2005－BIO－3（7）
2005／12／22

遺伝子発現プロファイルからのオペロン構造予測のための Path Consistency Algorithm の改良

齊藤 秀[1,2], 油谷 幸代[1], 堀本 勝久[1]

1 東京大学医科学研究所
2 インフォコム株式会社

概要 本研究では，遺伝子発現プロファイルから共発現する遺伝子を予測するシンプルなアルゴリズムを提案する．我々は, 因果グラフを推定するための手法である Path consistency (PC) algorithm に対し，遺伝子発現プロファイルへの適用のための改良および, オペロンに関する生物学的知見を反映させた改良を加えた. 改良アルゴリズムを大腸菌のオペロン構成遺伝子の発現プロファイルに適用した結果，既知オペロンについて検出率 90%以上, 極めて低い擬陽性率を示した.

Algorithm for predicting co-expressed genes from expression profiles by improvement of path consistency algorithm with biological knowledge

Shigeru Saito[1,2], Sachiyo Aburatani[1], Katsuhisa Horimoto[1]

1 Institute of Medical Science, University of Tokyo
2 INFOCOM CORPORATION

Abstract We design a simple algorithm to predict the co-expressed genes from expression profiles. The path consistency (PC) algorithm for inferring the causal graph, is modified by considering the nature of actual expression profiles, and further improved by interpolating the biological knowledge of operons. The algorithm was applied to the expression profiles of known operons in *Escherichia coli*, and then more than 90% of known operons were correctly detected with small errors.

1 Introduction

Monitoring of expression profile is a standard workflow for investigating transcriptional mechanisms for a large number of genes by experimental approach. After preprocessing of the monitored profiles, the clustering is usually performed for identifying the gene function in combination with the sequence analyses. Furthermore, as a challenging issue, the various kinds of biological networks such as the transcriptional unit and the regulatory relationship between transcriptional factors and their regulating genes are directly inferred from gene expression profiles by several methods.

The network inference from profiles is based on the two types of graphs. One is the undirected independence graph (UIG) for the association inference, and the other is the directed independence graph for the causal inference. In the former case, the graphical Gaussian model (GGM) [1] is one of the familiar models, and in the latter case, there are some models such as Boolean network and Bayesian network [2]. These models were successfully applied to expression analyses for inferring the regulatory network [3, 4, 5]. In particular, the models in the latter case are based on directed acyclic graph (DAG), and are constructed by two approaches: score-based approach and constraint-based approach [6]. Although the computational complexity is huge to

obtain strict solutions in both approaches, the constraint-based approach can be operated relatively at high speed, especially in a sparse network [6].

In this paper, we design an algorithm to predict the operons in which multiple open reading frames are transcribed from the same promoter to a single mRNA transcript and therefore genes are largely transcribed at the same levels. For this purpose, we first modify the algorithm of UIG construction in PC algorithm, which is one of the constraint-based methods, by considering the nature of expression profiles, and then interpolate the biological knowledge into the modified algorithm. The designed algorithm is applied to expression profile data sets of known operon genes in *Eschelicia coli* to evaluate the detection performances.

2 Materials and Methods

2.1 Overview

First, the relationship between two genes is statistically tested by calculating a correlation coefficient between their expression profiles, and then the relationship is further tested by calculating partial correlation coefficients between more than two genes, which enumerate the degree of conditional independence between them. Since the straightforward calculation needs huge computational time, we adopt a part for UIG construction in path consistency (PC) algorithm to complete the calculation within a reasonable time. In the UIG construction in PC algorithm (UIGPC), two points are modified by considering the characteristic features of actual expression profiles. Furthermore, the modified algorithm is specified to predict the operons in consideration with the biological knowledge. The modified algorithm (MUIGPC) in the present study is applied to the expression profile data of *Eschelicia coli*. The details of the MUIGPC are as follows.

2.2 UIG construction in PC algorithm (UIGPC)

In PC algorithm for constructing directed independence graph, the first part of the algorithm is responsible for constructing the UIG [6]. We call this part UIGPC (Fig 1). In the operon prediction, the causality inference in PC algorithm is not needed, since all genes are simultaneously transcribed.

Let $Adj(G, X) \setminus \{Y\}$ be the set of vertices adjacent to X except Y in the UIG.

1: $G \leftarrow$ complete UIG
2: $n = 0$
3: **repeat**
4: **for all** X such that $|Adj(G, X)| - 1 \geq n$ **do**
5: **for all** $Y \in Adj(G, X)$ **do**
6: **for all** subset $S \subseteq Adj(G, X) \setminus \{Y\}$ such that $|S| = n$ **do**
7: **if** $X \perp\!\!\!\perp Y | S$ **then**
8: delete edge between X and Y in G
9: **end if**
10: **end for**
11: **end for**
12: **end for**
13: $n = n + 1$
14: **until** $|Adj(G, X)| - 1 \leq n$, $\forall X$
15: **return** G

Fig 1: Algorithm for constructing the UIG in the PC algorithm (UIGPC)

2.3 Statistical Test for Independence

In the UIGPC, the independence and the conditional independence between the variables are tested statistically as follows.

We calculate the Pearson's correlation coefficient r_{ij} between two profiles, p_i and p_j, when $n = 0$, as follows:

$$r_{ij} = \frac{\sum_{k=1}^{m}(p_{ik} - \bar{p}_i)(p_{jk} - \bar{p}_j)}{\sqrt{\sum_{k=1}^{m}(p_{ik} - \bar{p}_i)^2 \sum_{k=1}^{m}(p_{jk} - \bar{p}_j)^2}} \tag{1}$$

where $\bar{p}_{i(j)}$ is the average of $p_{i(j)}$.

To estimate the conditional independence between more than two variables when $n \geq 1$, the partial correlation coefficient $r_{ij \cdot rest}$ between i and j is calculated under assumption that the variables are distributed according to the multivariate normal distribution, as follows:

$$r_{ij \cdot rest} = \frac{r^{ij}}{\sqrt{r^{ii} \cdot r^{jj}}} \tag{2}$$

where r^{ij} is the i-j element of inverse correlation coefficient matrix. By the Fisher's Z transformation of both correlation coefficients, i.e.,

$$Z = \frac{1}{2} log \frac{1 + r_{ij}}{1 - r_{ij}} \quad or \quad Z = \frac{1}{2} log \frac{1 + r_{ij \cdot rest}}{1 - r_{ij \cdot rest}} \tag{3}$$

Z is distributed approximately according to the following normal distribution:

$$N\left(\frac{1}{2} log \frac{1 + r_{ij}}{1 - r_{ij}}, \frac{1}{m - 3}\right) \quad or \quad N\left(\frac{1}{2} log \frac{1 + r_{ij \cdot rest}}{1 - r_{ij \cdot rest}}, \frac{1}{\{m - (c - 2)\} - 3}\right) \tag{4}$$

where c is the number of variables. Thus, we can test statistically the observed correlation coefficients under the following null hypothesis with the significance probability $\alpha/2$:

$$H_0 : Z(r_{ij}(r_{ij \cdot rest}), m) = 0 \quad if \quad Z(r_{ij}(r_{ij \cdot rest}), m) < Z(r_{ij}(r_{ij \cdot rest}), m)_{\alpha/2} \tag{5}$$

2.4 Modification of UIGPC (MUIGPC) for Expression Profile Analysis

We modify the UIGPC to analyze the actual expression profiles in two points. In the actual expression profile data, the number of measured points is frequently much smaller than the number of genes. In this situation, PC algorithm returns frequently unreasonable results in that the edges are underestimated from the true graph [7]. In PC algorithm, the edge is deleted if the conditional independence is detected in any cases (see Fig 1). In contrast, to escape the underestimation, we modified PC algorithm using Steck's method in the case when $n = 1$, and when three variables form a complete graph [7]. Consequently, the following algorithm is added in the edge deletion: if and only if $X \perp\!\!\!\perp Y|Z$ (i.e., X and Y are independent given Z) and $X \top\!\!\!\top Z|Y$ (i.e., X and Z are dependent given Y) and $Y \top\!\!\!\top Z|X$ (i.e., Y and Z are dependent given X), let X and Y are conditional independent given Z [7].

Another modification is also due to the nature of actual profile data. The profiles with the large number of genes and the small number of measured points share naturally similar patterns of profiles. Thus, many genes (variables) are highly correlated, and even the partial correlation coefficient is not obtained frequently in numerical calculation. In this case, only if all variables except the violated variables are dependent, then the violated variables are regarded to bet dependent. For example, among the partial correlation coefficient

of gene i with genes j, k, and l, if the partial correlation coefficients between i and j and between i and l are judged to be dependent, and if the partial correlation coefficient between i and k is not calculated due to their similar profiles, then genes i and k are regarded to be dependent. In addition, the dependence between the variables is kept in the higher-order correlation without calculation. This modification is also useful for escaping the violation in calculation of the inverse correlation coefficient matrix for higher-order correlation due to the highly correlated variables.

2.5 Interpolation of Operon Knowledge into MUIGPC

To specify the MUIGPC for predicting the operons, we consider the following three kinds of biological knowledge into MUIGPC. The first biological knowledge is that the number of genes in one operon is limited from the fact that most operon is composed of two genes and the operon size is ranged within about twenty genes in known operons. The second is that all genes of each operon are coded in the same strand. The third is that all genes of one operon is adjacent each other. Fortunately, the above biological knowledge can be easily interpolated into MUIGPC as the prior knowledge.

The first knowledge is realized by replacing the observed values with zero values in the elements which are positioned at more than userdefined elements, which is determined from the maximum size of operons, far from the diagonal in the correlation coefficient matrix. In the algorithm, the replaced elements keep zero values in proceeding operation; no edges are established. The second knowledge is realized by separately operating the algorithm for two data of genes in respective strands and by finally conjecturing the respective outputs. The third knowledge is realized by simply excluding the disordered genes from the outputs. Note that the interpolation of the three kinds of knowledge into the algorithm is ordered. The first two modifications are set in the initial step for calculation, while the last modification is set in the final step.

2.6 Definition of Operon Prediction Accuracy

We define the five values to evaluate the accuracy of operon prediction. The five values are as follows: NOE, the total number of established edges in the graph; GDR, the ratio of the total number of correctly detected genes to the total number of all operon genes; ODR, the ratio of the number of correctly detected operons to the number of all operons; OCR, the average ratio of the number of correctly detected genes within each operon to the number of all genes within each operon; NFC, the average number of genes falsely connected with the operon genes.

2.7 Gene Expression and Operon Data

The gene expression profile data analyzed in the present study is listed in Table 1. We compile the expression profiles measured for all genes in *Escherichia coli* from eight experiments, and totally, the number of genes and measured points are 4289 and 178 in the present data set.

The information of operons is cited from Ecocyc [16], and the operons that are composed of more than two genes whose profiles exist in the data set are selected. The total numbers of operons thus selected and of genes in the operons are 377 and 1163, and the frequency of gene numbers in each operon is listed in Table 2. As seen in the table, about half of operons are composed of two genes, and the maximum number of genes in an operon is 15.

Table 1: Expression profile data sets used in this work

Condition	Number of measured points	Reference
M9+glucose and LB media	16	[8]
Degradosome mutants	78	[9]
UV irradiation	15	[10]
Tryptophan regulation	27	[11]
DNA gyrase and topoIV regulation	21	[12]
RraA regulation	7	[13]
Rnag regulation	10	[14]
Adaptation to famine	4	[15]

Table 2: Gene frequency in operons of the present data set

No. of genes in operon	2	3	4	5	6	7	8	9	10	11	12	13	14	15
No. of operons	169	81	50	35	13	8	6	5	3	2	1	1	0	3

3 Results and Discussion

3.1 Selection by UIGPC

We counted the number of edges established in different correlation orders by UIGPC, before evaluating the detection accuracy by UIGPC and MUIGPC. In this analysis, the significance probability for independence test was set to 10^{-4}. Fig 2 shows the plot of the numbers of established edges by the UIGPC, against the correlation orders. Remarkably, the number of edges decreased drastically from the zeroth to the first-order correlations (from 550,262 to 104,752 edges), while the edge number decreased slightly from all connected case to the zeroth-order correlation (from 675,703 (=1163·1162/2) to 550,262 edges). Thus, the calculation based on the conditional independence shows possibility to discriminate powerfully true correlations from false ones.

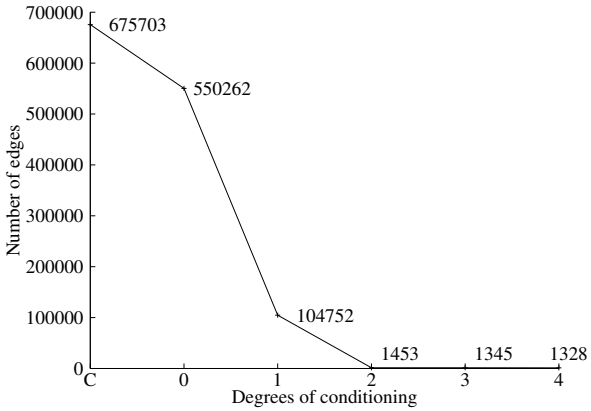

Fig 2: Gene selection by UIGPC. The horizontal axis is the order of correlation, and the vertical axis is the number of the edges established by the UIGPC in the corresponding order. The graph construction naturally stopped in the fourth-order correlation

The second remarkable feature is that the numbers of edges in more than the second-order correlation

were almost the same values. From Table 2, the maximum number of edges in operons is calculated to be 2477, when all genes in each operon are connected (complete graph), and the minimum number of edges is to 928, when they are connected linearly (chain graph). Interestingly, the number of detected edges was ranged between the above minimum and maximum numbers.

The third feature is that the algorithm stopped naturally in operating by the fourth-order correlation. In other words, the conditional independence between more than seven genes (the fifth-order correlation) was not needed in the present data set. This promises that the present algorithm is applicable to operon prediction within a reasonable computational time. At any rate, the UIGPC excludes effectively the noises included in actual profile data.

3.2 Operon Detection Accuracy

We evaluated the accuracy for detecting the operons by the UIGPC and MUIGPC, in terms of the five values defined in the preceding subsection.

Table 3 shows the detection accuracy by MUIGPC in comparison with that by UIGPC. As for the number of detected edges (NOE), the number by MUIGPC was slightly smaller than the actual number in known operons (928), while the number by UIGPC was much larger than the actual case. One of the striking features of our algorithm is high accuracy rate for detecting the operons. Indeed, the three values about the detection accuracy, GDR, ODR, and OCR, showed more than 90% in the case by MUIGPC, while the three values were far less than 90% in the case by UIGPC. Furthermore, the degree of false edge was only 0.36 in MUIGPC, in contrast to 5.46 in UIGPC. Thus, the present modifications improve successfully the operon detection performance.

Table 3: Comparison of detection accuracy between UIGPC and MUIGPC

Algorithm	NOE	GDR	ODR	OCR	NFC
UIGPC	1328	0.54	0.62	0.80	5.46
MUIGPC	826	0.92	0.94	0.97	0.36

To further evaluate the performance of the present algorithm, we investigated the detection accuracy for each operon. Fig 3 shows the relationship the fraction of number of accurately detected genes to total number of genes, the number of falsely detected genes for each operon, and their frequencies. As expected from $OCRs$ and $NFCs$ in Table 3, MUIGPC detected accurate operon structures, while UIGPC detected some false genes. Indeed, most operons by MUIGPC showed 100% accuracy and no false genes. In contrast, although many operons by UIGPC showed 100% accuracy, several false genes were included. Thus, the present modification of UIGPC operates successfully to predict the operons.

3.3 Comparison with Previous Related Methods

The detection of co-expressed genes is one of the issues that show new advances by simultaneously monitoring the expressions of many genes. Indeed, some methods have been proposed to detect the co-expressed genes [17]. In most cases, the similarity between expressions in two genes is measured by only the expression profiles of two genes such as correlation coefficient, by neglecting the effect of other genes. Thus, these methods need frequently some heuristic conditions to exclude the large noise also shown in this study. In contrast to the correlation coefficient, the partial correlation coefficient includes the information on other genes except the two corresponding genes. Thus, MUIGPC by using the partial correlation coefficient can be designed to detect the operon effectively.

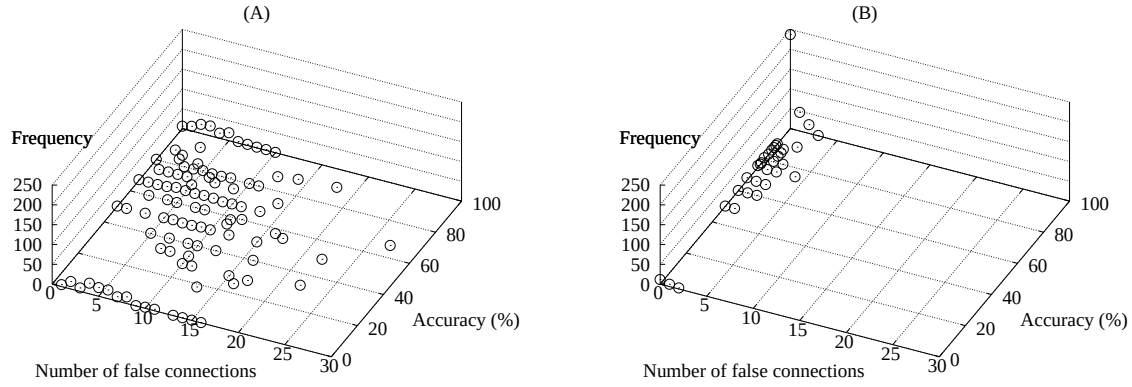

Fig 3: Relationship between detection accuracy and numbers of false edges for each operon. The two horizontal axes are the number of false edges in each operon and the detection accuracy that is calculated by dividing the number of the correctly detected genes by that of known genes for each operon. The vertical axis indicates the number of the operons in the corresponding points. A: UIGPC, B: MUIGPC

The partial correlation coefficient is utilized to detect the co-expressed genes in a few studies. The method based on the straightforward calculation of partial correlation coefficients faces frequently difficulty in the computational performance [18]. In contrast, MUIGPC stops automatically from the nature of profile data, while the previous method artificially stops in the predetermined correlation order.

The GGM is one of well-known methods to construct the UIG, and is also based on the conditional independence estimated by the partial correlation coefficients [1]. The GGM was successfully applied to estimate a framework of regulatory networks, in combination with the clustering, from a genome-wide expression profiles [19, 5]. The method infer the network between gene clusters (gene systems network), but can not infer the network between the genes (genetic network). The present MUIGPC is useful to investigate the details of the gene systems network by GGM.

3.4 Concluding Remarks

We designed a simple algorithm to predict the co-expressed genes from the expression profiles. The algorithm is modified from the PC algorithm by considering the nature of actual profile data and the biological knowledge of operons. Our algorithm, MUIGPC detected the known operons in *E. coli* from their profiles with high accuracy (about 90%) and with small error (about 0.5 genes per operon). Note that the present algorithm is naturally extended to infer the regulatory network by considering the latter part of PC algorithm. This study will appear in near future.

reference

[1] Whittaker, J. (1990) Graphical Models in Applied Multivariate Statistics, Wiley, New York.

[2] Pearl, J. (1988) Probabilistic reasoning in intelligent systems : Networks of plausible inference, Morgan Kaufmann Publishers, Palo Alto, CA.

[3] Akutsu, T., Miyano, S., and Kuhara, S. (2000) Algorithms for identifying Boolean networks and related biological networks based on matrix multiplication and fingerprint function. *J Comput Biol,* **7**(3-4), 331–343.

[4] Friedman, N., Linial, M., Nachman, I., and Pe'er, D. (2000) Using Bayesian networks to analyze expression data. *J Comput Biol,* **7**(3-4), 601–620.

[5] Toh, H. and Horimoto, K. (2002) Inference of a genetic network by a combined approach of cluster analysis and graphical Gaussian modeling. *Bioinformatics,* **18**(2), 287–297.

[6] Spirtes, P., Glymour, C., and Scheines, R. (2001) Causation, Prediction, and Search (Springer Lecture Notes in Statistics, 2nd edi-tion, revised), MIT Press, Cambridge, MA.

[7] Steck, H. and Tresp, V. (1996) In *In Proceedings of the 2. Workshop on Data Mining und Data Warehousing als Grundlage moderner entscheidungsunterstützender Systeme* University of Magdeburg,Germany: pp. 145–154.

[8] Bernstein, J. A., Khodursky, A. B., Lin, P.-H., Lin-Chao, S., and Cohen, S. N. (2002) Global analysis of mRNA decay and abundance in Escherichia coli at single-gene resolution using two-color fluorescent DNA microarrays. *Proc Natl Acad Sci U S A,* **99**(15), 9697–9702.

[9] Bernstein, J. A., Lin, P.-H., Cohen, S. N., and Lin-Chao, S. (2004) Global analysis of Escherichia coli RNA degradosome function using DNA microarrays. *Proc Natl Acad Sci U S A,* **101**(9), 2758–2763.

[10] Courcelle, J., Khodursky, A., Peter, B., Brown, P. O., and Hanawalt, P. C. (2001) Comparative gene expression profiles following UV exposure in wild-type and SOS-deficient Escherichia coli. *Genetics,* **158**(1), 41–64.

[11] Khodursky, A. B., Peter, B. J., Cozzarelli, N. R., Botstein, D., Brown, P. O., and Yanofsky, C. (2000) DNA microarray analysis of gene expression in response to physiological and genetic changes that affect tryptophan metabolism in Escherichia coli. *Proc Natl Acad Sci U S A,* **97**(22), 12170–12175.

[12] Khodursky, A. B., Peter, B. J., Schmid, M. B., DeRisi, J., Botstein, D., Brown, P. O., and Cozzarelli, N. R. (2000) Analysis of topoisomerase function in bacterial replication fork movement: use of DNA microarrays. *Proc Natl Acad Sci U S A,* **97**(17), 9419–9424.

[13] Lee, K., Zhan, X., Gao, J., Qiu, J., Feng, Y., Meganathan, R., Cohen, S. N., and Georgiou, G. (2003) RraA. a protein inhibitor of RNase E activity that globally modulates RNA abundance in E. coli. *Cell,* **114**(5), 623–634.

[14] Lee, K., Bernstein, J. A., and Cohen, S. N. (2002) RNase G complementation of rne null mutation identifies functional interrelationships with RNase E in Escherichia coli. *Mol Microbiol,* **43**(6), 1445–1456.

[15] Tani, T. H., Khodursky, A., Blumenthal, R. M., Brown, P. O., and Matthews, R. G. (2002) Adaptation to famine: a family of stationary-phase genes revealed by microarray analysis. *Proc Natl Acad Sci U S A,* **99**(21), 13471–13476.

[16] Keseler, I. M., Collado-Vides, J., Gama-Castro, S., Ingraham, J., Paley, S., Paulsen, I. T., Peralta-Gil, M., and Karp, P. D. (2005) EcoCyc: a comprehensive database resource for Escherichia coli. *Nucleic Acids Res,* **33**, D334–D337.

[17] Sabatti, C., Rohlin, L., Oh, M.-K., and Liao, J. C. (2002) Co-expression pattern from DNA microarray experiments as a tool for operon prediction. *Nucleic Acids Res,* **30**(13), 2886–2893.

[18] de laFuente, A., Bing, N., Hoeschele, I., and Mendes, P. (2004) Discovery of meaningful associations in genomic data using partial correlation coefficients. *Bioinformatics,* **20**(18), 3565–3574.

[19] Aburatani, S., Goto, K., Saito, S., Toh, H., and Horimoto, K. (2005) ASIAN: a web server for inferring a regulatory network framework from gene expression profiles. *Nucleic Acids Res,* **33**, W659–W664.