共分散選択とPageRankに基づく評価関数による遺伝子ネットワーク推定

安富祖 仁[†], 名嘉村 盛和[††], 岡崎 威生[††],

[†] 琉球大学大学院理工学研究科情報工学専攻　[††] 琉球大学工学部情報工学科

遺伝子間の影響を調べることができるDNAマイクロアレイデータから遺伝子ネットワークを推定するため、(1)DNAマイクロアレイデータから偏相関係数と逸脱度による共分散選択を用いて、直接的因果パスを抽出し、(2) 抽出された因果パスに対して、PageRankに基づいた評価関数を最大化する因果方向を探索する、という２つのステップからなる遺伝子ネットワーク推定法を提案した。また、因果方向探索に必要となる探索手法を比較した。

Genetic Network Estimation with Covariance Selection and Score Function based on PageRank

Hitoshi AFUSO[†], Morikazu NAKAMURA[††], and Takeo OKAZAKI[††]

[†] Graduate School of Engineering, University of the Ryukyus
[††] Faculty of Information Engineering, University of the Ryukyus

To estimate genetic network from DNA microarray data, we proposed the method constructed from two parts. First, extraction of direct causal path from DNA microarray data using partial correlation and covariance select with deviance. And second, search the orientation to extracted causal path such maximize score function based on PageRank. We also compared required search methods.

1 Introduction

As examples of genetic network estimation method, we can see the approach with boolean network[1)] , bayesian network, differential equasion system[2)] and petrinet[3)] . But these approaches have some difficulties. Afuso[4)] extracted direct causal paths among genes with covariance selection based on partial correlation and deviance, and proposed the path orientation algorithm such satisfies causal merge condition.

In this paper, We proposed a score function for orientations such shows fitness between orientations and given DNA microarray data[5)] . Also we implemented orientation search methods such maximize the score function and selected one such has highest orientation accuracy by experiments.

2 Outline of Proposal Genetic Network Estimation Method

Proposal estimation method is constructed from two parts. First, extraction of direct causal path from DNA microarray data, and search the path orientation such mixinize score function.

DNA microarray data[5)] contains information about influences among genes. The information has two kind, direct and indirect. It is difficult to distinguish them from DNA microarray data. To extract direct influence, we considered partial correlation that can be obtained from DNA microarray data as direct influence among genes and calculated it. Partial correlation between element i and j r_{xy} can be obtained by the formula below.

$$r_{xy} = \frac{-r^{ij}}{\sqrt{r^{ii}r^{jj}}} \qquad (1)$$

In this formula, r^{ij} denotes (i,j) element of inverse matrix $\mathbf{R^{-1}}$ of correlation matrix $\mathbf{R}$ from DNA microarray data.

Partial correlation values are calculated from DNA microarray data include unsignificant one. To remove unsignificant one, we executed covariance selection based on deviance. Deviance Dev denotes fitness between model and DNA microarray data and can be obtained by the formula below.

$$Dev = \log \frac{\left|\hat{\mathbf{R}}\right|}{|\mathbf{R}|} \qquad (2)$$

In covariance selection, we let the partial correlation value such has minimum absolute value be 0, and calculate deviance Dev. If the value of significance probability p of Dev is lower than the threshold, we consider the corresponding causal path as unsignificant and remove it. If not, we consider it as significant and leave it. we repeat this procedure until any causal paths can't be removed. By following above steps, we can extract direct and significant causal paths among genes.

3 Score Function for Orientation

To choose the orientation more fit to DNA microarray data, we need a score function for orientations.

Generally, it is difficult to observe true genetic network. So, direct comparison between affair network and true network is hard. By focusing to network characteristic such we can estimate from DNA microarray data, we proposed score function that reprisenting fitness between affair and given DNA microarray data.

In DNA microarray experiment, we control one gene and observe the change of expression level of other gene. This change of expression level is occured by influence of control of gene. We can say that gene such many influences come in from other genes has change of expression level in many DNA microarray experiments. This fact shows that reachability among genes has connectivity with DNA microarray data.

On the other hand, PageRank[6] is used in Web as network characteristics, and connected to reachability of user among Web pages.

By the analogy that DNA microarray data and PageRank is connected to reachability of each network, we assumed that PageRank can be estimated from DNA microarray data and defined the distance between estimated PageRank from DNA microarray data and calculated PageRank from affair network.

To estimate PageRank from DNA microarray data, we used mean of DNA microarray data among experiments.

$$\hat{pr}_i = \frac{\sum_j exp_{ij}}{N} \tag{3}$$

In this formula, $\hat{pr}_i$, exp_{ij}, and N denote estimated PageRank of gene i, observed value about gene i in experiment j, and the number of experiments. We used correlation between two PageRank values because PageRank is relative value. So, the score $Scr(\mathbf{A})$of affair $\mathbf{A}$ is obtaind by next formula.

$$Scr(\mathbf{A}) = Cor(epr, cpr) \tag{4}$$

In this formula, Cor is correlation function. epr and cpr coresspond to extimated PageRank from DNA microarray data and calculated PageRank from affair network.

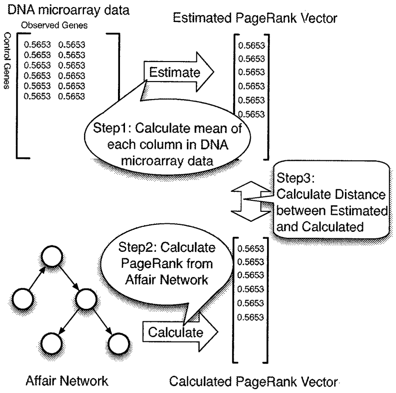

Figure 1: Score Calculation

By searching affair such has score near 1, we can get the network that is most fit to given DNA microarray data.

Orientations of extracted direct causal paths has another condition to satisfy. Causal path extraction with partial correlation has a property such if true network includes the part of causal merge structure, then we should observe pseudo causal path in the part. Causal merge structure is the structure that some variables influence to same variable.

As a result of the property, if the orientation to direct causal paths has some part of causal merge structure, then genes such construct causal merge structure should have a causal path each other. In other words, if two genes has no path to each other, they don't construct causal merge structure. We call this condition as causal merge condition(CMC), and we consider orientations such satisfy this condition as affair.

4 Search Method for Optimal Orientation

To find the orientation such has highest score, we need appropriate searching method. The orientation is formulated as conditioned combination optimization problem. As approach to combination optimization problem, we can see local search method and heuristics like genetic algorithm. In this paper, we proposed network neighborhood that is needed local search and chrosome expression of affair network to use genetic algorithm.

Because of that PageRank is relative value, we defined neighborhood among affair network by transformation such adjust the order of PageRank among genes. We show the transformation steps for affair network below. At first, compare order of estimated PageRank from DNA microarray data and calculated PageRank from affair network. Second, choose gene such has highest wrong order. Third, choose one path from coming ones to chosen gene. And finally, reverse the chosen path. we defined affair network such obtained by these steps as neighborhood.

This transformation may make affair network not to satisfy causal merge condition. So we need to more transformation method such adjust the network to satisfy the condition.

```
Procedure Network Adjustment
  Input : initial focused node $if$,
          initial direction node $id$
Begin
  Stack s($focused$, $direction$, $objective$);
  $f := if$;
  $d := id$;
  do
    lsing = checkCMC($f$, $d$);
    $o \in$ lsing;
    while(|lsing| > 0)
      s.push($d$, $f$, $o$);
      lsing := lsing - {$o$};
    end while
    $(d, f, o)$ := s.pop();
    ReverseEdge($f, o$);
    d := f;
    f := o;
  while(|s| != 0)
End
```

However, this network adjustment method may get loop and doesn't finish the adjustment. So, we need the method to determine adjustment loop.

```
Procedure Determination of Adjustment
Loop
  Input : initial state of orientation $iori$,
          initial focused node $if$,
          initial direction node $id$,
          current state of orientation $ori$,
          current focused node $f$,
          current direction node $d$
Begin
  if($if == f$ and $id == d$)
    $nd$ := CountDifferentPoints($iori, ori$);
    if($nd == 0$)
      Return(loop occured);
    else
      Return(loop doesn't occured);
    end if
  end if
End
```

We using these two method, local search can be done.

To search optimal by genetic algorithm, we need chrosome expression of affair network. we proposed two chrosome expression method. First, chrosome expression by path condition. Fig.2 shows the example of this chrosome expression.

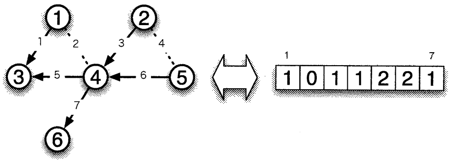

Figure 2: Chrosome Expression by Path Condition

In Fig.2, the number on each path denotes path label and array on right side is chorosome expression of network on left side. When we assume $i < j$, the value of locus in chrosome expression means as follows: 0 is cutted, 1 is oriented from gene i to j, and 2 is oriented from gene j to i. This chrosome expression has problem that it is difficult to consider causal merge condition. So we proposed chrosome expression by 3-cycle condition. 3-cycle means the cycle in affair network that has length 3. The value of each locus means label of cutted path. Because of that orientations have to satisfy causal merge condition, If cutted path is determined in 3-cycle, then the condition of other path in 3-cycle is also determined. But using this chrosome expression, it is occured that there are some paths such their condition is not determined. So we used local search method to search the orientation of such paths.

Other genetic algorithm conditions are shown next. As gene select method, we used roulette and 10% elite. As mixture method, we used uniform mixture. As mutation method, we used method that select one position randomly and change. As score function for each gene, we used $Score(A) + 1$. Using above chrosome expression, we can search the optimal orientation with genetic algorithm.

By applying of the causal merge condition strictly, we may get affair network such has small score value. So, we proposed the search methods that consider causal merge condition in the end of search. The search method that we proposed are summarized as follows: Local search method(LS1), local search method that apply causal merge condition in the end of search(LS2), search method with simple genetic algorithm(GA), composite search method with genetic algorithm and local search(GA+LS1), and composite search method that apply causal merge condition in the end of search(GA+LS2). Using these methods, we can search the optimal orientation.

5 Result of Comparison Experiment

To select the search method such has highest accuracy of orientation, we implemented some experiments.

We generated artificial data by two steps as follows. First, generate artificial genetic network randomly using Barabasi-Albert model[7] , and second, generate artifical data by following formula.

$$data_j = v_0 + \mathbf{A}v_0 + \cdots + \mathbf{A}^k v_0 \qquad (5)$$

In this formula, **A** dentes adjacency matrix of artificial genetic network and v_0 do initial expression vector.

To evaluate orientation accuracy for each method, we used four statistics; the value of score function(SCR), the number of incorrect orientation to correct causal path(DI), the number of orientation to pseudo causal path(DP), and the number of cutted edge incorrectly(CI).

Using these statistics, we implemented experiments to compare the score behavior of each methods. The experiment contains 50 trials that we observed SCR, DI, DP, and CI. For the methods with genetic algorithm, we assigned the number of generations to 500 and the number of individuals to 100. The experimental results for mean of four statistics and worst are showed in table.1 and table.2.

Table 1: Mean Value of Each Statistics

Method	SCR	DI	DP	CI
LS1	1.9230	0.1468	1.1426	0.0074
LS2	1.9293	0.1229	1.0934	0.0409
GA	1.9348	0.0465	0.9495	0.2688
GA+LS1	1.9502	0.0365	0.8654	0.0822
GA+LS2	1.9538	0.0239	0.7834	0.0988

Table 2: Worst Value of Each Statistics

Method	SCR	DI	DP	CI
LS1	1.7627	0.2549	1.5794	0.0277
LS2	1.8309	0.2258	1.5064	0.1290
GA	1.8814	0.0588	1.4054	0.4705
GA+LS1	1.9014	0.0530	1.2123	0.3558
GA+LS2	1.9253	0.0518	1.0540	0.4152

GA + LS2 method has highest accuracy in SCR, DI, and DP. The reason why CI value of LS1 method is highest is considered as LS1 method cut edge in the case only adjustment get loop. So, the total number of edge cut by LS1 method is smaller than others.

6 Conclusion

We proposed the score function for orientations to causal paths and search methods for the optimal orientation. And also, we implemented experiment to observe the orientation accuracy of each search method. The experimental result shows the GA+LS2 method has highest accuracy in five search methods.

References

1) T.Akutsu, S.Kuhara, O.Maruyama, S.Miyano, "Identification of gene regulatory networks by strategic gene disruptions and gene over-expressions", Proc 9th ACM-SIAM SODA, pp.695-702, 1998

2) E.V.Someren, L.Wessels, and M.Reinders, "Linear modeling of genetic networks from experimental data", ISMB, vol.18, pp.355-366, 2002

3) T.Hayashi, M.Nakamura, T.Okazaki, "A Petri Net Model of Gene Networks and Its Identification.", Proceedings of the Society Conference of IEICE, pp.162, 2002

4) H.Afuso, T.Okazaki, "Genetic Causal Network Estimation from Gene Expression Data", IPSJ SIG Technical Reports, BIO-6, pp.9-16, 2006

5) Brown, P.O. and D, Botstein, Exloring the new world of the genome with DNA microarrays, Nature Genetics, 21(Suppl. 1), pp.33-37, 1999

6) Brin S, Page L, "the Anatomy of a Large-Scale Hypertextual Web Search Engine", Computer Networks and ISDN Systems, vol30, pp.107-117, 1998

7) A.L. Barabasi, Z.N. Oltvai, "Network biology: understanng the cell's functional organization", Nat Rev Genet, vol5, pp.101-113, 2004