On the LZ distance for dereplicating redundant prokaryotic genomes

Raphael R. Leonard[1,2] , **Damien Sirjacobs**[2], **Eric Sauvage**[1], **Frederic Kerff**[1], **Denis Baurain**[2].

[1] Centre for Protein Engineering, University of Liege; [2] PhytoSYSTEMS, University of Liege

rleonard@doct.ulg.ac.be, denis.baurain@ulg.ac.be

The fast-growing number of available prokaryotic genomes, along with their uneven taxonomic distribution, is a problem when trying to assemble broadly sampled genome sets for phylogenomics and comparative genomics. Indeed, most of the new genomes belong to the same subset of hyper-sampled phyla, such as Proteobacteria and Firmicutes, or even to single species, such as *Escherichia coli* (almost 2000 genomes as of Sept 2015), while the continuous flow of newly discovered phyla prompts for regular updates. This situation makes it difficult to maintain sets of representative genomes combining lesser known phyla, for which only few species are available, and sound subsets of highly abundant phyla. An automated straightforward method is required but none are publicly available. The LZ distance, in conjunction with the quality of the annotations, can be used to create an automated approach for selecting a subset of representative genomes without redundancy. We are planning to release this tool on a website that will be made publicly available.

Methods

LZ complexity (Lempel and Ziv, 1976)

For each pair of genomes A and B, one out of four LZ distances is computed from the LZMA-compressed file sizes (`Compress::LZMA::External`) of the corresponding nucleotide assemblies $c(S)$ and $c(Q)$ and of their concatenations $c(SQ)$ and $c(QS)$. These distances, along with taxonomic information, are stored in a database. A clustering algorithm is then applied to regroup the similar genomes into a user-specified number of clusters. For each of these clusters, a representative genome is chosen based on the quality of the genomic assemblies (chromosomes rather than scaffolds) and of the protein annotations (e.g., few rather than many *unknown proteins*).

$$d(S,Q) = max\{c(SQ) - c(S), c(QS) - c(Q)\} \quad (1)$$

$$d(S,Q) = \frac{max\{c(SQ) - c(S), c(QS) - c(Q)\}}{max\{c(S), c(Q)\}} \quad (2)$$

$$d(S,Q) = c(SQ) - c(S) + c(QS) - c(Q) \quad (3)$$

$$d(S,Q) = \frac{c(SQ) - c(S) + c(QS) - c(Q)}{\frac{1}{2}[c(SQ) + c(QS)]} \quad (4)$$

Results

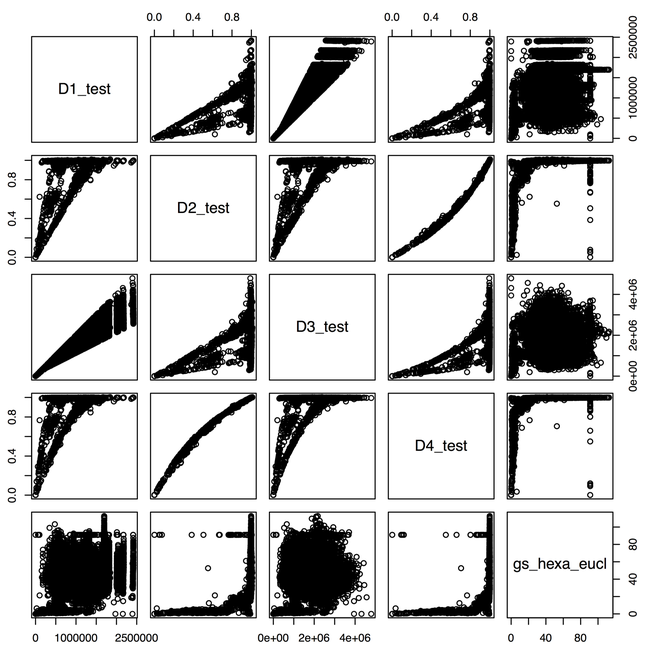
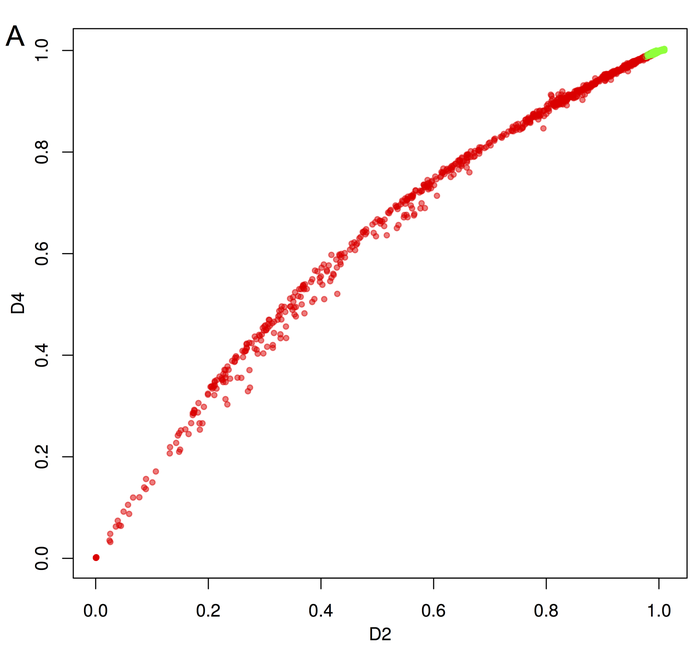


Figure 1: ***Correlation between LZ distances.*** *The correlation between the four LZ distances, and the Euclidean distance based on 6-mer odd ratios (genomic signatures), was studied on a sub-sample of 242 complete prokaryotic genomes (**bacteria_1_collection**) taken from Ensembl Bacteria release 28. Genome sequences were generated by collapsing all nucleotide sequences (chromosomes, plasmids, scaffolds, contigs) available for each assembly.*



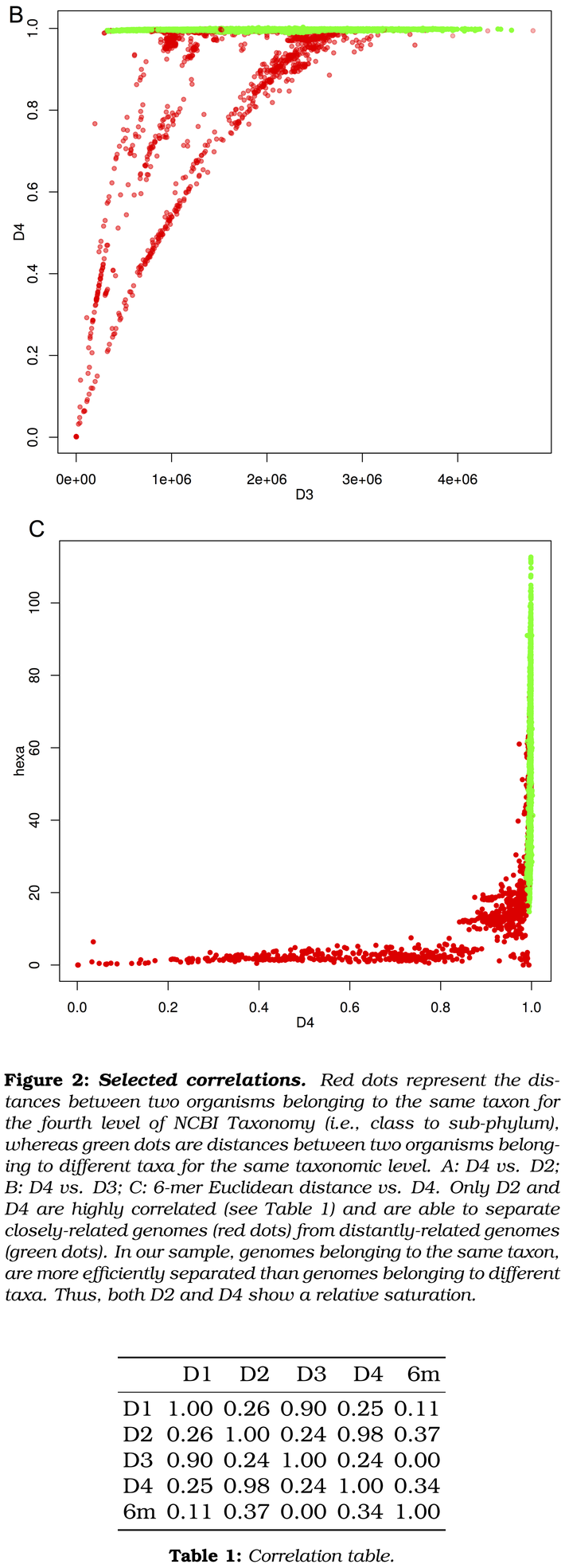


Figure 2: ***Selected correlations.*** *Red dots represent the distances between two organisms belonging to the same taxon for the fourth level of NCBI Taxonomy (i.e., class to sub-phylum), whereas green dots are distances between two organisms belonging to different taxa for the same taxonomic level. A: D4 vs. D2; B: D4 vs. D3; C: 6-mer Euclidean distance vs. D4. Only D2 and D4 are highly correlated (see Table 1) and are able to separate closely-related genomes (red dots) from distantly-related genomes (green dots). In our sample, genomes belonging to the same taxon, are more efficiently separated than genomes belonging to different taxa. Thus, both D2 and D4 show a relative saturation.*

	D1	D2	D3	D4	6m
D1	1.00	0.26	0.90	0.25	0.11
D2	0.26	1.00	0.24	0.98	0.37
D3	0.90	0.24	1.00	0.24	0.00
D4	0.25	0.98	0.24	1.00	0.34
6m	0.11	0.37	0.00	0.34	1.00

Table 1: *Correlation table.*

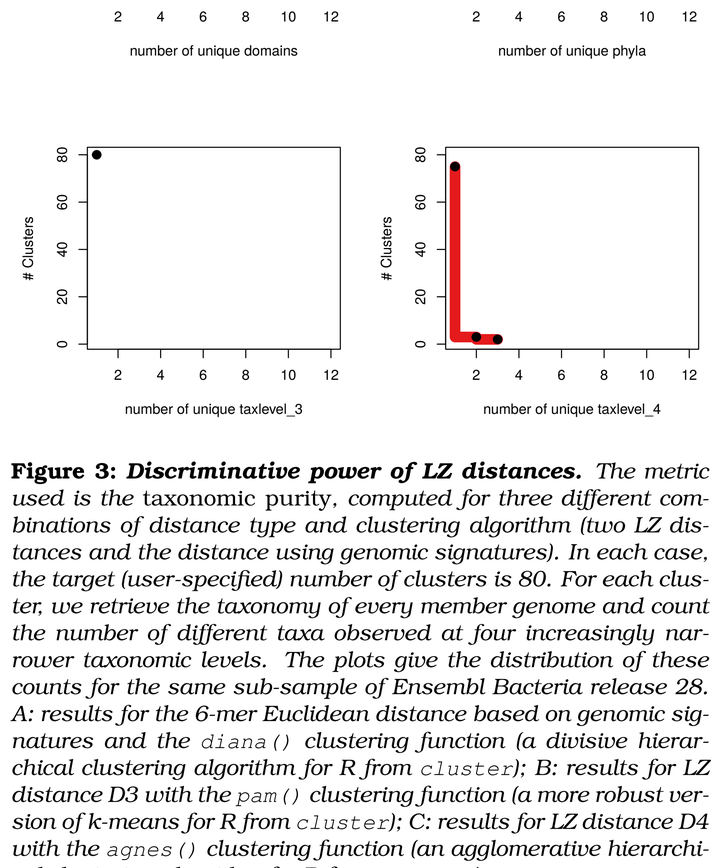


Figure 3: ***Discriminative power of LZ distances.*** *The metric used is the* taxonomic purity*, computed for three different combinations of distance type and clustering algorithm (two LZ distances and the distance using genomic signatures). In each case, the target (user-specified) number of clusters is 80. For each cluster, we retrieve the taxonomy of every member genome and count the number of different taxa observed at four increasingly narrower taxonomic levels. The plots give the distribution of these counts for the same sub-sample of Ensembl Bacteria release 28. A: results for the 6-mer Euclidean distance based on genomic signatures and the `diana()` clustering function (a divisive hierarchical clustering algorithm for R from `cluster`); B: results for LZ distance D3 with the `pam()` clustering function (a more robust version of k-means for R from `cluster`); C: results for LZ distance D4 with the `agnes()` clustering function (an agglomerative hierarchical clustering algorithm for R from `cluster`).*

Conclusions

1. The fourth variant of the Lempel-Ziv distance (D4, normalized by the mean LZ size of the concatenated genome pair) appears quite promising to cluster related prokaryotic genomes, at least for dereplication purposes.
2. This constrasts with other variants of the LZ distance, which are either only as discriminant as genomic signatures based on 6-mer odd ratios (D2, normalized by the maximum of the single LZ sizes for the genome pair) or worse (D1 and D3).
3. Unfortunately, widely available compression algorithms (e.g., `gzip`) use heuristics to speed up the compression process that reveal unsuitable for accurately computing LZ sizes. Only `lzma` (from **XZ Utils**) appears to yield consistent results but at the expense of a large computational burden (*ca.* 100 CPU hours for 242 genomes).

Benelux Bioinformatics Conference 2015, The University of Antwerp, Belgium, 7–8 December 2015

R.R. Leonard is a Ph.D. student funded by the F.R.S.-FNRS.