*Ec*MLST: an Online Database for Multi Locus Sequence Typing of Pathogenic *Escherichia coli*

Weihong Qi, David W. Lacher, Alyssa C. Bumbaugh, Katie E. Hyma, Lindsey M. Ouellette, Teresa M. Large, Cheryl L. Tarr, and Thomas S. Whittam
Microbial Evolution Laboratory, National Food Safety and Toxicology Center, Michigan State University
qiw@msu.edu

Abstract

In order to provide a portable and accurate typing system for the unambiguous characterization of pathogenic Escherichia coli isolates to the scientific community, we have constructed an online database for MultiLocus Sequence Typing of pathogenic E. coli (EcMLST) using current internet and open source technology. The system consists of an XML specification of the E. coli MLST system, and a set of perl modules defining the database tables and generating dynamic web pages for querying the database. It is implemented on a Sun server running the Apache web server. The underlying tier is the MySQL database system. Currently, the database contains nucleotide sequence data and annotated allelic profile data of 15 house-keeping genes for 600 representative E. coli isolates. Access to the central-held typing and epidemiology data is supported by parametric searching, full-text searching, as well as query interface links to the reference center of Shiga Toxin-producing E. coli (STEC, http://www.shigatox.net/stec). EcMLST has been used by public health laboratories and researchers for epidemiology and evolutionary studies. The system can be accessed at http://www.shigatox.net/mlst.

1. Introduction

Multi Locus Sequence Typing is a nucleotide sequence-based typing system that has been developed for tracking pathogenic bacteria with a global epidemiology perspective [1]. In this method, internal fragments of several (usually seven) house-keeping genes (loci) are sequenced. Each unique sequence is given a unique and arbitrary allele number. The combination of allele numbers at all loci is defined as the allelic profile. Each unique allelic profile is also assigned a unique and arbitrary number as the sequence type (ST) [2]. An organism-independent MLST database system (PubMLST) has been developed and applied to several pathogenic bacteria [3]. However, with the database structure and search algorithms of PubMLST, the system can only be accessed with sequences of loci that have been used to define the STs. This limits their applications in many clinical and research laboratories.

The pathogenic *Escherichia coli* MLST scheme developed in our laboratory uses internal fragments of seven house-keeping genes to define STs. Some isolates are also characterized using internal fragments of eight additional house-keeping genes [4]. We designed and developed a new MLST database system, *Ec*MLST, for the *E. coli* MLST scheme, which supports multiple avenues of access. For example, both Restriction Fragment Length Polymorphism (RFLP) data and sequence data of any 15 loci or any combination of 15 loci can be used to access the typing data. This allows more laboratories, especially those using RFLP for isolate characterizations, to make use of the database and the typing system.

2. Database design

The *Ec*MLST database system comprises locus tables, profile tables, and an isolate table.

There is one locus table for each gene (15 tables in total). Every locus table contains the

Proceedings of the 2004 IEEE Computational Systems Bioinformatics Conference (CSB 2004)
0-7695-2194-0/04 $20.00 © 2004 IEEE

sequences of all the alleles with the allele number as the primary key.

There are three different profile tables for easier data management. The first profile table contains complete allelic profiles of the 7 loci that define STs, ST being the primary key. The values of the allelic profile are linked to the locus tables. The second profile table contains both complete and incomplete allelic profiles of all 15 alleles and corresponding STs defined by 7-allele profiles. The values of the allelic profile are either linked to the locus tables, or are assigned allele number 99, representing unavailable sequence. Similarly, for each unique profile of 15 alleles, a unique and arbitrary identification number (ST15) is assigned as the primary key. Creation of the second profile table makes it possible to access the typing system with any or any combinations of the 15 alleles, not limited to the 7 alleles that define STs. The third profile table contains allelic profiles of all 15 alleles, in which the corresponding 7-allele profiles are incomplete. This profile table keeps track of the progress of the MLST project and isolate accession number is the primary key. Administrative information such as sender, curator, and date is also stored in all locus tables and profile tables.

The isolate table contains isolate accession number, ST, and ST15, with isolate accession number as the primary key. Through isolate accession number, the typing system is linked with the isolate resource database, which stores epidemiological and microbiological data of isolates.

3. Implementation

The system is implemented as a set of perl scripts for generating, managing and querying the database using perl DBI and CGI modules.

To make the system flexible enough to be applied to any bacteria, the database and web interface is generated based on an XML file describing the MLST system using the perl XML::Parser module. The database can be managed through an internal web interface using the WDBI system [5].

Available query methods include locus queries with RFLP data or sequence data. To support queries with RFLP data, the Bioperl module Bio::Restriction::Analysis, is used to predict cutting patterns of allele sequences stored in the database. Any or any combinations of 523 different restriction enzymes can be chosen for the prediction. Then the sizes of predicted internal fragments are matched with users' input data to find the closest related allele numbers. Once the allele numbers are known, allelic profile queries with any combinations of the 15 alleles, and sequence type queries with at least a given number of matches are useful for finding a group of closely related isolates. For all the methods, it is possible to submit data in batches.

The database can also be searched using customized queries, which allows users to combine all 27 searchable fields with AND or OR. For each field, queries can be made using =, >, <, <>, and LIKE.

For all searches, the query results are tabulated with hyperlinks to the complete information for each matched isolate and statistic analysis of all matched isolates.

Except database searching, there are also data analysis tools in the system, such as scripts for sequence concatenation and allele frequencies calculation. Additional data visualization and analysis tools are being developed and will be added into the web software.

4. Acknowledgements

This project has been funded in part with federal funds from the NIAID, NIH, DHHS, under Contract # N01-AI-30058.

5. References

[1] Urwin R, M.M., Multi-locus sequence typing: a tool for global epidemiology. *Trends Microbiol.*, 2003. 11(10): pp. 479-487.

[2] Maiden, M.C.J., et al., Multilocus sequence typing: A portable approach to the identification of clones within populations of pathogenic microorganisms. *PNAS*, 1998. 95(6): pp. 3140-3145.

[3] Chan, M.-S., M.C.J. Maiden, and B.G. Spratt, Database-driven Multi Locus Sequence Typing (MLST) of bacterial pathogens. *Bioinformatics*, 2001. 17(11): pp. 1077-1083.

[4] Whittam, T.S., et al., MultiLocus Sequence Typing (MLST) of Pathogenic Escherichia coli. http://www.shigatox.net/stec/mlst-new/index.html, 2002.

[5] Rowe, J., WDBI-Web Database Interface. http://emma-dev.ebi.ac.uk/wdbi/, 1999.