# Visualization and analysis of an interconnected network of genomic elements

Olivier Poirion, Bénédicte Lafay

# Visualization and analysis of an interconnected network of genomic elements

Olivier POIRION*[1] and Bénédicte LAFAY[1]

[1]Université de Lyon, CNRS UMR5005-Laboratoire Ampère, École Centrale de Lyon, 36 avenue Guy de Collongue, F-69134 Écully

June 24, 2013

**Abstract**

In the context of studying the relationships among bacterial replicons (*i.e.*, chromosomes and plasmids), we investigate various methodological approaches using a test dataset. Standard method fail to describe the complexity of the genetic events that occur in the evolution and adaptation of these elements. Given a set of genes of interest linked functionally and used as variables, the organization of beta-proteobacterial replicons was studied using several dimension reduction methods, graphs as well as supervized classification. Combinations of methods prove indispensable to characterize the relationships between replicons, permitting to identify both global trends among replicons as well as specific features of single replicon. Furthermore, our study underlines the inherent difficulty to explore exhaustively genomic data using a single tool.

**Key words**: comparative genomics, evolutionary biology, dimension reduction, clustering, graph.

## 1 Introduction

Bacterial genomes are constistuted of different types of replicons [Led98], separated into chromosomes and plasmids. The former are the essential component of the genome whereas the latter are dispensable to the host bacterium. Numerous inter- and intra-species DNA exchanges have been reported between chromosomes and plasmids [SGG$^+$09]. Interactions and recombinations between these different replicons are expected to result in a complex set of gene homologies and thus in the blurring of the inference of genome and organismal evolution. Furthermore, some bacteria harbour in their genome replicons that exhibit both chromosomal and plasmidic features [MKC04]. Because these elements may be plasmids adapted to the cell cycle [PCF$^+$12], we decided to investigate further the stabilization of mobile elements into the bacterial stable genome. We thus compared bacterial replicons irrespective of their type, using genes involved in the replicons replication and segregation to analyze the relationships between these elements.

For that purpose, it is essential to have tools and methodologies that allow the visualizing and efficient mining of the "network" of replicons as well as the evaluating of the role of the different genes in the organization of those genetic elements. As a pilot study to select and validate an appropriate analytical procedure, we focused on the Betaproteobacteria lineage that contains species (*Burkholderia, Ralstonia/Cupriavidus*) possessing three types of genomic elements: the primary or "true" chromosomes (chrI), secondary and third chromosomes (chrII and chrIII), as well as true plasmids [PCF$^+$12]. Because chrIIs and chrIIIs may actually be integrated plasmids, three types of genomic elements are thus expected to be differentiated. A clear separation between these elements would therefore reveal specific adaptations and genetic mechanism evolution involved in replication and cell cycle integration. We first describe the input dataset, then using different clustering procedures we evaluate their respective efficiency in separating the different types of replicons. Finally we assess the importance

*olivier.poirion@ec-lyon.fr

of the obtained classification and discuss on the use of additional tools: graphs, supervised classification and regression, to go deeper into the analysis of the data.

## 2 Dataset construction

Proteins involved in the replication and segregation of the replicons and the cell cycle were used to build annotated clusters of functional homologs using BLAST [CCA+09] and TRIBE-MCL [EVO02] clustering algorithm. A query dataset was constructed based on chosen proteins homolog families in ACLAME [LLMT10] and in KEGG [KGS+12] using KEGG BRITE hierarchy. The query set was then used as input in a BLAST analysis (10e-3 E-value cutoff) to identify putative homologous proteins among all Betaproteobacteria protein sequences available from the Genbank database [GMBG+10] on 30/11/2012. Then all-vs-all BLAST analysis was conducted and the resulting score matrix was used as input to TRIBE-MCL to form clusters of homologous proteins. A very high rate of functional homologs by cluster was obtained, making us confident in the MCL outputs (the relevance of the clusters will not be discussed here in detail). Each replicon was then considered as a vector of size $n$ with $n$ the number of clusters of proteins. For a given replicon and its corresponding vector $Xi$, the $j$-th element of this vector $Xi[j]$ is equal to the number of proteins assigned to the $j$-th cluster, which belongs to this replicon. Our final dataset contained **304** replicons and **501** variables.

## 3 Projection of the dataset

Several dimension reduction approaches were tested: Principal Component Analysis (PCA), Self-Organized-Network (SOM) [Koh82], multi-dimensional scaling (MDS) [TdSL00] and the ISOMAP [TdSL00]. The projection result for each method was coupled with a kmeans clustering procedure to retrieve the formed subsets. A relative high number of clusters, common to all procedures, was arbitrarily chosen as input to the kmeans algorithm to allow all the fragmentations to be observed as unique clusters on the projections. The consistency of the obtained clusters was evaluated using the homogeneity score introduced by Rosenberg and Hirschberg [**?**]. This score is part of the V-measure and evaluates the quality of a clustering solution with respect to a reference clustering. We assessed the homogeneity for each clustering result and for each type of genomic elements. Projections results are presented in Figure 1 and scores are reported in Table 1. All the projection algorithms used produced a clear separation of chrI and plasmids. Although the chrII/chrIII positioning could be somewhat ambiguous, they appeared to be more closely related to plasmids than to chrI. SOM performs better in term of biological coherence of the obtained clusters as well as in terms of bacterial taxonomy and species biology, and is thus expected to offer a more acurate two-dimensional representation of the data.

## 4 Graph

We tested an alternative data exploration procedure. The input dataset (replicons and protein clusters) was transformed into bipartite graphs considering only the links between replicons and variables and Gephi [BHJ09] was used to visualize the graphs. Spatialization algorithms were used to represent graphs according to the graph connections. The projected graph is presented in Figure 2a. Additionally, several clustering algorithms developed for graphs were tested. We analyzed the results as previously, using the three community detection algorithms, the random walk algorithm infomap [RB08], the label propagation algorithm [CGP11] and the Newman leading eighenvector algorithm [CGP11]. Results are presented in Table 3. Infomap performs best in separating the different types of replicons. It produces good results regarding cluster homogeneity although presenting twice more clusters than other methods. The graph approach thus appears to constitute an interesting clustering alternative to dimension reduction methods described above. Moreover, when coupled with graph visualization tools such as Gephi, various informations could be extracted from the graph visually. For instance, several clusters connecting chrI and chrII/chrIII (Figure 2a), as well as a putative horizontal gene transfer (HGT) between chrI and chrII could easily be identified (Figure 2b). Furthermore, using the graph structure, we were able to identify the 20 variables most connected to ChrII/chrIII replicons (Figure 2c). Strikingly, the majority of those variable are specific to chrII/chrIII or shared with chrI. However, even for a small dataset such as the Betaproteobacteria replicons, the high level of interconnections between the replicons makes any exhaustive visual analysis labor intensive.

## 5 Supervized classification

Despite specificities of chrII/chrIII, these replicons seem to share numerous similarities with plasmids.

| | chrI | chrII/chrIII | plasmid | number of clusters |
|---|---|---|---|---|
| ACP | 0.93 | 0.69 | 0.72 | 15 |
| MDS | 0.88 | 0.69 | 0.68 | 15 |
| ISOMAP | 0.93 | 0.75 | 0.70 | 15 |
| SOM | 0.94 | 0.80 | 0.70 | 15 |
| KMEANS | 0.93 | 0.59 | 0.68 | 15 |

**Table 1:** *Homogeneity score results for kmeans. The last row are the scores obtained using kmeans alone.*

| | chrI | chrII/chrIII | plasmid | number of clusters |
|---|---|---|---|---|
| infomap | 0.97 | 0.82 | 0.85 | 28 |
| Newman | 0.91 | 0.59 | 0.68 | 8 |
| Label propagation | 0.03 | 0.05 | 0.01 | 6 |

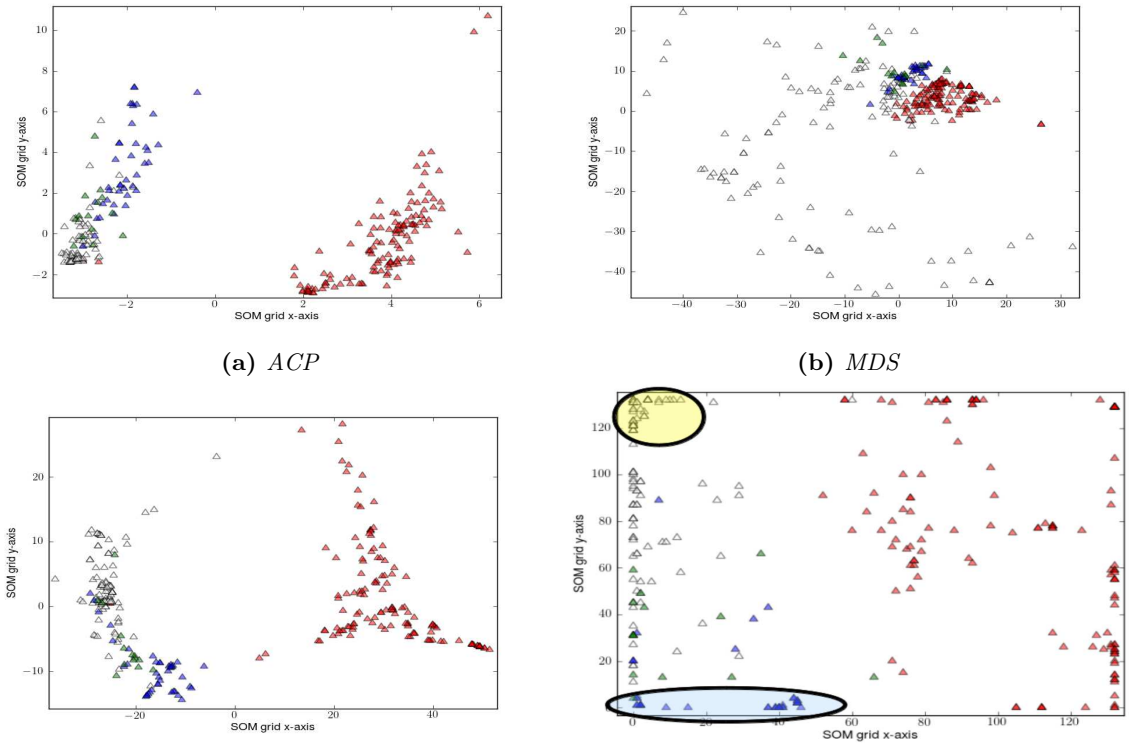
**Table 2:** *Homogeneity score results for graph community detection algorithms.*

Moreover, different degrees of specificities may occur among them. In particular, chrIII seem to be more closely related to plasmids than to chrII and, in reverse, some plasmids seem to be closer to chrII/chrIII. We thus attempted to detect ambiguous plasmids using supervized classification. First, a valid and coherent training set must be carefully selected to build a classification model. The difficulty here is to choose non ambiguous plasmids and chrII/chrIII. From the SOM results, we selected two unambigous groups of chrII/chrIII and one group of plasmids (Figure 1d). We then built an extremely-randomized-trees classifier [GEW06] trained using these groups. Compared to other learning algorithms (SVM), extremely-randomized-trees present several advantages. It has few parameters to fit, it easily estimates the class probabilities (the predicted class probabilities of a given replicon is computed as the mean predicted class of the trees in the forest) and it seems to give more biologically coherent results. This approach revealed that 3 of the 14 chrIII in our dataset are classified among plasmids while 14 of the 117 plasmids are classified among chrII/chrIII. The biological implications of these results will not be discussed here. Yet, the ambiguous replicons underlined by the classification were coherent with our expectations and allowed us to confirm and draw hypotheses on the origin and evolution of those replicons.

# 6 Discussion

Here we studied several analytical procedures aiming to characterize different types of genomic elements according to a given set of genes linked functionally. Working on a subset of bacteria: the Betaproteobacteria, known to harbor distinct types of replicons, we highlighted clear separations of those replicons indicating different distributions and specificities relatively to the set of genes chosen, *i.e.*, the genes involved in the replication and the segregation of the replicons. Graphs proved very useful in the visualizating and manipulating of the data, permitting to focus on particular replicons and/or gene cluster, and to identify visually trends and specificities. Finally, a supervised classification approach underlined bias and unusual replicons. However, our study, on such a small replicon subset, underlines the inherent difficulties to explore exhaustively the data using a single tool.

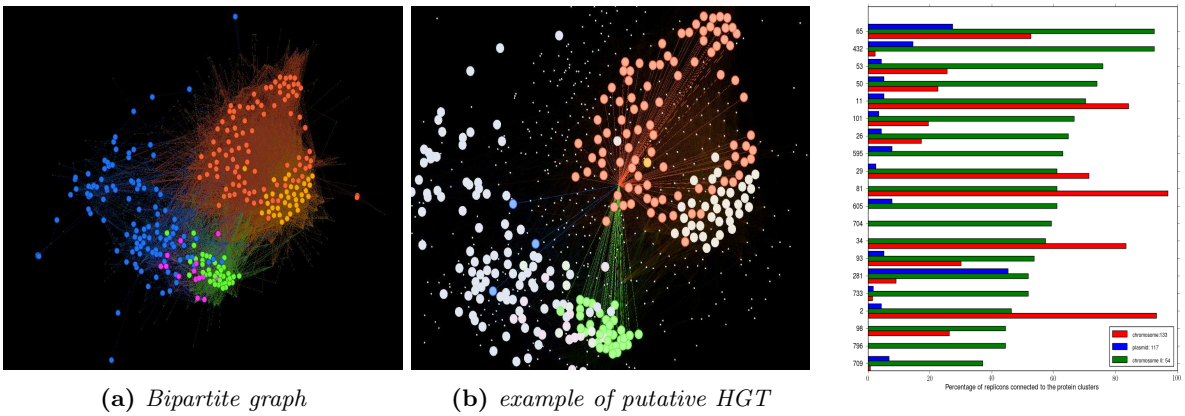
These different results provide converging evidence about the evolutionary origin of the various types of replicons in Betaproteobacteria. The existence of complex interconnections between replicons suggests that multiple and more or less specific genomic events have occurred and that every replicon is only witness to a small part of them. This study is part of a more ambitious project aiming to understand the stabilization of replicons in bacterial genomes. To analyze as exhaustively as possible the relationships between replicons, it is thus necessary to look at global trends as well as specificities of single replicons. However, we must point out the bias in the representation of bacterial species in terms of availability of sequenced genomes as well as in the incomplete knowledge of the biology of some species.




**(a)** *ACP* **(b)** *MDS*

**(c)** *ISOMAP* **(d)** *SOM*

**Figure 1:** *Projection results. Red, white, blue and green corresponds to chrI, plasmids, chrII and chrIII, respectively. In figure 1d, the two groups of replicons selected for the supervized classification analysis (section 5) are indicated in blue (chrII/chrIII groups) and yellow (plasmid group).*




**(a)** *Bipartite graph* **(b)** *example of putative HGT*

**(c)** *degrees*

**Figure 2:** *Graph results. a. Representation of the spatialized graph. chrI, plasmids, chrII and chrIII appear in orange, blue, green and pink, respectively. chrI of Burkholderia appear in light orange. b. Highlight of a particular gene cluster its specific replicons connexions. Color code is as before. In Betaproteobacteria, the gene is present on chrI except for all but one Burkholderia species. All other Burkhoderia species harbor this gene on their chrII and in one case on plasmids, thus suggesting that this gene transfered from chrI to chrII. c. Degree of the 20 clusters of proteins most connected to chrII/chrIII. Red, blue and green bars represent the average number of edges connecting a given cluster of proteins to chrI, plasmids and chrII/chrIII, respectively.*

# 7 Methods

## 7.1 Software

Python/C/C++ were used as programming languages and Mysql [Man10] as database solution. ISOMAP, MDS, the extremely-randomized-trees classifier and the homogeneity score were computed using the python library Scikit-klearn [PVG+11], whereas SOM, PCA and kmeans computation relied on part of codes from the Pycluster [dHINM04]. Graphs were visualized using Gephi [BHJ09]. The graph manipulation and community detection were performed using the igraph library [CN06].

## 7.2 Computation parameters

We chose cutoff of 10e-3 and 10e-2 for the first and the second BLAST procedures respectively. A granularity of 4 was chiosen for the TRIBE-MCL algorithm. Data were not normalized for PCA and ISOMAP computations. We used 10,000 iterations and size as defined in [Joc10] for the parameters of the grid used for the SOM. A modified euclidian distance was used for the MDS algorithm, setting a ceiled score (200) when two vectors have nothing in common (instead of 0). For the kmeans algorithm, we made 12,000 runs and selected the result with the lowest error. For the ISOMAP procedure, we used k=5, number of neighbors, which gives the best separation. For graphs visualization, we chose a graph spatialization algorithm Force Atlas 2 implemented in Gephi by default. The default parameters in igraph were used for the community detection algorithms. For the extremely-randomized-trees classifier, we chose a forest of 1000 trees and we fixed *max_feature, i.e.* the size of the random subsets of features to consider when splitting a node, to sqrt(501). All parameters were cross validated.

# References

[BHJ09] Mathieu Bastian, Sebastien Heymann, and Mathieu Jacomy. Gephi: An open source software for exploring and manipulating networks. In *International AAAI conference on weblogs and social media*, volume 2. AAAI Press Menlo Park, CA, 2009.

[CCA+09] Christiam Camacho, George Coulouris, Vahram Avagyan, Ning Ma, Jason Papadopoulos, Kevin Bealer, and Thomas L Madden. BLAST+: architecture and applications. *BMC Bioinformatics*, 10(1):421, 2009.

[CGP11] Michele Coscia, Fosca Giannotti, and Dino Pedreschi. REVIEW A Classification for Community Discovery Methods in Complex Networks. *Analysis*, 2011.

[CN06] Gabor Csardi and Tamas Nepusz. The igraph software package for complex network research. *InterJournal, Complex Systems*, 1695:38, 2006.

[dHINM04] Michiel JL de Hoon, Seiya Imoto, John Nolan, and Satoru Miyano. Open source clustering software. *Bioinformatics*, 20(9):1453–1454, 2004.

[EVO02] A J Enright, S Van Dongen, and C A Ouzounis. An efficient algorithm for large-scale detection of protein families. *Nucleic Acids Research*, 30(7):1575–1584, 2002.

[GEW06] Pierre Geurts, Damien Ernst, and Louis Wehenkel. Extremely randomized trees. *Machine learning*, 63(1):3–42, 2006.

[GMBG+10] Lewis Y Geer, Aron Marchler-Bauer, Renata C Geer, Lianyi Han, Jane He, Siqian He, Chunlei Liu, Wenyao Shi, and Stephen H Bryant. The NCBI BioSystems database. *Nucleic Acids Research*, 38(Database issue):D492–D496, 2010.

[Joc10] Barbara. P. Buttenfield Jochen Wendel. Formalizing Guidelines for Building Meaningful Self-Organizing Maps. *giscience2010.org*, 2010.

[KGS+12] Minoru Kanehisa, Susumu Goto, Yoko Sato, Miho Furumichi, and Mao Tanabe. KEGG for integration and interpretation of large-scale molecular data sets. *Nucleic acids research*, 40(Database issue):D109–14, 2012.

[Koh82] T Kohonen. Self-organized Formation of Topologically Correct Feature Maps. *Biological Cybernetics*, 43(1):59–69, 1982.

[Led98] Joshua Lederberg. Plasmid (1952–1997). *Plasmid*, 39(1):1–9, 1998.

[LLMT10] Raphaël Leplae, Gipsi Lima-Mendez, and Ariane Toussaint. ACLAME: A CLAssification of Mobile genetic Elements, update 2010. *Nucleic Acids Research*, 38(Database issue):D57–D61, 2010.

[Man10] Mysql Reference Manual. MySQL 5.0 Reference Manual. *Syntax*, page 3079, 2010.

[MKC04] Chris Mackenzie, Samuel Kaplan, and Madhusudan Choudhary. Multiple chromosomes. *Microbial Evolution*, pages 82–101, 2004.

[PCF+12] Fanny M Passot, Virginie Calderon, Gwennaele Fichant, David Lane, and Franck Pasta. Centromere binding and evolution of chromosomal partition systems in the burkholderiales. *Journal of bacteriology*, 194(13):3426–36, 2012.

[PVG+11] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[RB08] Martin Rosvall and Carl T Bergstrom. Maps of random walks on complex networks reveal community structure. *Proceedings of the National Academy of Sciences of the United States of America*, 105(4):1118–23, 2008.

[SGG+09] Steven C Slater, Barry S Goldman, Brad Goodner, João C Setubal, Stephen K Farrand, Eugene W Nester, Thomas J Burr, Lois Banta, Allan W Dickerman, Ian Paulsen, et al. Genome sequences of three agrobacterium biovars help elucidate the evolution of multichromosome genomes in bacteria. *Journal of bacteriology*, 191(8):2501–2511, 2009.

[TdSL00] J B Tenenbaum, V de Silva, and J C Langford. A global geometric framework for nonlinear dimensionality reduction. *Science (New York, N.Y.)*, 290(5500):2319–23, December 2000.