Graph Cut for Identification of Bacterial Clonal Complexes from Multi-Locus Sequence Types

Wasinee Rungsarityotin[1]

[1] Computational Molecular Biology, Max Planck Institute for Molecular Genetics
Ihnestraße 63–73, D-14195 Berlin, Germany, rungsari@molgen.mpg.de
Tel: +49-30-8413-1166 Fax: +49-30-8413-1152

1 Introduction

Microbiologists require tools that can group closely related organisms at various levels of resolution. They also require a method that can discern the phylogenetic relationship between groups. This is especially important for the study of microbial pathogens and their epidemics, where the origin and diversity of pathogenic strains is a central concern [1, 11, 12]. One method used for distinguishing between different strains of pathogenic bacteria is multi-locus sequence typing (MLST) [5, 7]. In brief, MLST consists of identifying specific loci on the genome that code for neutral (and hence conserved) house-keeping genes. For each locus, a fragment of approximately 500bp is sequenced, and each unique sequence is assigned an arbitrary allelic label. Hence, given m loci, each individual MLST entry consists of a vector **S** of length m (for example $m = 7$), whereby each vector component s_i is an integer corresponding to the allele number. An MLST data set consists of an ordered set of vectors of type **S**. Each unique vector **S** is also given a label and referred to as a sequence-type or ST (eg. ST1).

MLST has proven to be a reliable method for identifying different strains within a bacterial species, provided that the loci chosen for analysis exhibit a sufficient amount of allelic variation [12]. However the analysis of MLST data sets for discerning higher order phylogenetic relationships has proven to be more difficult due to the fact that bacterial populations exhibit a high rate of recombination, see Feil et al. [3]. An intermediate approach for discerning phylogenetic groupings has been to cluster closely related STs into distinct groups. Each group is referred to as an ST complex. However there is no standardized methodology for identifying ST complexes across bacterial species. Assignment to different ST complexes has been so far performed by the convergence of different heuristic methods, such as eBURST, split decomposition and consensus estimations by epidemiologists, for example see Jolley et al. [5]. The fact that ST complex assignment is performed by the selective use of different methodologies, points to the need for a single method that can identify ST complexes based on a set of prescribed objective criteria.

Our method is to model a set of STs as a connected weighted undirected graph and find a k-partition of a vertex set in a graph by successively solving bipartition problems. The objective functions considered in the next section will provide a simple quality measure of a cluster without the need for information about the evolution of prokaryotes which is difficult to reconstruct. In this paper, our goal is to investigate if there is a statistical significance of biological groups derived from our clustering method.

2 Method

We address the problem of identifying k groups of sequence types (STs) as a k-way graph cut in which k is not known. We first convert the set of m sequence types into a fully connected undirected graph. The

m	the number of locus
s_i	allele number $\in \mathbb{N}$, $i \in \{1, ..., m\}$
$\mathbf{S}_i$	an ordered vector for a sequence type, $i \in \{1, ..., n\}$
V	a vertex set, each vertex is $\mathbf{S}_i$

Table 1: **Notations.**

graph is defined as $G = (V, E)$ where V is the set of nodes representing all STs and E is a set of edges $e(i, j)$ with weights $w(i, j)$, which are set to the similarity score between two sequence types: $\mathbf{S}_i$ and $\mathbf{S}_j$. A node i in G represents a unique sequence type $\mathbf{S}_i$. The choice of the similarity function depends on the data set, but they are constrained to be non-negative and symmetric. In this view, the problem of clustering unlabeled STs into k groups can be reduced to partitioning a vertex set V into k partitions. The problem of finding a k-partition can now be formulated as the problem of partitioning V into k subsets, $V = \cup_{i=1}^{k} V_i$.

Let us consider the problem of finding a $k = 2$ partition, say $V = A \cup B$. This can be achieved by removing edges $\{i, j\}$ from E for which $i \in A$ and $j \in B$. Such a set of edges which leaves the graph disconnected is called a *cut* and the weight function allows us to quantify cuts by defining their weight *cut-value*, $\mathbf{cut}(A, B) := \sum_{\{i,j\} \in E, i \in A, j \in B} w(i, j)$. A natural objective is to find a cut of minimal value. A problem with this objective function is that sizes of partitions do not matter. One alternative measure is the normalized cut, denoted by $\mathbf{Ncut}(A, B)$. We introduce the so-called *association* value of a vertex set A denoted by $a(A, V) := \sum\limits_{i \in A} \sum\limits_{j \in V} w(i, j)$ and defining the normalized cut by

$$\mathbf{Ncut}(A, B) = \frac{\mathbf{cut}(A, B)}{a(A, V)} + \frac{\mathbf{cut}(A, B)}{a(B, V)}.$$

We observe that the cut value is now set into relation to the similarity of each partition to the whole graph. Vertices which are more similar to many data points are harder to separate. As we will see, the normalized cut is well suited as an objective function for minimizing because it keeps the relative size and connectivity of clusters balanced.

The min-cut problem can be solved in polynomial time for $k = 2$. Finding k-way cuts in arbitrary graphs for $k > 2$ is NP-hard proven by Dahlhaus et al. [2]. For the other cut criteria, already the problem of finding a 2-way cut is in NP, for proof see [9]. However, we can find good approximate solutions [6, 9] to the 2-way normalized cut by considering a relaxation of the problem. Instead of discrete assignments to partitions consider a continuous indicator for membership. As it turns out, the eigenvectors obtained from a suitable eigenvector problem for the Laplacian of the pairwise-similarity graph G can be interpreted for exactly that purpose. This so-called spectral method has been used for solving the k-partition problem directly as well as through successive computation of 2-partitions. The successive 2-way problem can be used to solve the k-partition problem with a loose bound on the correctness of membership assignment, see [6] for more details.

For solving the 2-partition problem, we are interested in the eigenvector y_2 for the second-smallest eigenvalue, similar to the algorithm in Kannan et al. [6], Shi and Malik [9]. In particular, we will inspect its sign structure and use the sign of an entry $y_2(i)$ to assign vertex i to one or the other vertex set. Similarly, for direct computation of k-partitions one can use all k eigenvectors to obtain k-dimensional indicator vectors. Previous approaches [8, 9] relied on k-means clustering of the indicator vectors to obtain k clusters in this space.

In our problem of finding groups of STs, the number of partitions k is not known in advance. Because we solve this problem as a successive 2-way normalized cut, k is an outcome governed by the *cut-value* and the minimum size of partitions. As a terminating criterion, the threshold on the maximum of Ncut values will control the depth of cutting hierarchy and thus the number of leaves which are the number of partitions. Because our parameters of interest are the **Ncut** values and sizes of k groups, we search for the minimum of the $\mathbf{Ncut}(A, B)$ over the sorted 2nd-smallest eigenvector y. At each successive partition if the **Ncut** value is larger than a threshold on the maximum **Ncut** and the size of the resulting terminal

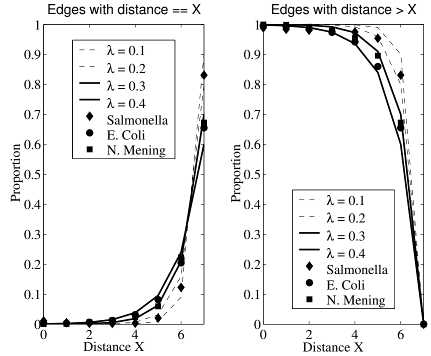

Figure 1: Observed distance and its distribution, compared to the estimation (dashed lines) from the geometric distribution $\mathbf{P}[X = d] = (1 - \lambda)\lambda^{(7-d)}$ with different rate λ.

set smaller than a certain size, we do not cut further. A hierarchy of cuts results in a tree whose edges are sets of collapsed edges in a cut and leaves are the terminal vertex sets. As we increase the threshold on the maximum **Ncut** we increase the depth of the cutting tree, but we have not yet explored if there is any biological meaning implied by the hierarchy of cuts.

3 Clustering MLST using the normalized cut

The goal of our clustering experiment is to recover known complexes and evaluate if the normalized cut leads to biologically meaningful groups. The dataset is unanimously curated by experts who define biological clonal complexes and other attributes. The feature set is a vector of seven alleles for the seven genes selected by the experts and dependent on the species. To assign a set of integer numbers to these seven alleles, biologists compare a sequence of a gene to the existing database of known alleles. A gene receives a new allele identification if its sequence does not completely match any other existing allele.

In our experiment with different weight functions, we investigated two choices: the Hamming distance and a probability that two STs share similar d or more alleles. We use the Hamming distance to count how many alleles are exchanged between two STs. Because epidemiologists believe that the distance 7 implies no relation, we prune all edges with distance $= 7$ from the graph. For the other weight, we use the fraction of all edges that have a larger distance than the given edge; the density is shown in Fig. 1.

3.1 Result and Discussion

We separately cluster the MLST data from each of three species: *E. coli*, *N. meningitis* and *Salmonella* because the gene sets are different. The MLST data from *E. coli* and *Salmonella* are a courtesy of our collaborator Mark Achtman and not yet publicly available. After clustering using various threshold values, to study how well the normalized cut results in biologically meaningful groups, we compute two quantitative measures: (1) the sensitivity and specificity and (2) the point-correlation between clusters and ST-Complex defined by experts. We also compute the statistical significance (p-value) for both measures. We cluster all STs and evaluate the response curve and significance on a subset with the annotation.

To test the specificity and sensitivity, we check whether cluster membership correctly predicts equality in ST-Complex. For example, for a true positive, belonging to the same cluster has to agree with being in the same ST-Complex. We count over all pairs of sequence types for which we have the gold standard available.

To test for correlation between clusters and ST-Complex, we apply the label permutation test, formally described by the Mantel's statistic (see e.g. Sokal and Rohlf [10]), for each biological property and compute the simulated p-value. Given two binary matrices X and Y, where X_{ij} is 1 when $\mathbf{S}_i$ and $\mathbf{S}_j$ are

ST-Complex		Methods					
		eBURST		**Ncut** $<$ Threshold			
		Size > 2	All	< 0.5	< 0.7	< 0.9	< 1.1
E. coli	Sens.	0.31	0.31	0.99	0.99	0.77	0.37
	Spec.	1.0	1.0	0.13	0.25	0.43	0.73
	k	21	38	4	6	15	30
Salmo.	Sens.	0.65	0.74	1.0	1.0	1.0	0.88
	Spec.	1.0	1.0	0.58	0.79	1.0	1.0
	k	4	18	11	16	27	30
N. mening.	Sens.	0.0	0.69	–	1.0	0.94	0.86
	Spec.	0.15	0.97	–	0.12	0.47	0.56
	k	134	223	1	9	30	449

Table 2: Specificity and sensitivity values on classification of ST-Complex from *E. coli, Salmonella and N. meningitis*. We performed clustering using all STs and computed the sensitivity and specificity on the subset with the expert's annotation on ST-Complex.

in the same cluster (or same class for Y_{ij}) and 0 otherwise, the Mantel's statistic Z is defined as:

$$Z = \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} X_{ij} Y_{ij}.$$

If clustering correlates with biological grouping, the observed Z should be high and distinct from the random Z-values. The computation of the distribution of Z-statistic is more complex than the Fisher's exact p-value. We are interested in the standardized form of Z-statistic defined by

$$\Gamma = \frac{Z - E_0[Z]}{\gamma},$$

where $E_0[Z]$ is the expected value of Z for a random observation and γ is its standard deviation. Because there is no closed form for computing $E_0[Z]$ and γ for random observations, we find them from a large number of simulated Z-values computed from random permutations of labels of X. To obtain a reliable p-value for Z, we need at least N^2 random permutations where N is the number of STs. We generated 10^4 random permutations and obtained significant results for all three species (p-value = 10^{-4}). We also computed the p-value at various thresholds of the **Ncut** values, but data not shown here due to limited space.

Comparison with eBURST clusters. We compare Ncut and eBURST with the expert's annotation and summarize the sensitivity and specificity as the ROC plot in Fig. 2. For each species, we perform clustering of all STs using eBURST and Ncut at various threshold values and evaluate the prediction power of both algorithms on a subset of annotated STs; therefore, discovery of new ST-complexes is not considered here. To be fair with eBURST, we consider two cases of clustering results: (1) all eBURST clusters including pairs and singletons and (2) only eBURST clusters larger than two STs. The result from eBURST appears to be consistent for *E. coli* and *Salmonella* regardless of cluster sizes, but for *N. mening.* eBURST clearly splits more ST-Complex than the Ncut method, hence low specificity and sensitivity when not considering pairs.

4 Conclusion

Identification of clonal complexes in bacterial species is a difficult problem due to recombination. Furthermore, it is still unclear if relying only on MLST from seven loci is sufficient. The expert's annotation

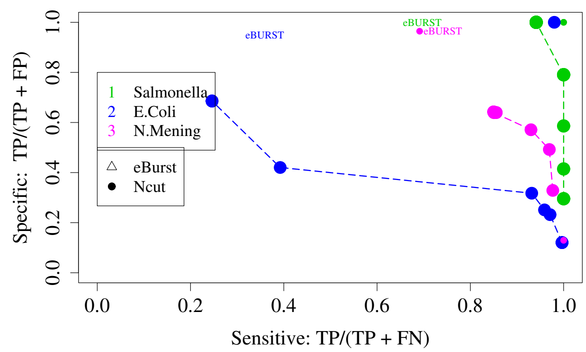

Figure 2: Receiver Operating Characteristic (ROC) of Ncut and eBURST. We can plot ROC for Ncut at various threshold values, but there is only one data point from eBURST.

data is also difficult to reconstruct because further knowledge involves in the annotation. Nevertheless, using only pairwise distance, we were able to obtain results that were comparable to manually curated data and consistent with previous methods such as eBURST. Our future work is to evaluate the mapping between clusters and phenotypes and compare the hierarchy of the cuts to the Split Decomposition [4].

Acknowledgements.
This publication made use of the Neisseria Multi Locus Sequence Typing web-site `http://pubmlst.org/neisseria/` developed by Dr. Man-Suen Chan and Dr. Keith Jolley and sited at the University of Oxford. The development of this site has been funded by the Wellcome Trust and the European Union. We acquired the version of the database in August, 2004 and the annotation may differ from the current version. We also acknowledge our collaborators: Homayoun C. Bagheri and Mark Achtman from Max Planck Institute for Infectious Biology for comment and discussion.

References

[1] M. Achtman. A phylogenetic perspective on molecular epidemiology. *Molecular medical microbiology*, 1: 485–509, 2002.

[2] E. Dahlhaus, D. S. Johnson, C. H. Papadimitriou, P. D. Seymour, and M. Yannakakis. The complexity of multiterminal cuts. *SIAM J. Comput.*, 23(4):864–894, 1994. ISSN 0097-5397.

[3] E. J. Feil, E. C. Holmes, D. E. Bessen, M. Chan, N. Day, M.C. Enright, R. Goldstein, D.W. Hood, A. Kalia, C.E. Moore, J. Zhou, and B.G. Spratt. Recombination within natural populations of pathogenic bacteria: short-term empirical estimates and long-term phylogenetic comparisons. *Proc Natl Acad Sci (U S A)*, 98: 182–187, 2001.

[4] D. Huson. SplitsTree: a program for analyzing and visualizing evolutionary data. *Bioinformatics*, 14:68–73, 1998.

[5] K. A. Jolley, M. S. Chan, and M. C. Maiden. mlstdbnet - distributed multi-locus sequence typing (mlst) databases. *BMC Bioinformatics*, 5(1):86, Aug 2004.

[6] Ravi Kannan, Santosh Vempala, and Adrian Vetta. On Clusterings: Good, Bad and Spectral. *Proceedings of IEEE Foundations of Computer Science*, 1999.

[7] M. C. Maiden, J. A. Bygraves, E. Feil, G. Morelli, J. E. Russell, R. Urwin, Q. Zhang, J. Zhou, K. Zurth, D. A. Caugant, I. M. Feavers, M. Achtman, and B. G. Spratt. Multilocus sequence typing: a portable approach to the identification of clones within populations of pathogenic microorganisms. *Proc Natl Acad Sci U S A*, 95 (6):3140–3145, Mar 1998.

[8] A. Y. Ng, M. I. Jordan, Y. Weiss., T. Dietterich, S. Becker, and Z. Ghahramani (Eds.). On spectral clustering: Analysis and an algorithm. *In Advances in Neural Information Processing Systems (NIPS) 14*, 2002.

[9] Jianbo Shi and Jitendra Malik. Normalized Cuts and Image Segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(8):888–905, 2000.

[10] Robert R. Sokal and F. James Rohlf. *Biometry: The Principles and Practice of Statistics in Biological Research*. W. H. Freeman, 1994.

[11] B. G. Spratt, E. Feil, and N. H. Smith. The population genetics of bacterial pathogens. *Molecular medical microbiology*, 1:445–484, 2001.

[12] R. Urwin and Martin C.J. Maiden. Multi-locus sequence typing: a tool for global epidemiology. *Trends in Microbiology*, 11(10):479–487, 2003.