社団法人 情報処理学会 研究報告
IPSJ SIG Technical Report

2006－BIO－5 (20)
2006／6／16

時系列マイクロアレイ実験デザインの発現解析に及ぼす影響

伊藤遼佑, 高橋弘喜, 大島拓, 小笠原直毅, Md.Altaf-Ul-Amin, 金谷重彦, 黒川顕
奈良先端科学技術大学院大学

本研究の目的は、連続的で信頼ある時系列マイクロアレイデータを得るための手法を提案することである。そのために、二種類の異なるデザインに基づくマイクロアレイデータを用意した。一つ目は、定点を固定するデザインで、二つ目は、隣同士を比較するデザインである。それにより、双方の利点を補完しあうことで、より精度の高いタイムコースデータを得ることが可能と考えられる。そこで、今回は、その足がかりとして、これら二つのデザインが実際にどのように異なるかを順位に基づく手法で示すことを試みた。その結果、二つのデザイン間で異なる発現変化を表すデータが多数存在することが示された。

Experimental design for time-series microarray analysis

Ryosuke Ito, Hiroki Takahash, Taku Oshima, Naotake Ogasawara, Md.Altaf-Ul-Amin,
Shigehiko Kanaya, Ken Kurokawa
Nara Institute of Science and Technology

The goal of our study is to propose a reliable method for consecutive time series data obtained from microarray experiments. For this purpose, we prepared the time series data based on different experimental designs. In first design, the control experiment is set to one representative point. Second design sets the control experiment to closest time point. It is considerd that these designs compensate the weak points of each other. In this paper, we discuss the difference between two designs with respect to rank of expression ratio. The results show that there are many contradictions between two designs.

1. INTRODUCTION

Microarrays can view the gene expression changes as snapshot on a large scale. Transcriptional analysis of bacterial system is carried out for examining gene expression response to environmental changes [1] and time-series responses to stress [2].

Normalization is performed to eliminate the system variation and bias caused by some artifact such as intensity measurement properties or cross hybridization. In cDNA microarray data, normalization is performed to detect genes with different expression levels in the case when expression levels are equivalent for most of the genes. This situation is assumed in the MA-plots developed by Dudoit [3], which is generally used for normalization in cDNA microarray (See section 2.4).

Generally, the expression levels are normalized between a representative point ($t = p$) and individual time points (t) in case that the differences of gene expression levels are examined for time series data consisting of T time points (t=1,2,...,T) called F-design (Control Fixed design). Second design is not commonly used. It sets the control experiment to closest time points, so the hybridization pair moves forward along with time proceeding one by one. We call this second design as N-design (Neighborhood design).

In F-design, expression change along with time is indirectly pursued by setting the control to one time point, so there is no confirmation that the data from this design could actually reflect true expression changes. Besides, the snapshot of expression given as control experiment is fixed, so the data from this design would produce the wrong outcome, especially at normalization process, if the expression of many genes is already

shifted. While in N-design, the control is altered one after another, so the normalization could be carried out on essential premise that almost all the genes do not change their expression state.

In the present study, we examine validation of time-series microarray data by comparing the expression differences between N-design and F-design using a rank-based method. The rest of the paper is organized as follows. Section 2 describes the experimental and methodological procedure. Section 3 explains the method we present in this paper. In Section 4, we discuss about the result of microarray experiments and the difference between two designs. Finally, Section 5 discusses our findings and futures.

2. EXPERIMENTAL PROCEDURE

2.1 Bacterial strains, medium, growth conditions and RNA extraction

The wile-type *E.coli* strain cells were grown in 2l Luria-Bertani(LB) medium, pH 7.4, in a 3l jar-fermenter. Culture was started by 250-fold. The temperature was maintained at 37℃ and the agitation speed was kept at 300 rpm. Growth was monitored by measuring the optical density (OD) at 600nm. Cells were harvested by centrifugation at 5,000 rpm after adding the RNA protect Bacteria Reagent (Qiagen), then stored at -80℃. **Table 1** shows harvesting time. Two independent samplings were performed. RNA was isolated using the Qiagen RNAprotect, RNeasy Mini Kit (Qiagen) and RNase-free DNase sets according to the manufacturer's instructions and stored at -80℃. Genomic contamination was estimated by gel electrophoresis.

Table 1: The culture time and sampling amount

Time(min)	Amount(ml)*	ID*	Time(min)	Amount(ml)*	ID*
60	90	1	240	8	-
90	75	-	250	8	8
100	50	2	300	8	-
120	50	3	360	5	-
130	40	*p*	370	5	-
135	15	4	420	5	9
140	15	-	480	5	10
150	10	5	490	5	-
160	10	-	600	5	-
170	10	6	610	5	-
180	8	-	720	5	23
190	8	7			

*Amount means sampling volume at corresponding time point and ID means that this sample was used for microarray experiments and that it is a serial index.

2.2 Labeling

For each labeling reaction, 15μg total RNA was used. First-strand cDNA synthesis was primed with 1.2 μg random primer (Invitrogen) in Nuclease free water (total volume is 31μl) by heating at 70℃ for 10 min and incubating at 25℃ for an additional 10 min. Reverse transcription were performed by SuperScript III (Invitrogen) in reverse-transcription buffer[1×1stStrand buffer, 10mM DTT] the presence of 5 mM dATP, 5 mM dUTP, 5mM dCTP, 0.25 mM dTTP and 0.25 mM AA-dUTP. Amino-allyl-labelled nucleotides were incorporated into the cDNA. This reactions were incubated at 25℃ for 10 min, 37℃ for 60 min ,42℃ over night, and quenched by heating at 70℃ for 10 min. RNA template was hydrolyzed by the addition of 20μl of 1N NaOH followed by heating at 65℃ for 30 min. Reactions were neutralized with 20μl of 1N HCl. cDNA was purified using CyScribe GFX Purification Kit (GE Healthcare) according to the manufacturer's directions. NHS ester of Cy3 and Cy5 dye was added to cDNA solution and incubated for 4 hours. Coupling reactions were quenched by the addition of 15μl of 4M hydroxylamine and incubated at room temperature

for 15 min in dark. Labeled cDNA was purified using CyScribe GFX Purification Kit again.

2.3 Hybridization and Spot detection

Prehybridization of the array slides was performed for 3h in filtered prehybridization solution [25% formamide, 5x SSC, 10mg l-1 BSA (fraction V), 0.1% SDS] at 42℃. Slides were briefly washed in miliQ water and 80% ethanol and dried by centrifugation at 1000g for 5 min. Hybridization of the probe was performed using hybridization solution (25% formamide, 5x SSC, 0.1% SDS, 0.1μg poly(A) ml-1, 1x Denhardt's solution and 100pmol Cy3 and Cy5 combined probe). The hybridization solution containing the Cy-Dye-labelled cDNA was heated to 95℃ for 3 min, Hybridization is performed by advalytix hybridaization macine (ArrayBooster) at 42℃ for 16 hours. After hybridization, The slides were washed and dried by centrifugation at 1000g for 5min, and then analyzed using an Fuji FLA-8000 scanner and Array gauge ver.2.0 software (Fuji film)

2.4 Microarray data normalization

Normalization is performed to eliminate the system variation and bias caused by some artifact such as intensity measurement properties or cross hybridization. The method of normalization of cDNA microarray data using MA-plots is developed by Dudoit [3]. MA-plots can reveal the spot artifacts globally and show the intensity-dependent logarithmic ratio of raw microarray data. In MA-plots, we calculate two parameters, average of logarithmic transferred intensity $A_s = \{\log(T_s) + \log(R_s)\}/2$ and logarithmic ratio of intensity $M_s = \log(T_s / R_s)$ Here, T_s and R_s are the intensity of target and control experiments for sth spot, respectively.

By plotting values of A_i on the abscissa and M_i on the ordinate of a coordinate system, it is possible to evaluate the bias error with respects to the average logarithmic intensities. Normalized log ratio $\tilde{M}_s$ is estimated as the difference between M_s and baseline $\hat{M}_s$. Here, using a relation between M_s and A_s, ($M_s = f(A_s) + \varepsilon_s, \varepsilon_s$ is the difference between M_s and $f(A_s)$ for gene s) by MA plot; the baseline for sth spot is estimated by $\hat{M}_s = f(A_s)$. The genes whose signal intensity is regarded as zero are eliminated in the present analysis. With this methodology, it is assumed that there is no large error due to expression intensity in the majority of the spots and that expression change does not occur on the majority of the spots.

2.5 Filtering based on intensity distribution

In general, a log transformed ratio of target and control intensities, while at least one with very low intensity level, is unreliable in terms of expression change estimation [4]. So, the robust method to find the optimal signal intensity thresholds to identify and to eliminate the unreliable low signal intensities has been developed by Gao et al. [6], which can distinguish truly expressed genes and the noise derived from background perturbation or cross hybridization (AD Method; Accumulated Distribution Based Threshold Method).

AD Method is based on the plots of cumulative distribution obtained from log transformed intensity as shown in **Fig.1**. Here, x-axis represents the logarithm value of signal intensity with the base of 2 and y-axis represents the ratio of the cumulative number of spots to the total number. One common feature of the cumulative distribution is that the shape of the curve is fairly stable in spite of variation of spot number of an array and regardless of image analysis software [6]. It can be seen that plots are approximately divided into three parts, the top part parallel to X-axis which represents the high intensity region, the central part, almost linear, and the bottom part of the low intensity spots parallel to X-axis. The bottom and the central parts of cumulative distribution function curve are picked out for linear fit. The intensity threshold value is determined by intersection point of two fitted line. In the present study, the fitting line of the low intensity part of cumulative distribution curve is approximately the equation $y = 0$. The central fitting line is estimated by the least square method in the interval between 0.2 and 0.8 of the Y-axis. Then, the threshold of intensity

is determined by the x-value in the linear equation by putting y=0.

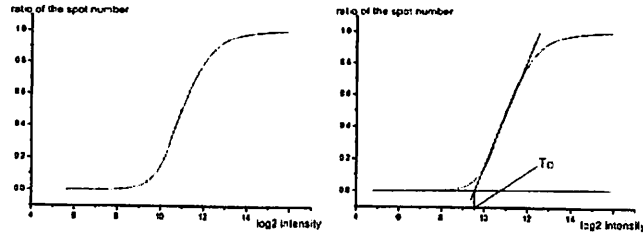

Fig.1. Accumulated distribution curve of log 2 intensity and Determination of T_D(Gao.X et al. 2003)

2.6 Averaging over duplicate spots

Each microarray used in this experiment has two or more replicated spots for each gene. The normalized logarithm ratio of intensities is caluculated for the spots whose intensities in target and reference conditions are larger than the threshold T_D. Then, those of the replicated spots are averaged. The genes that completely survive the filtering are only used in the following analysis.

3. METHODS

3.1 Experimental design and time series data

Consider the time-series experiment includes T time points denoted by t = 1, 2, ..., T. In N-design, the expression intensities are measured for the neighboring time points for each hybridization. N_{gt} denotes the expression intensity of gth gene (g=1, 2, ..., G) at time point t (t=2, 3, ..., T). In hybridization experiments the intensity for tth time point is measured between time points t-1 and t. In F-design, one time point is selected as reference which is referred to as p. Here, real expression levels from two of neighboring time points are denoted by $I_{real}(t\text{-}1)$ and $I_{real}(t)$, and one reference time point is denoted by $I_{real}(p)$.

Relation between N- and F-designs is represented by Eq. (1).

$$N_{real}(t) = F_{real}(t) - F_{real}(t-1) \tag{1}$$

where

$$N_{real}(t) = \log(I_{real}(t)/I_{real}(t-1)) \tag{2}$$

$$F_{real}(t) = \log(I_{real}(t)/I_{real}(p)) \tag{3}$$

$$F_{real}(t-1) = \log(I_{real}(t-1)/I_{real}(p)) \tag{4}$$

Practically, Eq.(1) does not hold because of an error caused by the handling of microarray experiments and operation of normalization (Eq.5).

$$N_{gt} \neq F_{gt} - F_{gt-1} \tag{5}$$

It is important to assess the difference between the left and right sides of equation 5 in time series data

because regarding cDNA microarrays, normalization of log-ratio based on MA-plot is the parameter with the highest reproducibility. In the present study, we examine differences between N- and F-designs based on ranking of gene expression levels.

3.2 Evaluation method based on gene expression ranking

The dyes generally used, Cy3 and Cy5, have some un-preferable properties, that is, these dyes are relatively unstable and may differentially influence the incorporation efficiencies during labeling, have different quantum efficiencies and are detected by the scanner with different efficiencies. The order of log-ratio at a time t is fundamentally important for comparing the experimental designs, N- and F-designs. When the arrangement of genes from the largest to the smallest in log-ratio of N_{gt} is identical to that of genes in F_{gt}- F_{gt-1}, the difference of expression profiles at time t-1 and t indirectly estimated for F-design can be regarded as identical to that for N-design, so we can estimate the difference of expression profile at time t-1 and t by indirect measurements of cDNA microarray. In the present study, we propose the validation method for the difference of expression profiles between N- and F-designs in time series data.

Initially, the order of genes in the log-ratio (N_{gt}) from the largest to smallest is denoted by rank $r_N(g)$, and that in the log-ratio (F_{gt}- F_{gt-1}) is denoted by rank $r_F(g)$, where g is an index for individual genes.

For a set of genes (S(g)) whose rank satisfy the range $r_N(g) - A/2 \le r_N(g') < r_N(g) + A/2$, local average of rank difference is calculated by Eq.(6).

$$R(g) = \frac{\sum_{g' \in S(g)} \{r_N(g') - r_F(g')\}}{A} \qquad (6)$$

Here, A represents the number of genes included in Set S(g).

Difference in log-ratio of indirect estimation (F-design) to direct estimation (N-design) in the degree of log-ratio can be easily comprehended by the relation between R(g) and N_{gt}.

4. RESULTS AND DISCUSSION

4.1 Cell cultivation results

As shown in **Fig. 2**, two series of growth profiles were similar, so expression profile assumed to be the same. Therefore we performed the microarray experiments using these two series.

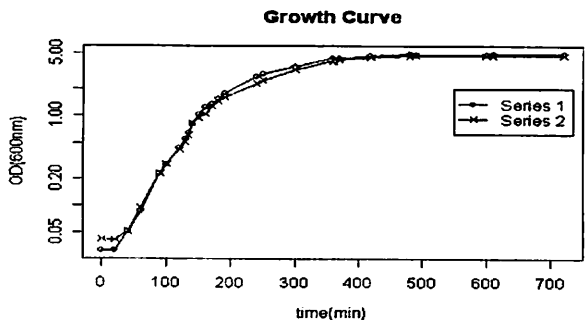

Fig. 2 Growth curve in the two experiments

4.2 Microarray data and Thresholds of intensity

Table 2 shows the microarray hybridization pairs (Name), the intensity threshold values of each fluorescent (Threshold of Cy3 and Cy5) and the number of survived data from filtering (Reliable). In

F-design, the control sample was always set by time point p. The threshold of Cy3 intensity was of the range from 2 to 6 approximately in series 1 and from 2 to 18 in series 2. While the threshold of Cy5 intensity was of the range from 1 to 6 in series 1 and from 2 to 14 in series 2. Microarray data which have rather large thresholds, for example, 7-8 s2, 8-9 s2, 9-10 s2 and 10-11 s2, may contain a lot of high intensity spots compared to other data. The spots actually detected the expression level and assured reproducibility may be marked as unreliable spots, especially at the neighborhood of thresholds as shown by the circle in **Fig. 3**. This information whether or not expression actually occurs needs to be examined in detail and only reliable data should be used in the following time series analysis and methods.

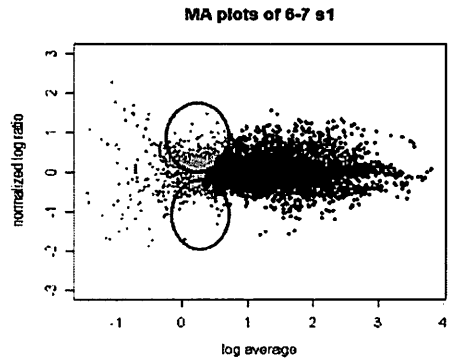

Fig. 3 MA plots of normalized data

Table 2 cDNA microarray experiments for two time-series data (Two time points for cDNA microarray experiments are described in Name column, e.g. 1-2 s1, suggesting A time point corresponds to the ID in **Fig. 1**. The number of genes selected by AD method is represented in Reliable column.)

	Series 1				Series 2			
	Name	Threshold of Cy3	Threshold of Cy5	Reliable	Name	Threshold of Cy3	Threshold of Cy5	Reliable
N-design	1-2 s1	4.75	3.26	3254	1-2 s2	2.33	3.05	3239
	2-3 s1	4.04	4.71	3603	2-3 s2	2.42	2.49	3187
	3-4 s1	4.81	2.97	3629	3-4 s2	4.24	3.08	3405
	4-5 s1	5.89	6.46	3478	4-5 s2	4.20	3.39	3557
	5-6 s1	6.17	4.09	3489	5-6 s2	5.98	4.79	3589
	6-7 s1	2.28	1.74	3404	6-7 s2	5.02	3.14	3619
	7-8 s1	5.47	4.56	3630	7-8 s2	10.83	11.24	3684
	8-9 s1	5.10	2.51	3595	8-9 s2	9.48	13.71	3607
	9-10 s1	3.96	1.51	3608	9-10 s2	11.44	6.96	3692
	10-11 s1	2.35	0.63	3538	10-11 s2	17.59	6.97	3730
F-design	p-1 s1	4.72	2.88	3554	p-1 s2	4.44	3.65	3565
	p-2 s1	2.96	1.41	3515	p-2 s2	7.55	4.31	3663
	p-3 s1	5.37	3.70	3636	p-3 s2	6.63	2.21	3624
	p-4 s1	4.00	1.57	3621	p-4 s2	5.71	4.15	3662
	p-5 s1	3.86	3.46	3498	p-5 s2	3.91	2.31	3581
	p-6 s1	4.57	2.58	3552	p-6 s2	6.66	5.73	3607
	p-7 s1	3.68	2.25	3567	p-7 s2	8.14	3.82	3618
	p-8 s1	2.88	2.39	3531	p-8 s2	7.46	3.05	3591
	p-9 s1	3.14	1.36	3544	p-9 s2	6.50	3.52	3560
	p-10 s1	3.01	0.72	3512	p-10 s2	7.19	1.73	3598
	p-11 s1	3.18	0.38	3508	p-11 s2	6.38	1.29	3569

4.3 Local average of rank

The results of evaluating the difference between two designs is shown in **Fig. 4**. If there is no difference between the order of expression changes obtained from two designs, the value of local average of rank becomes zero, but as a whole the average of rank difference is about 700~1000 around the center of x-axis, i.e. concerning the expression changes around zero value of the X-axis, suggesting that there is always some difference between two designs. These difference may be caused by random noise happened when Cy3 and Cy5 intensities are transformed to the logarithmic ratio. But especially at figure 1-2, 4-5, 5-6, 7-8 and 9-10, shape of rank difference has valley-like features around the center of X-axis. These mean that there is larger gap between information about how the genes expressed than the gap from random noise, because it is assumed that the value of local average of rank becomes certain amount if ranking alternation randomly occurs. For example, it may suggest that one design insists that the gene is up regulated, while another indicates that the expression change is stable. This contradiction would produce wrong interpretation on the time-series analysis. If these mistakes happen to both sides of X-axis in **Fig. 4**, in the region where expression changes occur, it would much influence the result. In addition, at 9-10 and 10-11, the rank differences tend to gradually increase from zero to outward region of the X-axis on both sides. It can be suggested that when expressions between two states is widely shifted, such as between log-phase and stationary-phase, analysis using F-design is opt to make mistakes with respect to expression changes.

5. CONCLUSION

In this paper we discuss the difference between results of analysis of microarray data using two different designs, N-design and F-design. As described at section 4.3, it is suggested that the features of expression change determined by different experimental designs are not quite the same in our method, called local average of rank. But this method can not exactly detect the genes that have contradictory aspects from the result of two designs, so we should examine the contradiction of the two designs in more detail. The problem we have to address more is the existence of not-expressed sets on control of F-design, because these genes cannot be detected due to their small intensity. But using N-design, it is detectable even if the genes express after control point of F-design. We plan to combine merits of these two designs and develop a more reliable method for constitutive time series data.

REFERENCES

1. Dharmadi Y, Gonzalez R: DNA Microarrays: Experimental issues, data analysis, and application to bacterial system. *Biotechnol* 2004, 20: 1309-1324.
2. Chang DE, Smalley DJ, Conway T: Gene expression profiling of Escherichia coli growth transitions: an expanded stringent response model. *Mol Microbiol* 2002, 45(2): 289-306.
3. Dudoit S, Yang YH, Callow MJ, Speed TP: Statistical methods for identifying differentially expressed genes in replicated cDNA microarray experiments. *Statistica Sinica* 2002, 12: 111-139.
4. Quackenbush J: Microarray data normalization and transformation. *Nat Genet* 2002, 32(12): 496-501.
5. Tseng GC, Oh MK, Rohlin L, Liao JC, Wong WH: Issues in cDNA microarray analysis: quality filtering, channel normalization, models of variations and assessment of gene effects. *Nucleic Acids Res* 2001, 29(12): 2549-2557.
6. Gao X, Fu X, Li T, Zi J, Luo Y, Wei Q, Zeng E, Xie Y, Li Y,Mao Y: Determining a detectable threshold of Signal intensity in cDNA microarray based on accumulated distribution. *J Biochem Mol Biol* 2003, 36(6): 558-564

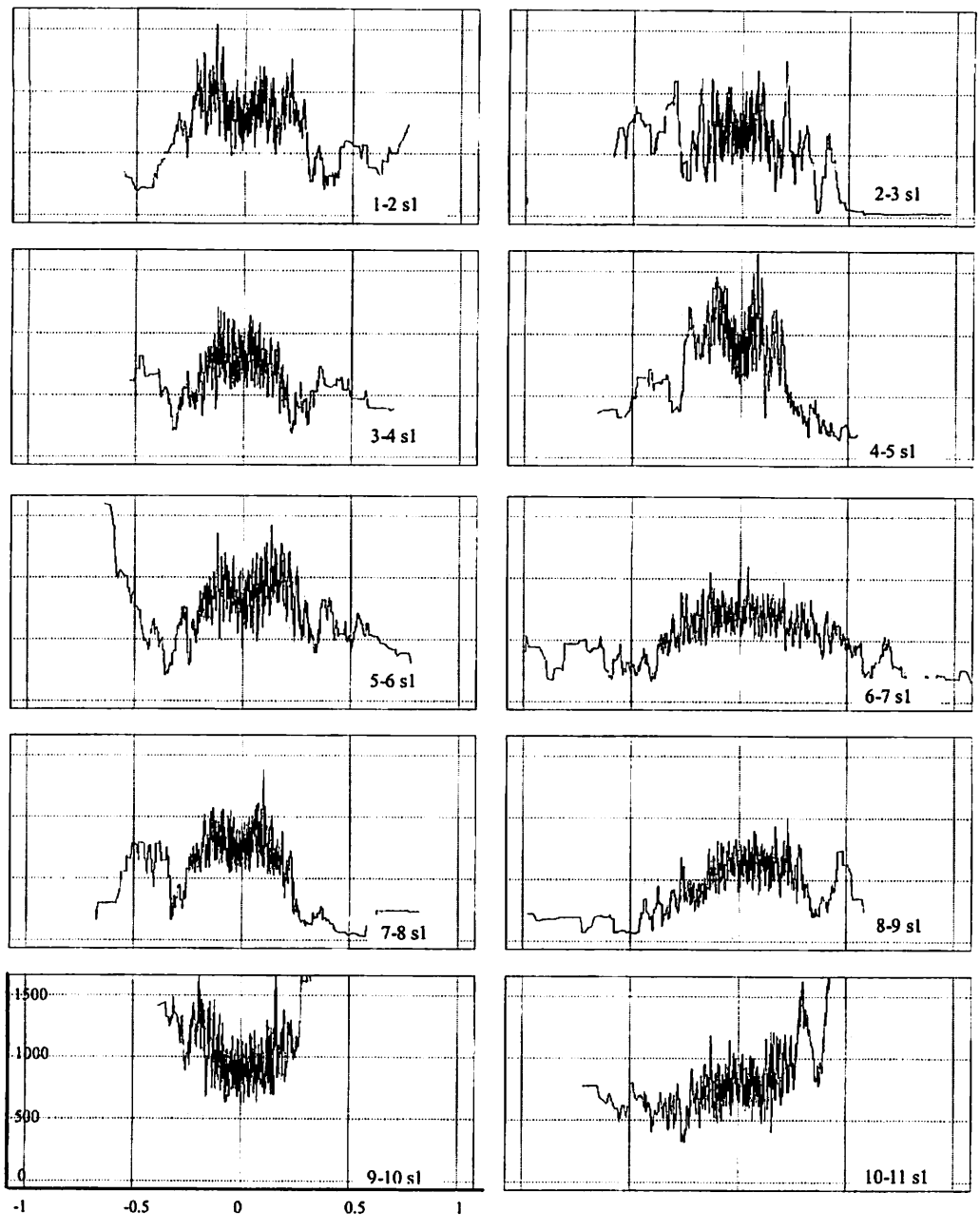

Fig. 4 Relation between rank difference and log-ratio in the 1st time-series experiment. R(g) is represented by X-axis, and the log-ratio for N-design is represented by Y-axis.