Regions of Unusual Statistical Properties as Tools in the Search for Horizontally Transferred Genes in *Escherichia coli*

Catherine Putonti[1,2], Sergei Chumakov[3], Arturo Chavez[3], Yi Luo[1], Dan Graur[2], George E. Fox[2] and Yuriy Fofanov[1,2]

[1]*Department of Computer Science, University of Houston, Houston, TX 77204-3058, USA*
[2]*Department of Biology and Biochemistry, University of Houston, Houston, TX 77204-5001, USA*
[3]*Department of Physics, University of Guadalajara, Guadalajara, Jalisco 44420, Mexico*

Abstract. The observed diversity of statistical characteristics along genomic sequences is the result of the influences of a variety of ongoing processes including horizontal gene transfer, gene loss, genome rearrangements, and evolution. The rate at which various processes affect the genome typically varies between different genomic regions. Thus, variations in statistical properties seen in different regions of a genome are often associated with its evolution and functional organization. Analysis of such properties is therefore relevant to many ongoing biomedical research efforts. Similarity Plot or S-plot is a Windows-based application for large-scale comparisons and 2D visualization of similarities between genomic sequences. This application combines two approaches wildly used in genomics: window analysis of statistical characteristics along genomes and dot-plot visual representation. S-plot is effective in detecting highly similar regions between two genomes. Within a single genome, S-plot has the ability to identify highly dissimilar regions displaying unusual compositional properties. The application was used to perform a comparative analysis of 50+ microbial genomes as well as many eukaryote genomes including human, rat, mouse, and drosophila. We illustrate the uses of S-Plot in a comparison involving *Escherichia coli* K12 and *E. coli* O157:H7.

Keywords: horizontal gene transfer, sequence composition, *Escherichia coli* K12, *Escherichia coli* O157:H7.
PACS: 87.23.–n

SIMILARITY PLOT (S-PLOT)

To assess the degree and pattern of similarity (or dissimilarity) between two 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 two neighboring windows) of size s. In the simplest case, $w=s$, i.e., 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

CP854, *Medical Physics: Ninth Mexican Symposium on Medical Physics,*
edited by H. Mercado-Uribe, G. García-Torales, J. L. Flores-Núñez, and E. Castillo-Corona

will be equal to zero or one, 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 two 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

$$O\left(M_1 + M_2 + \frac{M_1 M_2}{w^2}\right) \quad (1)$$

An application to generate S-plots using the C# language for Windows was created. It takes only 60-100 seconds to create an S-plot for a pair of microbial genomes of size ~5 Mb on a standard 1GHz PC.

A two-step procedure has been developed to identify regions that may have originated through horizontal gene transfer. First a genome is compared to itself in order to identify regions unusual with respect to the genome in which they are contained. The average correlation coefficient for each window (a_w) is calculated as is the average for the genome, the average of all windows (a_g). Through Monte Carlo simulations (results not shown), the values of a_w were found to follow a normal distribution. Windows with an average greater than or less than two standard deviations from the genomic mean are identified. We refer to such windows as "regions of unusual compositional properties" or RUCPs. 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. A RUCP in one sequence that does not have a corresponding RUCP in the other genome must have either been introduced through horizontal gene transfer into the first genome or precisely excised from the close relative after the divergence of the two species. Corresponding windows or a lack thereof can be determined by referencing the S-plot graphic and/or the matrix of correlation coefficients.

RESULTS

In the first step, S-plot was used to analyze the pathogenic *E. coli* O157:H7 genome against itself. The frequency distribution of 6-mers was considered using a window and step size of 5000 nucleotides in length (w=s=5000) in which both the original and complementary strands were considered (Figure 1A). As indicated by the color scale of this figure, regions of high similarity appear in red while regions of high dissimilarity appear in green or blue. The S-plot for w=s=1000 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 RUSPs found at a greater window size.

The non-pathogenic and pathogenic *E. coli* genomic sequences were then compared for 6-mers with and $w=s=5000$ (Figure 1B) and $w=s=1000$. From the S-plot, one can readily identify regions of high similarity or alignment in addition to numerous insertions within this alignment. These insertions suggest that either the windows in K12 corresponding to those present in the pathogenic O157:H7 genome have been lost or that the windows in O157:H7 have been gained, most probably from HGT, since the divergence of the two organisms. Regions of insertion were identified by analyzing the matrix of correlation coefficients for $w=s=5000$ and $w=s=1000$; 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.

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. 99 of the 244 RUCPs at $w=s=1000$ 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 HGT after its divergence from the non-pathogenic *E. coli*. Using the annotation files available from NCBI, we identified the genes located within each of these windows. 75% of the genes located within these windows are annotated as hypothetical or putative genes. As was expected, no structural proteins were found within these windows.

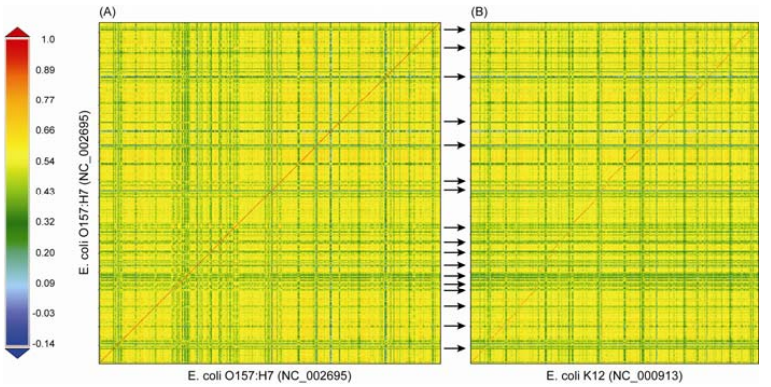

FIGURE 1. S-plot for pathogenic E. coli O157:H7 versus itself (A), and versus the non-pathogenic E. coli K12 (B). The black arrows indicate some of the RUCPs identified in A.

ACKNOWLEDGMENTS

We would like to express our gratitude to the Texas Learning and Computation Center (TLCC) for partial support of this work. CP's work 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).