9-17-2018

S-plot2: Rapid Visual and Statistical Analysis of Genomic Sequences

Laurynas Kalesinskas
Evan Cudone
Yuriy Fofanov
Catherine Putonti

Loyola University Chicago, cputonti@luc.edu

Recommended Citation

This Article is brought to you for free and open access by the Faculty Publications at Loyola eCommons. It has been accepted for inclusion in Biology: Faculty Publications and Other Works by an authorized administrator of Loyola eCommons. For more information, please contact ecommons@luc.edu.
Creative Commons License
This work is licensed under a Creative Commons Attribution-Noncommercial-No Derivative Works 3.0 License.
© The Authors 2018
S-plot2: Rapid Visual and Statistical Analysis of Genomic Sequences

Laurynas Kalesinskas¹,², Evan Cudone¹,³, Yuriy Fofanov⁴ and Catherine Putonti¹,²,⁵

¹Bioinformatics Program, Loyola University Chicago, Chicago, IL, USA. ²Department of Biology, Loyola University Chicago, Chicago, IL, USA. ³Department of Mathematics and Statistics, Loyola University Chicago, Chicago, IL, USA. ⁴Department of Pharmacology and Toxicology, The University of Texas Medical Branch at Galveston, Galveston, TX, USA. ⁵Department of Computer Science, Loyola University Chicago, Chicago, IL, USA.

ABSTRACT: With the daily release of data from whole genome sequencing projects, tools to facilitate comparative studies are hard-pressed to keep pace. Graphical software solutions can readily recognize synteny by measuring similarities between sequences. Nevertheless, regions of dissimilarity can prove to be equally informative; these regions may harbor genes acquired via lateral gene transfer (LGT), signify gene loss or gain, or include coding regions under strong selection. Previously, we developed the software S-plot. This tool employed an alignment-free approach for comparing bacterial genomes and generated a heatmap representing the genomes’ similarities and dissimilarities in nucleotide usage. In prior studies, this tool proved valuable in identifying genome rearrangements as well as exogenous sequences acquired via LGT in several bacterial species. Herein, we present the next generation of this tool, S-plot2. Similar to its predecessor, S-plot2 creates an interactive, 2-dimensional heatmap capturing the similarities and dissimilarities in nucleotide usage between genomic sequences (partial or complete). This new version, however, includes additional metrics for analysis, new reporting options, and integrated BLAST query functionality for the user to interrogate regions of interest. Furthermore, S-plot2 can evaluate larger sequences, including whole eukaryotic chromosomes. To illustrate some of the applications of the tool, 2 case studies are presented. The first examines strain-specific variation across the Pseudomonas aeruginosa genome and strain-specific LGT events. In the second case study, corresponding human, chimpanzee, and rhesus macaque autosomes were studied and lineage specific contributions to divergence were estimated. S-plot2 provides a means to both visually and quantitatively compare nucleotide sequences, from microbial genomes to eukaryotic chromosomes. The case studies presented illustrate just 2 potential applications of the tool, highlighting its capability to identify and investigate the variation in molecular divergence rates across sequences. S-plot2 is freely available through https://bitbucket.org/lkalesinskas/splot and is supported on the Linux and MS Windows operating systems.

KEYWORDS: comparative genomics, alignment-free, gene transfer, gene loss

Background

Modern sequencing technologies can quickly and affordably produce genomic sequences for species across the tree of life. Consequently, many new lineages and poorly resolved areas of the tree have been identified.¹-³ With tens of thousands of bacterial genomes now publicly available, comparative genomics has produced numerous insights into microbial life.⁴ Several tools are currently used to detect genome similarity through sequence alignment.⁵-⁸ In addition, tools employing a graphical “dot plot” approach, such as Gepard,⁹ Serolis,¹⁰ and SeqTools’ Dotter,¹¹ can highlight genomic similarities and rearrangements as well as gene duplications. These tools, however, have their limitations: Serolis¹⁰ is limited in the size of sequence it can analyze (4kb), and Dotter¹¹ is significantly slower than Gepard⁹ for larger sequences. Nevertheless, alignment-free approaches, including the aforementioned “dot plot” tools, have a significant advantage over alignment-based methods: they are less computationally expensive (regarding both time and resources) and impervious to synteny-related problems (see reviews of Vinga and Almeida¹² and Bonham-Carter et al¹³).

In addition to sequence similarities, the dissimilarities between genomic sequences can be equally informative.¹⁴ These dissimilarities can indicate strain-specific genes horizontally/laterally acquired rather than vertically inherited. Lateral gene transfer (LGT) is an important force in the evolution of prokaryotes,¹⁵ including the exchange of defense mechanisms and virulence factors.¹⁵,¹⁶ Although certainly less prevalent (and fiercely debated), LGT between eukaryotes and prokaryotes can also occur.¹⁷-²⁰ Disparities between genomic sequences can also be the result of gene loss, another pervasive and often significant driver of evolution in prokaryotic¹¹,²² and eukaryotic species²³ (see review Albalat and Cañestro²⁴). Moreover, recognition of substantial sequence divergence between orthologous gene sequences can signify genes under strong selection (see review Long et al²⁵). Such genes can provide insight into phenotypic differences between species.²⁶,²⁷

Previously we developed S-plot, a tool for the rapid analysis and visualization of bacterial genomic sequences.²⁸ This tool was applied to the examination of Escherichia,²⁸ Bacillus,²⁹ and Neisseria genomes, identifying regions of unusual nucleotide
Evolutionary Bioinformatics

composition corresponding to LGT events. Herein, we present the next generation of this tool: S-plot2. Similar to its predecessor, S-plot2 creates an interactive, 2-dimensional heatmap. Similar to the aforementioned dot-plot tools, S-plot2 captures the similarities in nucleotide usage between genomic sequences, but unique to this tool is the fact that it also captures the dissimilarities in nucleotide usage between genomic sequences. Through the examination of nucleotide usage, phylogenetic signals can be uncovered.31 In S-plot2, whole eukaryotic chromosomes and smaller prokaryotic genomes can be efficiently compared. Furthermore, the new version includes functionality to extract, analyze, and automate BLAST queries of regions of interest within the heatmap. This facilitates the investigation of quickly evolving coding regions, novel coding regions, and laterally transferred elements.

Implementation

Developed in Java, S-plot2 performs pairwise comparisons of genomic sequences (partial or complete) via a sliding window approach. Windows can be of a user-defined length (the "genome approach") or confined to annotated coding regions (the "gene-by-gene approach"). The "genome approach" permits windows to be either adjacent or overlapping. Regardless of the approach selected, each window’s $k$-mer (subsequence of length $k$) frequencies are enumerated. The similarity/dissimilarity between 2 windows is calculated based on these $k$-mer frequencies, using either the Pearson ($r$) or Spearman rank ($\rho$) correlation coefficient. The resulting values for each pairwise window comparison are then graphed as a 2-dimensional heatmap using Glimpse32 (eg, Figure 1A). Windows with a similar $k$-mer usage are represented in the heatmap using colors at one end of the color spectrum, whereas windows with dissimilar $k$-mer usage are represented by colors at the other end of the spectrum. Draft genome sequences that include several scaffold sequences can be examined using the "genome approach" in S-plot2. The scaffolds can be concatenated, separated by, eg, Ns, into a single FASTA sequence. S-plot2 does not calculate frequencies for windows in which greater than half of the sequence is not A, T, C, or G; thus, windows containing more than one scaffold will be ignored. The "gene-by-gene approach" is a new feature released in S-plot2, as is the Spearman rank correlation coefficient metric for sequence comparison.

Functionality has been developed in S-plot2 to aid in the interpretation of the heatmap. Users can specify regions of interest based on window coordinates or select windows meeting specific criteria (eg, regions exhibiting aberrant $k$-mer usage) and then output or BLAST33 these regions. For instance, a cluster of genes which appears in one genome and not the other can be identified and investigated.

Figure 1. Comparison of Pseudomonas aeruginosa PAO1 (x-axis) and PA7 (y-axis) genomes. (A) "Genome approach" comparison with a window and offset of 5000bp. (B) Genomic island present with the PA7 strain. (C) "Gene-by-gene approach" comparison of protein-coding gene sequences annotated for the 2 genomes in panel A (*.faa files). Here, the window size is equivalent to a single coding region and $k=3$ is evaluated (the same color bar as shown in panel A). The comparisons conducted here for both approaches were done using the Pearson correlation coefficient. Sequence similarity is measured by the frequency of shared $k$-mers, with green signifying low similarity and red signifying high similarity.
other (indicative of a gene loss/gain), such as that shown in Figure 1B, could be queried; in the case in which a gene was acquired via LGT, the putative source can be identified. Queries to National Center for Biotechnology Information's (NCBI) eUtils API were automated using JEuUtils. All BLAST queries in S-plot2 use the blastn algorithm and remotely query the NCBI nucleotide collection (nr/nt) database. Users can also output statistics computed for the heatmap as well as generate multi-FASTA format files for windows with an r or p value within a user-defined range. The heatmap image itself can be saved to file as a TIF file, implemented using the iCafe package.

An executable jar file, sample sequence data, and a tutorial are freely available through https://bitbucket.org/lkalesinskas/splot. S-plot2 was tested thoroughly on the Windows and Ubuntu operating systems. Due to the lack of support for compatibility profiles on MacOS, rendering and maneuvering within the S-plot2 heatmap are suboptimal (due to incompatibilities with the Glimpse visualization version used) on MacOS. Exploration of the S-plot2 heatmap (scrolling through a sequence, zooming in/out, etc.) was optimized for use with the mouse on Windows and Ubuntu.

As the similarity between windows is calculated based on the correlation (either the Pearson or Spearman rank correlation coefficient) of the frequency of shared k-mers, the condition $\frac{4^k}{w} \ll 1$ where w is the window size must be followed. If $w \geq 4^k$, then most k-mer frequencies will be 0 or 1 and thus unsuitable for the correlation analysis. Run time and memory usage are dependent on the number of windows. For a genome of size M, the number of windows, n, is $M/w$. Thus, for smaller window sizes, a larger heatmap will be generated. For each window, k-mer frequencies are enumerated for the original and reverse-complement sequences. Calculation of k-mer usage is linear and values are stored in a sorted array. The run time and memory usage estimate is $O(n^2)$. For instance, the heatmap generated in Figure 1A was generated in 42 seconds. It is important to note that for large sequences, the required RAM may exceed the RAM allocated or available for the Java Virtual Machine (particularly if the user has a 32-bit version installed) in which case the application will not execute. Nevertheless, a complete human chromosome can be compared using less than 8 GB of RAM in a matter of minutes; S-plot2's performance is significantly faster than other graphical alignment-free available graphical tools.

Results and Discussion

To illustrate the functionality and utility of S-plot2, we conducted 2 case studies. In addition to providing a visualization of the genomic sequences under investigation, the new functionality developed in S-plot2 can lead to a deeper understanding of the variation in molecular divergence rates across sequences.

Case study 1: exploring the evolution of bacterial genomes

The genomes of the opportunistic bacterial pathogen *Pseudomonas aeruginosa* are highly mosaic and include regions of genomic plasticity. The *P. aeruginosa* accessory genome exceeds that of its core genome. Figure 1 shows the pairwise comparison of the *P. aeruginosa* strains PAO1 (NC_002516) and the known "taxonomic outlier" for the species, PA7 (NC_009656). Two comparisons were conducted: the "genome approach" using a fixed window size (Figure 1A) and the "gene-by-gene approach" in which each window is an individual gene (Figure 1C). As even closely related *P. aeruginosa* strains can be distinguished by single-nucleotide polymorphisms, indels, and inversions, it is thus not surprising to observe genomic variation between the PAO1 and PA7 genomes (Figure 1A and C). The nucleotide sequence of the PA7 region shown in Figure 1B was investigated using S-plot2's automated BLAST functionality. This region includes numerous transposases and integrases as well as plasmid- and phage-associated genes. It corresponds to the previously identified genomic island RGP42 within the *P. aeruginosa* PA7 genome. The region shown in Figure 1B is but one of the many genomic islands within these 2 strains. Users can recognize windows of unusual composition visually via the "genome approach" or individual genes of interest via the "gene-by-gene approach" and BLAST the sequences. Furthermore, S-plot2 can automatically identify such regions and BLAST their sequences.

Recombination within *P. aeruginosa* species is frequent and previous research has found variation in the evolutionary histories of regions of the *P. aeruginosa* genome. To exemplify how S-plot2 can be used to investigate recombination, 7 genomes included in the comparative genomic study of Dettman et al were selected (Table 1) and pairwise comparisons were performed. Sequence similarity was assessed for each window size of 5000 bp (base pairs) for $k=6$ using the Pearson correlation coefficient. Figure 2 (panels B, C, and D) shows the pairwise comparisons for PAO1 and C3719, LESB58, and PACS2, respectively. These heatmaps illustrate the presence/absence of unique regions within the genomes and, most notably, rearrangements. The matrices generated by S-plot2 were saved and contiguous 0.2 Mbp regions along the PAO1 genome were evaluated. Thus, an alignment-free approach was used to identify and quantify similarity/dissimilarity between homologous regions of the PAO1 genome and other *P. aeruginosa* strains. As shown in Figure 2A, different regions of the PAO1 genome are represented by different topologies. Consistent with prior alignment-based analyses, we find that the evolution of the *P. aeruginosa* genome is not uniform across the entire genome sequence. In this fashion, S-plot2 can provide evidence of evolution across a genome sequence both visually and quantitatively.
### Table 1. Seven *Pseudomonas aeruginosa* genomes examined.

<table>
<thead>
<tr>
<th>STRAIN</th>
<th>GENOME SIZE, MBP</th>
<th>NO. OF SCAFFOLDS</th>
<th>NO. OF CODING REGIONS</th>
<th>ASSEMBLY</th>
</tr>
</thead>
<tbody>
<tr>
<td>PAO1</td>
<td>6.26</td>
<td>1</td>
<td>5572</td>
<td>GCA_000006765</td>
</tr>
<tr>
<td>LEB58</td>
<td>6.60</td>
<td>1</td>
<td>6041</td>
<td>GCA_00026645</td>
</tr>
<tr>
<td>C3719</td>
<td>6.22</td>
<td>1</td>
<td>5648</td>
<td>GCA_000152525</td>
</tr>
<tr>
<td>PACS2</td>
<td>6.49</td>
<td>1</td>
<td>5913</td>
<td>GCA_000168335</td>
</tr>
<tr>
<td>JD316</td>
<td>6.19</td>
<td>1882</td>
<td>6590</td>
<td>GCA_000506125</td>
</tr>
<tr>
<td>JD317</td>
<td>6.49</td>
<td>2043</td>
<td>6979</td>
<td>GCA_000506145</td>
</tr>
<tr>
<td>JD320</td>
<td>6.41</td>
<td>2038</td>
<td>6876</td>
<td>GCA_000506165</td>
</tr>
</tbody>
</table>

Sequences were retrieved for genomes (*_genomic.fna.gz*) and coding sequences (*_cds_from_genomic.fna.gz*).

**Figure 2.** Evolution of the *Pseudomonas aeruginosa* chromosome. (A) Comparison of cluster topologies based on sequence similarity based on 6-mer usage for window size = offset size = 5000 bp over 0.2 Mbp regions of the PAO1 genome. Heatmaps for (B) PAO1 vs C3719, (C) LESB58, and (D) PACS2. The same color scale as Figure 1 is used here: sequence similarity is measured by the frequency of shared k-mers, with green signifying low similarity and red signifying high similarity.
Case study 2: exploring the evolution of primate chromosomes

S-plot2 is also capable of evaluating whole eukaryotic chromosomes. As such, it can be used to estimate chromosome-specific molecular divergence rates, estimate lineage specific contributions to divergence, and identify regions that are significant contributors to observed divergence. As a case study of S-plot2, we performed pairwise comparisons for all homologous human, chimpanzee, and rhesus autosomes (window size = offset size = 100 Kbp for \( k = 6 \) using the Pearson correlation coefficient). Each chromosome was also compared with itself using the same window size, offset size, and \( k \). This self-sequence comparison provides a baseline for the variation within a chromosome relative to that observed between species (see Supplemental File 1). Prior whole genome comparison studies between human and chimpanzee found \( \approx 1.4\% \) sequence divergence\(^43\) and 23 inversions,\(^44\) as well as other differences (for a review, see the work by Kehrer-Sawatzki and Cooper\(^45\)). Sequence analysis of human-chimpanzee chromosome pairs suggests that recombination, proximity to telomeres, bias in repair mechanisms, and GC content are all exerting influence on genetic variation.\(^46\text{–}50\)

Here, we present a comparison between human chromosome 17, chimpanzee chromosome 17, and rhesus chromosome 16. As the heatmaps in Figure 3 show, the pericentric inversion previously found between these sequences\(^44\) can be identified through the pairwise comparisons of the human, chimpanzee, and rhesus autosomes. The heatmaps for these 3 pairwise comparisons, however, do not readily present how
these chromosomes are evolving. For instance, the differences observed between the homologous human and chimpanzee chromosomes may be the result of changes within the chimpanzee chromosome or changes within the human chromosome. Comparisons of both chromosomes to the rhesus chromosome let us distinguish between these 2 scenarios. If we oversimplify the process of species divergence to a single point in time (thus ignoring subsequent gene flow), one could assume that the chromosomal sequences are essentially identical. Thus, for a window in the human chromosome, its homologous window in the chimpanzee genome would have the same sequence (and thus nucleotide composition). As such, the heatmap for an individual chromosome compared with itself would be indiscernible from the comparison of the chromosome to its homolog. Post-speciation, the 2 genomes would begin to diverge and this divergence can be quantified by the cross-species comparison value (eg, human vs chimpanzee) relative to the intraspecies comparison (eg, human vs human). The matrices of \( r \) values were retrieved for each of the plots shown in Figure 3 and used to calculate the divergence between species (see Supplemental File 1 for details regarding this calculation). The inlay in Figure 3 shows the results of this calculation for human vs chimpanzee (red), chimpanzee vs rhesus (yellow), and human vs rhesus (blue). In this figure, the \( x \)-axis is representative of the divergence calculated for a window relative to its GC content. As shown in the inlay in Figure 3, regions in the human genome with a GC content \( \approx 45\% \) are the most divergent windows from chimpanzee; these regions are evolving within the human lineage.

Conclusions

S-plot2 provides a means to visually and quantitatively compare genomic sequences ranging from microbial genomes to eukaryotic chromosomes. These comparisons can be generated in a matter of seconds to minutes (depending on the size of the sequence under consideration). S-plot2 includes functionally to aid in the analyses of genomic sequences, allowing users to quickly investigate their data and test hypotheses based on either observed patterns or statistics capturing both the similarities and dissimilarities of sequences. The case studies presented highlight just some of the applications of S-plot2. Furthermore, the analyses performed for the \emph{Pseudomonas} genomes and human-chimpanzee-rhesus autosomes illustrate the variation in molecular divergence rates across sequences.

Acknowledgements

The authors would like to thank MrArya Mehrtash and Matt Lewczuk for their assistance in early software development and also thank Dr Heather Wheeler and Jon Brenner for their feedback during development. The authors also greatly appreciate the many conversations with Dr Michael Travisano on the analyses of the primate chromosomes.

Author Contributions

LK and EC implemented the application. YF and CP designed the project. LK, EC, and CP performed the analyses of the case studies. CP was a major contributor in writing the manuscript. All authors read and approved the final manuscript.

REFERENCES


