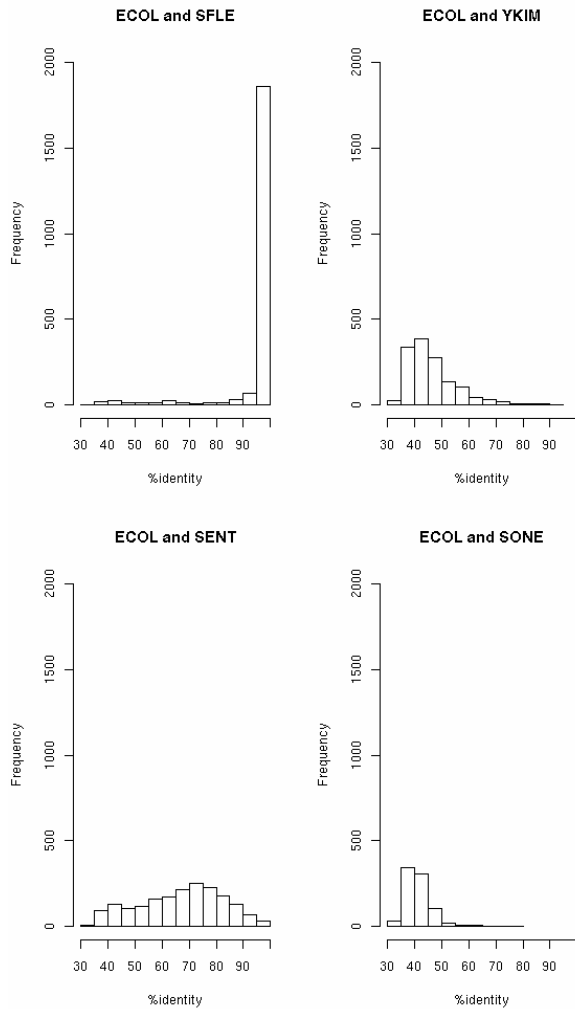


Supplementary data for:

Software to perform automated microbial
species comparisons using pairwise percent
identities

Sean Conlan and Lee Ann McCue

F1 - Histograms of percent identities between *E. coli* and other species.



T1 - Analysis of ECOL alignments

seq_pair	N	mean	s.d.	Q1	Median	Q3	IQR	skew
ECOL_SFLE	2106	95.84	11.22	97.98	99.17	100.00	2.02	-0.18
ECOL_SENT	1875	66.85	15.45	55.71	68.86	78.46	22.75	-0.16
ECOL_YKIM	1367	46.54	9.45	39.84	44.20	50.30	10.46	0.17
ECOL_SONE	817	41.31	4.84	38.20	40.48	43.33	5.13	0.11

The values in T1 were produced by the **analyze.identity.pl** script and illustrate how the quartile skewness coefficient (see E1) can be used as an indicator of asymmetry. Negative values of skew are "tail left" and positive values of skew are "tail right". See random sequence alignments below (F3 and T2) for comparison.

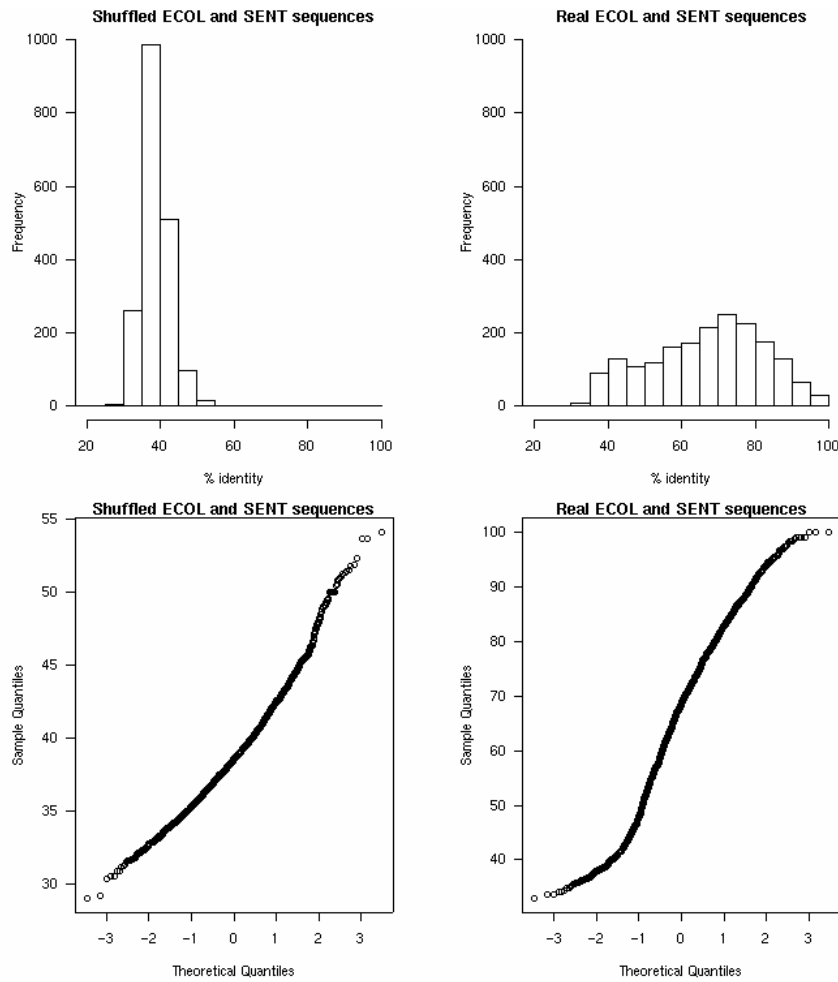
$$\text{skew} = \frac{(Q_3 - Q_2) - (Q_2 - Q_1)}{(Q_3 - Q_1)} \quad \text{E1}$$

where Q_s denote the interquartile ranges

The comparison of ECOL and SENT orthologous intergenic regions demonstrates that

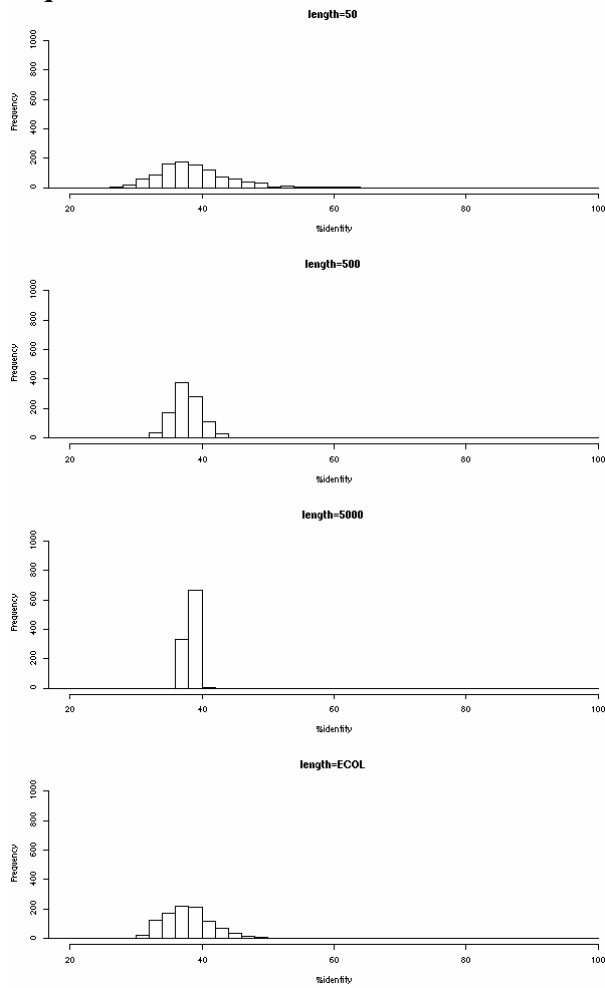
the summary statistics can suggest overdispersion of the data (note the large s.d.) and how a visual inspection of the % identity histogram could be used to detect the possibility of more than one distribution.

F2 - Shuffled and real percent identity distributions visualized using histograms or quantile-quantile plots.



The quantile-quantile plot of the percent identities for real sequences vs a normal distribution (F2, lower right plot) indicates the possibility of two distributions by the apparent change in slope around the -1 theoretical quantile.

F3 - Histograms of percent identity distributions for alignments of random sequences

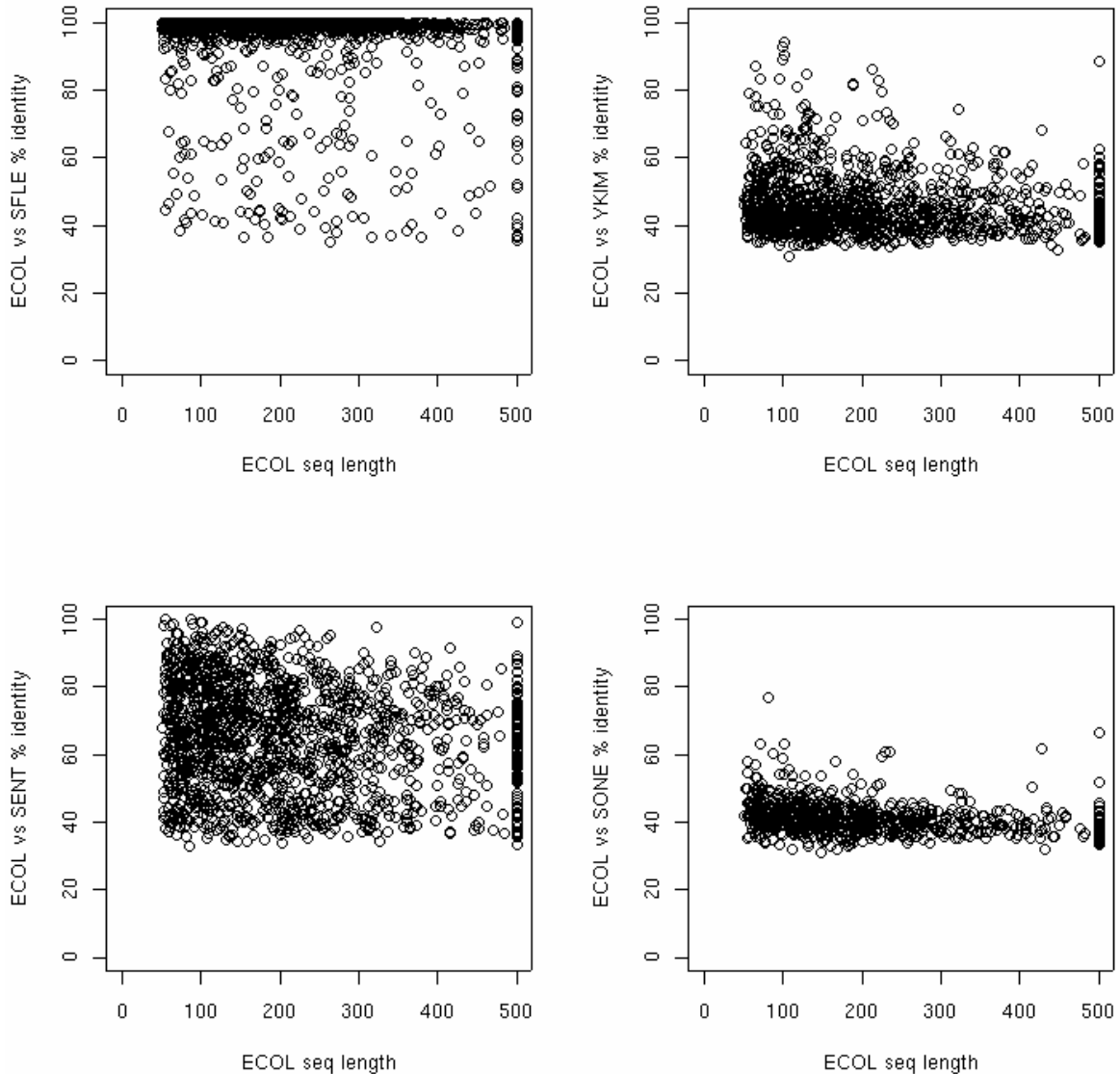


T2 - Summary statistics for alignments of random sequences

Length	N	mean	s.d.	Q1	Median	Q3	IQR	skew
50	1000	38.88	5.33	35.42	38.09	41.86	6.44	0.17
500	1000	37.74	2.13	36.33	37.55	39.05	2.72	0.10
5000	1000	38.31	0.70	37.84	38.30	38.80	0.96	0.04
ECOL	1000	37.82	3.57	35.20	37.66	40.00	4.80	-0.02

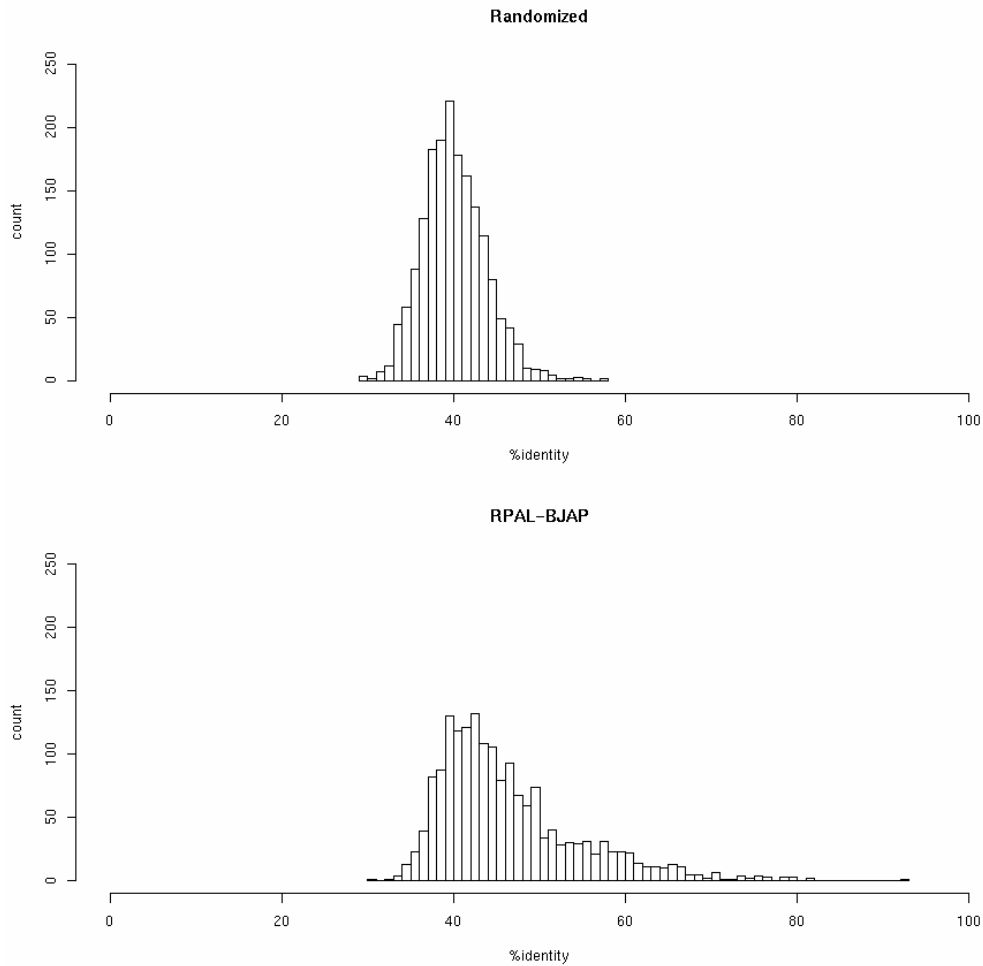
For comparison, sets of random sequences with nucleotide compositions similar to *E. coli* (30% A, 30% T, 20% G, 20% C) were generated with fixed lengths (50,500,5000) or with lengths drawn from the real *E. coli* sequence lengths. These were then aligned with Gap and the distribution of percent identities was analyzed. These data demonstrate how the width (standard deviation and IQR) and skew vary with sequence length. The median and mean, however, do not vary significantly as a function of sequence length and this is further shown in F4.

F4 - Percent identity between sequences as a function of sequence length



These plots illustrate that there is not a strong correlation between the length of the *E. coli* sequence and mean percent identity of the alignment with another species. Measurements pile up at 500 nucleotides because all sequences were truncated to a maximum length of 500 nucleotides. The *collect.identity.pl* program produces a *lengths.txt* file, containing the sequence lengths, that can be used to evaluate whether the percent identities for a given set of sequences show a dependence on length.

F5 - Example of using the percent identity distributions in a comparative study



In this example, alignments of orthologous intergenic sequences from *Rhodopseudomonas palustris* and *Bradyrhizobium japonicum* show that the sequences may be drawn from two distributions, centered near 40% and 60% identity (bottom panel). Shuffled sequences (top panel) show the expected distribution, based on composition and sequence length alone. In Conlan et al. (1), *cis*-regulatory motifs were predicted using a Gibbs sampling strategy, and when motifs were found in sequences from these two species, they were held to more stringent significance criteria if the input sequences showed correlation (>50%).

1. Conlan et al. Applied and Environmental Microbiology 71(11):7442-52 (2005)