Bootstrapping Cluster Analysis: Assessing the Reliability of Conclusions from Microarray Experiments

M. Kathleen Kerr
The Jackson Laboratory
Bar Harbor, Maine 04609 U.S.A.
mkk@jax.org

Gary A. Churchill[1]
The Jackson Laboratory
Bar Harbor, Maine 04609 U.S.A.
garyc@jax.org
207-288-6189 (voice)
207-288-6077 (fax)

Abstract

We introduce a general technique for making statistical inference from gene expression microarray data. The approach utilizes an analysis of variance model to achieve normalization and estimate differential expression of genes across multiple conditions. Statistical inference is based on two applications of a randomization technique, bootstrapping. Bootstrapping is used to obtain confidence intervals for differential expression estimates from individual genes, and then to assess the stability of results from a cluster analysis. We illustrate the technique with a publicly available data set and draw conclusions about reliability of clustering results in light of variation in the data. The bootstrapping procedure relies on experimental replication. We discuss the implications of replication and good design in microarray experiments.

[1] Corresponding author

Microarray technology [1] is a revolutionary high-throughput tool for the study of gene expression. The ability to simultaneously study thousands of genes under a multitude of conditions is an exciting advancement. It is also a huge challenge to comprehend and interpret the resulting mass of data. Early research with cDNA microarrays that demonstrated the promise of the technology has also influenced the direction of research to answer this challenge [2, 3]. Specifically, a variety of clustering techniques have been developed and applied to identify groups of genes with similar patterns of expression [3, 4, 5, 6]. A great deal of effort has gone into identifying the best clustering techniques for microarray data. However, there is a second question at least as important that has received less attention. Namely, how does one make statistical inference based on the results of clustering? The input into any clustering technique is a set of estimates of relative gene expression from a microarray experiment. In current practice, these estimates are taken to be precisely known quantities, ignoring the fact that every estimate has a margin of error. Consider two genes that cluster together. Are the patterns of expression for these genes sufficiently similar beyond any reasonable doubts raised by the noise in the data, or could these genes have clustered together by chance? Here, we propose a bootstrap method to assess the reliability of clustering results in a statistically quantifiable manner [7]. We illustrate our approach with the clustering technique used by Chu *et al.* [2], but bootstrapping can be applied to any clustering technique.

Details of the Chu *et al.* [2, 8] experiment are summarized briefly here. Spotted cDNA microarrays containing 97% of the known genes of Saccharomyces cerevisiae (yeast) were used to study gene expression during meiosis and spore formation. Yeast cells were transferred to a nitrogen-deficient medium to induce sporulation and mRNA samples were taken at seven timepoints: 0, 30 minutes, and 2, 5, 7, 9 and 12 hours. For each of the seven timepoints, the scientists prepared a "red"-labeled cDNA pool. In addition, they prepared a "green"-labeled cDNA pool from the time 0 sample. Seven microarrays were used in the study, one for each of the seven timepoints. Each array was probed with the green-labeled sample mixed with one of the seven red-labeled samples. In effect, time 0 serves as a reference for all of the samples. This experimental setup has some peculiar consequences for analysis we will discuss later.

For any particular spot representing a particular gene there are four readings: green signal, green background, red signal, red background. As their estimate of relative expression of a gene at time k compared to time 0, Chu *et al.* use the background-corrected ratio (red signal − red background)/(green signal − green background) from the array containing red-labeled cDNA from time k and green-labeled cDNA from time 0. There is variability in spot size and the concentration of DNA in each spot, and further variation in the hybridization efficiency of each probe. Thus the meaningful interpretations of microarray data are in terms of relative comparisons, e.g. the relative expression of gene g at timepoint A compared to timepoint B.

Chu *et al.* are particularly interested in genes induced (as opposed to repressed) during sporulation. The authors identify seven temporal patterns or "profiles" of induced transcription of special interest. Their clustering method matches genes to these profiles based on the 7-vector of log ratios. First, they create a model profile based on the average pattern of expression for a hand-picked set of 3 to 8 genes per profile. Second, they filter out about 80% of the genes that do not increase relative to time 0. Third, they calculate correlation coefficients for each induced gene with each of the seven model profiles and match each gene to the profile with which it has highest correlation. Of about 1000 genes that pass their filter, about 450 are assigned to one of the seven profiles.

We modify the Chu *et al.* clustering methodology to incorporate two fundamental changes. The first is the use of ANOVA (analysis of variance) estimates of the relative expression between samples. The second is to evaluate the reliability of clustering results by bootstrapping.

We base our estimates of relative expression on fitting a linear model designed to capture the multiple sources of variation in microarray data [9]. For the sporulation data, genes and timepoints are not the only sources of variation. There are also 7 arrays, each containing over 6000 spots, and two dyes. Systematic differences occur across arrays, spots, and dyes that need to be taken into account. Our general approach is to correct for these sources of variation in a systematic manner via a statistical model rather than using a "pre-processing" approach to normalization. Let y_{ijkg} be the natural logarithm of the background-corrected measurement from array i for dye j and gene g representing time k. Consider the model

$$y_{ijkg} = \mu + A_i + D_j + T_k + G_g + (AG)_{ig} + (TG)_{kg} + \epsilon_{ijkg}, \tag{1}$$

where $i = 1, \dots, 7; j = 1, 2; k = 1, \dots, 7;$ and $g = 1, \dots, 6118$. The A_i terms in this model account for "array effects" — overall variation in fluorescent signal from array to array. Such variation arises if, for example, hybridization conditions vary from array to array leading to some arrays having greater overall signal. The D_j terms account for overall differences between the dyes and the T_k terms are timepoint effects that capture differences in the overall concentration of mRNA in the samples from the

Source	SS	df	MS
Array,Dye,Timepoint	6896.24	13	530.48
Gene	48329.71	6117	7.90
TG,AG	22907.16	73404	0.31
Residual	89.18	6117	0.0146
Adjusted Total	78222.28	85651	

Table 1: Analysis of variance for sporulation data. SS=Sum of Squares; df=degrees of freedom; MS=mean square

different timepoints [10]. The gene effects G_g capture the average levels of expression for genes across the arrays, dyes, and timepoints. The array-by-gene interactions $(AG)_{ig}$ represent the signal contribution due to the combination of array i and gene g. In effect, the AG terms are the "spot" effects, capturing differences due to varying sizes and concentrations of spots on arrays. None of the main effects or the spot effects are of particular interest, but amount to a normalization of the data for ancillary sources of variation. The effects of interest are the interactions between genes and timepoints, $(TG)_{kg}$. These terms capture differences from overall averages that are attributable to the specific combination of a timepoint k and gene g. These timepoint-by-gene interactions play the role of ratios in our framework. Table 1 gives the analysis of variance [11, 12].

Instead of using ratios to estimate differential expression we estimate the relative difference in gene expression for gene g at time k compared to time 0 with $\widehat{(TG)}_{kg} -$

$\widehat{(TG)}_{0g}$ (a $\widehat{}$ over an effect means the least-squares estimate). In addition, we use bootstrapping [13] to construct 99% confidence intervals for these estimates [9]. We chose the bootstrap to construct confidence intervals to avoid making distributional assumptions about the random error. In detail, we produced a set of simulated datasets y^*_{ijkg}, where

$$y^*_{ijkg} = \hat{\mu} + \hat{A}_i + \hat{D}_j + \hat{V}_k + \hat{G}_g + \widehat{(AG)}_{ig} + \widehat{(VG)}_{kg} + \epsilon^*_{ijkg}.$$

The ϵ^*_{ijkg} are drawn independently from the studentized residuals [14] from the original fit of the model [15]. For each simulated data set, we refit the model (1), so for 10,000 bootstrap data sets we obtained 10,000 bootstrap estimates $\widehat{(TG)}_{kg} - \widehat{(TG)}_{0g}$. We take the limits of the middle 99% of these estimates as a 99% confidence region for $(TG)_{kg} - (TG)_{0g}$. Figure 1 shows estimated profiles for select genes.

With these estimates and accompanying confidence intervals, we proceed with a modified version of the Chu *et al.* clustering method. We created model profiles based on the same representative genes identified by Chu *et al.* (Figure 2) [16]. As our filter, we exclude any gene that does not satisfy the following criteria: for at least one timepoint k not zero, $\widehat{(TG)}_{kg} - \widehat{(TG)}_{0g} > 0$ and the 99% confidence interval for $(TG)_{kg} - (TG)_{0g}$ does not contain 0. Thus we attempt to mimic the filter used by Chu *et al.* but with a statistically based criterion. Our filter is not as stringent as that in [2] and passes almost twice as many genes, close to 2000. For each gene g passing our filter we calculate the correlation coefficient r_{gp} for that gene and the $p = 1, \ldots, 7$ profiles. Gene g is assigned to profile p if $r_{gp} > 0.9$ and r_{gp} is larger than r_{gq} for the remaining 6 profiles q. From columns (a) and (b) in Table 2, we see that the number of genes clustering to each profile is somewhat larger here than for Chu *et al.*, except for profile 2 (Early I). We

	Clustering Method			
	(a) Chu *et al.*	(b) Nominal	(c) 95% Stable	(d) 80% Stable
Profile 1	52	65	3	8
Profile 2	61	51	7	11
Profile 3	45	74	3	11
Profile 4	95	151	12	27
Profile 5	158	241	86	120
Profile 6	61	145	17	36
Profile 7	5	15	2	6

Table 2: Number of genes matching to each profile for (a) Chu *et al.* clustering method, (b) modified clustering method with no reliability measure, (c) modified clustering method requiring 95% stability, (d) modified clustering method requiring 80% stability. Column (d) is included because our choice of 95% for stability is somewhat arbitrary.

suppose the difference is due to the greater number of genes passing our filter.

The next step is to assess the reliability of the clusters. We do this with a second application of the bootstrap. We create 499 bootstrap data sets y^*_{ijkg}, as described above. For each simulated data set, we construct a bootstrap temporal pattern based on the estimates $(\widehat{TG})^*_{kg} - (\widehat{TG})^*_{0g}$ for each gene, and repeat the filtering and clustering steps with these bootstrap estimates. The result is 500 clusterings, 1 based on the actual data and 499 bootstrap simulated clusterings. The match of a gene to a profile is declared "significant" if it occurs in the analysis of the actual data and in at least 95%

of the bootstrap clusterings. Column (c) of Table 2 shows the much smaller numbers of genes that prove to be reliable matches to the clusters at the 95% level. We refer to these as stable genes. Figure 3 plots the profiles of these genes. For the most part, the stable genes are a subset of the Chu *et al.* genes [17]. The greatest difference in the clustering methodologies is the exclusion of genes that do not prove to be reliable matches.

This method of clustering is based on correlations. When profiles are themselves highly correlated, one can expect that genes with high correlation to one profile will also have high correlation to the other. Table 3 gives the correlations between the seven model profiles. We see that profiles 4 and 5 have correlation 0.95. This leads us to suspect that in the bootstrap, if the magnitude of error is large enough some genes will sometimes match to profile 4 and sometimes to profile 5 and thus, in the end, fail to be a reliable match to either. Figure 4 shows this to be the case. Consider the genes that initially match to profile 4. Figure 4(a) shows the percentage of bootstraps in which these genes match to profile 5. All genes to the right of the dotted line fail to match to profile 4 at the 95% confidence level simply because of the presence of profiles 5. The story is similar for genes initially matching to profile 5, as seen in Figure 4(b). Given the level of noise in the data, these two profiles are too similar to be readily distinguished.

The importance of replication in microarray experiments has been noted in several recent publications [9, 18, 19]. Replication is a fundamental principle of good experimental design and serves two purposes. First, replication increases the precision of estimated quantities. Second, and perhaps most important, it provides information

	Profile					
	2	3	4	5	6	7
1	.65	.19	.03	-.14	-.39	-.41
2		.73	.77	.60	.40	.11
3			.84	.78	.46	.01
4				.95	.84	.44
5					.83	.36
6						.75

Table 3: Pairwise correlations among the seven profiles.

about the uncertainty of estimates [20]. Only with an appropriately designed experiment that includes replication can statistically valid conclusions be drawn. In the yeast sporulation experiment that is re-analyzed here, a kind of replication is achieved by making a self-comparison of the time 0 sample. Although this is adequate for obtaining a variance estimate and providing residuals for bootstrap analysis, it is not an ideal situation. All of the non-zero residuals from the ANOVA analysis come from the self-comparison array — all other data points are fit exactly because they are not replicated. If the self-comparison array is not typical of the experiment as a whole, one can be misled in imputing the same level of variation to the other arrays.

Although perhaps counterintuitive, it is possible to replicate all samples without using additional arrays. For example, samples could be arranged in a loop as shown in Figure 5, so that samples from each timepoint appear on two arrays. Fitting model (1) with this design, residuals are obtained from every array. In addition to the built-in replication, the timepoint (T) factor is balanced with respect to the dye (D) factor. This balance has certain advantages for the data analysis [19], although there is additional cost associated with the number of labeling reactions required. With the loop design, estimates of $(TG)_{k+1,g} - (TG)_{kg}$ for adjacent timepoint have variance 85.7% as large as estimates of $(TG)_{kg} - (TG)_{0g}$ with the design used by Chu *et al.* [21]. This increased precision, balance among design factors, and the fact that residuals are obtained from every array make this design one alternative worthy of consideration.

In scientific experimentation, results depend on experimental designs that yield precise estimates of quantities of interest as well as estimates of the precision achieved. Furthermore, the design should allow for the assumptions of analysis to be verified. Microarray experiments are no exception. It is certainly an interesting exercise to run a clustering algorithm on gene expression data. However, without an assessment of the reliability of the clusters one cannot make valid inferences about co-regulated genes. Whatever clustering algorithm is chosen, it is imperative to assess whether the results are statistically reliable relative to the level of noise in the data. Bootstrapping is a straightforward way to do this.

References

[1] *Nature Genetics Supplement* 21 (January 1999).

[2] S. Chu *et al.*, *Science* 282, 699 (23 October 1998).

[3] M. Eisen, P.T. Spellman, P.O. Brown, D. Botstein, *Proc. Natl. Acad. Sci. U.S.A.* 95, 14863 (8 December 1998).

[4] P. Tamayo *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* 96, 2907 (March 1999).

[5] T. Hastie *et al.*, *Genome Biology* 1, (4 August 2000). http://genomebiology.com/2000/1/2/research/0003.

[6] L. Lazzeroni, A. Owen, submitted.

[7] J. Felsenstein, *Evolution* 39, 783 (1985). The bootstrap is widely accepted as a method to assess the reliability of phylogenetic reconstruction. Our application to gene expression clusters was inspired by the work of J. Felsenstein.

[8] http://cmgm.stanford.edu/pbrown/sporulation/

[9] M.K. Kerr, M. Martin, G.A. Churchill, submitted. http://www.jax.org/research/churchill

[10] The confounding structure of this experimental design is such that including A, D, and T effects accounts for all two- and three-way interactions of the factors array, dye and time. For example, effects of the labeling reactions would be reflected in DT interaction terms, but these are indirectly accounted for through confounded with A and T.

[11] We assume the error ϵ has mean 0 and variance σ^2 but do make any other distributional assumption.

[12] Before settling on this model, we considered several alternative models. Omitting the spot effects $(AG)_{ig}$ gives a residual mean square of 0.1475, ten times larger than for model (1), so we concluded that spot effects are in fact a significant source of variation. Dye-by-gene interactions are a phenomenon we have seen in other microarray data. Replacing the spot effects $(AG)_{ig}$ with dye-by-gene interactions $(DG)_{jg}$ gives a residual mean square of 0.1575, showing no evidence for dye-by-gene effects in this experiment. We note that it is possible to fit a model that includes both AG and DG effects. However, due to the lack of replication in the experimental design it is not possible to evaluate the fit of such a model because there are no remaining degrees of freedom to estimates the error term.

[13] B. Efron, R.J. Tibshirani, *An Introduction to the Bootstrap* (Chapman and Hall, London, 1994).

[14] N.R. Draper, H. Smith, *Applied Regression Analysis* (Wiley, New York, ed. 3, 1998), p.207. The h_{ii} components of the studentized residuals were obtained by the method outlined on p.215.

[15] C.F.J. Wu, *Annals of Statistics* 14, 1261 (1986) also used studentized residuals in bootstrapping.

[16] Two genes, MRD1 and NAB4, for profile 3 and two genes, KNR4 and EXO1, for profile 4 could not be found in the publicly available data file. We constructed profiles 3 and 4 with the remaining genes.

[17] Seven stable genes matched to profiles 5 6, or 7 but did not match to any profile by Chu *et al.* One of these, YPL280W, is in Figure 1. Like YPL280W, the other

6 genes have fairly flat profiles, so the difference is likely due to the less stringent filter.

[18] M.-L.T. Lee, F.C. Kuo, G.A. Whitmore, J. Sklar, *Proc. Natl. Acad. Sci. U.S.A.* 97, 9834 (29 August 2000).

[19] M.K. Kerr, G.A. Churchill, *Biostatistics*, in press.

[20] R.A. Fisher, *The Design of Experiments* (Oliver and Boyd, Edinburgh, ed. 6, 1951).

[21] With the loop design, the variance of gene-specific differences in timepoints depends on the relative position of the corresponding samples in the loop. Since adjacent timepoints are estimated most precisely, it will be most efficient to estimate profiles using those comparisons rather than using time 0 as a fixed reference point.

[22] M.K.K. is supported by a Burroughs-Wellcome postdoctoral fellowship from the Program in Mathematics and Molecular Biology. G.A.C. is supported by an NIH grant.

Figure 1. Temporal profiles for select genes. The solid line gives the temporal profile estimated using model (1). Around each line are error bars according to bootstrap-estimated 99% confidence intervals. For comparison, the dotted line gives the temporal profile estimated with ratios, re-scaled to have the same standard deviation as the solid line.

Figure 2. The seven model profiles used for clustering. The profiles are re-scaled to have standard deviation 1. This adjustment does not affect the clustering results because clustering is based on correlations.

Figure 3. Stable genes for the seven model profiles based on 500 bootstrap clusterings. The plotted profiles have been re-scaled to have standard deviation 1.

Figure 4. Bootstrap behavior of genes with nominal match to profile 4 or 5. Genes with high correlation to profile 4 tend to have high correlation with profile 5 and vice versa. There were 151 genes that initially match to profile 4. In 500 bootstrap clusterings, each gene matched to profile 5 in some percentage of the clusterings. Figure (a) shows the distribution of those percentages for the 151 genes. Figure (b) shows the percentages of bootstraps in which genes that initially matched to profile 5 matched to profile 4. The histograms show that many genes fail to be stable matches to profile 4 because of the presence of profile 5, and vice versa.

Figure 5. An alternative experimental design for the sporulation experiment. Here the experimental layout is represented as a directed graph. The boxes represent the mRNA samples and the arrows represent microarrays. The tail of an arrow is, say, the “red” dye and the head of an arrow is the “green” dye. Thus the arrow from the time 0 sample to the half-hour sample means to probe an array with red-labeled time

0 mRNA and green-labeled mRNA from the half-hour sample. Such a design may not be possible if too little mRNA is available from some samples for two arrays. On the other hand, such a design has advantages over the plan used by Chu *et al.* in balance, precision of estimates, and distribution of residuals.

Figure 1

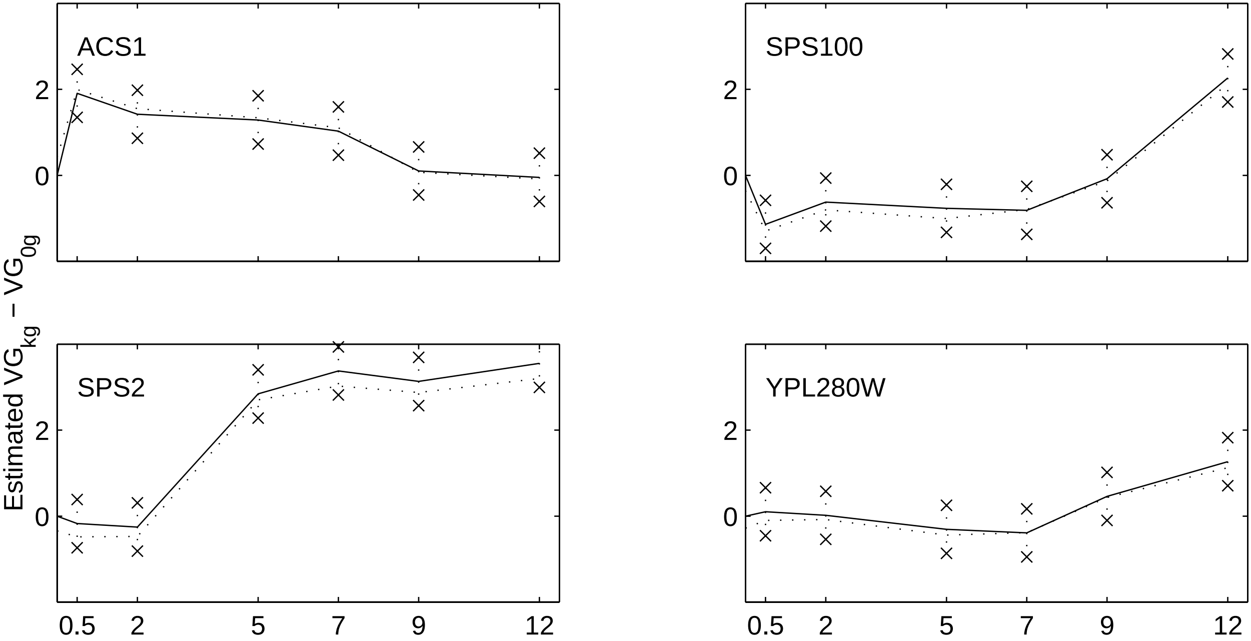
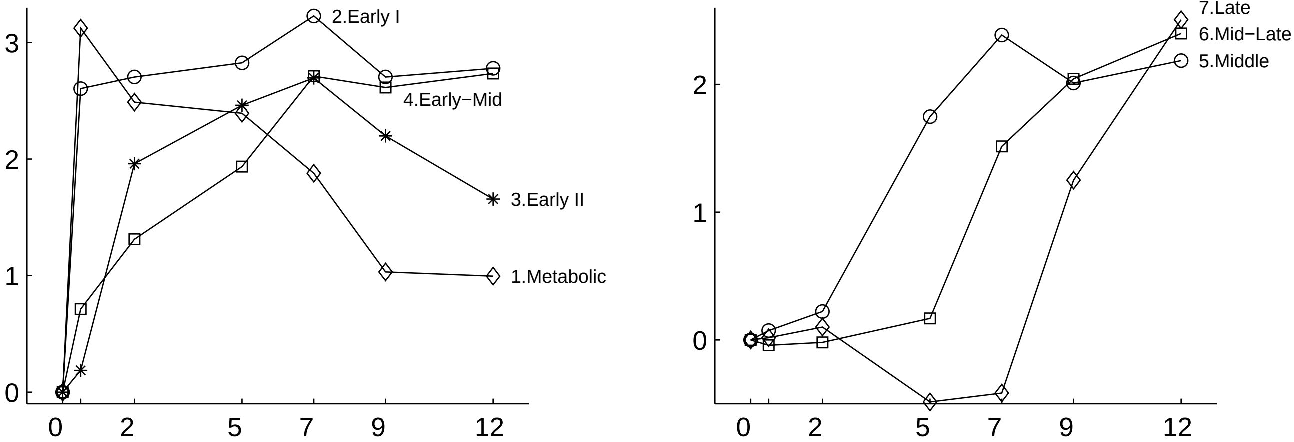

ACS1
SPS100
SPS2
YPL280W
Estimated $VG_{kg} - VG_{0g}$
2
0
2
0
2
0
2
0
0.5
2
5
7
9
12
0.5
2
5
7
9
12

Figure 2



2.Early I
4.Early−Mid
3.Early II
1.Metabolic
3
2
1
0
0
2
5
7
9
12
7.Late
6.Mid−Late
5.Middle
2
1
0
0
2
5
7
9
12

Figure 3

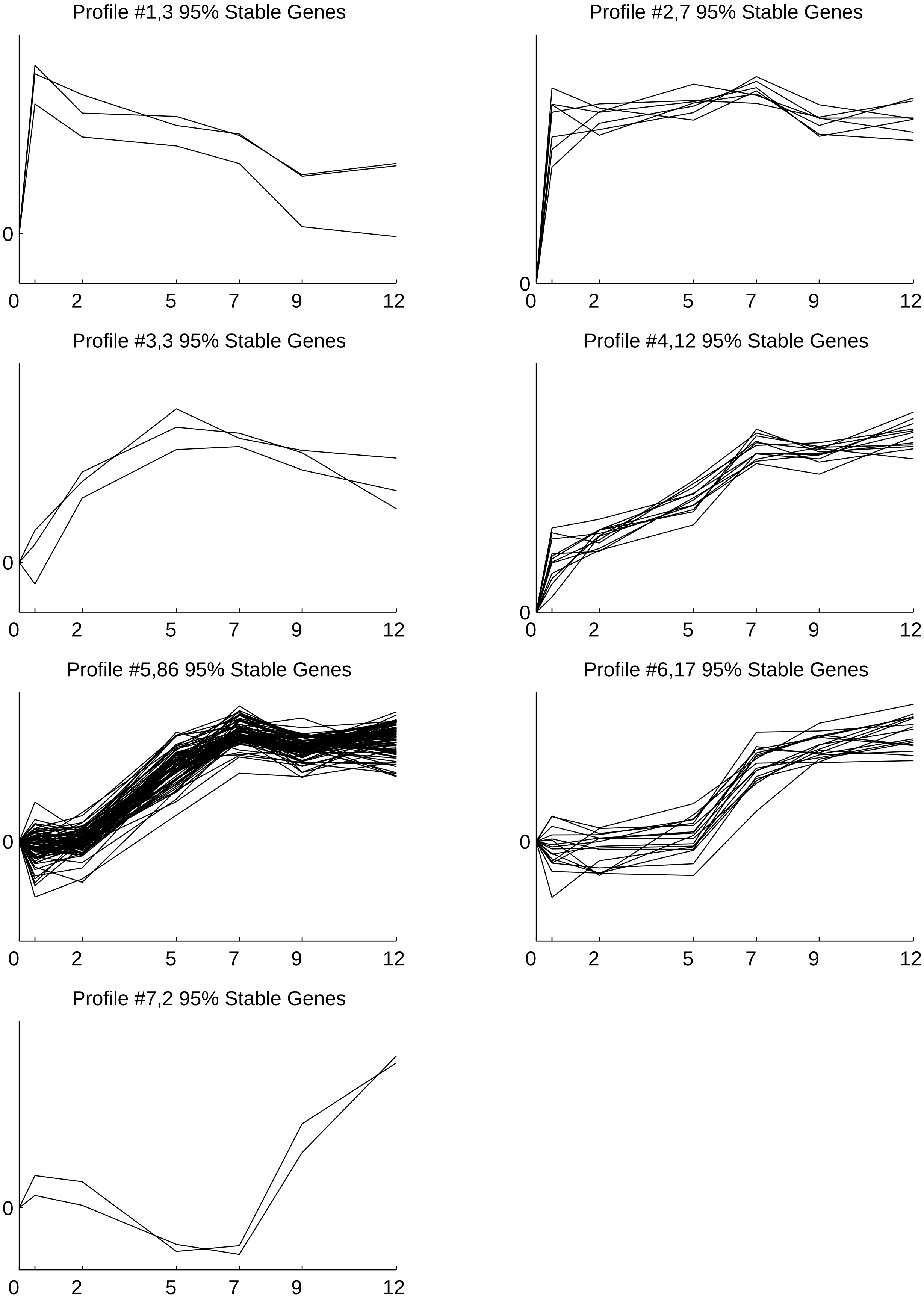

Profile #1,3 95% Stable Genes
0
0
2
5
7
9
12
Profile #2,7 95% Stable Genes
0
0
2
5
7
9
12
Profile #3,3 95% Stable Genes
0
0
2
5
7
9
12
Profile #4,12 95% Stable Genes
0
0
2
5
7
9
12
Profile #5,86 95% Stable Genes
0
0
2
5
7
9
12
Profile #6,17 95% Stable Genes
0
0
2
5
7
9
12
Profile #7,2 95% Stable Genes
0
0
2
5
7
9
12

Figure 4

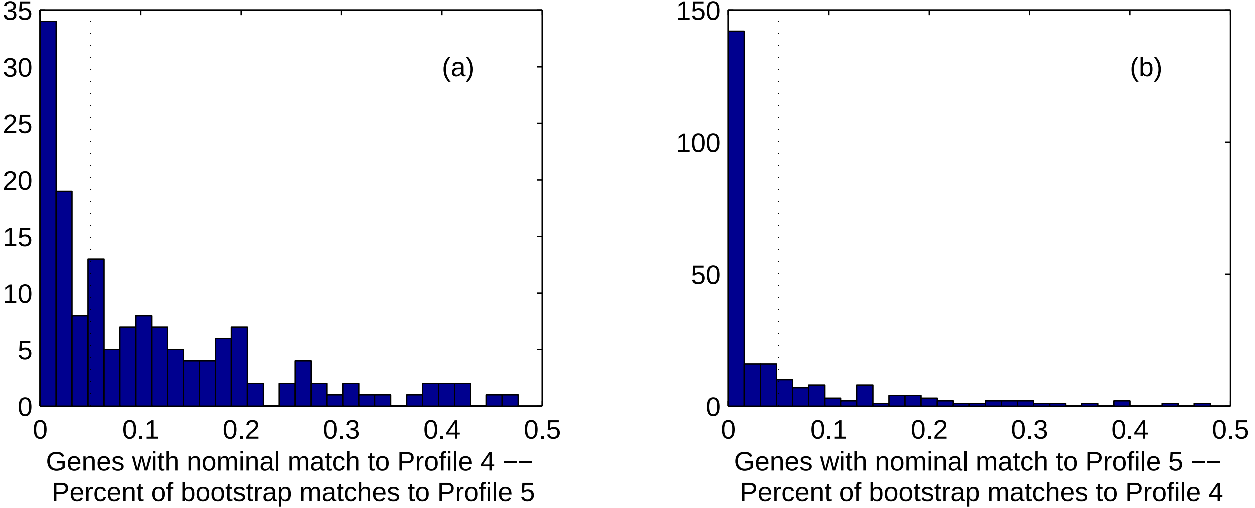
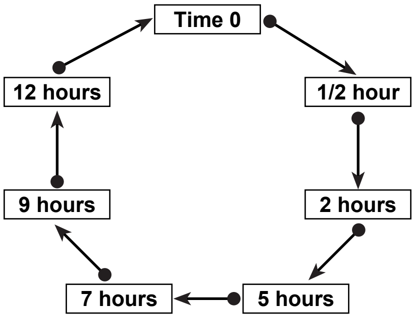

35
30
25
20
15
10
5
0
(a)
0
0.1
0.2
0.3
0.4
0.5
Genes with nominal match to Profile 4 −−
Percent of bootstrap matches to Profile 5
150
100
50
0
(b)
0
0.1
0.2
0.3
0.4
0.5
Genes with nominal match to Profile 5 −−
Percent of bootstrap matches to Profile 4

Figure 5


Time 0
1/2 hour
2 hours
5 hours
7 hours
9 hours
12 hours