Efficient Method for Inferring Hierarchy of Clonal Complexes from Multi-Locus Sequence Types

Wasinee Rungsarityotin [1], Mark Achtman [2], Homayoun Bagheri-Chaichian [3], Alexander Schliep [4]

Keywords: genetic structure, molecular sequence types, evolution, graph algorithms

1 Introduction.

In 1998 multi-locus sequence typing (MLST) was proposed as a nucleotide sequence based approach that could be applied to many bacterial pathogens [4]. 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 $\mathbf{S}$ of length m (for example $m = 7$ for E. coli), 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 $\mathbf{S}$. Each unique vector S is also given a label and referred to as a sequence-type or ST (eg. ST1). High-throughput sequencing technology facilitates large scale collection of MLST data and thus causes a need for a portable, reproducible, and scalable typing system that reflects the population and evolution of bacterial species. Perfect phylogeny may not work on MLST data because in practice there are not enough loci. Even if nucleotide sequences of isolates are available, it is still difficult to reconstruct a perfect phylogeny due to a high rate of recombination in bacterial pathogens [1, 5].

2 Algorithm.

Existing programs such as BURST [2], though simple to implement and visualize, do not provide an analytical method to infer relationship between groups of sequence types — clonal complexes. In this paper we examine an algorithm for finding k-way partitioning of a fully connected weight undirected graph which can be efficiently approximated with a generalized eigenvalues problem of the Laplacian matrix [3]. This methodology allows us to infer groups of clonal complexes, using only pairwise similarity. In brief, the algorithm is testing some objective functions for splitting, perform recursive bipartition of a vertex set of a graph. To decide which partition can be cut further, we order splits by their significance, considering the cost of the cut and number of clusters. The order in which the successive splits occur imposes a hierarchy of groupings, allowing to infer relations among complexes at varying levels of resolution. To confirm this hypothesis, more comparative result between simulation data and real data from bacterial species will appear in a future paper.

[1] Computational Molecular Biology, Max Planck Institute for Molecular Genetics, Ihnestraße 63–73, D-14195 Berlin, Germany. E-mail: rungsari@molgen.mpg.de

[2] Max Planck Institute for Infection Biology, Schumann Straße 21/22, D-10117 Berlin, Germany. E-mail:achtman@mpiib-berlin.mpg.de

[3] Max Planck Institute for Infection Biology, Schumann Straße 21/22, D-10117 Berlin, Germany.

[4] Computational Molecular Biology, Max Planck Institute for Molecular Genetics, Ihnestraße 63–73, D-14195 Berlin, Germany.

References

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

[2] The Multi Locus Sequence Typing website. http://www.mlst.net.

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

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

[5] Wang, L., Zhang, K. and Zhang, L. 2001. Perfect phylogenetic networks with recombination. *J. of Comp. Biology*, 8:69-78, 2001.