Cheng et al. BMC Bioinformatics 2011, 12:302
http://www.biomedcentral.com/1471-2105/12/302

BMC Bioinformatics

RESEARCH ARTICLE Open Access

# Bayesian semi-supervised classification of bacterial samples using MLST databases

Lu Cheng[1], Thomas R Connor[3], David M Aanensen[2], Brian G Spratt[2] and Jukka Corander[1*]

## Abstract

**Background:** Worldwide effort on sampling and characterization of molecular variation within a large number of human and animal pathogens has lead to the emergence of multi-locus sequence typing (MLST) databases as an important tool for studying the epidemiology and evolution of pathogens. Many of these databases are currently harboring several thousands of multi-locus DNA sequence types (STs) enriched with metadata over traits such as serotype, antibiotic resistance, host organism etc of the isolates. Curators of the databases have thus the possibility of dividing the pathogen populations into subsets representing different evolutionary lineages, geographically associated groups, or other subpopulations, which are defined in terms of molecular similarities and dissimilarities residing within a database. When combined with the existing metadata, such subsets may provide invaluable information for assessing the position of a new set of isolates in relation to the whole pathogen population.

**Results:** To enable users of MLST schemes to query the databases with sets of new bacterial isolates and to automatically analyze their relation to existing curated sequences, we introduce here a Bayesian model-based method for semi-supervised classification of MLST data. Our method can use an MLST database as a training set and assign simultaneously any set of query sequences into the earlier discovered lineages/populations, while also allowing some or all of these sequences to form previously undiscovered genetically distinct groups. This tool provides probabilistic quantification of the classification uncertainty and is highly efficient computationally, thus enabling rapid analyses of large databases and sets of query sequences. The latter feature is a necessary prerequisite for an automated access through the MLST web interface. We demonstrate the versatility of our approach by anayzing both real and synthesized data from MLST databases. The introduced method for semi-supervised classification of sets of query STs is freely available for Windows, Mac OS X and Linux operative systems in BAPS 5.4 software which is downloadable at http://web.abo.fi/fak/mnf/mate/jc/software/baps.html. The query functionality is also directly available for the *Staphylococcus aureus* database at http://www.mlst.net and shortly will be available for other species databases hosted at this web portal.

**Conclusions:** We have introduced a model-based tool for automated semi-supervised classification of new pathogen samples that can be integrated into the web interface of the MLST databases. In particular, when combined with the existing metadata, the semi-supervised labeling may provide invaluable information for assessing the position of a new set of query strains in relation to the particular pathogen population represented by the curated database. Such information will be useful both for clinical and basic research purposes.

## Background

The widespread availability of DNA sequencing technology over recent years has lead to the widely adopted practice of routinely characterizing bacterial samples in terms of molecular variation over a set of core genes that have been established by the international research community for the organism in question [1,2]. Given success of the technologies behind these community-based efforts, there are now Multi-Locus Sequence Typing (MLST) databases available for many bacterial species, most hosted at http://www.mlst.net and http://www.pubmlst.org. These provide access to a vast amount of information about many important pathogens. More recently, geographical tools such as Google Maps, have been integrated into the databases for quick access and visualization of spatial data related to strain

* Correspondence: jukka.corander@helsinki.fi
[1]Department of Mathematics and statistics, P.O.Box 68, University of Helsinki, 00014, Finland
Full list of author information is available at the end of the article

distribution. For examples of these advances, see http://www.spatialepidemiology.net/ and http://maps.mlst.net/. Another example of the evolution of these tools is the portal http://www.emlsa.net/, which provides access to electronic taxonomy of bacteria, through a common format and software for assigning strains to species via the Internet. Nevertheless, there is still substantial potential for global advances in pathogen epidemiology as the community using these tools keeps increasing and new functionality will be added on a continuous basis.

Thus far, the MLST information content displayed through the web access to either spatial or non-spatial data is based on relatively light procedures when considered from a statistical and/or computational perspective. This is reasonable, since the majority of more advanced model-based statistical methods for analyzing such data would not be scalable to provide real-time online access to results for users. However, provided that a statistical method for analyzing the MLST data meets the requirement of reasonable scalability, it may become a highly useful epidemiological tool and gain popularity very rapidly within this research community. The eBURST program available at http://eburst.mlst.net/ is an example of such a success story [3], making evolutionary snapshots of relatedness among sampled strains of pathogens.

Currently, MLST databases can be queried in various ways, including comparison of DNA sequences for a new set of samples with those previously existing in the database. However, when samples contain strains not currently present in the curated database, a user does not have an automated access to information which enables assessment of the relation of these samples to the earlier detected evolutionary groups. Such information is useful for various epidemiological and clinical purposes, in particular when considering the virulence and resistance characteristics of the strains. To enhance the querying features of the databases, we introduce here a statistical method for providing rapid access to probabilistic assignment of new strains to either predetected or earlier unseen evolutionary groups. The method is based on extending a Bayesian unsupervised classification method for MLST data [4] to a semi-supervised setting, where the existing curated MLST database plays the role of training data. To be able to handle the computational challenge of doing inference for the semi-supervised classification model, we adopt the computational strategies based on a stochastic optimization algorithm for unsupervised classification which are implemented in the BAPS software [5-7]. In contrast to more conventional Markov chain Monte Carlo (MCMC) methods for Bayesian inference, our algorithm is able to handle the computational issues more efficiently, such that the method can be applied to online use for MLST database queries. Alternative computationally fast approaches could also be developed by considering some of the recent advances in methodology for the analysis of genetic population structure [8,9], based on principal component and discriminant analysis.

## Methods

### Bayesian semi-supervised classification model

Standard MLST databases contain DNA sequences for 7 housekeeping genes shared by a pathogen species or a species group. Typical lengths of these genes vary in the range 350-500 basepairs. Let $g = 1, ..., 7$, denote the index of a single MLST gene and $\mathbf{x}_{ig}$ the observed DNA sequence for gene $g$ in strain $i$. It is assumed that the sequences $\mathbf{x}_{ig}$ are aligned and of the same length $d_g$ for all considered strains. The total set of sequences for each strain is written as $\mathbf{x}_i$. Each element $x_{ijg}$ in $\mathbf{x}_{ig}$ belongs to the finite alphabet $\mathcal{X} = \{A, C, G, T\}$, which is uniquely mapped to a set of integers such that we get the sample space $\mathcal{X} = \{1, \ldots, 4\}$ for each site $j$, $j = 1, ..., d_g$. However, to obtain a less parameter-heavy classification model, we define the sample spaces for all 3-mers in these sequences in a more parsimonious manner (for details see below).

Corander and Tang [4] introduced a Bayesian second-order Markov model for unsupervised classification of MLST sequence data, which aims at a balance between a parsimonious parametrization and an adequate representation of dependencies in observed nucleotide frequencies among neighboring sites. Such standard Markovian structures are ubiquitous in statistical modeling of DNA sequences. Here we adapt this modeling framework to a semi-supervised setting, where training data are used to pre-specify a finite set of $k_1$ possible distinct sources of new test strains, while not excluding the possibility that some (or even all) of these have emerged from a previously unseen evolutionary group. Let $V_g = \{1, ..., d_g\}$ denote the index set of the site variables $x_{ijg}$ and $G_g = G_g(V_g, E_g)$ an undirected graph on the node set $V_g$ with the edges in set $E_g$. The edge set is determined by a second-order Markov structure where for any pair $\{j, j^*\}$ of site indices $\{j, j^*\} \in E_g$ if and only if $|j - j^*| < 3$. Given the standard properties of decomposable graphical models [10], such a dependence structure leads to an explicit factorization of the joint probability distribution of site patterns given a joint classification of the training and query data. To define the factorization we let $cl(G_g)$ and $sep(G_g)$ denote the sets of cliques and the set of separators of the cliques of graph $G_g$, respectively. The cliques correspond to all the triplets of consecutive site indices, whereas the separators correspond to all the pairs of consecutive site indices, except the first and last pairs for each gene.

Assuming there are in total $n$ strains in a particular query, we index them by the set of integers $N = \{1, ..., n\}$. The observed sequence data for any subset $s \subseteq N$ of query strains is given by the collection $\mathbf{x}^{(s)} = \{\mathbf{x}_i : i \in s\}$, and hence $\mathbf{x}^{(N)}$ represents the entire query data set. The sequence types (STs) existing in a curated MLST database are used as labeled training data, indexed by $M = \{1, ..., m\}$. The labels are assumed to be specified by an earlier analysis of the database contents, which divides the $m$ STs into $k_1$ distinct evolutionary groups using, for instance, an unsupervised classification with the BAPS software. The labeling $T$ of the training data is a joint classification of the $m$ STs into $k_1$ classes and we use $\mathbf{z}_{ig}$, $\mathbf{z}_i$, $\mathbf{z}^{(s)}$, $\mathbf{z}^{(M)}$ for training data in a notation analogous to the query data as defined above.

For a set $a_g$ of sequence sites indexed by $V_g$, such that the cardinality $|a_g|$ equals three, we let $\mathbf{x}_{iag}$, $\mathbf{z}_{iag}$ be the corresponding 3-mers observed in gene $g$ for strain $i$ in the query and training data sets. Further, we let $r_{a_g}$ equal the total number of distinct 3-mers observed at the sites $a_g$ in the joint collection of query and training data: $r_{a_g} = |\{\mathbf{x}_{iag} : i \in N\} \cup \{\mathbf{z}_{jag} : j \in M\}|$.

Let $S$ denote a joint classification of the $n$ query STs into the $k_1 \geq 1$ sources labeled by training data and $k_2 \geq 0$ putative novel sources. Thus, $S = (s_1, ..., s_{k_1}, s_{k_1+1}, ..., s_{k_1+k_2})$ defines a (possibly) partially labeled partition of $N$ (semi-supervised classification), such that $\bigcup_{c=1}^{k_1+k_2} s_c = N$ and $s_c \cap s_{c'} = \emptyset$, for all pairs $\{c, c'\}$ ranging between 1, ..., $k_1 + k_2$. The partition is completely labeled (supervised classification) when $k_2 = 0$, that is when no query STs are assigned into previously unknown sources. We use $\mathcal{S}$ to denote the space of possible values of the semi-supervised classification structure $S$, conditional on a user-specified upper bound for $k_1 + k_2$.

The joint conditional likelihood of query data given the classification $S$ and the training data labeling $T$ is under our Markov model defined as

$$p(\mathbf{x}^{(N)}|\theta, S, T) = \prod_{c=1}^{k_1+k_2}\prod_{g=1}^{7}\frac{\prod_{a_g\in cl(G_g)} p(\mathbf{x}_{a_g}^{s_c})}{\prod_{b_g\in sep(G_g)} p(\mathbf{x}_{b_g}^{s_c})} = \prod_{c=1}^{k_1+k_2}\prod_{g=1}^{7}\frac{\prod_{a_g\in cl(G_g)}\prod_{t=1}^{r_{a_g}} p_{cga_gl}^{n_{cga_gl}}}{\prod_{b_g\in sep(G_g)}\prod_{t=1}^{r_{b_g}} p_{cgb_gl}^{n_{cgb_gl}}}, \quad (1)$$

where $b_g$ and $r_{b_g}$ are defined for subsets in $sep(G_g)$ analogously to the subsets in $cl(G_g)$, $p_{cga_gl} > 0$, $p_{cgb_gl} > 0$ are the probabilities of observing the $l$th 3-mer and 2-mer, respectively, in class $c$, and $n_{cga_gl}$, $n_{cgb_gl}$ are sufficient statistics corresponding to the observed counts of the $l$th 3-mer and 2-mer in class $c$. Parameter $\theta$ in (1) is used as a joint abbreviation for all the continuous parameters in the expression, which correspond to probabilities of observing the particular site patterns within the classes. Notice that the probabilities $p_{cgb_gl}$ and counts $n_{cgb_gl}$ are unambiguously determined by marginalization from $p_{cga_gl}$ and $n_{cga_gl}$, since each $b_g$ is a subsequence of a $a_g$ with cardinality equal to two, which follows from the order of the Markov model. Since the probabilities $p_{cga_gl}$ are unknown parameters, the training data are used for learning them for the $k_1$ *a priori* known classes, whereas only non-informative prior distributions are used for inferences about the remaining $k_2$ classes. Furthermore, since these probabilities are nuisance parameters regarding the classification task, they should be integrated out when making inferences about the classification $S$.

Assuming standard Dirichlet $(\{\lambda\}_{l=1}^{r_{a_g}})$ prior distributions which are factorized with respect to the graphs $Gg$ for all components of $\theta$ [10,11],we can derive an analytical expression for the posterior probability $p(S|\mathbf{z}^{(M)}, \mathbf{x}^{(N)}, T)$ of $S$. The conjugate Dirichlet prior is widely adopted in particular in bioinformatics applications due to the computational advantage provided by analytical marginalization over frequency (nuisance) parameters in multinomial models. The posterior of $S$ equals

$$p(S|\mathbf{z}^{(M)}, \mathbf{x}^{(N)}, T) = f(\mathbf{z}^{(M)}, \mathbf{x}^{(N)}, T)^{-1} \cdot \prod_{c=1}^{k_1}\int_\Theta p(\mathbf{x}^{(s_c)}|\theta, S, T)p(\theta|\mathbf{z}^{(M)}, S, T)d\theta \cdot \prod_{c=k_1+1}^{k_1+k_2}\int_\Theta p(\mathbf{x}^{(s_c)}|\theta, S)p(\theta|S)d\theta p(S|T), \quad (2)$$

where $p(S|T) > 0$ is the prior probability of $S$ and $f(\mathbf{z}^{(M)}, \mathbf{x}^{(N)}, T)$ is a normalizing constant equal to the sum

$$f(\mathbf{z}^{(M)}, \mathbf{x}^{(N)}, T) = \sum_{S\in\mathcal{S}}\prod_{c=1}^{k_1}\int_\Theta p(\mathbf{x}^{(s_c)}|\theta, S, T)p(\theta|\mathbf{z}^{(M)}, S, T)d\theta \cdot \prod_{c=k_1+1}^{k_1+k_2}\int_\Theta p(\mathbf{x}^{(s_c)}|\theta, S)p(\theta|S)d\theta p(S|T). \quad (3)$$

In the expressions below $m_{cga_gl}$ is the observed count of the $l$th 3-mer from the training data on class $c$ and $m_{cgb_gl}$is the corresponding marginalized count. The first one of the two above integrals can be written in detail as

$$\prod_{c=1}^{k_1}\int_{\Theta} p(\mathbf{x}^{(s_c)}|\theta,S,T)p(\theta|\mathbf{z}^{(M)},S,T)d\theta = \prod_{c=1}^{k_1}\prod_{g=1}^{7}\frac{\prod_{a_g\in cl(G_g)} f_{a_g}(\mathbf{z}^{(M)},\mathbf{x}^{(s_c)},S,T)}{\prod_{b_g\in sep(G_g)} f_{b_g}(\mathbf{z}^{(M)},\mathbf{x}^{(s_c)},S,T)}, \quad (4)$$

where the term $f_{a_g}(\mathbf{z}^{(M)},\mathbf{x}^{(s_c)},S,T)$ equals

$$f_{a_g}(\mathbf{z}^{(M)},\mathbf{x}^{(s_c)},S,T) = \int_{\Theta_{a_g}}\cdots\int\prod_{l=1}^{r_{a_g}} p_{cga_gl}^{n_{cga_gl}+m_{cga_gl}+\lambda_{cga_gl}-1} dp_{cga_g1}\cdots dp_{cga_gr_{a_g}}, \quad (5)$$

which further simplifies to the expression

$$\frac{\Gamma\left(\sum_{l=1}^{r_{a_g}}\lambda_{cga_gl}\right)}{\Gamma\left(\sum_{l=1}^{r_{a_g}}\lambda_{cga_gl}+n_{cga_gl}+m_{cga_gl}\right)}\cdot\prod_{l=1}^{r_{a_g}}\frac{\Gamma(\lambda_{cga_gl}+n_{cga_gl}+m_{cga_gl})}{\Gamma(\lambda_{cga_gl})} \quad (6)$$

where $\Theta_{a_g}=\{p_{cga_gl}>0:\sum_{l=1}^{r_{a_g}}p_{cga_gl}=1\}$. Correspondingly, $f_{b_g}(\mathbf{z}^{(M)},\mathbf{x}^{(s_c)},S,T)$ equals

$$f_{b_g}(\mathbf{z}^{(M)},\mathbf{x}^{(s_c)},S,T) = \int_{\Theta_{b_g}}\cdots\int\prod_{l=1}^{r_{b_g}} p_{cgb_gl}^{n_{cgb_gl}+m_{cgb_gl}+\lambda_{cgb_gl}-1} dp_{cgb_g1}\cdots dp_{cgb_gr_{b_g}}, \quad (7)$$

which in turn simplifies to

$$\frac{\Gamma\left(\sum_{l=1}^{r_{b_g}}\lambda_{cgb_gl}\right)}{\Gamma\left(\sum_{l=1}^{r_{b_g}}\lambda_{cgb_gl}+n_{cgb_gl}+m_{cgb_gl}\right)}\cdot\prod_{l=1}^{r_{b_g}}\frac{\Gamma(\lambda_{cgb_gl}+n_{cgb_gl}+m_{cgb_gl})}{\Gamma(\lambda_{cgb_gl})}, \quad (8)$$

where $\Theta_{b_g}=\{p_{cgb_gl}>0:\sum_{l=1}^{r_{b_g}}p_{cgb_gl}=1\}$.We set all the Dirichlet hyperparameters in $\{\lambda\}_{l=1}^{r_{a_g}}$ equal to the reference value $1/r_{a_g}$, which is generalization of the Jeffreys' prior and reflects *a priori* symmetry with respect to the 3-mer values. For a detailed discussion about such reference priors see [11]. The prior distribution of $S$ is set equal to the uniform distribution in $\mathcal{S}$, which is defined as

$$p(S|T)=|\mathcal{S}|^{-1}, \quad (9)$$

where $|\mathcal{S}|$ refers to the cardinality of the space $\mathcal{S}$. Similarly, the second integral can be written as

$$\prod_{c=k_1+1}^{k_1+k_2}\int_{\Theta} p(\mathbf{x}^{(s_c)}|\theta,S)p(\theta|S)d\theta = \prod_{c=k_1+1}^{k_1+k_2}\prod_{g=1}^{7}\frac{\prod_{a_g\in cl(G_g)} f_{a_g}(\mathbf{x}^{(s_c)},S)}{\prod_{b_g\in sep(G_g)} f_{b_g}(\mathbf{x}^{(s_c)},S)}, \quad (10)$$

where again

$$f_{a_g}(\mathbf{x}^{(s_c)},S) = \frac{\Gamma\left(\sum_{l=1}^{r_{a_g}}\lambda_{cga_gl}\right)}{\Gamma\left(\sum_{l=1}^{r_{a_g}}\lambda_{cga_gl}+n_{cga_gl}\right)}\prod_{l=1}^{r_{a_g}}\frac{\Gamma(\lambda_{cga_gl}+n_{cga_gl})}{\Gamma(\lambda_{cga_gl})}, \quad (11)$$

and

$$f_{b_g}(\mathbf{x}^{(s_c)},S) = \frac{\Gamma\left(\sum_{l=1}^{r_{b_g}}\lambda_{cgb_gl}\right)}{\Gamma\left(\sum_{l=1}^{r_{b_g}}\lambda_{cgb_gl}+n_{cgb_gl}\right)}\prod_{l=1}^{r_{b_g}}\frac{\Gamma(\lambda_{cgb_gl}+n_{cgb_gl})}{\Gamma(\lambda_{cgb_gl})}, \quad (12)$$

since the previously unseen sources lack the training data observations.

We define our joint semi-supervised classifier as the classification structure $\hat{S}$ corresponding to the posterior mode over the distribution specified in (2)

$$\hat{S}=\arg\max_{S\in\mathcal{S}} p(S|\mathbf{z}^{(M)},\mathbf{x}^{(N)},T). \quad (13)$$

Given $\hat{S}$, one may calculate the conditional posterior distribution over possible assignments of the $n$ query STs according to

$$p(i\in s_c|\mathbf{z}^{(M)},\mathbf{x}^{(N)},T) = \frac{p(\mathbf{x}^{(N)}|\hat{S}(i\in s_c),\mathbf{z}^{(M)},T)}{\sum_{c=1}^{k_1+k_2}p(\mathbf{x}^{(N)}|\hat{S}(i\in s_c),\mathbf{z}^{(M)},T)}, \quad (14)$$

where $\hat{S}(i\in s_c)$ is the mode classification with $i$th query strain re-assigned to class $c$. These probabilities reflect the local posterior uncertainty about the possible sources of the query STs and they can be calculated in a simple manner using the above analytical expressions. In the next section it is shown how fast stochastic optimization can be used to obtain a plausible estimate of $\hat{S}$ in the online setting considered here.

### Inference algorithm

A standard Bayesian supervised classifier, for example the naive Bayes classifier [12], would treat each query ST separately and assign it to the class maximizing the posterior probability among the $k_1$ known alternative sources. Such an approach has very modest computational complexity and it can be easily extended to the

semi-supervised classification task where any single query ST is allowed to be assigned to an additional class lacking training data. However, considering the query STs individually has the disadvantage that when multiple STs are assigned to a previously unknown evolutionary group, the classifier provides no information about whether they should be interpreted as a single group or eventually be split into multiple novel lineages. In addition, when compared to a simultaneous classification, separate classification of all query STs offers lower statical power to detect strains from novel groups which are only modestly distinct from the $k_1$ groups in the training data. On the other hand, simultaneous semi-supervised classification of the query STs is computationally substantially more challenging than a separate classification, since the search operators must allow for the presence of multiple novel subsets of strains. Standard Bayesian computational tools, such as the Gibbs sampler [13], provide a straightforward way to implement a simultaneous semi-supervised classifier. However, due to their notoriously slow convergence for mixture models, they do not offer a highly versatile solution for an online application where query assignments are expected to be provided on a nearly real-time basis. Hanage et al. [14] analyzed a large MLST database for which they concluded that a Gibbs sampler based approach did not converge with a reasonable computational effort (~3 days on a single CPU). The same convergence issue was also explored for a different data type in [15], where Gibbs sampler and a stochastic greedy search algorithm were compared. Therefore, we use for semi-supervised classification the same efficient non-reversible stochastic search operators that are used for unsupervised classification of MLST data in the BAPS algorithm.

Given a set of query data and a preprocessed MLST database in which STs are divided into $k_1$ groups, it is necessary to determine first the total number $r_{a_g}$ of distinct 3-mers observed in the joint collection of query and training data for all collections of sites $a_g$ over the genes. This requires a linear scan of the observed sequences in the query data. Additionally, pairwise Hamming sequence distances are calculated for all pairs of query STs, as these are used to guide the stochastic search of the optimal classification. Notice that the unnormalized posterior probability distribution over the possible assignments $S$ of query strains is uniquely determined by the sufficient statistics $n_{cgb_gl}$, $m_{cgb_gl}$ and the Dirichlet prior hyperparameters $\lambda_{cgb_gl}$. Therefore, an efficient algorithm for searching the classification maximizing the posterior can be efficiently constructed by book-keeping changes in the sufficient statistics implied by re-assignments of subsets of query strains. The search operators in $\mathcal{S}$ that are used for improving any current state $S$ of the simultaneous assignment of query STs work as follows:

1. In a random order relocate each single ST to the class in $S$ that leads to the maximal increase in the posterior probability (2). This operator considers explicitly the assignment of each ST into a new singleton class, unless that would increase the number of classes $k_2$ beyond the user-specified upper bound.
2. Merge STs in the two classes of $S$ which leads to the maximal increase in the posterior probability (2). If no putative merge increases the probability, the state of $S$ is not altered. Notice that this operator applies to all classes irrespectively of their size, thus including any potential singleton classes introduced by the first or third operator.
3. In a random order, split each class into two maximally homogeneous subclasses using the complete linkage clustering algorithm with Hamming distances between the query STs. If a classification $S^*$ after split is associated with higher posterior probability than the current classification $S$, the split is accepted and otherwise it is rejected.
4. In a random order over the classes of $S$, simultaneous relocation of several STs from each class is attempted. The STs in a class are first sorted into a decreasing order with respect to the improvement in posterior probability (2) when they are assigned one-by-one into some other class, that is the ST associated with the largest improvement is placed first in the sorting etc. A candidate for new classification structure $S^*$ is then formed by relocating STs in this order to the class which leads to the largest increase in (2) or to the smallest decrease if no positive changes are possible. The relocation is continued either until the the total change in (2) becomes positive, in which case the candidate $S^*$ is set as the next state of the search algorithm, or until all STs in the class are relocated and the total change remains negative, in which case the candidate is rejected.

The search algorithm uses each of the above operators in varying combinations until no improvement in (2) is achievable after two consecutive attempts. Given its efficient implementation, even in an online application the algorithm can be independently run multiple times such that the globally best classification over the runs is chosen as the final estimate of the posterior mode classification. Multiple independent searches will reduce the probability that the best classification identified among them will be considerably suboptimal, representing a local peak in the posterior distribution. Since any two classification structures can be analytically compared, the searches can even be performed on separate

processors and results later combined using the batch mode interface of the BAPS software.

## Results

We have implemented the semi-supervised classification algorithm for MLST data in the BAPS software version 5.4 which is available for Windows, Mac OS × and Linux operative systems. It can be accessed both through the graphical user interface or the batch mode interface, which simplifies automation of the use of the tool in MLST web interfaces. In this section we demonstrate the performance of the semi-supervised classification tool using data from two MLST databases. The first database http://pubmlst.org/bcereus/ is for the pathogen species *Bacillus cereus* [16] and the second database http://saureus.mlst.net/ is for the pathogen species *Staphylococcus aureus* [17].

We extracted multilocus DNA sequences for 515 and 1404 STs from the two databases, respectively. The housekeeping genes used in typing of the *B. cereus* are: *glpF* (glycerol uptake facilitator protein), *gmk* (guanylate kinase), *ilvD* (dihydroxy-acid dehydratase), *pta* (phosphate acetyltransferase), *pur* (phosphoribosylaminoimidazolecarboxamide), *pycA* (pyruvate carboxylase), *tpi* (triosephosphate isomerase). The housekeeping genes used in typing of the *S. aureus* are *arc* (Carbamate kinase), *aro* (Shikimate dehydrogenase), *glp* (Glycerol kinase), *gmk*, *pta*, *tpi*, *yqi* (Acetyle coenzyme A acetyltransferase). The lengths of the MLST loci for *B. cereus* vary between 348-504 basepairs and the total concatenated length of the sequences equals 2829 basepairs. For *S. aureus* the lengths vary between 402-516 basepairs, the total concatenated length being 3198 basepairs.

A number of simulation experiments were performed using the real *B. cereus* and *S. aureus* data as the basis. Firstly, we divided the two databases into distinct groups of STs using an unsupervised classification (clustering) analysis option available in BAPS software for MLST type data. This resulted in 11 and 6 groups for the *B. cereus* and *S. aureus* data, respectively. For *B. cereus* the group sizes varied between 9-127 STs and for *S. aureus* between 9-444 STs.

In the first experiment we chose randomly 30% of the database STs as query data and the remaining 70% were used as training data. The training data were pre-classified into the groups identified by the earlier unsupervised analysis and the query data were analyzed assuming that there are at most 10 novel groups present in it. This setup was replicated 10 times and we calculated for each random data configuration how well the labels of the query STs matched the pre-classification labels using the adjusted Rand Index (ARI) [18]. The average ARI over the replicates (with std.dev. in parenthesis) is 1.000 (0.000) and 0.999 (0.003) for the *B. cereus* and *S. aureus* data, respectively.

In the second experiment the database STs were not randomly chosen into the query data as such, but we selected instead randomly 30% of the database ST groups as query data (3/11 and 2/6 groups), while leaving the remaining groups as training data. Notice that in the first experiment every class that was previously identified from the database had approximately 30% of its STs included in the test data, and thus, the same underlying classes were present both in the training and test data sets. In contrast, in the second experiment the test data consisted of groups of STs which did not correspond to any groups present in the training data, and thus, the training and test data sets were completely non-overlapping in terms of underlying groups.

The corresponding ARI values as in the first experiment are now 0.988 (0.029) and 0.966 (0.045) for *B. cereus* and *S. aureus* databases, respectively. To illustrate the data in the simulation experiments we made two Neighbor-Joining (NJ) trees annotated with pre-classification and novel labels. The trees were created with MEGA 4 software [19] using the maximum composite likelihood option. In Figure 1, the semi-supervised labeling is shown for one of the *S. aureus* database replicates in the second experiment. Here there are two novel groups of STs in the test data and only a single ST in one of them is mislabeled (uncolored in Figure 1) in the semi-supervised analysis. Note that BAPS groups may occasionally deviate from the groups derived from a phylogenetic tree, primarily due to presence of recombinant alleles in the data. For instance, all the long branches present in Figure 1 are due to a strongly deviating allele at a single locus, or even at two loci for some STs. We detected these cases by using the BRAT software [20] to screen the entire database (exact results not shown). When the deviating alleles were removed, all the long branches present in Figure 1 vanished, such that the corresponding STs closely resemble strains present in the remaining lineages. For instance, the single red-labeled ST with a very long branch had at one locus an allele with closest match to another species (*S. epidermidis*) when its DNA sequence was queried at the NCBI nucleotide collection, which could represent either a result of genuine inter-species recombination or a case of DNA contamination in the laboratory. Since BAPS recognized that the ST in question had very close resemblance to other red-labeled STs at the remaining six loci, the probabilistic query did not yield a label indicating separate origin.

It should also be noted that a small number of STs labeled as red reside in the NJ tree among green labeled taxa and conversely, a small number of STs labeled as green reside among red labeled taxa. Such a deviance

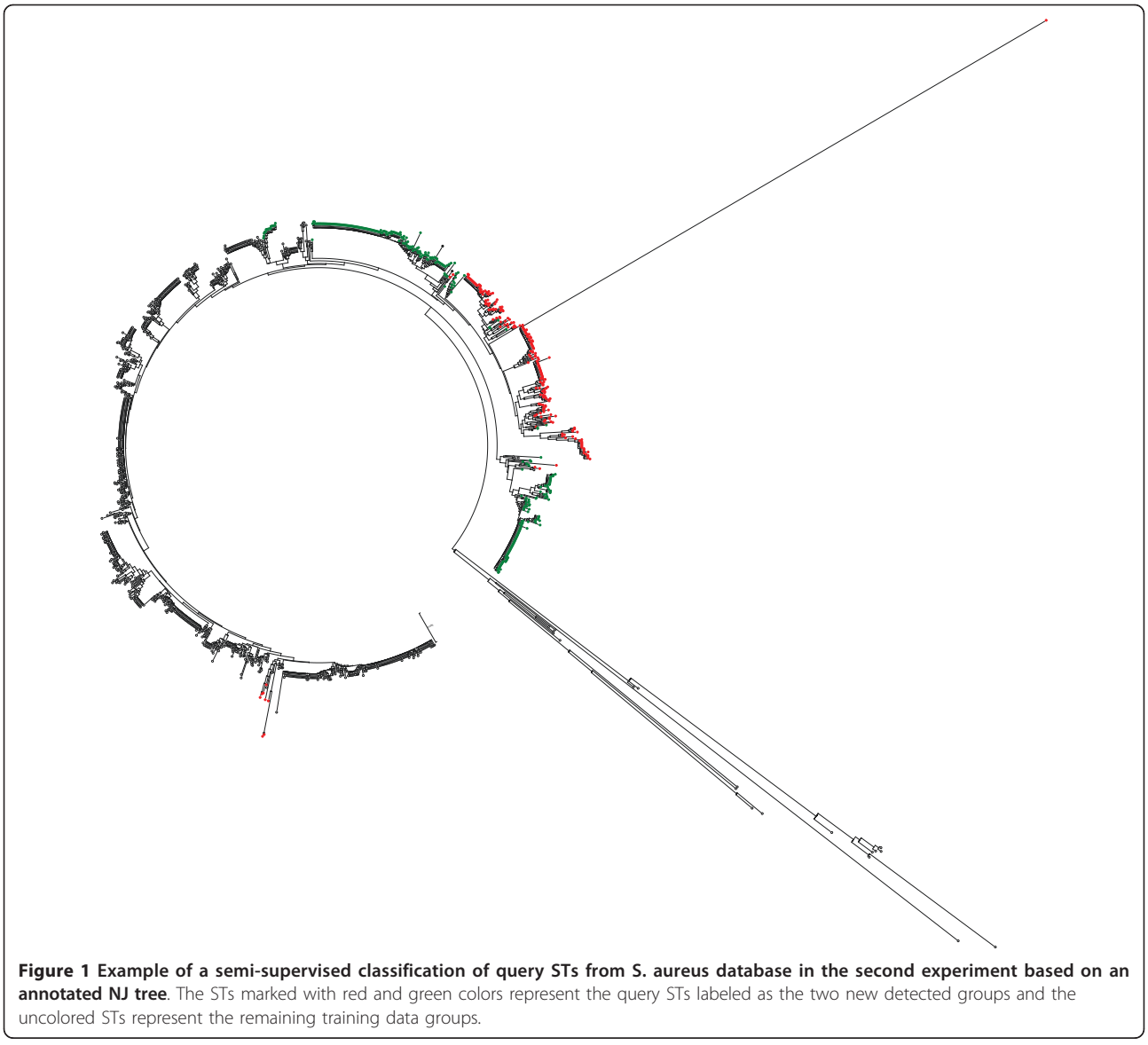

**Figure 1 Example of a semi-supervised classification of query STs from S. aureus database in the second experiment based on an annotated NJ tree**. The STs marked with red and green colors represent the query STs labeled as the two new detected groups and the uncolored STs represent the remaining training data groups.

has several possible explanations. Firstly, the labeling of these strains by the population genetic assignment may be erroneous, such that the tree correctly displays their origin. Secondly, due to the small evolutionary distances among these groups of strains, the NJ tree itself may provide a distorted view of their origin. In particular, under limited molecular resolution, the population genetic approach gains in a relative sense more statistical power to correctly detect lineage boundaries from a large sample in the presence of a small number of sites with highly characteristic nucleotides for a particular lineage, compared to a tree-based approach. This is primarily because the population genetic model directly compares nucleotide frequencies at sites within and between putative pools of samples and aims at answering a considerably simpler statistical question than a tree-based approach. We have obtained additional support for this tendency by examining data for *Burkholderia pseudomallei* STs, for which a large number of additional loci were available (exact data and results not shown). In general, in our bootstrap experiments BAPS assigned strains significantly to the correct lineage with a considerably smaller number of loci compared with a tree-based approach.

In the third experiment we chose the *B. cereus* database and introduced random mutations in the sequences of the query STs. Two types of query STs were generated to mimic a situation where some new strains represent previously detected lineages, whereas the others are sampled from multiple unseen lineages. To create novel

strains representing the first scenario, we chose randomly 25 STs from the database and introduced 1% of random mutations into their sequences. In addition, to create strains corresponding to the latter scenario, we randomly sampled 5 STs from the database and introduced 5% of random mutations to their sequences. Thereafter, 5 independent test strains were generated from each of these mutated STs by introducing further 1% of mutations to the sequences. The test data thus contains 50 query STs in total. Figure 2 illustrates the semi-supervised labeling of these data by showing simultaneously the training and test samples in an NJ tree. All 25 test STs representing previously sampled lineages, as well as all the five groups of STs from previously unsampled lineages were correctly labeled according to the group they were generated from.

The final experiment was performed to investigate the computational cost of applying our method to an online probabilistic query for an MLST database. We chose the following four sizes of query sets of STs to represent a wide range of typically expected queries: 5, 10, 50 and 100 STs. In each replicate of the experiment, the indicated number of STs were randomly chosen as test data and excluded from the database, while the remaining STs were used as training data. Independent point mutations were introduced to the sequences of test STs before submitting them as a query, such that on average nucleotide values at 1% of the sites were changed for the *B. cereus* STs and at 0.5% of the sites for the *S. aureus* STs. In total 10 replicates were performed on a PC with a 2.66 GHz processor and the mean time in seconds (SD within parenthesis) from the query submission to the final estimates of posterior assignment probabilities was for *B. cereus*: 0.320 (0.091), 3.887 (0.596), 56.175 (13.967), 181.705 (18.294), for the four distinct query set sizes,

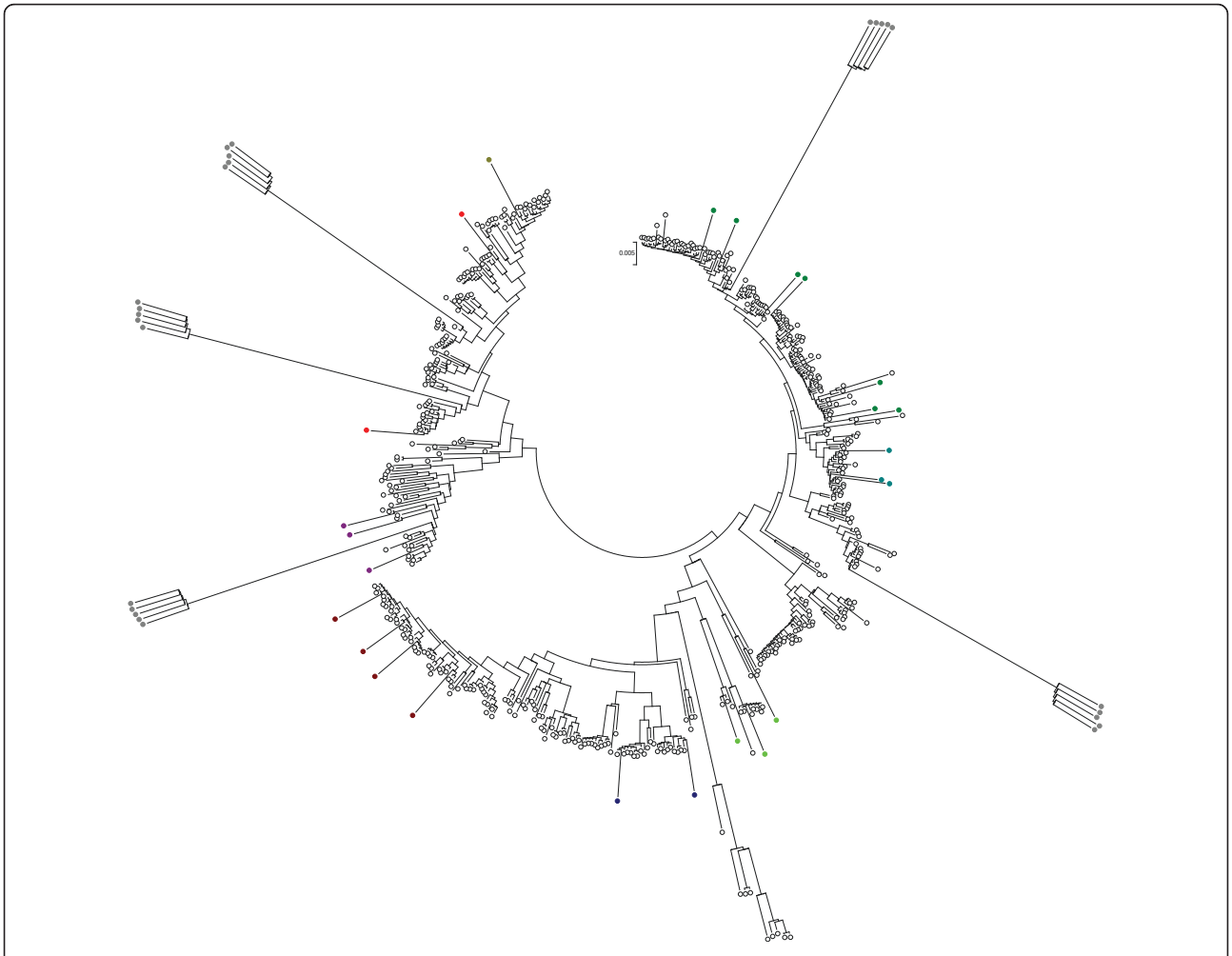


**Figure 2 Example of a semi-supervised classification of query STs from B. cereus database in the third experiment based on an annotated NJ tree**. The STs marked with grey colors are the new detected groups. The uncolored STs represent the STs in training data groups and the remaining colored STs are the 25 query STs that were correctly labeled by their respective groups.

respectively. For *S. aureus* the corresponding computation times were: 0.920 (0.034), 10.945 (2.102), 95.812 (12.966), 334.339 (84.523). These results illustrate that our method can easily be applied in an online query setting, as the required computation time is at most a couple of minutes even for large query sets. It is also worth noticing that the query sets are not expected to be that large in a majority of cases within clinical applications of MLST.

## Discussion

The epidemiological research community has with its combined efforts enabled a major leap forward in the understanding of the dynamics and evolution of major human and animal pathogens through the MLST web software. As all the MLST databases are continuously increasing in size and the popularity of these typing schemes continues to grow, the need of additional tools for rapidly simultaneously interfacing both previously curated and new data has emerged as well. Our example experiments based on real MLST databases illustrate that the model-based approach provides high accuracy in correctly labeling both strains from groups existing in the curated database as well as strains representing previously unseen lineages. In addition, our method provides a probabilistic characterization of the assignment uncertainty in terms of posterior probabilities calculated over the possible putative sources in the estimated mode classification structure. A classification framework where each query ST is labeled independently of other strains would provide a much simpler solution to the assignment problem in computational terms, however, on the other hand it is a more statistically coherent approach to handle all the query strains within a joint modeling framework to increase statistical power to detect samples from previously unseen evolutionary groups. It is worth noticing that since there is no other probability-based method available that would be tailored to MLST type data, we have not considered the semi-supervised classification task in a comparative fashion.

## Conclusions

We have introduced a model-based tool for automated semi-supervised classification of new pathogen samples that can be integrated into the web interface of the MLST databases. In particular, when combined with the existing metadata, the semi-supervised labeling may provide invaluable information for assessing the position of a new set of query strains in relation to the particular pathogen population represented by the curated database. Such information will be useful both for clinical and basic research purposes.

**Acknowledgements**

The authors would like to thank three anonymous reviewers and the participants of the 2nd Permafrost workshop for helpful comments and discussions on this work. This work was supported by a grant from Sigrid Juselius Foundation to JC, by Finnish Graduate School in Population Genetics (LC) and the Wellcome Trust (DMA and BGS).

**Author details**

[1]Department of Mathematics and statistics, P.O.Box 68, University of Helsinki, 00014, Finland. [2]Department of Infectious Disease Epidemiology, Imperial College London, Norfolk Place, London W2 1PG, UK. [3]Sanger Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge, CB10 1SA, UK.

**Authors' contributions**

TC introduced the original idea of using a model-based approach to semi-supervised classification of novel MLST sequence types. LC and JC designed and implemented the classification model, the stochastic inference algorithm and the computational experiments. TC and DMA developed the interface to MLST databases. BGS provided expertise on the clinical importance and use of MLST database information. All authors contributed to writing of the manuscript. All authors read and approved the final manuscript.

**Competing interests**

The authors declare that they have no competing interests.

Received: 18 April 2011 Accepted: 26 July 2011 Published: 26 July 2011

**References**

1. Maiden M, Bygraves J, Feil E, Morelli G, Russell J, Urwin R, Zhang Q, Zhou J, Zurth K, Caugant D, Feavers I, Achtman M, Spratt B: **Multilocus sequence typing: a portable approach to the identification of clones within populations of pathogenic microorganisms.** *Proceedings of the National Academy of Sciences of the United States of America* 1998, **95(6)**:3140-3145.
2. Spratt B: **Multilocus sequence typing: molecular typing of bacterial pathogens in an era of rapid DNA sequencing and the internet.** *Current opinion in microbiology* 1999, **2(3)**:312-316.
3. Feil E, Li B, Aanensen D, Hanage W, Spratt B: **eBURST: inferring patterns of evolutionary descent among clusters of related bacterial genotypes from multilocus sequence typing data.** *Journal of bacteriology* 2004, **186(5)**:1518-1530.
4. Corander J, Tang J: **Bayesian analysis of population structure based on linked molecular information.** *Mathematical biosciences* 2007, **205**:19-31.
5. Corander J, Marttinen P: **Bayesian identification of admixture events using multilocus molecular markers.** *Molecular ecology* 2006, **15(10)**:2833-2843.
6. Corander J, Marttinen P, Sirén J, Tang J: **Enhanced Bayesian modelling in BAPS software for learning genetic structures of populations.** *BMC bioinformatics* 2008, **9**:539.
7. Tang J, Hanage W, Fraser C, Corander J: **Identifying currents in the gene pool for bacterial populations using an integrative approach.** *PLoS Computional Biology* 2009, **5(8)**:e1000455.
8. Lee C, Abdool A, Huang C: **PCA-based population structure inference with generic clustering algorithms.** *BMC bioinformatics* 2009, **10(S1)**:S73.
9. Jombart T, Devillard S, Balloux F: **Discriminant analysis of principal components: a new method for the analysis of genetically structured populations.** *BMC genetics* 2010, **11**:94.
10. Lauritzen S: *Graphical models* Oxford: Oxford University Press; 1996.
11. Bernardo JS, Smith AFM: *Bayesian Theory* Chichester: Wiley; 1994.
12. Bishop C: *Pattern recognition and machine learning* New York: Springer; 2007.
13. Robert C, Casella G: *Monte Carlo statistical methods* New York: Springer; 2005.
14. Hanage W, Fraser C, Tang J, Connor T, Corander J: **Hyper-recombination, diversity, and antibiotic resistance in pneumococcus.** *Science* 2009, **324(5933)**:1454-1457.
15. Marttinen P, Myllykangas S, Corander J: **Bayesian clustering and feature selection for cancer tissue samples.** *BMC bioinformatics* 2009, **10**:90.
16. Jolley K, Chan M, Maiden M: **mlstdbNet - distributed multi-locus sequence typing(MLST) databases.** *BMC bioinformatics* 2004, **5**:86.

17. Enright M, Day N, Davies C, Peacock S, Spratt B: **Multilocus sequence typing for characterization of methicillin-resistant and methicillin-susceptible clones of Staphylococcus aureus.** *Journal of clinical microbiology* 2000, **38(3)**:1008-1015.
18. Hubert L, Arabie P: **Comparing partitions.** *Journal of classification* 1985, **2**:193-218.
19. Tamura K, Dudley J, Nei M, Kumar S: **MEGA4: molecular evolutionary genetics analysis (MEGA) software version 4.0.** *Molecular biology and evolution* 2007, **24(8)**:1596-1599.
20. Marttinen P, Baldwin A, Hanage W, Dowson C, Mahenthiralingam E, Corander J: **Bayesian modeling of recombination events in bacterial populations.** *BMC bioinformatics* 2008, **9**:421.

doi:10.1186/1471-2105-12-302
**Cite this article as:** Cheng *et al.*: Bayesian semi-supervised classification of bacterial samples using MLST databases. *BMC Bioinformatics* 2011 **12**:302.