Validation of predicted operons against N-starvation expression data

May 5, 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

The role of the transcriptional regulator NtcA in shaping the response to changes in N availability by examining the expression levels of genes using NtcA target sites predicted by Su et.al. (2005). For consistency of analysis, we also used the operons predicted by Su. et. al. (2005) in conducting motif searches in upregulated and downregulated gene clusters (see main text and Supplemental Information). Su et.al. (2005) predicted operons on the basis of the stringent assumption that tandem genes within 45 bp of each other were in an operon. We therefore sought to check the validity of this definition using the N-starvation expression data.

METHODS

We tested the Su et.al. (2005) predicted MED4 and MIT9313 operons by averaging Pearson correlation coefficients (ρ) of N starvation time series expression levels for genes in positions 1 and genes in position *n* of predicted operons (n = 2, 3, 4, 5), and comparing these against 2500 averages of same-sized sets of correlations of randomly chosen pairs of all predicted operon genes. For instance, there were 340 operons of length >=2 predicted by Su et.al. (2005) for MED4, and the average ρ of the genes in positions 1 and 2 of these operons across the N starvation series was .406 (Table 1); we estimated a *P* value by computing a set of 2500 averages of 340 ρ values for randomly picked pairs of genes and calculating the fraction of these 2500 averages that were > .406. We performed a similar calculation for tandem genes that were not in the same Su et.al. (2005) predicted operons, choosing random pairs from this same gene set. Operons in the Su et.al. (2005) predictions that contained genes not represented on the MD4-9313 array were excluded from these calculations.

RESULTS

Results are summarized in Table 1. For MED4, all operon position pairs are statistically significantly co-expressed better than chance. For MIT9313, only operon positions 1 and 2 are statistically significantly co-expressed better than chance; however, positions 1 and 3, and 1 and 4, exhibit elevated co-expression that does not achieve the level of statistical significance. These results indicate that the operon predictions of Su et. al. (2005) likely capture many real operons.

However, Table 1 also shows that tandem genes that are not in the same Su et. al. (2005) operons were also found to be statistically significantly co-expressed at better than chance levels, suggesting that the stringent Su et.al. (2005) prediction criteria may be missing some real operons.

Figures 1-3 show the actual average correlations for all predicted operon and tandem non-operon correlations plotted against histograms of the average correlations of corresponding randomly picked gene pairs.

Aside from validating the Su et.al. (2005) predictions, these results provide support for the general validity of the N starvation expression data by showing that these data contain a detectable signal of a general biological phenomenon: operons.

		Operon position pairs				Tandem non-operon genes
		1 & 2	1 & 3	1 & 4	1 & 5	
MED4	mean ρ	0.406	0.216	0.278	0.420	0.156
	N	340	139	57	27	390
	P value	0*	0*	0.0004*	0*	0*
MIT9313	mean ρ	0.263	0.131	0.202	0.134	0.242
	N	415	147	58	22	668
	P value	0*	0.0416	0.0228	0.2488	0*

Table 1: Genes in predicted operons are co-expressed at better than chance levels. Pearson correlation coefficients (ρ) were calculated using the expression levels for the six time points of the N starvation time series for all genes in pairs of positions in all operons defined by Su et. al. (2005). *N* = the number of gene pairs; mean ρ = average Pearson correlation coefficient; *P* value = probability that an average correlation > the actual mean ρ may be obtained from random pairs of the same genes (see Methods). * *P* value significant at < .005 (Bonferroni correction).

REFERENCES

Su, Z. Olman, V., Fenglou, M., Xu, Y., 2005, Comparative genomics analysis of NtcA regulons in cyanobacteria: regulation of nitrogen assimilation and its coupling to photosynthesis. *Nucleic Acids Res* 33(16): 5156-71

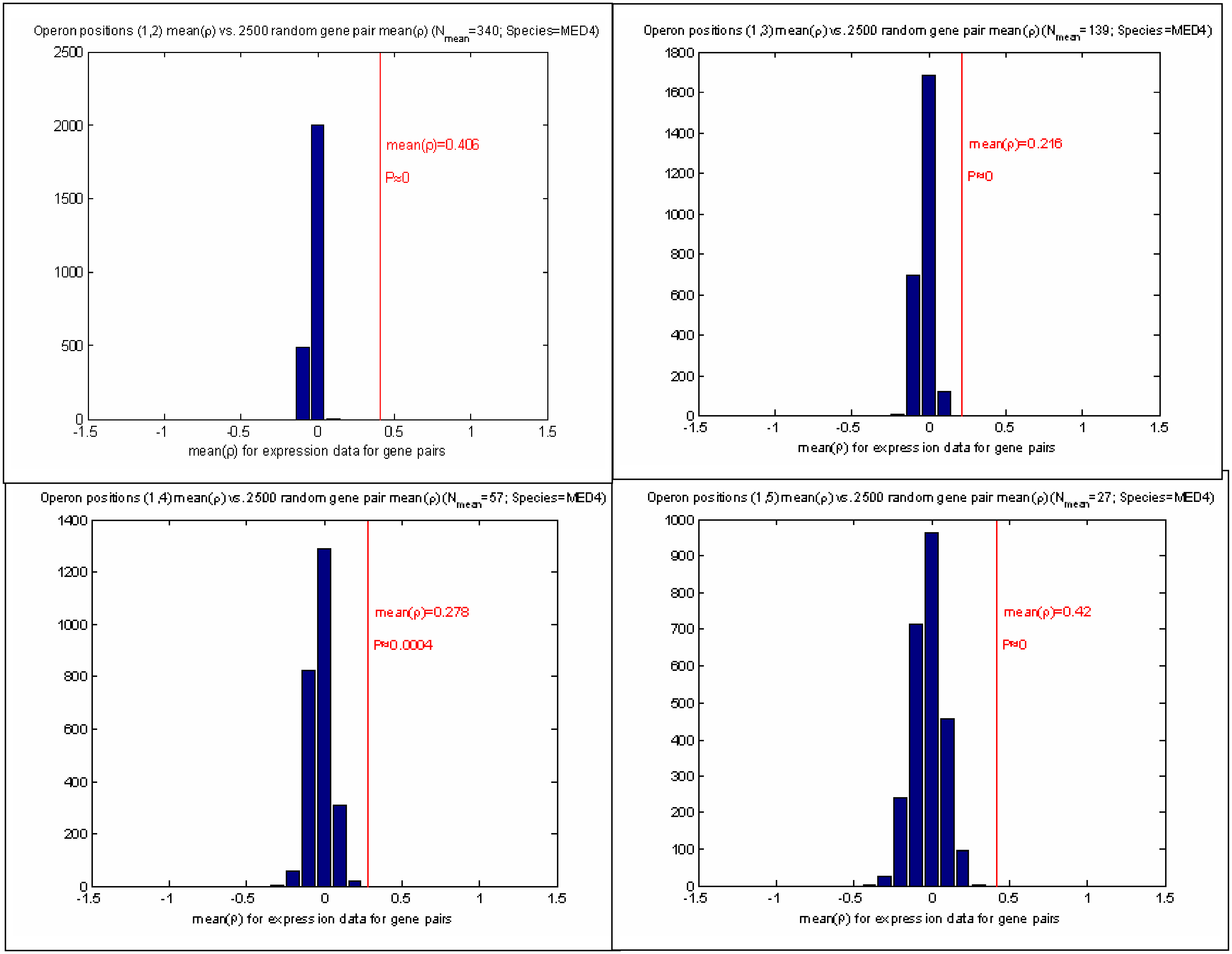

Figure 1: Average Pearson correlation coefficients of MED4 predicted operon positions 1 and *n* (*n*=2,3,4,5) *vs* 2500 average correlations of same-sized sets of random gene pairs.

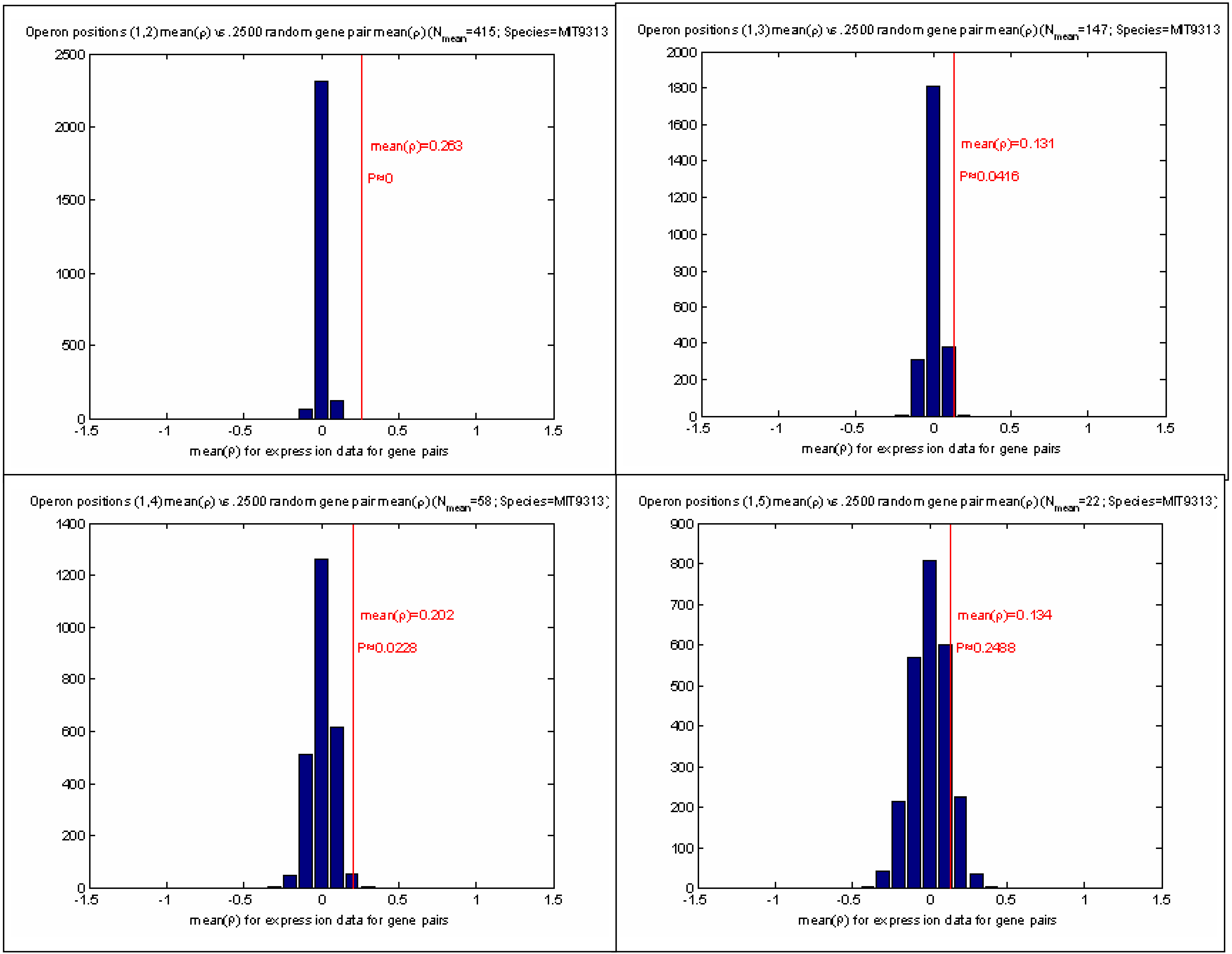

Figure 2: Average Pearson correlation coefficients of MIT9313 predicted operon positions 1 and *n* (*n*=2,3,4,5) *vs* 2500 average correlations of same-sized sets of random gene pairs.

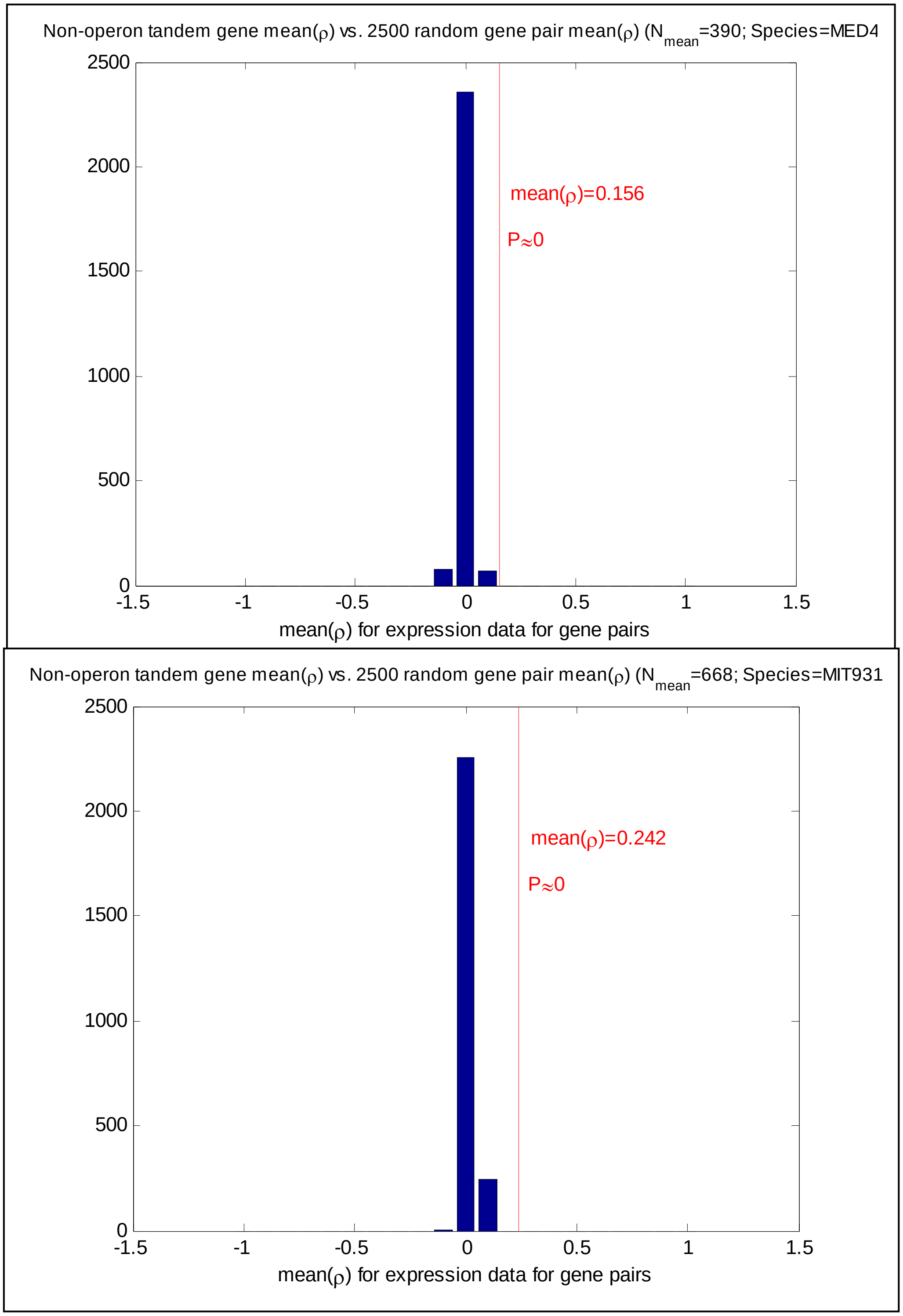

Figure 3: Average Pearson correlation coefficients for tandem genes not in predicted operons in MED4 (top) and MIT9313 (bottom) *vs* 2500 average correlations for same-sized sets of random gene pairs.