A Computational Tool for the Genomic Identification of Regions of Unusual Compositional Properties and Its Utilization in the Detection of Horizontally Transferred Sequences

Catherine Putonti,† Yi Luo,* Charles Katili,* Sergey Chumakov,*‡ George E. Fox,†§ Dan Graur,† and Yuriy Fofanov*†*

*Department of Computer Science, University of Houston; †Department of Biology and Biochemistry, University of Houston; ‡Department of Physics, University of Guadalajara, Guadalajara, Jalisco, Mexico; and §Department of Chemical Engineering, University of Houston

Similarity Plot (S-plot) is a Windows-based application for large-scale comparisons and 2-dimensional visualization of compositional similarities between genomic sequences. This application combines 2 approaches widely used in genomics: window analysis of statistical characteristics along genomes and dot-plot visual representation. S-plot is effective in identifying highly similar regions between genomes as well as regions with unusual compositional properties (RUCPs) within a single genome, which may be indicative of horizontal gene transfer or of locus-specific selective forces. We use S-plot to identify regions that may have originated through horizontal gene transfer through a 2-step approach, by first comparing a genomic sequence to itself and, subsequently, comparing it to the genomic sequence of a closely related taxon. Moreover, by comparing these suspect sequences to one another, we can estimate a minimum number of sources for these putative xenologous sequences. We illustrate the uses of S-plot in a comparison involving *Escherichia coli* K12 and *E. coli* O157:H7. In O157:H7, we found 145 regions that have most probably originated through horizontal gene transfer. By using S-plot to compare each of these regions with 277 completely sequenced prokaryotic genomes, 1 sequence was found to have similar compositional properties to the *Yersinia pseudotuberculosis* genome, indicating a transfer from a *Yersinia* or *Yersinia* relative. Based upon our analysis of RUCPs in O157:H7, we infer that there were at least 53 sources of horizontally transferred sequences.

Introduction

The identification of unusual compositional patterns within a genome is of great importance in evolutionary studies, because the existence of such regions may be indicative of either horizontal (or lateral) gene transfer or exceptional selective forces operating on particular molecular entities (e.g., Karlin 1998; Campbell et al. 1999; Kaper and Hacker 1999; Eisen 2000; Hacker and Kaper 2000; Karlin and Mrázek 2000; Ochman et al. 2000; Karlin 2001). The rate at which various processes affect the genome typically varies between different genomic regions. Consequently, local statistical characteristics are frequently associated with particular evolutionary events. For instance, genomic islands and pathogenicity islands have been discovered on the basis of unusual compositional characteristics within a genomic sequence, for example, di- and trinucleotide frequencies, G + C content, and amino acid biases, using sliding windows (Karlin 2001).

Dot-plot analysis (Gibbs and McIntyre 1970) was the first method in the literature intended to illustrate similarity or dissimilarity between biological sequences. The points on the 2-dimensional dot-plot matrix indicate identity of the subsequences in sliding windows along the 2 sequences. Using a small window size necessitates a very large matrix resulting in images that are extremely difficult to interpret and whose biological significance is difficult to assess due to background noise and random matches. Conversely, when a large window size is used, dot-plots can be generated much faster but such plots may fail to represent true sequence similarity because of possible gaps, inversions, and other rearrangements. Although available dot-plot applications (Sonnhammer and Durbin 1996; Junier and Pagni 2000; Rice et al. 2000; Huang and Zhang 2004) can quickly compare short sequences, longer sequences (>1.5 Mbp) require sacrificing accuracy and/or specificity. A fast, accurate dot-plot analysis of genomes with a length greater than 10 Mbp on a standard desktop computer remains impractical due to the computational limitations (Huang and Zhang 2004).

Here, we introduce Similarity Plot (S-plot), an application combining 2 approaches widely used in genomics: window analysis of statistical characteristics along genomes and dot-plot visual representation. Through a 2-step approach of using S-plot first to compare a genomic sequence with itself and later to compare it with a closely related species, we identify regions that may have originated through horizontal gene transfer. Moreover, by comparing these suspect sequences to one another, we can estimate a minimum number of sources for these putative xenologous sequences.

We illustrate the uses of S-plot in a study involving *Escherichia coli* K12 (NC_000913) and *E. coli* O157:H7 (NC_002695). The famous laboratory strain *E. coli* K12 is a nonpathogenic facultative anaerobe that mostly colonizes the lower gut of mammals. In comparison, *E. coli* O157:H7 is an enterohemorrhagic bacterium, which is responsible for tens of thousands of cases of food-borne gastroenteritis and hemolytic uremic syndrome each year. The genome of the pathological O157:H7 strain has been previously shown to contain numerous clusters of genes suspected to be of xenologous origin (Perna et al. 2001). More generally, it has recently been shown that horizontal gene transfer is an extremely important mechanism driving microbial genome diversification, with at least 14% of all open reading frames in 116 prokaryotic genomes being derived through horizontal transfer (Nakamura et al. 2004).

Key words: horizontal (lateral) gene transfer, sequence composition, *Escherichia coli* K12, *Escherichia coli* O157:H7.

E-mail: dgraur@uh.edu.

Mol. Biol. Evol. 23(10):1863–1868. 2006
doi:10.1093/molbev/msl053
Advance Access publication July 7, 2006

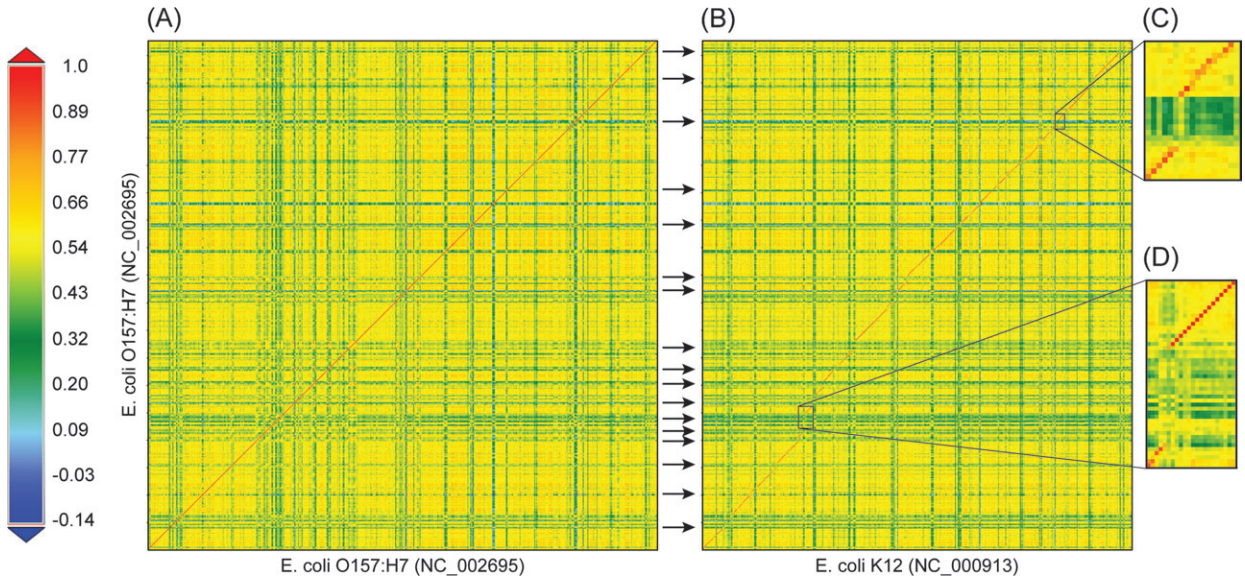

FIG. 1.—S-plot for pathogenic *Escherichia coli* O157:H7 versus itself (*A*) and versus the nonpathogenic *E. coli* K12 (*B*). The black arrows indicate some of the RUCPs identified in (*A*). Several regions that appear unusual with respect to the O157:H7 genome are not present in K12 and are most probably recent insertions (e.g., *C* and *D*).

Materials and Methods

Similarity Plot

To assess the degree and pattern of similarity (or dissimilarity) between 2 genomic sequences of size M_1 and M_2, we divide the genomes into windows of length w and slide these windows along each genome with steps (the distance between the start of 2 neighboring windows) of size s. In the simplest case, $w = s$, that is, the windows do not overlap one another, and we have approximately M_1/w and M_2/w different windows for genomes 1 and 2, respectively. As a measure of similarity, we use the Pearson correlation coefficient between the frequencies of n-mers (short subsequences of length n). The distribution $P(S)$ of appearances of all possible n-mers inside a given window is $P(S)=N_S/(w-n+1)$, where N_S and $w - n + 1 \approx w$ are, correspondingly, the number of appearances of n-mer S and the total number of n-mers in a window. The total number of all possible n-mers is 4^n. If $w = 4^n$, most N_S will be equal to 0 or 1, and in this case the frequencies of appearance may be unsuitable for the correlation analysis. Therefore, to collect representative statistics, one has to impose the condition $w > 4^n$.

In the application described here, the Pearson correlation coefficient is used to quantify the degree of similarity between the distributions of short n-mers present in the 2 genomes. To visualize the similarity, we plot the matrix of correlation coefficients $C(j,k)$ between the distributions of n-mers, where j is a window in the first genome and k is a window in the second genome. In figure 1, the vertical and horizontal coordinates represent the location of windows j and k, respectively. Different correlation coefficients are represented on the plots by different colors.

The estimated time complexity of this approach is $O[M_1+M_2+(M_1M_2)/w^2]$. An application to generate S-plots using the C# language for Windows was created. It takes only 60–100 s to create an S-plot for a pair of microbial genomes of size ~5 Mb on a standard 1-GHz PC (for performance analysis, see Fig. S1 in Supplementary Material online).

A 2-Step Procedure for Identifying Regions Acquired through Horizontal Gene Transfer

A 2-step procedure has been developed to identify regions that may have originated through horizontal gene transfer (HGT). By first comparing a genome against itself, the degree of homogeneity in a genome (a_g) can be determined as the average correlation value of the matrix C. The degree of similarity of each window with respect to its own genome (a_w) can also be calculated as the average of the correlation coefficients for the window's particular row (or column, because both averages are equivalent) in matrix C. Likewise, the standard deviations for these averages, σ_g and σ_w, can be determined. It was previously shown through studies of di- and trinucleotide frequencies (Karlin 1998) that windows of the same genomic sequence will have similar biases toward the presence/absence of particular di- and trinucleotides. Thus, one may expect that the frequency of presence of longer n-mers will follow a pseudo–random distribution. Through Monte Carlo simulations (results not shown), the values of a_w were found to follow a normal distribution. Variation from this distribution is expected due to the fact that the presence/absence of some n-mers is correlated to a particular function. In such a case, one would expect to see far more highly correlated windows than lowly correlated windows. Thus, very few, if any, windows are expected to have an a_w value smaller than or greater than 2 or 3 standard deviations from the genomic mean, a_g. Because it is our intent to identify foreign DNA within a genomic sequence, windows that are unusually dissimilar to the rest of the genome into which they are embedded ($a_w \leq a_g - 2\sigma_g$) are of particular interest. We refer to such windows as "regions of unusual compositional properties" or RUCPs.

Table 1
Number of Windows with an Average outside the Genomic Average for *Escherichia coli* O157:H7 versus Itself

	$<a_g - 3\sigma_g$	$<a_g - 2\sigma_g$	$<a_g - \sigma_g$	$>a_g + \sigma_g$	$>a_g + 2\sigma_g$	$>a_g + 3\sigma_g$
$w = s = 5{,}000$	28	60	151	70	0	0
$w = s = 1{,}000$	26	218	850	816	37	0

In the second step of this method, an S-plot for the comparison of the first genomic sequence with a closely related genome is generated. The RUCPs may or may not be present in the genomic sequence of the closely related species. Those RUCPs that appear in both genomic sequences may or may not have been introduced through horizontal gene transfer. In the case of the latter, the horizontal gene transfer event is presumed to have preceded the speciation event; thus, it is not possible to determine how long prior to speciation the region had been acquired. An RUCP in one sequence that does not have a corresponding RUCP in the other genome must have been either introduced through horizontal gene transfer into the first genome or precisely excised from the close relative after the divergence of the 2 species. Corresponding windows or a lack thereof can be determined by referencing the S-plot graphic and/or the matrix of correlation coefficients.

Results and Discussion

Test Case: RUCPs in *E. coli* O157:H7

In the first step, S-plot was used to compare the pathogenic *E. coli* O157:H7 genome against itself. The frequency distribution of 6-mers was first considered. Although the S-plot application has the ability to compare sequences with a window size of 500–50,000 nt, to uphold the condition $w > 4^n$, we first used a window and step size of 5,000 nt in length ($w = s = 5{,}000$), in which both the original and complementary strands were considered (fig. 1*A*). As indicated by the color scale of this figure, regions of high similarity appear in red whereas regions of high dissimilarity appear in green or blue. The S-plot for $w = s = 1{,}000$ for 6-mers was also generated. Although the condition $w > 4^n$ is no longer satisfied, by reducing the window or step size, we can more precisely determine the area of unusual compositional patterns in the RUCPs found at a greater window size. Both a_w and σ_w were calculated for each of the 1,099 windows ($w = s = 5{,}000$) and 5,498 windows ($w = s = 1{,}000$), as well as the average and standard deviation for the *E. coli* O157:H7 genome (a_g and σ_g). Windows with a_w greater than $a_g + \sigma_g$, $a_g + 2\sigma_g$, and $a_g + 3\sigma_g$ and windows with a_w less than $a_g - \sigma_g$, $a_g - 2\sigma_g$, and $a_g - 3\sigma_g$ were identified for both S-plots (table 1). None of the 1,000 nt sub-windows ($w = s = 1{,}000$) found within the 60 larger windows have an a_w value greater than a_g. In fact, the majority of the 218 windows ($w = s = 1{,}000$, $a_w < a_g - 2\sigma_g$) are found within one of the 60 larger windows ($w = s = 5{,}000$) with an a_w value smaller than 2 standard deviations away from the mean. This observation is expected due to the method in which the S-plot application accesses the similarity between windows; because each window where $w = 5{,}000$ contains 5 windows of $w = 1{,}000$ ($s = 1{,}000$), any subwindows of $w = 1{,}000$ that are unusual will impact the average distribution of the larger window for which it is contained. Thus, for the *E. coli* O157:H7 genome, all windows in which $a_g - 2\sigma_g \leq a_w$ are considered RUCPs. The genes encoded within the RUCP regions include genes with several functional or putative associations, hypothetical genes, and the highly conserved structural RNAs (tRNAs and rRNAs).

Regions Unique to the O157:H7 Genome

The nonpathogenic and pathogenic *E. coli* genomic sequences were then compared for 6-mers with $w = s = 1{,}000$ and $w = s = 5{,}000$. Figure 1*A* shows the S-plot of *E. coli* O157:H7 versus itself, seen next to figure 1*B* that shows the S-plot of *E. coli* O157:H7 versus *E. coli* O157:H7 for $w = s = 5{,}000$. From the S-plot, one can readily identify regions of high similarity or alignment in addition to numerous insertions within this alignment, such as those in figure 1*C* and *D*. These insertions suggest that either the windows in K12 corresponding to those present in the pathogenic O157:H7 genome have been lost or the windows in O157:H7 have been gained, most probably from horizontal gene transfer after the divergence between the 2 organisms from a common ancestor. Regions of insertion were identified by analyzing the matrix of correlation coefficients for $w = s = 5{,}000$ and $w = s = 1{,}000$; if a window in the *E. coli* O157:H7 genome did not have a corresponding window (correlation > 0.7) in the *E. coli* K12 genome, the window was considered to be an insertion in the O157:H7 genome. A comparison of the S-plot generated for the 2 *E. coli* genomes at $w = s = 5{,}000$ revealed 76 separate insertions in the alignment consisting of 340 windows. At $w = s = 1{,}000$, there are 1,483 windows comprising 147 insertions, some only 1,000–2,000 nt long flanked by highly similar windows. These insertion regions correspond closely to those previously identified (Hayashi et al. 2001).

Many windows contained within the insertion regions are RUCPs (some of which are indicated by black arrows) identified from the comparison of *E. coli* O157:H7 to itself. Ninety-nine of the 244 RUCPs at $w = s = 1{,}000$ have a corresponding window in the *E. coli* K12 genome. The remaining 145 windows, thus, are unusual to both *E. coli* genomes and most likely were obtained by the pathogenic *E. coli* from another organism through horizontal gene transfer after its divergence from the nonpathogenic *E. coli*. Using the annotation files available from National Center for Biotechnology Information (NCBI), we identified the genes located within each of these windows. The complete listing of these genes is available in Table S1 in Supplementary Material online. Seventy-five percent of the genes located within

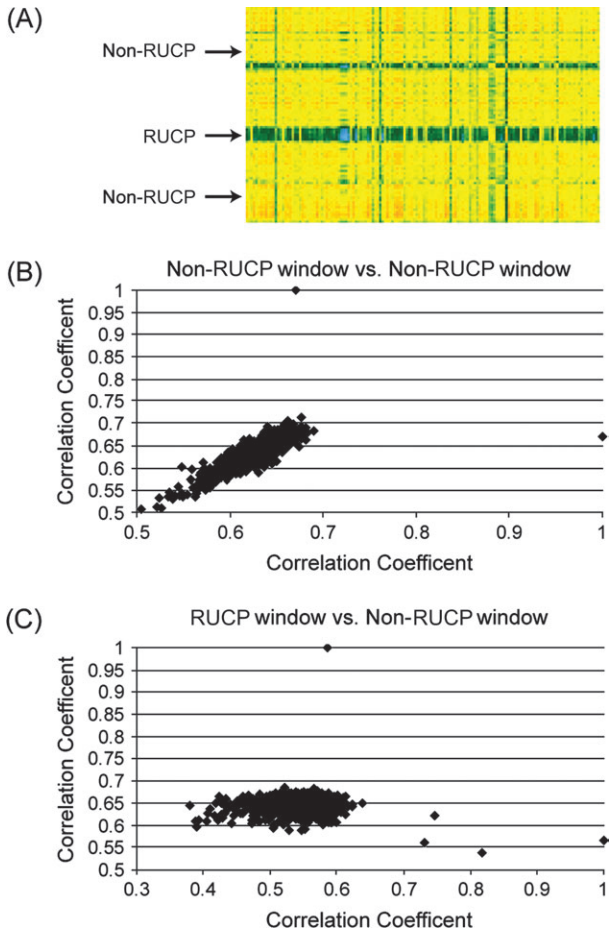

FIG. 2.—Similarity of windows from the same source and from different sources. In S-plot (*A*) one can observe that the 2 rows labeled "non-RUCP" in this small portion of the *Escherichia coli* O157:H7 versus itself ($w = s = 5{,}000$) look similar to each other and dissimilar to the row labeled "RUCP." The set of correlation coefficients for the first non-RUCP window (row) is plotted against the set of correlation coefficients for the second non-RUCP window (row) in (*B*) showing essentially a one-to-one correspondence in their correlations. In (*C*), we show a visualization of the correspondence between the first non-RUCP window and the RUCP window. The one-to-one correspondence observed in (*B*) is no longer observed. (The 2 outlying points in both [*B*] and [*C*] with correlation coefficients of 1 indicate when the window is compared with itself vs. the comparison of the two windows.)

these windows are annotated as hypothetical or putative genes. As was expected, no structural RNAs were found within these windows as such genes appear unusual due to high sequence conservation rather than HGT acquisition.

The locations of the 145 RUCPs were compared with the insertion sequence (IS), prophage, and prophage-like elements identified by Hayashi et al. (2001). Sixteen of the 80 IS elements identified by Hayashi et al. (2001) were contained within the RUCPs, and an additional 45 were within 10 Kb off the start/end positions of an RUCP. Of the 18 prophages, only 1 prophage, Sp7 (Hayashi et al. 2001), was not included in an RUCP, but because this element is the only one for which no comment was provided, it was impossible for us to determine the reason for it not being contained within an RUCP. All 6 of the prophage-like elements were contained within RUCPs. A table listing the locations of the IS, prophage, and prophage-like elements within the 145 RUCPs is provided as Table S1 in Supplementary Material online. Thirty RUCPs that had not been identified in Hayashi et al. (2001) have been Blasted against the entire GenBank database. The results are shown in Table S3 in Supplementary Material online.

Determining the Source of an RUCP

To determine the sources of the RUCP windows, we first compared each of the 145 RUCPs with 277 complete prokaryotic genomes (24 archaean and 253 bacterial) available from the NCBI database using S-plot. A listing of the 277 genomes is included in Table S2 in Supplementary Material online. The majority of the RUCPs had moderate to low (<0.5) correlation with the windows of the microbial genomes. One RUCP (position 270000–270999 bp), however, has a 0.9558 correlation to the first 1,000-bp window of the *Yersinia pseudotuberculosis* genome. This RUCP contains the *Rhs*G element (Wang et al. 1998), but the annotation of the corresponding window in *Y. pseudotuberculosis* does not list a homologous gene. The nucleotide sequence of the RUCP window extracted from the *E. coli* O157:H7 genome sequence and the window in *Y. pseudotuberculosis* were aligned using an in-house dot-plot application, and partial sequence similarity was observed. The RUCP sequence was then Blasted against the GenBank sequence database revealing that this 1,000-bp sequence is unique to the *E. coli* O157 strain based upon the observation that that the alignment between the 1,000-bp region and other *E. coli* strains are confined to the *vgrG* gene sequence. An S-plot of *Y. pseudotuberculosis* against itself was generated. Because the window that is highly correlated to the *E. coli* RUCP does not appear unusual with respect to the *Y. pseudotuberculosis* genome, that is, the RUCP sequence still carries the *n*-mer compositional attributes of the *Y. pseudotuberculosis* genome rather than *E. coli*, it is most likely that the RUCP sequence was acquired from an organism belonging to the *Y. pseudotuberculosis* species cluster or a close relative thereto, but, of course, we cannot exactly deduce where it came from. The existence of another bacterial species with a similar 6-mer distribution in its genome may be ruled out on probabilistic grounds.

Because none of the other 144 RUCPs were found to have a corresponding window in the 277 genomes considered, one of two conclusions can be drawn. First, it is possible that the remaining 144 were acquired from an organism whose genome is not available in the NCBI database and, thus, not included in our comparisons. On the other hand, if the region was introduced through horizontal gene transfer a long time ago, the selective forces applied to both the RUCP and the donating organism may have introduced significant variation such that the RUCP and its corresponding window have become less correlated and, thus, unidentifiable.

We can, however, predict the minimum number of sources from which RUCPs were acquired. The ability to do so is based upon the assumption that windows from the same source will be relatively similar/dissimilar to the same windows. When identifying RUCPs, our attention was focused on those windows appearing as dark green

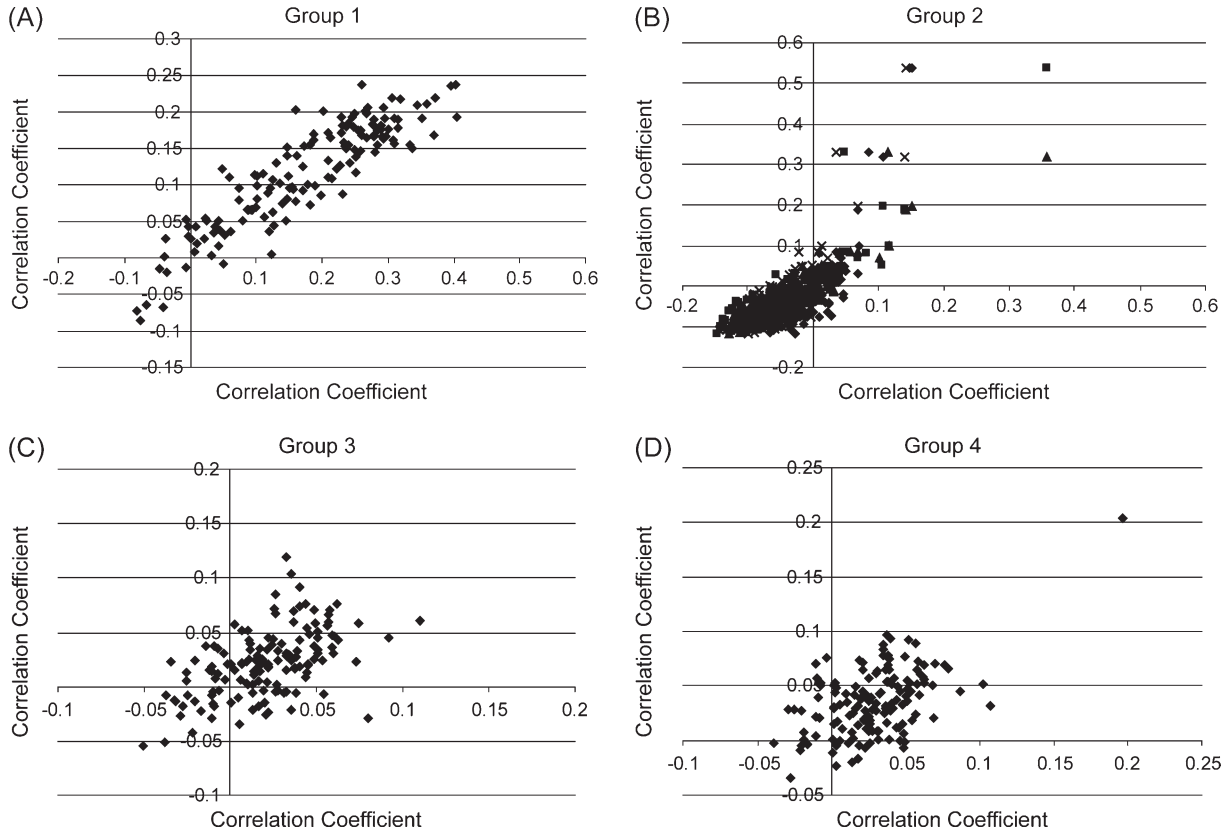

FIG. 3.—Four clusters of RUCPs, each predicted to be from the same source. Group 1, which comprises 87 RUCPs, is represented in (*A*) by comparing the set of correlations for one member of the group against the set of correlations for another member. (*B*) Group 2 shows 4 of its members against the fifth member. (*C*) Group 3 and (*D*) Group 4 each contains just 2 RUCPs.

or blue lines traversing the S-plot. Inspection of the S-plot also reveals that there are several rows (or columns) that appear similar to several other rows (or columns), that is, a yellow pixel is followed by a green pixel in both rows (or columns) followed by another yellow pixel and so on. In figure 2*A*, a small section of the *E. coli* O157:H7 versus itself S-plot ($w = s = 5{,}000$) is shown. The 2 rows indicated as "non-RUCP" windows look very similar, that is, the column in which one non-RUCP window is green and the other non-RUCP window is also green. Although there are several other rows, which also resemble these 2 non-RUCP windows, the row labeled RUCP looks dramatically different. The similarity/dissimilarity between any 2 rows can be visualized by taking the set of correlation coefficients for one row (or column) and graphing it against the set of correlation coefficients for another row (or column), as shown in figure 2*B* and *C*. As can be observed in these 2 graphs, the non-RUCP windows have a more similar set of correlations whereas a non-RUCP and RUCP window have less similar sets of correlations. We can quantify this similarity by determining the r^2 value of the linear regression on this graph (the correlation of correlations). The correlation for the 2 non-RUCP windows (fig. 2*B*) is 0.6318, whereas the correlation of the non-RUCP versus the RUCP windows (fig. 2*C*) is 0.0012. Because we believe that the RUCP window is from a source other than *E. coli* O157:H7, we can postulate that windows from the same source, as is the case with the 2 non-RUCP windows, will have a higher correlation of correlations than windows from different sources.

From the comparison of *E. coli* O157:H7 versus itself, a matrix C^R was created from the matrix of correlation coefficients C, in which only the intersections of the 145 RUCP windows were included. (If the 1,000-bp sequences of the 145 RUCP windows were to be concatenated and compared with itself using S-plot, the resulting matrix of correlation coefficients C would be identical to the matrix C^R described here.) The correlation of the correlations for each pair of RUCP windows was then calculated, the average correlation of correlations being 0.2555. Clusters of pairs were identified revealing 4 distinct groups comprising 96 of the 145 RUCPs. Group 1 is the largest cluster with 87 of the RUCP windows, which are spread throughout the entire *E. coli* O157:H7 genome. Figure 3*A* shows the plot of the set of correlations for one member of this group against another member of the group. The average correlation of correlations for all members of Group 1 is 0.5948. Group 2 has 5 RUCPs (fig. 3*B*) with an average correlation of correlations of 0.5480. Included in this group is the RUCP found to be highly similar to the *Y. pseudotuberculosis* window. The other 4 members of this group, however, do not have a corresponding window in the *Y. pseudotuberculosis* genome (maximum correlation with any window of *Y. pseudotuberculosis* genome is 0.3270). It is possible that these 4 RUCPs were also acquired from *Yersinia* through horizontal gene transfer but due to some reason, for

example, the regions are no longer contained within *Y. pseudotuberculosis* genome or selective pressure has impacted the 4 regions altering their base composition more than the other RUCP, S-plot cannot identify the corresponding windows within the *Yersinia* genomes. Both Groups 3 (fig. 3*C*) and 4 (fig. 3*D*) consist of just 2 RUCPs with 0.5896 and 0.7443 correlations, respectively. The remaining 49 RUCP windows show no similarity to the other RUCP windows. The set of correlations for several of these windows were plotted against the set of correlations of other RUCP windows resulting in graphs very similar to that shown in figure 2*C* (average correlation of 0.0947). From the results of our analysis, we predict that there should be at least 53 sources from which *E. coli* O157:H7 acquired pieces DNA through horizontal gene transfer.

Conclusions

S-plot provides an attractive solution for visualizing statistical similarity or dissimilarity within and between genomes. S-plot comparisons have been conducted for several different pairs of genomic sequences within a genus as well as across genera, for example, comparisons of *E. coli* and *Shigella flexneri* or *E. coli* and *Salmonella enterica*. The closer the 2 genomes are in sequence similarity and nucleotide composition, the more informative the S-plot image and underlying matrix will be. The genomic sequences of more distant relatives may result in an S-plot with little or no highly correlated windows, for example, *E. coli* and *Bacillus subtilis*; thus, the S-plot comparison of 2 such genomic sequences is not informative. Due to the performance of the application, it is easy for the user to explore the possible genomic sequences for which informative comparisons can be obtained outside the genus of the organism of interest. S-plots for all of these 3 comparisons can be found in Figure S2 in Supplementary Material online.

The 2-step process suggested here can be used to identify regions introduced through horizontal gene transfer in completely sequenced genomes, for which the complete sequence of a closely related taxon is available. If an RUCP appears unusual in both the genome of interest and its close relative, the region may have been acquired prior to their divergence. The possibility of 2 separate events for each of the 2 genomes is, of course, less likely. Comparison with other closely related genomes may provide further insight into such regions as well as assist in inferring their phylogeny. The S-plot Windows-based application is freely available for download along with user documentation at www.bioinfo.uh.edu/splot.

Supplementary Material

Supplementary Figures S1 and S2 and Tables S1, S2, and S3 are available at *Molecular Biology and Evolution* online (http://www.mbe.oxfordjournals.org/).

Acknowledgments

We would like to express our gratitude to the Texas Learning and Computation Center for their support. C.P. was supported in part by a training fellowship from the Keck Center for Computational and Structural Biology of the Gulf Coast Consortia (NLM Grant No. 5T15LM07093). G.E.F. was supported in part by grants from National Aeronautics and Space Administration (NNG05GN75G) and the Institute of Space Systems Operations. D.G. was supported in part by the Grants to Enhance and Advance Research Program at the University of Houston. We thank Drs Ricardo Azevedo, Audrey Hart-van Tassell, Bogdan Nowicki, Stella Nowicki, and Petri Urvil for their comments.

Literature Cited

Campbell A, Mrázek J, Karlin S. 1999. Genome signature comparisons among prokaryote, plasmid and mitochondrial DNA. Proc Natl Acad Sci USA 96:9184–9.

Eisen AJ. 2000. Horizontal gene transfer among microbial genomes: new insights from complete genome analysis. Curr Opin Genet Dev 10:606–11.

Gibbs AJ, McIntyre GA. 1970. The diagram, a method for comparing sequences. Its use with amino acid and nucleotide sequences. Eur J Biochem 16:1–11.

Hacker J, Kaper J. 2000. Pathogenicity islands and the evolution of microbes. Annu Rev Microbiol 54:641–79.

Hayashi T, Makino K, Ohnishi M, et al. (22 co-authors). 2001. Complete genome sequence of enterohemorrhagic Escherichia coli O157:H7 and genomic comparison with a laboratory strain K-12. DNA Res 8:11–22.

Huang Y, Zhang L. 2004. Rapid and sensitive dot-matrix methods for genome analysis. Bioinformatics 20:460–6.

Junier T, Pagni M. 2000. Dotlet: diagonal plots in a Web browser. Bioinformatics 16:178–9.

Kaper JB, Hacker J, editors. 1999. Pathogenicity islands and other mobile virulence elements. Washington, DC: ASM Press.

Karlin S. 1998. Global dinucleotide signatures and analysis of genomic heterogeneity. Curr Opin Microbiol 1:598–610.

Karlin S. 2001. Detecting anomalous gene clusters and pathogenicity islands in diverse bacterial genomes. Trends Microbiol 9:335–43.

Karlin S, Mrázek J. 2000. Predicted highly expressed genes of diverse prokaryotic genomes. J Bacteriol 182:5238–50.

Nakamura Y, Itoh T, Matsuda H, Gojobori T. 2004. Biased biological functions of horizontally transferred genes in prokaryotic genomes. Nat Genet 36:760–6.

Ochman H, Lawrence JG, Groisman EA. 2000. Lateral gene transfer and the nature of bacterial innovation. Nature 405:299–304.

Perna NT, Plunkett G III, Burland V, et al. (28 co-authors). 2001. Genome sequence of enterohaemorrhagic *Escherichia coli* O157:H7. Nature 409:529–33.

Rice P, Longden I, Bleasby A. 2000. EMBOSS: the European Molecular Biology Open Software Suite. Trends Genet 16: 276–7.

Sonnhammer ELL, Durbin R. 1996. A dot-matrix program with dynamic threshold control suited for genomic DNA and protein sequence analysis. Gene 167:GC1–10.

Wang Y-D, Zhao S, Hill CW. 1998. Rhs elements comprise three subfamilies which diverged prior to acquisition by Escherichia coli. J Bacteriol 180:4102–10.

Takashi Gojobori, Associate Editor

Accepted June 14, 2006