Supplementary Table 1: Agreement of our unsupervised and supervised predictions with experimentally identified operon and not-operon pairs in *E. coli* and *B. subtilis*. AOC is the area under the operating curve (e.g., Figure 3A), or the probability that an operon pair will have a better score than a not-operon pair if both pairs are chosen at random. Default sensitivity (fraction of known operon pairs which are correctly predicted) and specificity (fraction of known not-operon pairs which are correctly predicted) are computed with a threshold of predicted $p>0.5$, and maximum accuracy is the maximum over all possible thresholds of the average of sensitivity and specificity. The unsupervised microarray-based predictions, which are shown only in this table, use a logistic regression of the microarray data (rank of Pearson r, total intensity, and total absolute change of log-levels for the pair, with pairwise interactions) versus the usual unsupervised predictions (thresholded at 0.5).

For comparison, we show results from our supervised predictions, from Salgado *et al.* 2000 for *E. coli* (using distance and Monica Riley's functional classification, or just distance), from Sabatti *et al.* 2002 for *E. coli* (using correlation in microarray data and/or distance as features, on a somewhat different training set), from Bockhorst *et al.* 2003b for *E. coli* (distance-only or distance plus microarrays and further sequence-based features), from Moreno-Hagelsieb and Collado-Vides 2002 for *B. subtilis* (using a distance model trained in *E. coli*), and from De Hoon *et al.* 2004 for *B. subtilis* (using distance and/or microarray correlation, and a much larger unpublished training set). We do not show the results of Bockhorst *et al.* 2003a because they report accuracy for predicting transcripts, not individual pairs of genes.

Measure	**AOC**	**Max. Acc.**	**Def. Sens.**	**Def. Spec.**
E. coli				
Unsupervised (Sequence-only)	0.920	0.852	0.883	0.799
Distance-only	0.886	0.829	0.794	0.857
Unsupervised with microarrays	0.925	0.863	0.890	0.817
Microarray-only	0.820	0.750	0.834	0.660
Supervised (Sequence-only)	0.919	0.859	0.865	0.850
Salgado *et al.* 2000	–	0.87	–	–
Distance-only	–	0.82	–	–
Sabatti *et al.* 2002	–	0.88	0.88	0.88
Distance-only	–	0.83	0.84	0.82
Microarray-only	–	0.76	0.82	0.70
Bockhorst *et al.* 2003b	0.929	–	0.78	0.90
Distance-only	0.915	–	–	–
B. subtilis				
Unsupervised (Sequence-only)	0.888	0.815	0.909	0.710
Distance-only	0.882	0.863	0.825	0.863
Unsupervised with microarrays	0.885	0.844	0.922	0.727
Microarray-only	0.748	0.692	0.804	0.545
Supervised (Sequence-only)	0.907	0.868	0.877	0.847
Moreno-Hagelsieb & Collado-Vides 2002	–	0.82	–	–
de Hoon *et al.* 2004	–	0.884	0.888	0.879
Distance-only	–	0.856	0.821	0.890
Microarray-only	–	0.796	0.801	0.791

Supplementary Table 2: Statistical tests of differences between *E. coli*'**s distance model and those of** *Halobacterium NRC-1* **and** *Helicobacter pylori***.** To confirm differences in distance models, we tested same-strand pairs separated by 20-49 base pairs (*E. coli* vs. *Halobacterium*) or by 50-99 base pairs (*E. coli* vs. *H. pylori*). We compared how often these pairs were conserved within 5 kb in a distant genome, relative to other pairs in the same genome. We show the 90% confidence intervals of the odds ratios from the Fisher exact test. In both cases the odds ratio in *E. coli* is higher, indicating significantly greater conservation at these separations (*p*<0.05).

		Conserved within 5 kb		
Genome	**Range (bp)**	**In-range pairs**	**Other pairs**	**Odds Ratio**
Halobacterium	20–49	12/194 (6.2%)	173/1017 (17.0%)	0.18–0.55
E. coli	20–49	127/324 (39.4%)	956/2681 (35.7%)	0.95–1.4
H. pylori	50–99	15/143 (10.5%)	314/1083 (29.0%)	0.17–0.46
E. coli	50–99	117/426 (27.5%)	966/2,579 (37.5%)	0.52–0.77

Supplementary Table 3: Comparison of "strand-wise" and "strand-naive" models for estimating P(Operon—Same). The strand-wise estimate leads to significantly more accurate unsupervised predictions in *B. subtilis*. The poor agreement between both estimates and the *E. coli* distance model-based method of Moreno-Hagelsieb and Collado-Vides (2002) probably reflects the biologically meaningful variation in the distance distributions of different genomes (Rogozin *et al.* 2002).

Issue	**Measure**	**Strand-wise**	**Strand-naive**	**p**
# Operons in *B. subtilis*	% same-strand pairs that are within operons	51.7%	41.3%	–
Accuracy on known operons in *B. subtilis*	Area under the operating curve	0.888	0.864	$<10^{-5}$, test of DeLong *et al.* 1988
Agreement with micro-array data for *B. subtilis*	Spearman correlation of $P(Operon \mid AllFeatures)$ with microarray similarity r	0.461	0.433	$<10^{-10}$, two-sided t-test of correlation between $rank(r)$ and differences in $rank(p)$
Agreement of estimated # operons with *E. coli*-based estimates	Spearman correlation, 124 genomes	0.363	0.223	0.04, correlation test of ranked differences