sum of pheromone at every edge is $f_i^{(t)}$. Then there is ratio of pheromone of route $r_i^{(t)}$, it is defined as

$$p_i^{(t)} = \frac{f_i^{(t)}}{\sum_{j=1}^{m} f_j^{(t)}} .$$

$p_i^{(t)}$ is called as pheromone probability. The bigger $p_i^{(t)}$ is, the more possibly route $r_i^{(t)}$ is selected by other ants at next iteration step.

Since there are m ants and each ant associates with pheromone probability $p_i^{(t)}$, there are set of pheromone probability

$$P\_Set^{(t)} = \{p_1^{(t)}, p_2^{(t)}, \cdots, p_i^{(t)}, \cdots, p_m^{(t)}\} .$$

Entropy of pheromone is defined as

$$H(t) = -\sum_{i=1}^{m} p_i^{(t)} \ln p_i^{(t)} \qquad (1)$$

The study [16] shows that sequence $\{H(1), H(2), ..., H(t), ...\}$ is convergent, and the convergence shows the statistics feature of ACO convergence. According to this convergence property, the criterion of entropy convergence $\frac{|H(t)-H(t-1)|}{H(t)} < \varepsilon$ is used as the termination criterion of ACO, by which t is estimated, where ε is a threshold value.

The method of applying the above criterion to calculate gene order is proposed as below:

```
Step1: Construct Graph for Genes: As it is introduced
    in section 1, a gene is associated with a vector. The
    vector is regarded as a virtual city in
    multi-dimensional space. Calculate the distances
    between two arbitrary virtual cities, construct a
    completed graph, and the weight of edge is the
    distance between two virtual cities.
Step2: Initialize pheromone on edges and other
    parameters on constructed graph. Suppose $t$ denotes
    the $t$-th iteration step, and let $t=0$. And set a sufficient
    value as the initial entropy value $H(0)$.
Step3: do
  {
    Step3.1. $t = t + 1$
    Step3.2. Every ant selects its route according to the
        rule of ACO.
    Step3.3. Update the pheromone of the shortest route
        traveled by ants.
    Step3.4. Calculate entropy $H(t)$ by formula (1)
  } while ( $\frac{|H(t)-H(t-1)|}{H(t)} \ge \varepsilon$ )
Step4: Select the shortest route traveled by ants as
solution of TSP route, and the solution is a gene order.
```

The above proposed method is named ACO-Entropy in this paper.

III. Experiments and Results

To test the performance of ACO-Entropy, three other algorithms (ACO, GA and an Improved GA (IGA) [18]) is used as reference in this section, where ACO refers to often used algorithm Ant-Cycle [19]. And their runtimes and qualities of gene order (i.e., solution) are compared.

The source data is downloaded from GEO Datasets, NCBI [20]. And four cases of control, incipient, moderate and severe data are provided in this original microarray data. Experiment results of 9 times of control are put together to form a matrix (table), in which every column corresponds to a different experiment result. Then the size of matrix is 22283 lines and 9 columns, and each line vector (virtual city) is 9-dimensional. Using the same method, the experiment data of incipience, moderateness and severeness are organized as different matrix respectively.

In addition, according to usual practice, all data of AD gene is log-transformed and all columns of selected data are normalized before the applications.

The experiments of this paper are done on personal computer, CPU (2): 2.99GHZ, 3.0GHZ; Memory: 1.0GB.

The parameters of ACO are set as below:

$\alpha = 1$, $\beta = 2$, $\rho = 0.7$, $Q = 100$, $\tau_{ij}(0) = 1$, $m = 50$, $t_{\max} = 100$.

(Note: the mean of these symbols are shown in ref. [5])

The parameters of ACO-Entropy are set as below:

$\alpha = 1$, $\beta = 10$, $\rho = 0.7$, $Q = 100$, $\tau_{ij}(0) = 1$, $m = 50$, $\varepsilon = 0.00001$

The parameters of GA are set as below:

$t_{\max}$ =500, N =400.

, where $t_{\max}$ and N represents the maximum number of iteration and the size of populations respectively.

The parameters of IGA are set as below:

$t_{\max}$ =2500, N =1500.

All experiment results are shown in Table 1 and Table 2. From the test results, the following conclusion is got:

Compared with ACO, GA and IGA, ACO-Entropy proposed in this paper costs the minimum average time and the quality of gene order (solution) is good. The larger the data size is, the clearer the superiority of ACO-Entropy is.

IV. Conclusion

Alzheimer's disease (AD) is the most common form of dementia. To find a way of cure it, gene study is necessary; and in which gene clustering is the base. Gene order is a new conception of gene clustering presented recently, and it is a permutation of genes in which similar genes are ordered together. Optimal gene order can be abstracted a shortest TSP route, then computation of optimal gene order is equal to calculate a shortest TSP route approximately. Currently, only two types of tool are reported to calculate gene order, which are Genetic Algorithm (GA) and ACO. In these two types, one bottleneck of computation is that their runtime is too long while gene data is too large. To weaken the bottleneck, in this paper, the feature of entropy convergence of ACO is used as the criterion of ACO convergence to speed up ACO and to calculate gene order.

To check the performance of the method proposed in this paper, common used ACO (Ant-Cycle), GA, an excellent improved GA (IGA) are tested under different conditions. Experiment shows that the method proposed in this paper has obvious advantage on runtime and quality of solution.

Acknowledgment

The work is supported by the open fund of key lab. of modern communication and information process of Shenzheng city and the fund of key developing program of technology innovation from Education Dept. of Sichuan Province (No. 09zz028). The authors appreciate the help from Prof. J. Zhang and Prof. J. Zhou. The authors appreciate the help from the other members of gene computation team: W. Hu, C.-B. Wang, X. Li, H. Liu, L.-J. Ye, J.-L. Zhou, P. Shuai, Y. Li, Z.-C. Dan, M.-X. Chen, D.-X. Zhang, L. Yang, and X.-P. Xia.

REFERENCES

[1] R. Jakob-Roetne and H. Jacobsen, “Alzheimer’s Disease: from Pathology to Therapeutic Approaches”. Medicinal Chemistry, 2009. 48(17): pP. 3030-59, doi: 10.1002/anie.200802808.

[2] L.E. Hebert, P.A.Scherr, J.L. Bienias, D.A. Bennett and D.A. Evans, “Alzheimer Disease in the US Population”. Arch Neurol, 2003. 60: p. 1119-1122.

[3] R.E. Tanzi and L. Bertram, “Twenty Years of the Alzheimer’s Review Disease Amyloid Hypothesis: A Genetic Perspective”. Cell, 2005. 120: p. 545-555, doi: 10.1016./j.cell.2005.02.008.

[4] C.-Y. Pang, W. Hu, B.-Q. Hu, Y. Shi, C.R. Vanderburg, J.T. Rogers and et al, “A Special Local Clustering Algorithm for Identifying the Genes Associated With Alzheimer’s Disease”. IEEE transactions on nanobioscience, 2010. 9: p. 44-50.

[5] C.-Y. Pang, G. Jiang and B.-Q. Hu. “Quality of Gene Order Calculated by Ant Colony Algorithm is Sensitive to Distance Formula”. in 2009 3rd International Conference on Genetic and Evolutionary Computing. IEEE Press and CPS Press, Oct. 2009, pp. 781- 785, doi: 10.1109/WGEC.2009.63.

[6] M.B. Eisen, P.T. Spellman, P.O. Brown and D. Botstein. “Cluster analysis and display of genome-wide expression patterns”. Proceedings of the National Academy of Sciences, National Academy of Sciences, USA. 1998.

[7] P. Tamayo, D. Slonim, J. Mesirov, Q. Zhu, S. Kitareewan, E. Dmitrovsky, and et al. “Interpreting patterns of gene expression with self-organizing maps: Methods and application to hematopoietic differentiation”. in Proceedings of the National Academy of Sciences. 1999.

[8] A.A. Alizadeh, M.B. Eisen, R.E. Davis, C. Ma, I.S. Lossos, A. Rosenwald, and et al, “Distinct types of diffuse large b-cell lymphoma identified by gene expression profiling”. Nature, 2000. 403(6769): p. 503-511.

[9] J. Herrero, A. Valencia and J. Dopazo, “A Hierarchical Unsupervised Growing Neural Network for Clustering Gene Expression Patterns”. Bioinformatics, 2001. 17: p. 126-136.

[10] P. Merz and A. Zell. “Clustering gene expression profiles with memetic algorithms”. Proceedings of the 7th International Conference on Parallel Problem Solving from Nature, 2002, pp. 811-820, doi: 10.1007/3-540-45712-7.

[11] H.K. Tsai, J.M. Yang and C.Y. Kao. “Applying genetic algorithms to finding the optimal order in displaying the microarray data”. Proceedings of the Genetic and Evolutionary Computation Conference. 2002.

[12] H.K. Tsai, J.M. Yang and C.Y. Kao. “A genetic algorithm for traveling salesman problems”. Proceedings of the Genetic and Evolutionary Computation Conference. 2001.

[13] S.-K. Lee, K. Yong-Hyuk and M. Byung-Ro, “Finding the Optimal Gene Order in Displaying Microarray Data”. Lecture Notes in Computer Science, Springer Berlin / Heidelberg, 2003: p. 1611-3349.

[14] M. Dorigo and L.M. Gambardella, “Ant colony system: A cooperative learning approach to the traveling salesman problem”. IEEE Transactions on Evolutionary Computation, 1997. 1: p. 53-66.

[15] G.-L. Qin and J.-B. Yang, “An improved ant colony algorithm based on adaptively adjusting pheromone”. Information and Control, 2002. 31(3): p. 198-201.

[16] C.-Y. Pang, C.-B. Wang and B.-Q. Hu, “Experiment Study of Entropy Convergence of Ant Colony Optimization”. Unpublished, 2009, arXiv:0905.1751v4.

[17] F. Djannaty and S. Doostdar, “A Hybrid Genetic Algorithm for the Multidimensional Knapsack Problem”. International Journal of Contemporary Mathematical Sciences, 2008. 3: p. 443-456.

[18] J. Kirk. “Traveling Salesman Problem-Genetic Algorithm”. Available from:http://www.mathworks.com/matlabcentral/fx_files/13680/2/tsp_ga.zip.

[19] H.-B. Duan, “Ant Colony Algorithms: Theory and Applications”. 2005, Bei Jing: Science Press.

[20] E.M. Blalock, J.W. Geddes, K.C. Chen, N.M. Porter, W.R. Markesbery and P.W. Landfield, “Incipient Alzheimer’s disease: Microarray correlation analyses reveal major transcriptional and tumor suppressor responses”. PNAS, 2004, 101: p. 2173-2178, doi: 10.1073/pnas.0308512100.

[21] Hameed Al-Qaheri, Abhijit Mustafi, and Soumya Banerjee, “Digital Watermarking using Ant Colony Optimization in Fractional Fourier Domain,” Journal of Information Hiding and Multimedia Signal Processing, Vol. 1, July, 2010, pp. 179-189.

[22] S. C. Chu, John F. Roddick and J. S. Pan, "Ant Colony System with Communication Strategies", Information Sciences, Vol. 167, Issues 1-4, 2004, pp. 63-76.

[23] Shu-Chuan Chu, John F. Roddick, Che-Jen Su and Jeng-Shyang Pan, "Constrained Ant Colony Optimization for Data Clustering", 8th Pacific Rim International Conference on Artificial Intelligence, LNAI 3157, 2004 , pp. 534-543.

TABLE I. THE PERFORMANCE COMPARISON OF THE FOUR ALGORITHMS

Data	***Indexes***	***ACO-Entropy***	***ACO***	***IGA***	***GA***
Control AD (See Table 2)	Average Quality	249.4558	243.4009	336.4084	4009.4750
	Average Time	17.5563	337.6572	412.1270	346.5300
Incipient AD	Average Quality	192.8771	187.1551	280.2695	3155.9156
	Average Time	16.4431	337.3377	413.0560	347.8631
Moderate AD	Average Quality	214.7265	209.8339	288.3581	3574.7114
	Average Time	17.1515	337.8827	412.5678	347.3183
Severe AD	Average Quality	256.0042	238.6089	356.2051	3234.6133
	Average Time	16.1681	337.2706	412.7362	347.8523
Conclusion: The speed of ACO-Entropy is faster than the other three algorithms by factor of 19 at least, and its solution error rate is less than the best solution obtained by the other three algorithms by 7.29%. In addition, the larger the data scale is, the more obvious the superiority of ACO-Entropy is.					

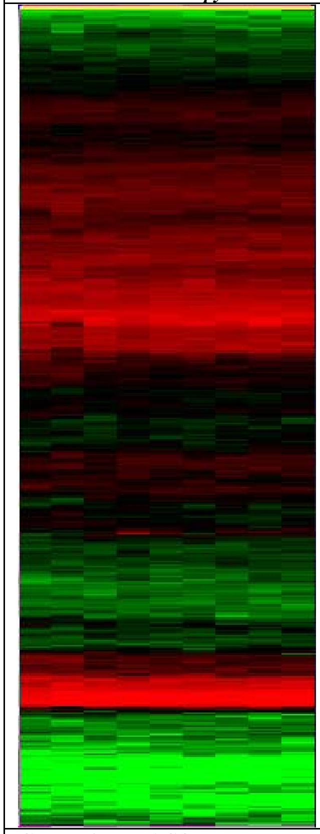
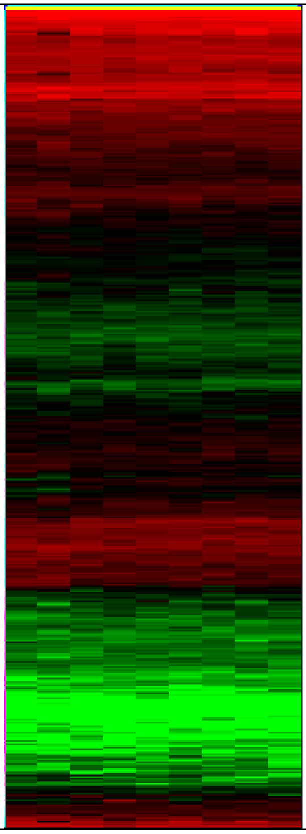
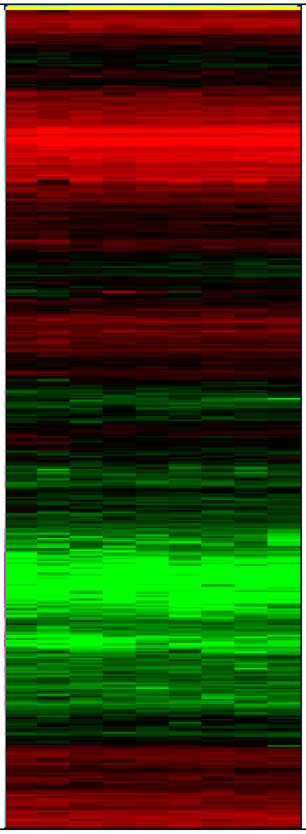
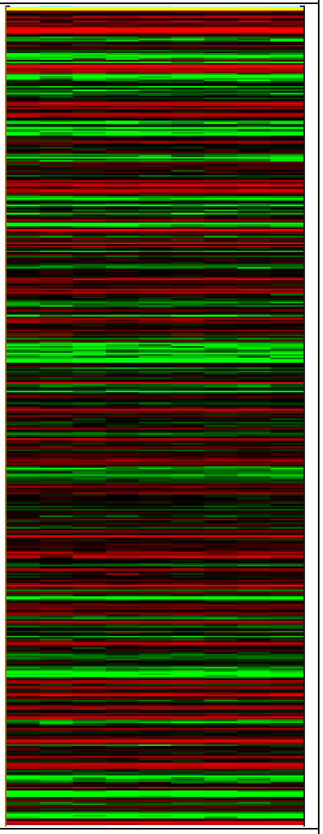
a. The tested data are selected from records 1st-500th of original data [20] (if number of records is too large, much more runtime will be cost).

b. The selected data are normalized before applications.

c. 40 tests are done on the selected data sets and the data listed above are the average results.

d. The vector distance formula is Squared Euclidean Distance, and it is used to calculate Fitness Function [5] too.

TABLE II. THE HEAT MAPS OF CONTROL MAN CALCULATED BY THE FOUR ALGORITHMS

ACO-Entropy	***ACO***	***IGA***	***GA***
			
Conclusion: ACO-Entropy generates the best quality of gene order in general, and GA generates the worst quality.			

a. All heat maps are generated by TreeView downloaded from http://rana.lbl.gov/downloads/TreeView/TreeView_vers_1_60.exe.

b. Each heat map is corresponding to the solution that has the best fitness value of the 40 tests.