Supplemental material on clustering

May 4, 2006
John Aach

Supplemental material for

Global gene expression of *Prochlorococcus* ecotypes in response to changes in nitrogen availability

Andrew C. Tolonen, John Aach, Debbie Lindell, Zackary I. Johnson, Trent Rector, George M. Church, Sallie W. Chisholm

This document gives additional information on the clustering method used to analyze the expression data.

Gibbons and Roth (2002) describe a mutual information Z-score (MIZ) measure that compares gene functional annotations with clustering results, and use it to assess the performance of various clustering algorithms and for determining of the number of clusters. They demonstrate the measure on four yeast expression data sets using Gene Ontology categories for annotation (The Gene Ontology Consortium (2000)). K-means clustering based on Euclidean distance was found to be one of the best algorithms for clustering for ratio-style expression data. Based on these results, we applied K-means clustering to our N-starvation series expression data, using Cyanobase level 1 functional categories, and used the Gibbons and Roth (2002) MIZ to determine the number of k-means clusters to generate.

Before doing the K-means clustering, the expression data for each organism was filtered by accepting only those N starvation expression profiles for which at least one of the time points 3, 6, 12, 24, or 48 had a q value < .05 as determined from GoldenSpike-normalized mean (log[experiment/control]) ratios (see Methods in text and also Choe et.al. (2005)). Time point 0 was excluded from consideration in both the filtering and in the subsequent clustering under the assumption that any apparently significant difference between experimental and control expression levels at this time point must represent measurement noise rather than biological function. This filtering reduced the number of genes in the clustering to 410 for MED4 and 559 for MIT9313. (One MED4 gene among the 410 that passed these filters is actually represented by two expression profiles because it has two feature sets on the Affymetrix array.)

Meanwhile, some of the 16 Cyanobase level 1 functional categories were eliminated from the computation for MIZ, namely the "Hypothetical" and "Other categories" categories (on the grounds that these are not biologically informative), and any category represented by fewer than 10 genes among the filtered gene sets (following Gibbons and Roth (2002)). Thus, MIZes were computed with 8 functional categories for MED4 and 10 for

MIT9313. Categories excluded at this point were also excluded from scoring for functional category enrichment by hypergeometric statistics (see below).

Clustering and MIZ calculation were done in MatLab. K-means clustering was performed for all *k* between 2 and 20 inclusive. For each *k*, the MatLab K means function was used with 'sqEuclidean' distance (the default) and with the following options: 'replicates',500,'maxiter',100,'emptyaction','singleton'. Because of random initial cluster seeding, K-means is a non-deterministic algorithm so that identical executions of the same kmeans command will not generally yield the same clustering or, in consequence, the same MIZ. The 'replicates' option repeats a kmeans clustering a specified number of times and picks from among them the tightest clustering as measured by sums of all distances to cluster centroids. Prior experimentation had shown that increasing replicates resulted in lower MIZ variance (data not shown).

The MIZ scores found for MED4 are shown in Figure 1, and those for MIT9313 are shown in Figure 2, along with the values of *k* derived from these scores. In the case of MIT9313, the *k* with the absolute maximum value MIZ was used. For MED4, the k with the best local MIZ value was used because using the absolute maximum would have led to a simple dichotomy of MED4 genes rather than a multi-clustering.

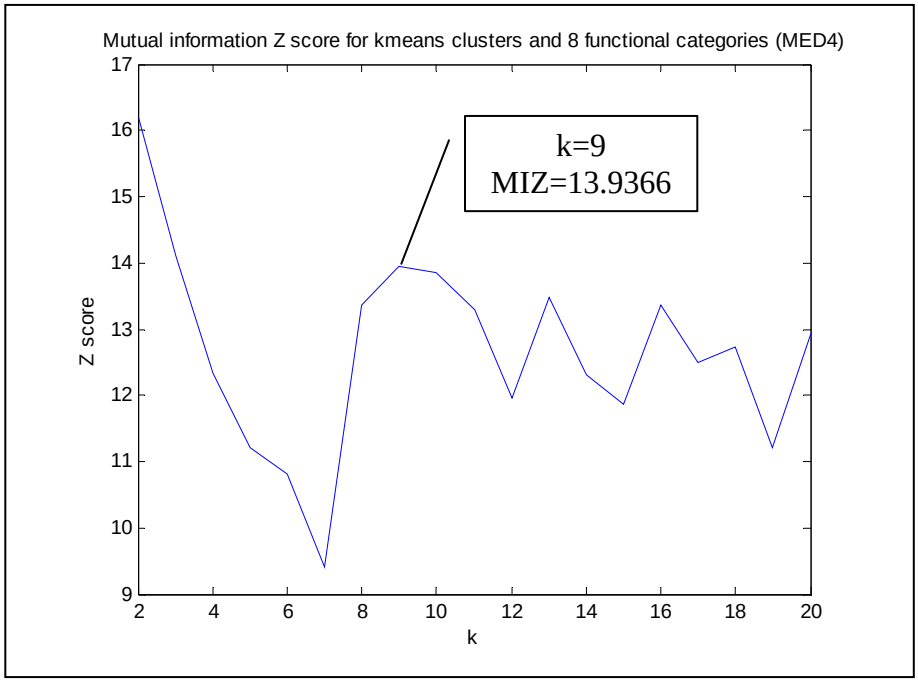

Figure 1: Mutual information Z score (MIZ) for MED4 kmeans clusterings for k between 2 and 20. The *k* with the best *local* maximum score (*k* =9) was used for the final MED4 clustering rather than the *k* with absolute maximum *k* =2), because a 2 cluster analysis would be uninteresting while *k* values between 3 and 8 would be less informative about functional categories than *k*=2.

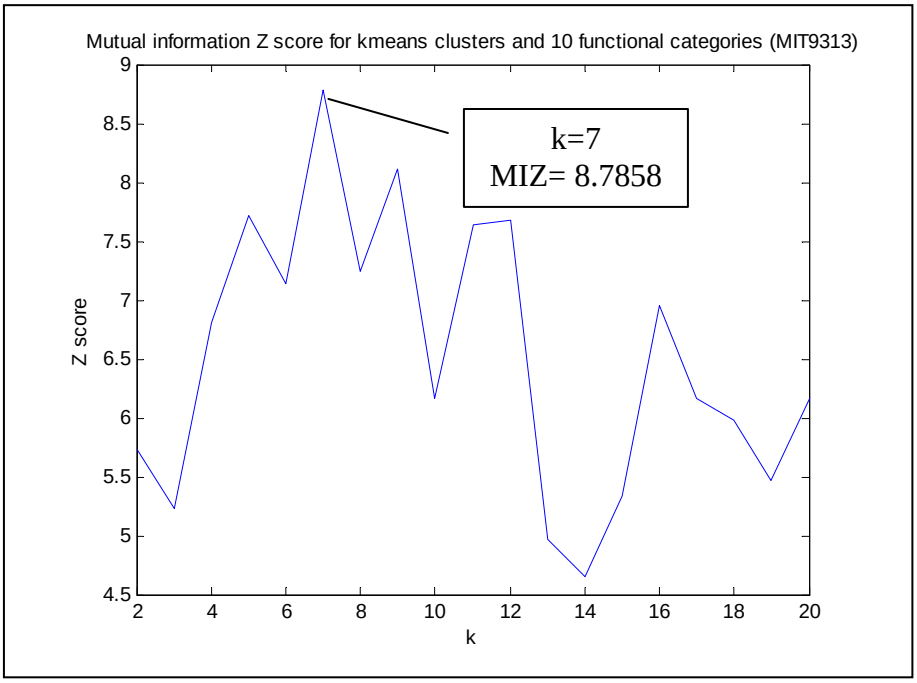

Figure 2: Mutual information Z scores (MIZ) for MIT9313 kmeans clusterings for k between 2 and 20. The k with the absolute maximum score (k =7) was used for the final MIT9313 clustering.

Figure 3 and 4 show the cluster mean and individual gene expression level profiles of the k = 9 MED4 clustering. Figures 5 and 6 show these profiles for the k = 7 MIT9313 clustering.

With the genes clustered, the functional categories represented in each cluster were then analyzed to identify the functional category most enriched in the cluster. Functional category enrichment analysis was based on hypergeometric P-values using the standard formula:

$$CategoryEnrichmentPvalue(i,c) = \sum_{x=O_{ic}}^{\min(L_i,N_c)} \frac{\binom{L_i}{x}\binom{G-L_i}{N_c-x}}{\binom{G}{N_c}}$$

where G = total number of genes in the clustering, L_i = number of genes in c*L*uster i, N_c = number of genes in category c, and O_{ic} = number of genes in the intersection (overlap) between cluster i and category c. These values were computed in MatLab via the hygecdf function as 1 – hygecdf(O_{ic}–1, G, L_i, N_c), where $O_{ic} > 0$.

To evaluate these P values for statistical significance requires consideration of multiple hypotheses at two levels. For the first, within the clustering at hand there are k clusters and C categories, where C=8 for MED4 and 10 for MIT9313 (see above). Since there are

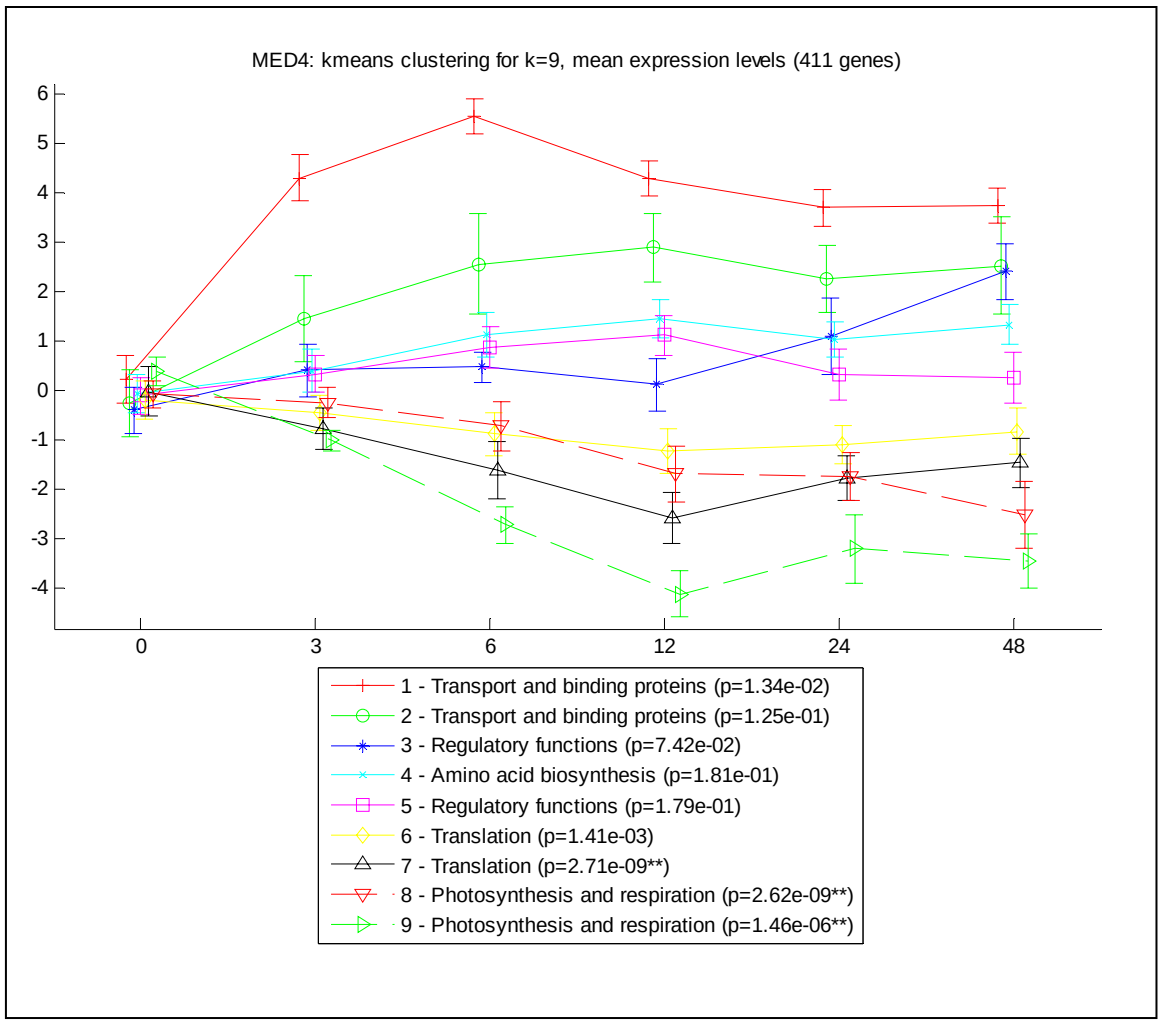

Figure 3: Cluster means for the MED4 *k*=9 clustering across the N-starvation time series. Error bars represent standard deviations. X-coordinate is slightly jittered to enhance visibility. The most enriched functional categories for each cluster are identified along with their P-values and statistical significance (** = "stringent" threshold, * = "permissive" threshold; see text).

$k \cdot C$ tests for functional category enrichment within this clustering, this yields a significance threshold $P<.05/ k \cdot C$. We refer to this as the "permissive" threshold.

However, an issue with this threshold is that in picking *k*, a range of possible clusterings between *k*=2 and 20 were evaluated with respect to the mutual information between the clusterings and functional categories. Picking a *k* that optimizes this value is thus likely to pick a *k* for which at least some clusters are already enriched for some categories, so that the hypergeometic P value will be biased to lower values. Thus we also considered a conservative "stringent" threshold for significance for which $P</(209 \cdot C)$. The value 209 represents *all* of the clusters for all of the clusterings for *k*=2,3,…,20; i.e., 209 = 2+3+ … + 20. The "stringent" threshold thus compensates for the bias in P values by treating the effect of using of mutual information in picking *k* as if it were equivalent to picking the *k* which yielded the cluster with the minimum *CategoryEnrichmentPvalue* over all the 209

clusters. The P values and significance thresholds for the most enriched function in each the MED4 k = 9 and MIT9313 k = 7 are shown in Figures 3-6.

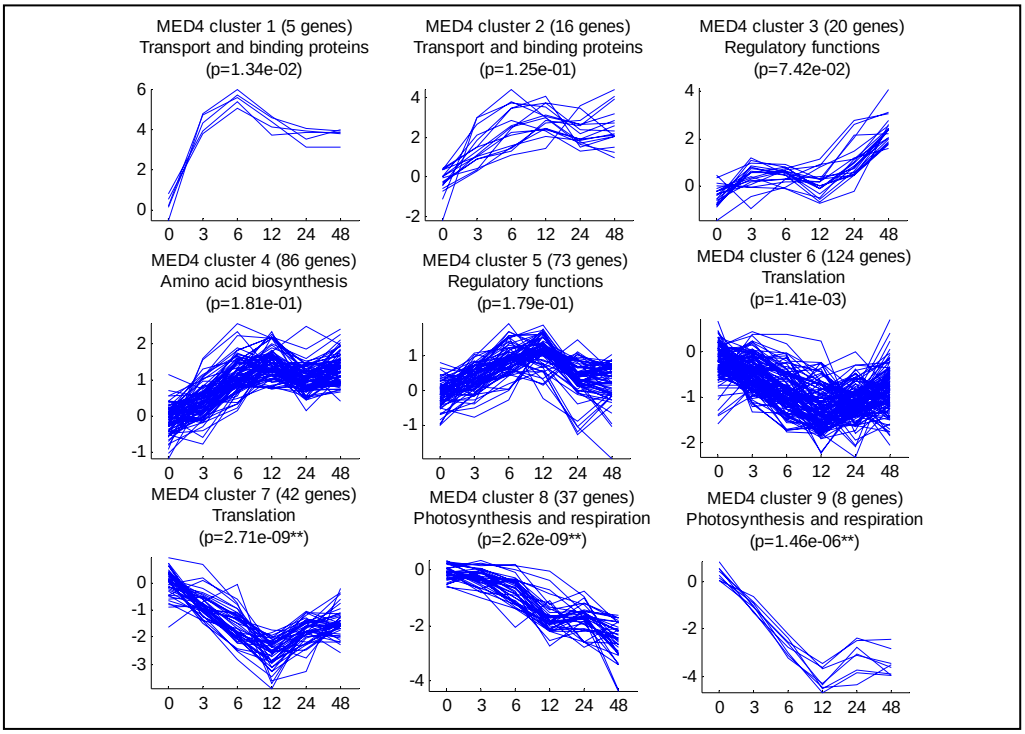

Figure 4: Individual gene expression profiles for each cluster in the MED4 k=9 clustering.

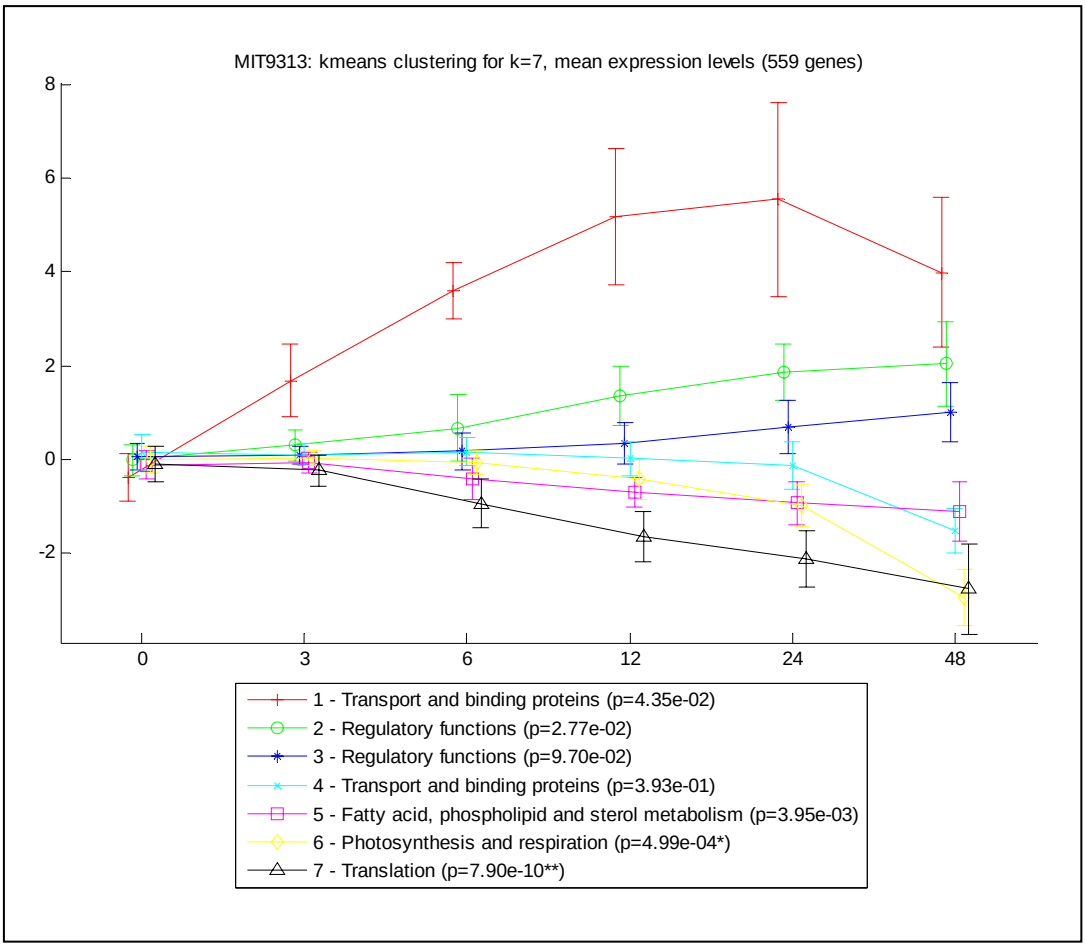

Figure 5: Cluster means for the MIT9313 *k*=7 clustering across the N-starvation time series. Error bars represent standard deviations. X-coordinate is slightly jittered to enhance visibility. The most enriched functional categories for each cluster are identified along with their P-values and statistical significance (** = "stringent" threshold, * = "permissive" threshold; see text).

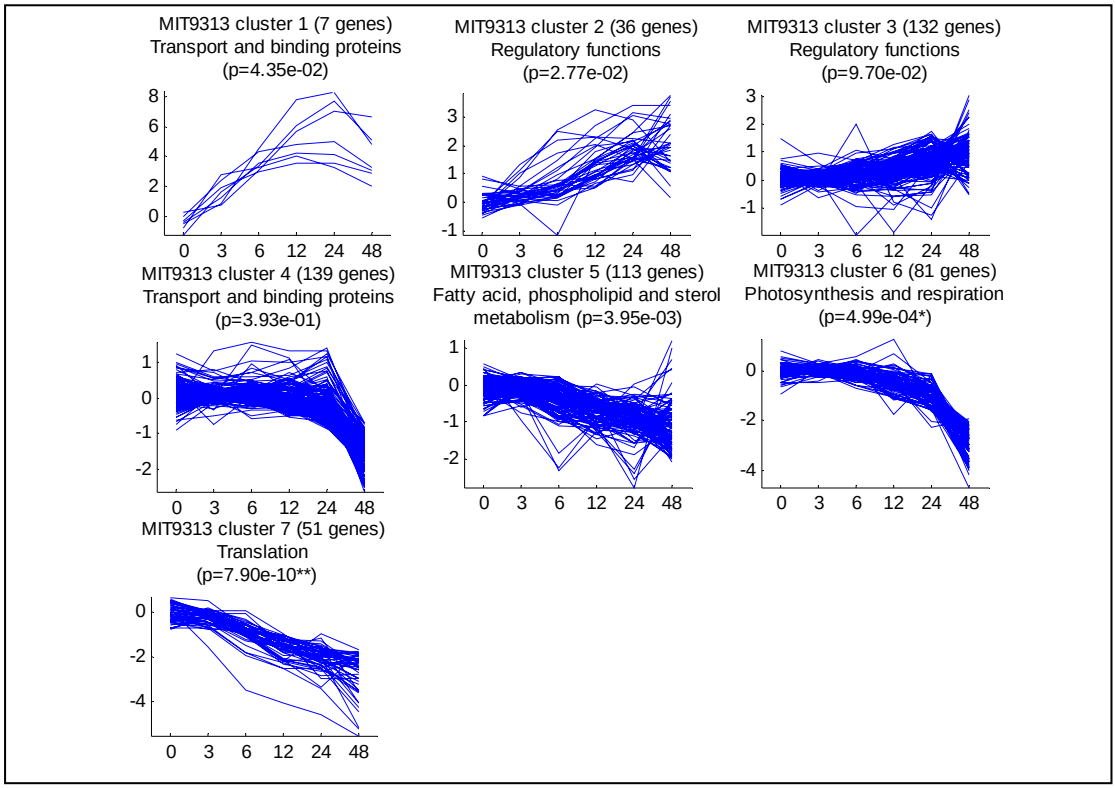

Figure 6: Individual gene expression profiles for each cluster in the MIT9313 *k*=7 clustering.

REFERENCES

Choe, S.E., Boutros, M., Michelson, A.M., Church, G.M., Halfon, M.S., 2005, Preferred analysis methods for Affymetrix GeneChips revealed by a wholly defined control dataset. *Genome Biol* 6(2): R16

Gibbons F.D., and Roth F. P., 2002, Judging the Quality of Gene Expression-Based Clustering Methods using Gene Annotation, *Genome Res.* 12:1574-81

Nakamura Y, Kaneko T, Hirosawa M, Miyajima N, Tabata S., 1998, CyanoBase, a www database containing the complete nucleotide sequence of the genome of *Synechocystis* sp. strain PCC6803. *Nucleic Acids Res.* 26(1):63-7.

The Gene Ontology Consortium, 2000, Gene Ontology: tool for the unification of biology. *Nature Genet.* 25: 25-29.