Topological representation of the within-species evolutionary structure of bacterial populations: the SLV graph

Pedro T. Monteiro, Alexandre P. Francisco,
Mário Ramirez, João A. Carriço
{ptgm,aplf}@kdbio.inesc-id.pt

Last modified: June 18, 2014

Abstract

The representation of relationships between haplotypes in a population is performed normally through the use of trees. Frequently, the selection of a single tree implies the application of maximum parsimony principles, being prone to interpretation errors. The adoption of a graph structure, rather than a tree-like structure, avoids making simplifying assumptions leaving open the number of possible interpretations. In particular, it allows for the representation of homoplasies which can represent recombination events, frequently ignored by existing methods.

Here, we propose a graph structure, denominated SLV (Single Locus Variant) graph, as an alternative to the BURST family of algorithms. This structure, connects individuals if and only if they share a single allelic difference in their MLST profile. We further explore the topological properties of such structure and propose biological interpretations for each property.

1 Introduction

Prokaryotes reproduce asexually by binary fission where each individual gives rise to two identical descendants. The existence and accumulation of mutation events allows for the introduction of variability, *i.e.*, creation of new individuals. These mechanisms create lineages that share a common ancestry, which have been typically represented and modeled using a tree-like structure.

The knowledge of how a set of bacterial isolates relate to each other, *i.e.*, their phylogenetic tree, is an important tool for their characterization. This characterization can then be used to infer the phenotypical properties of new isolates and study the formation of groups according to different characteristics, which is particularly useful for evolutionary and epidemiological studies.

With the diversification of sequencing techniques and their cost reduction, typing methods based on sequences have become the standard for the study of microbial populations and their epidemiological surveillance.

One of such methods is Multilocus Sequence Typing (MLST), which takes into consideration the nucleotide sequence variations at specific housekeeping genes, believed to accumulate little or no mutations since these *loci* are assumed to be under purifying selection. In most MLST schemes, the nucleotide sequences of approximately 450 bp fragments internal to (normally) seven housekeeping *loci* are sequenced, for each isolate. A new allele number is attributed to each new sequence found at each gene. Then, to each new combination of seven of these numbers, corresponding to the seven housekeeping *loci*, a new sequence type (ST) number is attributed.

Several methodologies have been developed for the analysis of the MLST data, by looking only at the difference between the allele numbers, aiming at the accurate reconstruction of the evolutionary history of the population through time. Dendograms, such as UPGMA [4] aim at obtaining a rooted tree, and algorithms such as eBURST [6] and goeBURST [7] aim at obtaining an unrooted tree. However, most of these methodologies make simplifying assumptions, not taking horizontal gene transfer and recombination into account. But, more often than not, existing data cannot be explained by mutation alone [9]. Accepting the existence of recombination events is hard, since a tree-like structure will prove insufficient to represent this exchange of alleles between isolates [16]. This is particularly true in bacterial species like *Neisseria spp.*, where recombination is known to play a dominant role [14]. This problem has been tackled by Bandelt and Dress [1] and developed in the tool Splitstree [13]. Splitstree computes unrooted phylogenetic networks from molecular sequence data, based on the split decomposition method. A special case of such networks is the reticulate network, which is capable of representing reticulate events such as hybridization, horizontal gene transfer, or recombination.

In this paper, we propose an alternative to the BURST family of algorithms to represent the relationships between different individuals in a graph-like structure, whilst making a minimal assumption. This graph, denominated *SLV graph* (Single Locus Variant graph), connects two individuals if and only if they share only one allelic difference in their MLST profile. Additionally, we show that the structure of the SLV graph is robust to the increasing acquisition of data through time and to possible sampling problems. Moreover, we show that it is possible to infer biological events through the structure of the SLV graph. Finally, we argue that the advantages of the SLV graph are not limited to MLST, and can prove useful for other typing methods such as Multiple-Locus Variable number tandem repeat Analysis (MLVA).

2 Methods

In this section, we describe the algorithms for the construction of the goeBURST unrooted tree and the construction of the SLV graph. Also, we enumerate the biological data sets considered for the validation of this study.

2.1 Construction of the goeBURST unrooted tree

The goeBURST algorithm [7], a globally optimized implementation of the eBURST algorithm [6], aims at the identification of the relationships between isolates, assuming essentially that bacterial populations are dominated by diversification of a few dominant clones. The identification of these relationships is performed by following a specific set of rules which divide the data into several clusters of related strains, designated clonal complexes. The existing relationships between the STs belonging to a given clonal complex, aim at representing the most parsimonious pattern of evolutionary descent between those STs. If two STs are directly linked in the tree this means that the genotype of those STs differ only by one housekeeping gene, called a single *locus* variant (SLV).

The goeBURST algorithm identifies the pattern of evolutionary descent within a given set of STs $\mathcal{S}$ by applying the following set of rules: a) it computes the number of SLVs existent among all of the STs in the set $\mathcal{S}$; b) it then chooses the links between STs with higher number of SLVs; c) in case of a tie in choosing a given SLV link, it computes the number of DLVs (Double Locus Variant) and chooses the link between STs with higher number of DLVs; d) in case of a tie at this level too, it proceeds for the TLV (Triple Locus Variant) disambiguation, and in case of persistence of the tie, it disambiguates using the ST frequency and then the ST identifier as tie-breaker. The eBURST [6] algorithm implements the same set of rules with an heuristic optimization, whereas goeBURST performs a global optimization taking into consideration all possible ties at all levels between STs in the set $\mathcal{S}$.

It is worth noting that running the algorithm multiple times for the same set of STs will result in the same set of unrooted trees. However, a drawback of the goeBURST algorithm (or any other tree-like algorithm) is the fact that the shape of the resulting tree may change upon the inclusion of new STs, forcing the displacement of a given set of STs from one part of a given eBURST group to another part of the same eBURST group, effectively suggesting a different evolutionary history for the affected group of STs.

2.2 Construction of the SLV graph

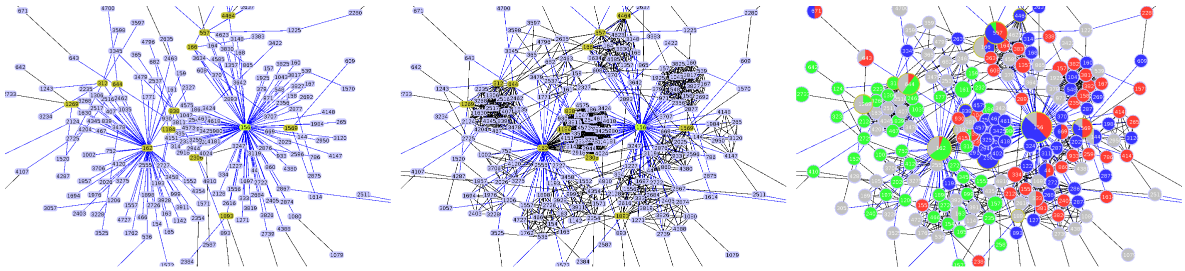
The implementation of the algorithm for the construction of the SLV graph follows Occam's principle of parsimony. It proceeds by following only the first rule of the goeBURST algorithm: for a set of STs in the set $\mathcal{S}$, it computes the allelic differences between all the STs in $\mathcal{S}$. It then creates a link between every pair of STs $u, v \in \mathcal{S}$ that share one allelic differences between themselves.

Topologically, by drawing every link between each pair of STs that are SLVs between themselves, the structure looses its tree-like shape and becomes a graph (see Figure 1 a) and b)). The rationale is that any assumption made with the attempt of recovering the true pattern of evolutionary descent, may in fact obscure it, because since we are dealing with events at a micro-evolutionary scale, this possibly can never be achieved with certainty. The SLV graph structure will therefore consist of an overlay of all possible trees. The impact and the advantages of representing the microbial typing data using such a structure is described in the following sections.

2.3 Biological data sets

We downloaded several complete bacterial allelic profiles available on MLST data sets from `http://pubmlst.org` and `http://www.mlst.net`[1]. Most of the available data sets have a relatively small number of isolates, and frequently even a smaller number of unique

[1] The biological data sets considered are from January 16, 2014.



a) goeBURST tree b) SLV graph c) Penicillin data overlay

Figure 1: Visual representations of the (second group/clonal complex of the) *S. pneumoniae* data set. Subfigure a) represents the goeBURST unrooted tree containing the founders ST 156 and ST 162. Subfigure b) represents the SLV graph containing all the SLV relations between all of the STs, including the ones represented by goeBURST. Subfigure c) represents the overlay information concerning the resistance to penicillin for each ST by color, where green means susceptible and blue means resistant. The shared subclique suggests the group of STs where the resistance to penicillin was first acquired.

ST profiles. Due to this scarcity of data, we only considered some of the data sets with more than 500 unique STs in the complete SLV graph, in order to adequately compare the types of structures observed in the bacterial MLST data sets with the types of structures observed in the simulated MLST data sets.

In total, we considered 9 bacterial species for this study (see Table 3): *Campylobacter jejuni*, *Pseudomonas aeruginosa*, *Neisseria spp.* and *Streptococcus agalactiae* from `http://pubmlst.org`; and *Burkholderia pseudomallei*, *Enterococcus faecium*, *Haemophilus influenzae*, *Staphylococcus aureus* and *Streptococcus pneumoniae* from `http://www.mlst.net/`.

3 Results

The problem of recovering the true phylogeny from a set of STs is a challenging one. Algorithms from the BURST family aim at recovering the true phylogeny with the hypothesis that a given set of STs can be represented using a tree-like structure. This objective is immediately hampered by several constraints, like the considered model, the selected data structure, or even potential sampling problems. The latter may be due to geographical factors or even due to temporal discontinuities during data acquisition. Moreover, the problem of recovering the true phylogeny in bacterial populations includes an additional obstacle: the existence and influence of recombination events, keeping distinct lineages from completely diverging through allelic exchange. This is more evident in highly recombinogenic species like *Neisseria* where a “bifurcating tree-like phylogeny is not an appropriate model” [12].

3.1 Basic topological structure of the SLV graph

The topological structure of the SLV graph is characterized by several types of basic motifs (see Table 1). We proceed by describing the existing types and the conditions for their occurrence.

The trivial case is to have a *singleton*, corresponding to a single ST which is not a SLV of any other ST in set $\mathcal{S}$ (Type 1 of Table 1). Then two STs can be SLVs between themselves without being a SLV of any other ST in set $\mathcal{S}$ (Type 2 of Table 1), forming a *doubleton*. Three STs can form a *clique*[2], if they are all are SLVs between themselves fully (Type 3a of Table 1), and they are not a SLV of any other ST in set $\mathcal{S}$; or a *linear chain*, if they are SLVs between themselves pairwise (Type 3b of Table 1). It is clear from this definition that a doubleton is a particular case of a linear chain.

Generally speaking, whenever a subset of STs $\mathcal{L} \subseteq \mathcal{S}$ is considered, containing more than two STs, they will be completely linked, forming a *clique*, if the differences between them are all in the same *locus* (*i.e.*, they form a set of complete SLVs). Whenever this condition is not verified, several substructures are possible, due to differences among the STs at distinct *loci* (Type 4b-e of Table 1). Whenever each ST $s \in \mathcal{L}$ is a SLV to at most two other STs $t, u \in \mathcal{L}$ and they do not share a SLV with any of the other STs in set $\mathcal{S}$, then the subset of STs $\mathcal{L}$ is represented by a *linear chain* (Type 4e of Table 1).

3.2 Basic topological structure of the goeBURST unrooted tree

The topological structure of the goeBURST unrooted tree is actually the represention of one particular tree contained in the SLV graph topological structure. However, one can also find and characterized several types of basic motifs (see Table 1) in the structure of a goeBURST unrooted tree. The two trivial cases in the goeBURST unrooted tree, are the *singleton* or a *doubleton*, which occur under the same conditions as those described for the SLV graph (Type 1 and 2 of Table 1). However, unlike the SLV graph, when considering three STs which are SLVs between themselves, the goeBURST always forms a *linear chain*, without making the distinction whether the STs are SLVs in the same *locus* or at two different *loci* (Type 3 of Table 1).

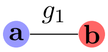
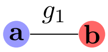
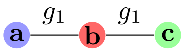
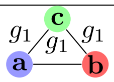
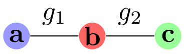
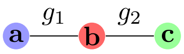
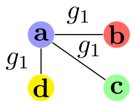
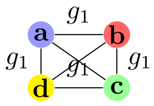
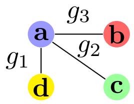
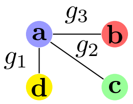
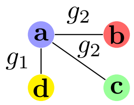
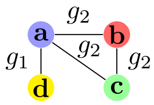
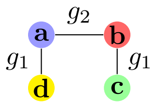
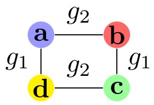
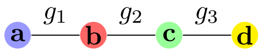
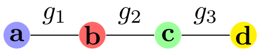
Generally speaking, when considering any subset of STs $\mathcal{L} \subseteq \mathcal{S}$, containing at least four distinct STs, if each ST in $\mathcal{L}$ is a SLV of at most two other STs and it is not a SLV of any other ST in set $\mathcal{S}$, then the subset of STs $\mathcal{L}$ is represented by a *linear chain* (Type 4d-e of Table 1). Alternatively, if every ST in $\mathcal{L}$ is a SLV of at least another ST in $\mathcal{L}$ and is not a SLV of any other ST in set $\mathcal{S}$, then the subset of STs $\mathcal{L}$ is represented by a *star-like tree* (see Type 4a-c of Table 1).

3.3 High-level topological structures of the SLV graph

Due to the nature of the biological data, one can observe the emergence of higher-level topological structures in the SLV graph, which would otherwise be invisible when considering solely the goeBURST unrooted tree. These structures emerge due to creation of all possible links between a set of STs whenever they share allelic differences at a single *locus* (see Type 4-a of Table 1). The founder, is inferred to have increased in the population, and then progressively diversified under the effects of mutation and recombination, forming a cluster of phylogenetically closely related isolates. In MLST, this diversification is

[2] In graph theory, a *clique* is an undirected graph $G = (V, E)$, such that for every two vertices $s, t \in V$, it exists an edge $e \in E$ connecting the two. Here, a *clique* is only observed whenever a set of STs are SLVs between themselves in the same gene.

Table 1: Comparison between basic topological structures found in a goeBURST group with their equivalent representation in the SLV graph. Each color node represents a distinct ST (from a to d) and g_n represents an allelic difference at gene n. Combinations of up to four STs are represented in the table, considering the possible variations between the number of different genes in which two STs may differ.

Type	**goeBURST tree**	**SLV graph**
1		
2		
3-a		
3-b		
4-a		
4-b		
4-c		
4-d		
4-e		
...	...	...

usually seen as changes in the allelic sequence at any of (the typically seven) distinct *loci*.

Let us consider a set of STs sharing allelic changes at a single *locus*. One can easily see that the SLV graph will represent a link between all the STs forming a clique (see Type 1 of Table 2), which is represented by goeBURST as a simple star-like structure. Additionally, if we consider a larger set of STs, which includes the previous one, sharing an allelic change at all the possible MLST *loci*, the goeBURST structure will still represent thhis set as a star-like structure, since the founder ST f is a SLV of all the other STs. On the other hand, the SLV graph takes into account all the possible relationships between all the STs. This resulting graph, a *SLV star-like subclique* structure, not only contains all the links between the founder ST f and the rest of the STs in the set, but also explicitly represents the relationships between all the STs that are STs between themselves, forming a set of subcliques. It should not come as a surprise to see that this *SLV star-like subclique* structure yields the same number of subcliques as the number of *loci* considered in the MLST data set, usually seven in the currently used schemes as indicated above, since each clique contains the subset of STs that are SLVs at a particular *locus*.

One can now imagine a given ST d, which is not only a DLV of the founder ST f but it is also SLV of two other STs, t and u, each belonging to two distinct subcliques (see Type 3 of Table 2). Whenever this situation occurs, the goeBURST algorithm will always decide on a single ST to be linked with ST d, either ST t or ST u. On the other hand, the SLV graph allows for the creation of both links $d - t$ and $d - u$ without any constraints. We denominate this substructure, composed of four STs, a *SLV square*, for reasons that will become apparent in the next subsection.

Finally, it is also possible to imagine two distinct *SLV star-like subcliques* structures sharing a subclique. This shared subclique is composed of a set of STs that are all SLVs between themselves, but are also SLVs of both founders of the two distinct *SLV star-like subcliques*. The goeBURST representation in this case, is to represent all the STs as two connected star-like trees around each of the two founders, where the subset of STs belonging to a shared subclique would be distributed between the two star-like trees acording to the goeBURST rules. Interestingly, the SLV graph representation avoids making this choice of distribution, connecting all the STs belonging to this shared subclique between themselves, as well as to both of the founder STs. The resulting high-level topological structure is shown in Type 4 of Table 2.

3.3.1 Inferring biological events from the topological properties of the SLV graph

Giving a set of STs $\mathcal{S}$, the corresponding SLV graph does not offer a single descendence path, since it represents the superposition of all possible trees between those STs, reflecting the uncertainty in the identification of the correct phylogeny in micro-evolutionary studies. The basic topological properties described in Table 1 can give rise to four high-level structures in a SLV graph. In the following subsections we describe each of these high-level structures and correlate them with possible biological events.

SLV clique: The first high-level structure that can be observed in the SLV graph is the *SLV clique* (see Type 1 of Table 2). It is characterized by a set of STs $\mathcal{S}$ that are all SLVs between themselves, sharing all their differences in the same *locus*, forming a clique. A direct consequence of this observation is that every triangle observed in the SLV graph, represents a set of three STs that are SLVs in the same *locus*.

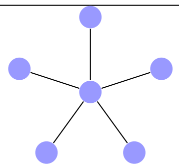
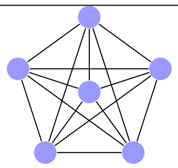
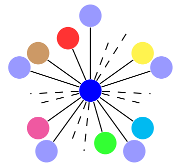
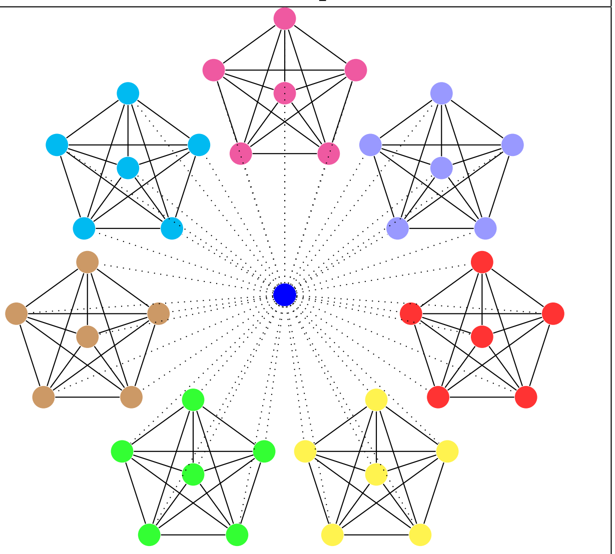
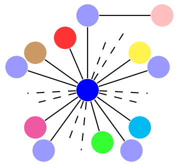
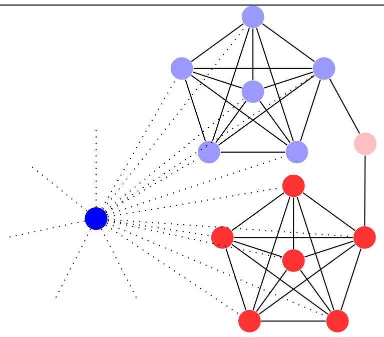
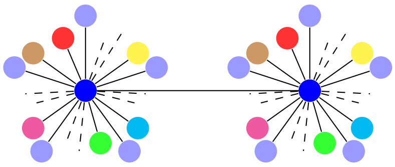
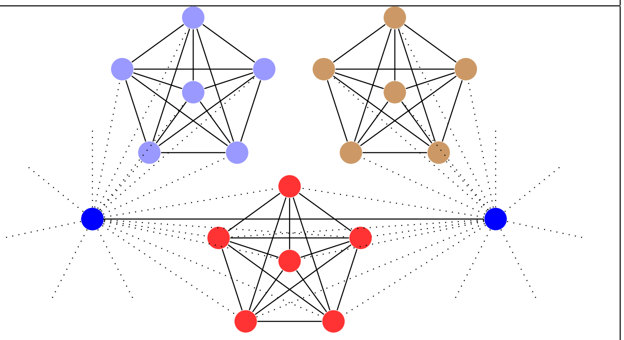
In contrast, the goeBURST algorithm follows a set of rules will select a single ST to become the center of a star-like tree, representing the perspective of the founder clone, forcing all the edges to connect to it. This will therefore suggest a single phylogeny ignoring the existence of multiple equally-probable descendence paths when relating individuals that are SLVs between themselves.

Biologically, whenever a group of closely related individuals forms a *SLV clique*, it suggests that we cannot identify the true path of evolutionary descent between them without making simplifying assuptions.

SLV star-like subcliques: The second high-level structure observed in the SLV graph, and undoubtedly the most important one, is the *SLV star-like subcliques*. It is characterized by a set of STs $\mathcal{S}$, divided into several subcliques connected to a common central ST, where each individual subclique corresponds to the previously described *SLV clique*.

An obvious property of the structure, is that the number of subcliques connected to single central ST, will be maximally bounded by the number of housekeeping genes

Table 2: Comparison between the high level topological structural representations of the hypothetical phylogenetic relations among a set of STs, obtained by the goeBURST unrooted tree and the SLV graph.

Type	**goeBURST**	**SLV graph**
1	 Star-like tree	 SLV clique
2	 Dense star-like tree	 SLV star-like subcliques
3	 DLV of founder	 SLV square
4	 Connected star-like trees	 Shared subclique

considered by the typing method (see Type 2 of Table 2 and Figure 1), since each subclique contains STs sharing differences in the same *locus*. Another interesting property is that the central ST will simultaneously belong to all the subcliques, since it shares a SLV with every other ST in each connected subclique.

In contrast, the goeBURST algorithm would be unable to distinguish this case from the previous *SLV clique*, where the algorithm also outputs a star-like tree (see Table 2 and Figure 1). Nevertheless, these two cases represent radically different events, with only the latter representing a variation at all *loci* compatible with the clonal expansion originally proposed by Maynard Smith *et al.* [17]. In this *SLV star-like subcliques* structure, the central ST is naturally identified as the founder of this group of related STs.

The *SLV star-like subcliques* can be clearly seen when computing the SLV graph for the *S. pneumoniae* data set (see Figure 1), where two distinct structures can be identified sharing a *SLV clique*. Here, we clearly see that one of the *SLV star-like subcliques* is penicillin-resistant while the other is penicillin-susceptible, with the some profiles of the shared *SLV clique* being penicillin-resistant and others penicillin-susceptible.

SLV square: The third high-level structure observed in the SLV graph, and one of the most interesting ones, is the *SLV square*. It is characterized by a ST which is a DLV of the central ST of a *SLV star-like subcliques*, through two other STs, each belonging to a distinct subclique, which are themselves SLVs of the central ST.

The identification of this type of structure, illustrated by Type 3 in Table 2, permits the identification of biological events in which some form of homoplasy must be invoked, as illustrated in Figure 4. Considering a genotype composed of two genes $g_i - g_j$, one can obtain a SLV graph square: either through a back-mutation at gene g_i or at gene g_j (Figure 4 b); or through a recombination event where an allele is inserted at position g_j of the descendants of a and b, originating STs d and c (Figure 4 c).

In order to study the number of SLV squares on different species, we have considered the SLV graph for all the MLST data sets represented in Table 3, and computed the corresponding number of SLV squares. The data indicates that different species present a wide range of SLV squares.

However, when considering the number of unique STs present in each SLV graph (and that each SLV square is necessarily composed of four distinct STs), one must conclude that the number of SLV squares higher, for most species, than it would be expected considering non-overlapping STs. This means that, for many of these species, the SLV squares must somehow overlap, sharing some of their STs. One can easily imagine possible ways for the creation of overlapping SLV squares[3], like those illustrated in Figure 2.

Despite the existence of different sequence of events capable of generating SLV squares, these occur with different probabilities. In order to discriminate between back-mutation and recombination events, we looked at the pairwise nucleotide differences between the allelic sequences of the four STs of an SLV square. Since these four STs share two *locus* variants between themselves, only two pairwise comparisons are necessary between the nucleotide sequences of the differing alleles. For each of the SLV squares of each of the species described in Table 3, we performed the two pairwise comparisons. Considering the

[3] In graph theory, the number of k-cubes contained in a minimal complete n-cube is given by the formula $C_k^n.2^{n-k}$. Considering seven housekeeping genes in MLST, the minimal 7-dimensional cube, has $C_0^7.2^{7-0} = 128$ vertices (0-dim cube), containing $C_2^7.2^{7-2} = 672$ squares (2-dim cube).

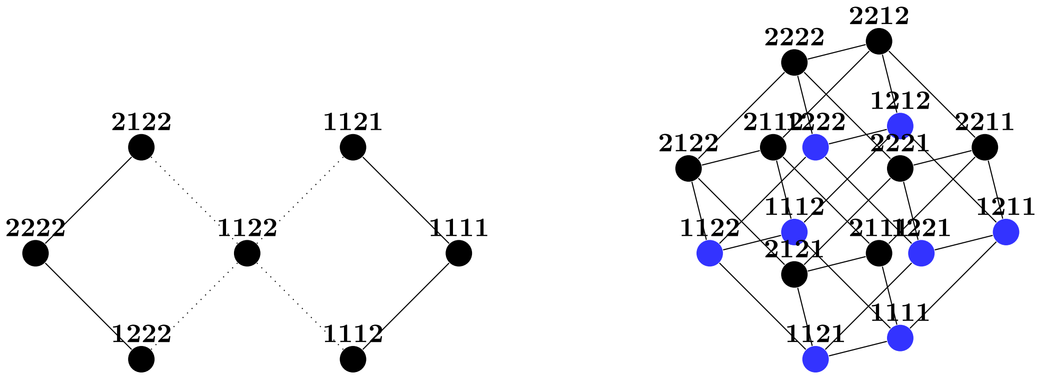


Figure 2: Structural configurations of a SLV squares. Subfigure A illustrates the formation of two SLV squares by a single recombination event of ST 1122. Subfigure B illustrates an hypothetical case having adjacent SLV squares forming a complete hypercube of dimension N_g, where N_g is the number of housekeeping genes used by the typing method.

number of nucleotides differing at each comparison, we denoted it maximal and minimal. Figure 3 shows the data for the set of SLV squares of each dataset, sorted first by the maximal (represented in red) and then by the minimal (represented in blue) nucleotide differences.

Table 3: Data sets statistics (downloaded on 2014-01-16). We present the number of isolates and the number of unique profiles in each species, as well as the number of STs, links, and SLV squares in the corresponding SLV graph (singletons are not considered). Also, the compactness and clustering coefficients are computed in the whole SLV graph and restricted to the biggest clonal complex of the corresponding species.

Data sets		**SLV Graph**							
Species	**Profiles**	**STs**	**STs/**	**Links**	**Homoplasy**	**Compactness**		**Clustering**	
			Profiles		**squares**	**Total**	**BigCC**	**Total**	**BigCC**
C. jejuni	6.972	5.629	81%	17.892	2.854	0,127	0,004	0,495	0,599
P. aeruginosa	1.610	977	61%	1.009	5	0,566	0,042	0,282	0,303
Neisseria spp.	10.642	8.511	80%	40.468	3.594	0,135	0,007	0,573	0,627
S. agalactiae	676	639	95%	2.848	98	0,064	0,019	0,660	0,680
B. pseudomallei	1.096	756	69%	1.526	551	0,172	0,008	0,233	0,286
E. faecium	886	723	82%	1.984	530	0,106	0,011	0,425	0,460
H. influenzae	1.301	957	74%	1.613	13	0,397	0,056	0,419	0,652
S. aureus	2.602	2.216	85%	12.837	144	0,120	0,014	0,730	0,791
S. pneumoniae	9.346	7.451	80%	22.185	1.063	0,206	0,007	0,524	0,647

By sectioning the distribution, one can clearly observe three distinct regions. In the first region, there is a high maximal nucleotide difference, suggesting that such difference is more likely to account for recombinations events. On the other hand, in the tail of the distribution, there is a low maximal nucleotide difference (indistinguishable from the minimal nucleotide difference), suggesting that back-mutation could account for that specific region. Finally, the middle region of the distribution, corresponds to SLV squares where we cannot distinguish between back-mutation and recombination events. In general, it is widely accepted that in regions with more than five nucleotide differences, recombination is more likely to be the main event [5]. But, one cannot clearly distinguish between back-mutation or recombination events whenever two to four nucleotide differences are observed in a given sequence. The nucleotide differences between the divergent allelic sequences of SLV squares for some of the species described in Table 3 is illustrated in Figure 3. By comparing the shape of each of the distributions it is possible to observe variations in the rate of mutation and recombination between the different species.

Additionally, we have used the an Infinite Allele Model (IAM) simulator to generate synthetic datasets with varying mutation and recombination rates ranging between 0-100. The result is illustrated in Figure 5 where we can observe that for low values of mutation (interval between 0-10), the recombination (interval between 0-30) is very efficient in the creation of new SLV squares, suggesting that for low values of mutation, the majority of the edges in the SLV graph are created by recombination. On the other hand, as the rate of mutation increases, the probability of a recombination event before a subsequent mutation occurs becomes lower. This means that for high values of mutation, the generation of new individuals becomes so fast that the increase of recombination is unable to generate a significant number of SLV squares.

Assuming the previously described IAM model with varying recombination and mutation, by comparing the number of SLV squares obtained through the synthetic data sets (Figure 5) with the ones obtained from the MLST data sets (Table 3), we can infer that most of the considered species of the MLST data sets do not fall into this area of high recombination, except for *Neisseria spp.* and *C. jejuni.* However, we can also observe that some of the species, like *B. pseudomallei* and *E. faecium*, contain a significantly high number of SLV squares relative to the number of STs in the corresponding SLV graph.

The shared subclique: The shared subclique, occurs whenever a subclique of STs are SLVs of two central STs of *SLV star-like subcliques*, that are SLVs between themselves. The goeBURST algorithm would simply output two adjacent star-like trees without adding much information. In contrast, the SLV graph representation of all the SLV connections, will show that a given central ST f_1, is a SLV of several subcliques, and one of these subcliques is also SLV of another central ST f_2 (see Type 4 in Table 2).

We have used PHYLOViZ [8] to load the allelic profile of the *S. pneumoniae* data set from `http://www.mlst.net` on 2014-01-16. We then added the isolate information from 2014-01-16, considering only the resistance to penicillin. Figure 1 illustrates ST 156 (represented in purple), susceptible to penicillin, and ST 162 (represented in green), resistant to penicillin. Each of these central STs are SLVs of seven surrounding subcliques, one of which is shared by both central STs. It is therefore reasonable to assume that at least one ST belonging to the shared subclique acquired resistance to penicillin and generated some descendency. However, at the micro-evolutionary time scale, one cannot accurately infer the true pattern of evolutionary descent. Nonetheless, this situation cannot occur with the SLV graph, since it performs the minimal assumption, by connecting all of the STs belonging to the shared subclique. Then, one of these descendant STs, either by a fitness advantage or random drift, increased in frequency in the population and diversified, generating a new graph *SLV star-like subclique* structure resistance to penicillin. The corresponding figure using the goeBURST representation can be seen in Figure 6, where closely related STs, represented in the shared subclique of the SLV graph, are

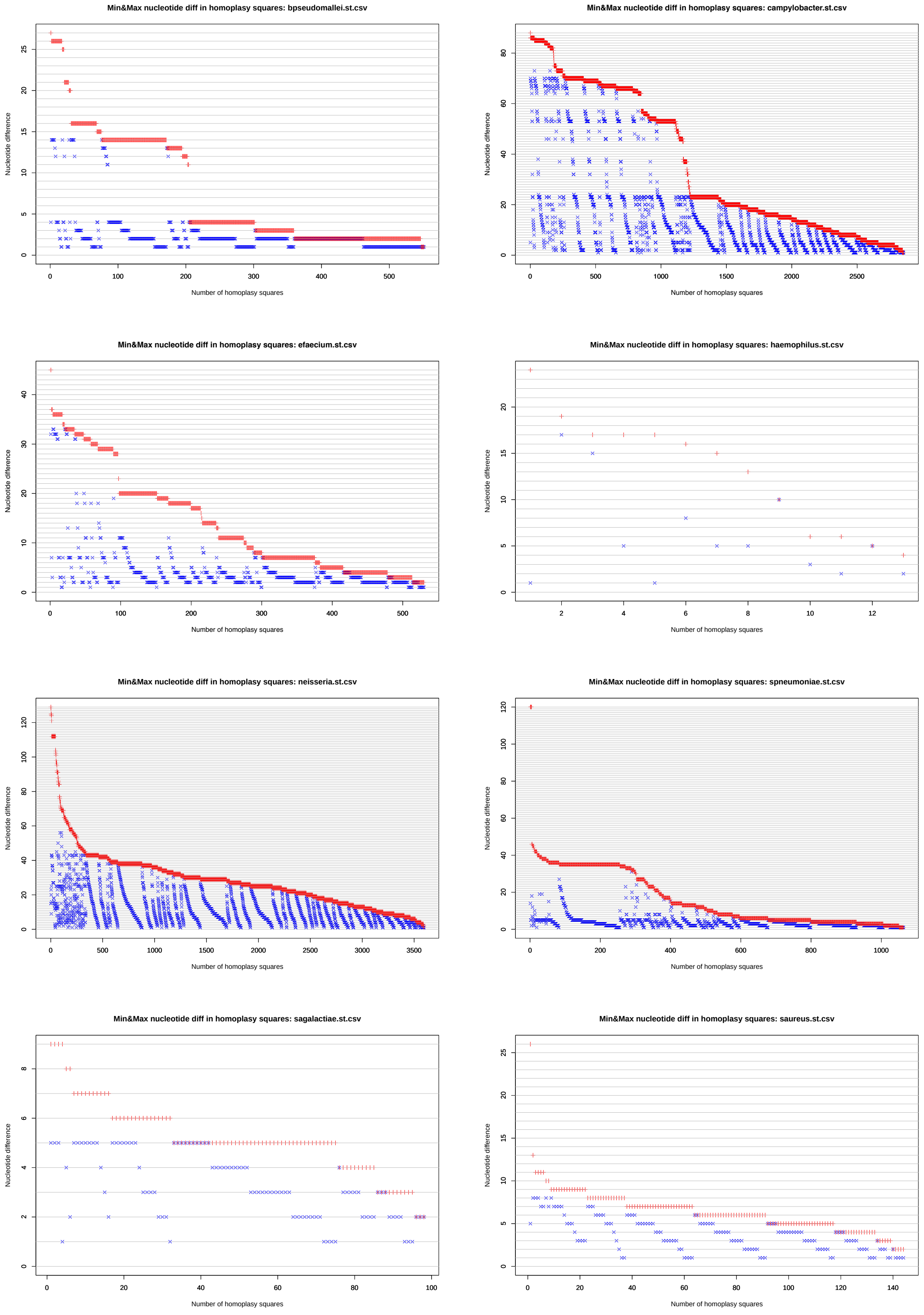


Figure 3: Max/min nucleotide differences between the pairwise allelic sequences. Obtained from two divergent *locus* of each homoplasy square (see Table 3).

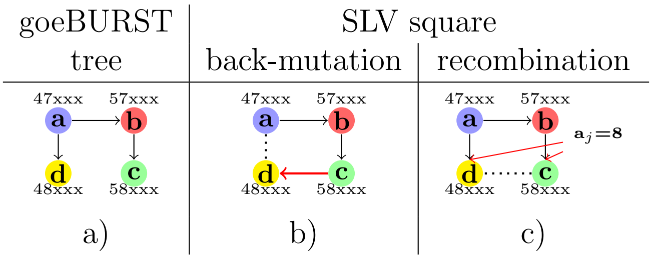


Figure 4: SLV square representation on a SLV graph composed of two genes $g_i - g_j$. Subfigure a) illustrates the SLV connections between four STs, where ST d is a SLV of both a and c, but only one of these two connections is drawn. Subfigure b) illustrates the creation of an SLV square through back-mutation of *locus* i (allele 5 → allele 4). Subfigure c) illustrates the creation of an SLV square through recombination, where allele 8 is inserted at *locus* j of STs c and d, which are descendants of STs a and b, respectively.

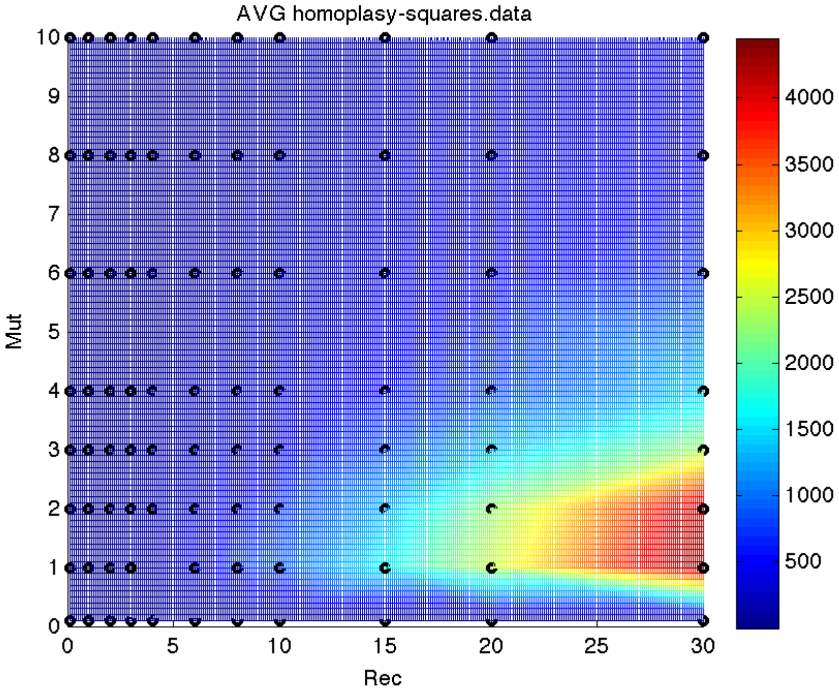


Figure 5: Heat map representation of the number of homoplasy squares contained in last 50 generations of the synthetic datasets. Each point is an average of 50 random simulations with the same set of parameters: population of 1.000 individuals, simulated throughout 10.000 generations and with a profile of 7 *loci*.

now represented apart from each other, connected with a single SLV connection to the corresponding founder.

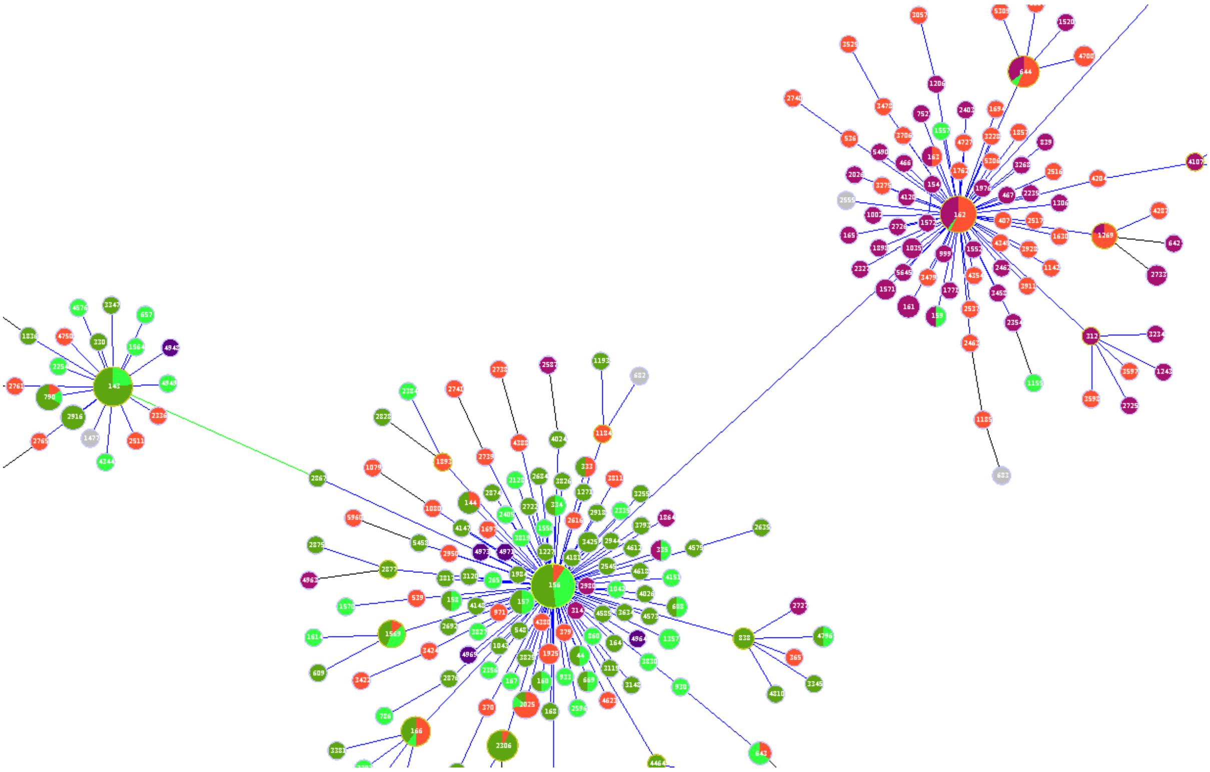


Figure 6: Visualization of the *S. pneumoniae* data set using the goeBURST algorithm illustrating the resistance to penicillin for each ST by color, where green means susceptible and purple means resistant. The set of closely related STs where the resistance of first acquired cannot be identified, contrary to the SLV graph representation in Figure 1.

3.4 Implementation and visualization of the SLV graph

The construction and spatial arrangement of a tree-like structure is not difficult, since one can always think on the naïve approach of propagating the successors of each node of the tree committing to a single direction. This is however not the case when constructing and spatially arranging graph-like structures, since we can always make connections to previously existing nodes, loosing the planar representation. In general, the problem of finding a planar representation of a graph minimizing the number of crossing edges is NP-COMPLETE [11], and it is usually tackled through the use of heuristics or greedy algorithms. Additionally, problems like the minimization of the area used by the graph also need to be taken into account. An alternative approach for the spatial arrangement of graphs is the use of force-directed physical models [3, 15]. The force-directed approach proceeds by computing the forces acting on each vertex, and then applying an optimization algorithm until the net force acting on each vertex reaches zero.

The use of force-directed approaches for the spatial arrangement of graphs presents several advantages in what concerns visualization capabilities and, given their complexity, running at $O(max(m, n \log n))$ per iteration, where n is the number of vertices and m is the number of edges in the graph, they are suitable for the visualization of reasonable large graphs [2, 10]. Since the SLV is a new type of structure, not previously visualized, it would be hard to find the best topological arrangement following metric approaches. The use of a force-directed approach permits the immediate visualization of the numerous subcliques surrounding a given founder ST, corresponding to the different genotypes of the (usually) seven housekeeping genes analyzed through MLST.

The visualization of the SLV graph was implemented in the context of PHYLOViZ. PHYLOViZ [8] is a platform for the integrated analysis of sequence-based typing methods and associated epidemiological data. Additionally, it allows for the visual representation of the possible evolutionary relationships between STs provided by the goeBURST algorithm. The visualization and spatial organization of the data is performed using the Prefuse toolkit for information visualization[4], which implements the force-directed graph layout. Due to PHYLOViZ modularity, we have easily implemented a plug-in to compute the SLV graph edges from a given set of STs, where the housekeeping *loci* from the MLST data sets are immediately identified as distinct subcliques surrounding a founding genotype (see Figure 1).

4 Discussion

Traditionally, the depiction of the relationships between haplotypes in a population is through the use of trees. Frequently, the application of maximum parsimony principles to haplotype data results in representations with cycles through graphs. In order to simplify this representation, assumptions are made allowing the break of these cycles and the representation of these relationships as a tree. This apparent simplifying assumption, imposes a time-ordering restriction on the data, severely hindering the number of its possible interpretations. The use of the SLV graph structure, relaxes this time ordering restriction by letting STs be connected instead by a graph-like structure. This aspect is particularly important when mapping additional genotypic or phenotypic information on this structure, since the existence of homoplasy is frequently inferred ignoring that the assumptions of the underlying model that generated the tree may themselves be violated.

The minimum spanning tree computed by the goeBURST algorithm may not represent the shortest path between a given set of STs. The SLV graph, by representing all the SLV relationships between STs, ensures that not only it contains the same goeBURST minimum spanning tree, but that it also contains the shortest (and the longest) evolutionary distance between any two STs. In general, the minimum spanning tree given by the goeBURST algorithm is always greater or equal than the minimum spanning tree contained in the corresponding SLV graph.

Additionally, when recovering the true phylogeny, sampling can be one of the biggest problems. Despite the fact that the amount of data is becoming increasingly available as time progresses, it is impossible to obtain new data from the past. Additionally, the amount of data acquisition is limited to the availability of resources. The representation of MLST data, using a data structure dependent on data availability, can severely obstruct our capacity to correctly interpret it. The SLV graph structure, can represent new acquired relationships without changing the already existing ones, being better suited for the representation of continuously increasing MLST data, than tree-like algorithms. In particular, as the sampling augments, the structure tends to become saturated

[4] Prefuse is available at `http://prefuse.org`.

without changing its structure. Moreover, associated to this topological property, is the non-commitment to a particular phylogenetic ordering of events, focusing instead on representing the relationships between all the individuals satisfying the minimal level of vicinity, the SLV.

It is also interesting to observe that the SLV graph structure presents topological properties that can be explored in order to suggest reasonable interpretations for biological events. This data structure endows the user with a simple visual method for the identification of subsets of STs that are SLVs in the same *locus*, and for the visual identification of the founder of a set of related STs. Additionally, we have identified a structural motif between four STs, presented a relation between the existence of topological squares in the SLV graph and the number of potential homoplasies in a given data set. We have also performed a sequence based analysis, for an important collection of data sets available in public repositories, in order to better distinguish between the likelihood of a back-mutation or a recombination event, on the formation of a particular SLV square. Lastly, it is now possible to visually identify the subset of STs that acquired a given characteristic, like the resistance to penicillin, carrying it as they diverge and form other subcliques.

Acknowledgments

The authors thank Francisco R. Pinto and Claudio Donati for helpful discussions. This work was supported by the European Commission under the FP-Health-2007-B/223111 (project CAREPNEUMO) and by Fundação para a Ciência e a Tecnologia with the grants Pest-OE/EEI/LA0021 and PTDC/EIA-CCO/118533/2010 (project NetDyn).

References

[1] Hans-Jürgen Bandelt and Andreas W. M. Dress. A canonical decomposition theory for metrics on a finite set. *Advances in Mathematics*, 92(1):47–105, 1992.

[2] Josh Barnes and Piet Hut. A hierarchical O(N log N) force-calculation algorithm. *Nature*, 1986.

[3] Peter Eades. A heuristic for graph drawing. *Congressus Numerantium*, 42:149–60, 1984.

[4] Mark C. Enright and Brian G. Spratt. A multilocus sequence typing scheme for streptococcus pneumoniae : identification of clones associated with serious invasive disease. *Microbiology*, 144:3049–60, 1998.

[5] Edward J. Feil. Small change: keeping pace with microevolution. *Nature Reviews Microbiology*, 2:483–495, 2004.

[6] Edward J. Feil, Bao C. Li, David M. Aanensen, William P. Hanage, and Brian G. Spratt. eBURST: Inferring patterns of evolutionary descent among clusters of related bacterial genotypes from multilocus sequence typing data. *Journal of Bacteriology*, 186(5):1518–1530, 2004.

[7] Alexandre P. Francisco, Miguel Bugalho, Mário Ramirez, and Jo ao Carriço. Global optimal eBURST analusis of multilocus typing data using a graphic matroid approach. *BMC Bioinformatics*, 10:152, 2009.

[8] Alexandre P. Francisco, Cátia Vaz, Pedro T. Monteiro, José Melo-Cristino, Mário Ramirez, and Jo ao A. Carriço. PHYLOViZ: Phylogenetic inference and data visualization for sequence based typing methods. *BMC Bioinformatics*, 13:87, 2012.

[9] C. Fraser, William P. Hanage, and Brian G. Spratt. Recombination and the Nature of Bacterial Speciation. *Science*, 315(5811):476–480, January 2007.

[10] Thomas M.J. Fruchterman and Edward M. Reingold. Graph drawing by force-directed placement. *Software: Practice and experience*, 21(11):1129–1164, 1991.

[11] Michael R. Garey and David S. Johnson. Crossing number is NP-complete. *SIAM Journal of Algebraic and Discrete Methods*, 4(3):312–6, 1983.

[12] Edward C. Holmes, Rachel Urwin, and Martin C. J. Maiden. The influence of recombination on the population structure and evolution of the human pathogen *Neisseria meningitidis. Molecular Biology and Evolution*, 16(6):741–9, 1999.

[13] Daniel H. Huson and David Bryant. Application of phylogenetic networks in evolutionary studies. *Molecular Biology and Evolution*, 23(2):254–67, 2006.

[14] K.A. Jolley, D.J. Wilson, P. Kriz, G. Mcvean, and M.C.J. Maiden. The influence of mutation, recombination, population history, and selection on patterns of genetic diversity in *neisseria meningitidis. Molecular Biology Evolution*, 22(3):562–9, 2005.

[15] Tomihisa Kamada and Satoru Kawai. An algorithm for drawing general undirected graphs. *Information Processing Letters*, 31:7–15, 1989.

[16] David Posada and Keith A. Crandall. Intraspecific gene genealogies: trees grafting into networks. *Trends in Ecology & Evolution*, 16(1):37–45, 2001.

[17] J.M. Smith, E.J. Feil, and N.H. Smith. Population structure and evolutionary dynamics of pathogenic bacteria. *BioEssays*, 22(12):1115–1122, December 2000.