Inference of a gene regulatory network by means of interactive evolutionary computing

Hitoshi Iba a, Atsushi Mimura b,*

a Graduate School of Frontier Sciences, The University of Tokyo, Hongo 7-3-1, Bunkyo-ku, Tokyo, Japan
b Sony Corporation Limited, 6-7-35 Kitashinagawa, Shinagawa-ku, Tokyo 141-0001, Japan

Received 5 June 2001; received in revised form 30 December 2001; accepted 27 April 2002

Abstract

Inferring a gene regulatory network is one of the challenging topics in the field of Bioinformatics. In order to infer a network structure effectively, the new approach that allows human intervention and strategic data acquisition in the inference process seems to be necessary. In this paper, we will propose an effective approach for interactively inferring gene regulatory networks using gene expression data from DNA microarrays. We will also establish the system that realizes our approach by GA-based interactive algorithm. Experimental results show that our method can infer the network structure accurately with a relatively small amount of expression data.

© 2002 Elsevier Science Inc. All rights reserved.

1. Introduction

Recent studies have been pursued energetically to infer gene regulatory networks from expression time series data obtained with DNA microarrays [1,3,10,11].

*Corresponding author.
E-mail addresses: iba@miv.t.u-tokyo.ac.jp (H. Iba), amimura@miv.t.u-tokyo.ac.jp (A. Mimura).
It is difficult to infer gene networks using continuous-value models [2,4], because a number of solution candidates generate similar expression patterns. To infer networks efficiently while avoiding numerous local solutions, human intervention such as employing heuristics and adding sufficient search restriction or data is necessary. A computer can be used interactively for the evaluation – this seems to be an effective approach instead of conventional top-down approaches (Fig. 1).

This paper presents an effective approach for interactively inferring gene regulatory networks. Our method requires a reasonable amount of gene expression time series data obtained with DNA microarrays for the inference. The system requests the user to add other data or search restriction between specific genes, whose control is ambiguous. This is realized by comparing the same parameters in several solution candidates. In accordance with this request, the user restricts the control between the specified genes using a new experimental verification and the inference proceeds. By repeating this interaction, the system infers a network accurately with the smallest amount of data.

2. Inference method

2.1. Structure of GA chromosomes

When inferring a network, Genetic Algorithms (GAs) optimize model parameters that generate an expression pattern that is most similar to the given one [5,6]. When inferring the gene that controls the expression of $i$th gene $X_i$ in a network of $n$ genes, the GA chromosome is configured by arranging the model parameters related to $X_i$, which includes the $i$th row of an influence matrix $M[n,n]$ (see Fig. 3 for example). Fig. 2 shows the correspondence between an influence matrix $M[n,n]$ and its network structure.
2.2. Generation of population groups

In a larger network, there are more local solutions, making it difficult to accurately infer the network configuration.

To eliminate this difficulty, a step is taken to generate several GA population groups and determine the parameters for optimization in each group (which we call an “exon”) and the other parameters consistently fixed at zero (which we call an “intron”), as shown in Fig. 3.

The number of “exons”, i.e., genes that regulate $X_i$, is limited to be 5 or so. This reflects the fact that the number of the regulators that directly activate or inhibit the expression of one mRNA is not so large.

If there is some initial restriction on the network structure by a pre-treatment, “exons” and “introns” can be determined on the basis of the restriction, i.e., structures near to the initial settings are generated from probability.

Fig. 4 shows the process of the proposed search. GA population groups are produced in accordance with user-defined restrictions. Individuals rated best are picked up from each group after the specified number of generations. Then they are compared to previous solution candidates, where poor candidates are replaced and better individuals are reserved as current solution candidates. After that, new GA population groups with fresh structures are produced and this process is executed until some terminal condition is satisfied.

![Fig. 2. Correspondence between influence matrix and network topology.](image)

![Fig. 3. Population generation of different structures.](image)
2.3. Evaluation of individuals

From the GA chromosomes, parameters are decoded and an expression time series is produced using the network model. For the accurate calculation, expression data are interpolated by the spline method.

The fitness value of each individual is defined as the average of square errors between the calculated time series data and experimental data

\[ F_i = \frac{\sum_{c=0}^{\text{#c}} \sum_{t=0}^{\text{#DF}} (x_{c,j}(\text{exp})(t) - x_{c,j}(\text{calc})(t))^2}{\text{#c}}, \]

where \( x_{c,j}(\text{exp})(t) \) is experimental expression data under the condition of \( c \) (e.g., disruption of gene \( X_j \)), and \( x_{c,j}(\text{calc})(t) \) is calculated expression data using the network model in accordance with the experimental condition. Thus, individuals with lower fitness values are better.

3. Interactive inference system

We have constructed the GUI environment to infer a gene regulatory network interactively using an algorithm like the one above (Fig. 5).

The user can edit and save network configurations and model parameters through the GUI. He or she can also give restrictions to the inference.

Network inference proceeds as follows:
Step 1. The user selects a gene \( X_i \) for the inference and gives the initial structure and search restrictions for the network in accordance with the preliminary results or heuristics.
Step 2. The GA mentioned above is executed to produce solution candidates that differ in the network structure.

Step 3. The system shows solution candidates, and calculates the regulation reliability score $S_{ij} \subset S_i$ of a gene $X_j$ over the gene $X_i$ by comparing values of the same parameters. The user is then requested to give data and/or search restrictions concerning the relationship between $X_j$ and $X_i$ with the lowest score (Fig. 6).

Step 4. In accordance with the request, the user experimentally verifies whether $X_j$ has any effect on the expression of $X_i$ or not (Fig. 7).

Step 5. On the basis of the experimental results, the user adds the search restriction and/or new expression data about $X_i$. Then he or she executes Step 2 again.

Step 6. Repeating this, the system identifies the gene that controls the expression of the gene $X_i$ and acquires a proper network structure (Fig. 8).

Step 7. Repeating the above process for all genes, the system estimates the overall structure of the network (Fig. 9).
In the above Step 3., the score $S_{i,j}$, i.e., the reliability of the effect of $X_j$ on search node $X_i$, is calculated as follows:

$$S_{i,j} = \frac{\text{Max}(V) - \text{Avr}(V)}{\sum V - \text{Avr}(V)}.$$  

(2)
where \( V = \{ V_0, V_+, V_- \} \) and

\[
V_0 = \sum_{j=1, M_0=0}^{\#\text{Sol}} \left( 1 - \frac{F_j}{\sum_i F_i} \right),
\]

and so on.

As is shown in Fig. 10, restrictions are requested by checking the consistency of positive, negative, and zero nature of the values for the influence matrix among the solution candidates. This is because values devoid of consistency are considered to be trapped in a local solution.

4. Experimental results

4.1. Assays

In order to show the effectiveness of the proposed approach, the following comparative experiments were carried out.

We used randomly generated networks composed of 10 nodes and 20 edges for the inference targets. An example is shown in Fig. 11.

For the network, a model based on the power-law formalism known as S-system [7,8] was used with the parameters restricted below

\[
\frac{dx_i}{dt} = x_i \prod_{j=1}^{n} x_j^{M_{ij}} - \beta_i x_i \quad (i = 1, 2, \ldots, n),
\]

where \( \alpha_i \) and \( \beta_i \) is set to be 1.0 for simplicity. The values of the influence matrix \( M[n,n] \) is derived from Fig. 11. Others are all set to be 0.0. It is assumed that wild-type expression data are available for each node. We used artificial data...
generated by the above equation. Each time series is composed of 10 points with 0.5 time intervals and its initial value is consistently set to be 0.10.

We applied our proposed method with the assumption that there was no initial search restriction. The inference process is described as follows:

For each node $X_i$ ($i = 1, 2, \ldots, 10$) apply the following process.

**Process 1** The GA algorithm (Section 2) is applied to each node $X_i$. The maximum number of “exons” in a chromosome is set to 5.

**Process 2** Calculate the reliability score $S_{i,j}$ for each node using the parameters of the solution candidates of which the fitness value is less than $\theta$ times of that of the best candidate (maximum five candidates).

**Process 3** If $\min(S_i) = S_{i,j} \leq \sigma$ ($0 \leq \sigma \leq 1$), the search restriction from the true relationship of $X_j$ over $X_i$ (positive, negative or none) is given. Besides, the expression time course data of gene $X_i$ when $X_j$ is disrupted is obtained, if possible. Both of them are assumed to be available here. Then go back to Process 1.

**Process 4** If $\sigma < \min(S_i)$, the regulatory relationship as the final solution of $X_j$ over $X_i$ is determined by using the values of $\max(V)$ (positive, negative or none).

When evaluating individuals in the GA operation, we need to calculate the expression pattern of a target node in accordance with the equation in (4).

Because data are given at discrete time points, the differential equation is approximated by the following difference approximation:

$$\frac{dx_i}{dt} \approx \frac{x_{i,(\text{calc})}(t + \Delta t) - x_{i,(\text{calc})}(t)}{\Delta t}.$$  (5)

When calculating the expression pattern of target node $X_i$, expression values of the other nodes can be used as shown below.
\[
\frac{x_{i}(\text{calc})(t + \Delta t) - x_{i}(\text{calc})(t)}{\Delta t} \simeq f_{i}(x_{1}(\exp)(t), \ldots, x_{i-1}(\exp)(t), x_{i}(\text{calc})(t), x_{i+1}(\exp)(t), \ldots, x_{n}(\exp)(t)),
\]

where \(x_{j}(\exp)(t)\), and \(x_{i}(\text{calc})(t)\) represent the observed expression value at \(t\), and the calculated expression value at \(t\), respectively. \(f_{i}\) represents the right-hand side of the equation in (4). This method enables us to infer each node’s regulators one by one, independently from the inferred structures of the other nodes.

4.2. Results

Table 2 shows the experimental results for the target network given in Fig. 11. The network shown in Fig. 12 was acquired by the proposed method, and the almost identical values of \(M[n, n]\) are obtained.

We used a Pentium 4 1.4 GHz PC with the memory of 512 MB.

To confirm the effectiveness of our approach, the performances of the following methods are compared:

(a) applying GA with a single population (all parameters are treated as “exon”);
(b) skipping Process 3 and applying Process 4 to the first generated solution candidates;
(c) giving the search restriction only in accordance with the requirement of Process 4;
(d) giving both the search restriction and the disruption data of randomly determined parts of the target gene, which is not necessarily consistent with the requirement of Process 4;
(e) our method proposed in this paper.

The averaged Sensitivity (SN) and specificity (SP) are defined by the following equations:

\[
SN = \frac{TP}{TP + FN},
\]

and

\[
SP = \frac{TN}{TN + FP},
\]

where TP, TN, FP and FN means true positive, true negative, false positive and false negative, respectively.

GA parameters are shown in Table 1.

In comparison with (a), the result of (b) shows that inferring method using population groups even with limited parameters effectively helps avoiding local minima. The result by the proposed method, in comparison with (d), shows
that adding the information in accordance with the requirements leads to inferring more accurate network structures with the small amount of data. In other words, our proposed method realizes the effective inference of the network structure.

The result of (c) implies that expression data of specific genes are needed for the accurate inference.

---

**Table 1**

Operation parameters

<table>
<thead>
<tr>
<th>Operation parameters</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td># of Population groups</td>
<td>5 (1 for (a))</td>
</tr>
<tr>
<td># of Generations of new groups</td>
<td>5</td>
</tr>
<tr>
<td># of Individuals in one pop.</td>
<td>300 (2000 for (a))</td>
</tr>
<tr>
<td># of Generations</td>
<td>60 (200 for (a))</td>
</tr>
<tr>
<td>Crossover rate</td>
<td>0.90</td>
</tr>
<tr>
<td>Mutation rate/elite rate</td>
<td>0.02/0.01</td>
</tr>
<tr>
<td>Search range</td>
<td>[-2.0, 2.0]</td>
</tr>
<tr>
<td>Threshold to be 0</td>
<td>[-0.05, 0.05]</td>
</tr>
<tr>
<td>θ/σ Values</td>
<td>3/0.80</td>
</tr>
</tbody>
</table>

**Table 2**

Experimental results for network in Fig. 11

<table>
<thead>
<tr>
<th>Approach</th>
<th>SN/SP</th>
<th>#RR</th>
<th>ST</th>
</tr>
</thead>
<tbody>
<tr>
<td>(a) Single population</td>
<td>0.600/0.769</td>
<td>–</td>
<td>74</td>
</tr>
<tr>
<td>(b) First candidates</td>
<td>0.700/0.837</td>
<td>–</td>
<td>28</td>
</tr>
<tr>
<td>(c) No. disruption data</td>
<td>0.783/0.844</td>
<td>24</td>
<td>97</td>
</tr>
<tr>
<td>(d) Random data</td>
<td>0.825/0.863</td>
<td>19</td>
<td>123</td>
</tr>
<tr>
<td>(e) Proposed method</td>
<td>0.925/0.944</td>
<td>12</td>
<td>71</td>
</tr>
</tbody>
</table>

RR: required restriction/data, ST: search time (min).

---

![Fig. 12. Inferred network by proposed method.](image-url)
Next, we have conducted experiments with noisy expression data. The noisy data were generated by the following equation with a random value $r$

$$e_{ix_i}(t) = \frac{x_i(t)}{C_0 R^6 (1 + r)};$$

(9)

where $R$ is the maximum noise rate.

Target networks were synthesized randomly at each trial.

Experimental results are shown in Table 3. The left part shows the SN/SP scores by the proposed method, whereas in the right part the number of restriction and data given to each node was limit to be 2.

The high rate of their sensitivities compared with their specificities implies that there is room for the improvement in our GA method (e.g., extending chromosome structures or operation parameters). However, more restrictions/data may be needed for the accurate inference from noisy data. Thus, the robust inference method will be required for the practical use.

### 5. Discussion

In adopting the reverse engineering approach as in the case of inferring a gene regulatory network, the major challenges are: (1) how we can save the search space to avoid the local minima and (2) how we can acquire the costly experimental expression data that are needed to infer the network in an effective way.

By using our proposed method, the user can select biologically proper solutions from solution candidates. Also the user can save the cost for acquiring the experimental data, because the system only requires the restriction/data needed to specify an ambiguous part of the network.

We adopted a GA-based method to realize this approach. However, making use of other methods (e.g., Genetic Programming, Neural Network, etc.) would avoid particular errors of the inference caused by using GA alone.

As a first step to the interactive inference approach, we showed experimental results using synthesized network and expression data. Network model was also simplified.
In these experiments, we added restrictions and expression data to only a part that the system required. If we can acquire the expression data of a gene under the condition that the other genes are disrupted or over-expressed, the same expression data of other genes may be available in this condition. Also, not only dynamic expression analysis but also static expression analysis can be helpful to give some restriction to the more accurate inference.

6. Conclusion

In this paper, we proposed an interactive approach to infer a gene regulatory network and showed the possibility of the proposed method in achieving an efficient network inference with a very little amount of the expression data.

The study will be continued with applying actual expression data with the consideration of other appropriate network models [9,12], expanding the functions of the inference system for more practical situation, and raising the performance of the inference process.

References