

# Rapid Assignment of Nucleotide Sequence Data to Allele Types for Multi-Locus Sequence Analysis (MLSA) of Bacteria Using an Adapted Database and Modified Alignment Program

M.A. Diggle<sup>1</sup> and S.C. Clarke<sup>\*1,2</sup>

<sup>1</sup>Scottish Meningococcus and Pneumococcus Reference Laboratory, North Glasgow University Hospital NHS Trust Department of Microbiology, Stobhill Hospital, Balornock Road, Glasgow, G21 3UW, UK

<sup>2</sup>Faculty of Biomedical and Life Sciences, University of Glasgow, UK

## Abstract

**A novel database and modified alignment program is described which provides a fast and accurate procedure for assigning nucleotide sequences to allele types for multi-locus sequence analysis (MLSA). The database has between 40 and 160 alleles per organism including *Neisseria meningitidis*, *Streptococcus pneumoniae*, *Staphylococcus aureus* and *Haemophilus influenzae*. The database directly compares the query nucleotide sequence against all alleles within the database and this system reduces the time taken for the analysis of nucleotide sequence data and assignment of alleles for subsequent sequence analysis.**

High-throughput nucleotide sequencing requires data handling systems which can maximise the use of a vast amount of information accurately and efficiently. MLSA utilises sequence data from a selection of housekeeping genes and antigenically variable genes from which sequence types for each allele are obtained (Clarke *et al.*, 2001a; Clarke *et al.*, 2001b). This genotyping tool has been successfully implemented as a method for better understanding the population biology of pathogenic bacteria and for the differentiation of strains that appear identical by standard phenotypic methods (Clarke *et al.*, 2001b; Dingle *et al.*, 2001; Enright and Spratt, 1999; Enright *et al.*, 2000; Enright *et al.*, 2001; Maiden *et al.*, 1998). It has also been used during outbreaks of high profile disease-causing organisms (Feavers *et al.*, 1999) and is now used for the characterisation of bacterial isolates received at a number of National Reference Laboratories (Feavers *et al.*, 1999; Clarke *et al.*, 2001a; Clarke *et al.*, 2001b). Such recent advances in nucleotide sequence methodology and technology generate vast amounts of information. We have therefore developed a nucleotide sequence database and modified alignment program based

on the DiscoverIR program (LI-COR Biosciences, Cambridge, UK). This has been achieved by using the main alignment system without using the necessary fixed set of compatibility sample files for sequence comparisons. The modified program only uses files in FASTA format and thus the program has greater flexibility as all DNA sequencers have a FASTA output format. The database stores housekeeping and virulence gene alleles for a number of medically important bacteria including *Neisseria meningitidis*, *Streptococcus pneumoniae*, *Staphylococcus aureus* and *Haemophilus influenzae*. Although there are other systems available for sequence analysis and characterisation, to our knowledge there is no simple tool available that can assign alleles for MLST with minimum time and effort. Alternative systems can execute various stages of this system although no system presently performs the full collection of tasks that are generally required such as the identification of nucleotide differences and allele assignment. To address this, we developed this adapted database and modified alignment program, which permits storage and retrieval of a variety of different multi-locus nucleotide sequences.

The database was developed to handle the large amount of nucleotide sequence data generated by our semi-automated nucleotide sequencing system (Clarke *et al.*, 2001a; Clarke *et al.*, 2001b). This program permits the comparison of sequence data in simple text format and does not discriminate against any individual sequence system, as all modern sequencers can perform base-calling and download into FASTA format. The nucleotide sequence of a particular organism in FASTA format is compared with the appropriate gene sequence database and the system automatically compares every query nucleotide sequence against known alleles. During alignment, the program examines the nucleotide base calls at each position of the sequence data and calculates the consensus for that position using a number of parameters. These include mismatch "penalty" whereby a penalty value is assigned to a mismatch when comparing two sequences. The larger the negative number, the more heavily matches are weighted. Open Gap penalty is the "cost" associated with existing gaps or insertions at a location in order to align two sequences. The algorithms were originally developed elsewhere (Huang, 1994). A list of sequence types is generated which is prioritised by the closest match, using a system which scores the closest match as the variant with the fewest nucleotide differences. A list of five sequence types are displayed and easily accessible, with the appropriate nucleotide differences automatically

\*For correspondence. Email [stuart.clarke@northglasgow.scot.nhs.uk](mailto:stuart.clarke@northglasgow.scot.nhs.uk); Tel. +44 141 201 3836; Fax. +44 141 201 3836.

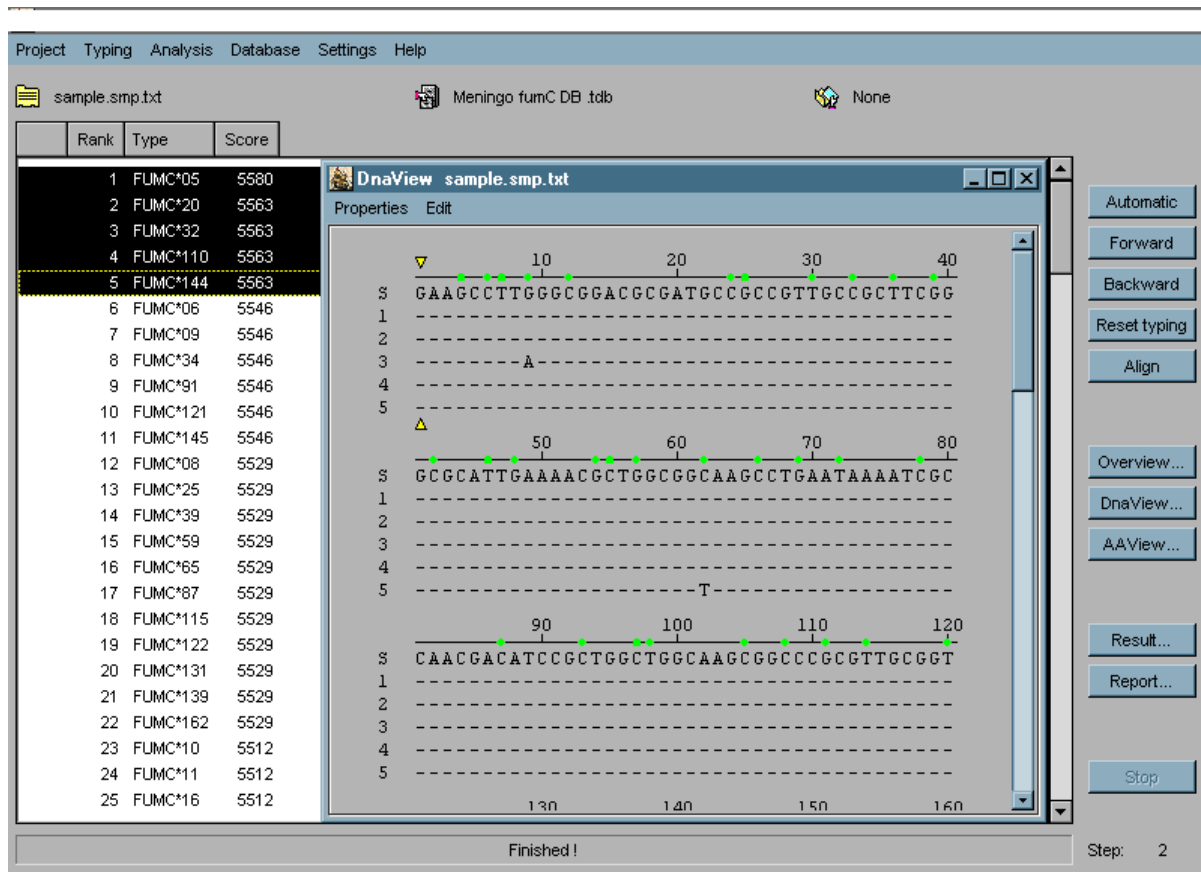


Figure 1. Example screen produced from the MLSA database. This view shows the analysis of the *fumC* gene from *Neisseria meningitidis*. The best five allele matches are highlighted on the left part of the screen and given a score, which in the case of the best match shown is 5580. Identity within the top five alleles matched is indicated by a '-' whereas variation is indicated by the nucleotide change. Variation at any nucleotide within any alleles contained in the database is indicated by a green dot above the respective nucleotide on the right hand portion of the screen.

displayed for detailed analysis. All variant types are freely available and downloaded from the multi-locus sequence typing (MLST) web-site (<http://www.mlst.net>). Our local database is updated each week with any new variants from the MLST web-site or from the *porA* web-site, which is used to discriminate different sero-subtypes of meningococci based on the analysis of two hypervariable regions (<http://neisseria.org/nm/typing/porA>) whilst others can be manually entered for individual projects.

The program is regularly used to characterise bacteria by MLSA within Scotland and, if an unknown sequence is found, the base-calling is checked manually on the DNA sequencer for any anomalies and the main nucleotide sequence databases are re-checked for new variants. If confirmed as a new variant, the sequence is subsequently forwarded to the main nucleotide sequence databases. One important key feature of this system is the acceptance of any length and quality of sample sequence and therefore a user is not at a disadvantage if not familiar with any aspect of a specific nucleotide sequence. The program can use "rough" data because it automatically eliminates sequence external of that being compared and concentrates on the discriminatory sequence for characterisation. This allows

any appropriate user to analyse the sequence and obtain a sequence type. For the purpose of MLSA, once all the allele types are known, a detailed report is produced which contains all the gene alleles and the resultant MLSA type which is then used for epidemiological studies and public health management (Figure 1).

This adapted database and modified alignment program is now used by our laboratory for MLSA of a number of important bacteria within Scotland. The database is held on the local network and accessed through Windows operating software. The DiscoverIR software is available through LI-COR Biosciences and costs approximately £3,000 for a two-user licence. The database modification is performed as follows:

- Install DiscoverIR10b software using manufacturers instructions.
- Within the software there are eight main folders, two of which are of importance, namely (1) Database and (2) Sample.
- Within the Database folder there are **tdb** files which are used to hold the information for the database .
- Create new **tdb** file(s) as required.

- Within each **tdb** file create your own list of alleles by precursing annotation for that sequence (a descriptive tag for that sequence type) e.g. `DATA; abcZ* 01;`. This annotation gives the type of information (data), the gene name (*abcZ*) and the sequence type (01).
- Follow this annotation with the actual sequence directly after the semi-colon.
- Repeat this for each allele starting on a new line for each.
- For each new gene create a new **tdb** file and repeat the same process.
- Once this is completed, a database of different files will exist containing alleles for MLST or other allele identification

The software can perform alignment and assignment at great speed as it is not reliant on remote systems for data analysis. Although the program is initially reliant on external data held on remote systems such as Genbank and MLST website databases, our experience indicates that remote real-time data analysis is time-consuming and access speed is dictated by the time of day and the number of users accessing the Internet. The database does not require specialist knowledge of the sequence data being analysed, does not require the exact size of gene fragment for analysis, and contains all the known variants which are available. This system is tailored towards facilities requiring a flexible, fast and accurate alignment program that is user-friendly and relatively inexpensive as an unmodified package.

Currently, this sequence database system is used for rapid characterisation and MLSA of a number of medically important bacteria. However, it can be used for any sequence project where unknown alleles require comparison against a database of known alleles. Importantly, this software can be modified for any organism and gene of interest so that it can be used in other national reference laboratories and research facilities for sequence analysis.

#### Acknowledgements

This publication made use of the Multi Locus Sequence Typing Website ([www.mlst.net](http://www.mlst.net)) developed by Dr Man-Suen Chan and David Aanensen and funded by the Wellcome Trust. We are also indebted to Martin Maiden, University of Oxford for his support and those who submit data. Robotic liquid handling systems and DNA sequencers used during MLSA at the SMPRL were kindly funded by the Meningitis Association (Scotland) and National Services Division of the Scottish Executive.

#### References

- Clarke, S. C., Diggle, M. A., and Edwards, G. F. 2001a. Semiautomation of Multilocus Sequence Typing for the Characterization of Clinical Isolates of *Neisseria meningitidis*. *J. Clin. Microbiol.* 39: 3066-3071.
- Clarke, S. C., Diggle, M. A., Reid, J. A., Thom, L., and Edwards, G. F. 2001b. Introduction of an automated service for the laboratory confirmation of meningococcal disease in Scotland. *J. Clin. Pathol.* 54: 556-557.
- Dingle, K. E., Colles, F. M., Wareing, D. R., Ure, R., Fox, A. J., Bolton, F. E., Bootsma, H. J., Willems, R. J., Urwin, R., and Maiden, M. C. 2001. Multilocus sequence typing system for *Campylobacter jejuni*. *J. Clin. Microbiol.* 39: 14-23.
- Enright, M. C., and Spratt, B. G. 1999. Multilocus sequence typing. *Trends Microbiol.* 7, 482-487.
- Enright, M. C., Day, N. P., Davies, C. E., Peacock, S. J., and Spratt, B. G. 2000. Multilocus sequence typing for characterization of methicillin-resistant and methicillin-susceptible clones of *Staphylococcus aureus*. *J. Clin. Microbiol.* 38: 1008-1015.
- Enright, M. C., Spratt, B. G., Kalia, A., Cross, J. H., and Bessen, D. E. 2001. Multilocus sequence typing of *Streptococcus pyogenes* and the relationships between emm type and clone. *Infect. Immun.* 69: 2416-2427.
- Feavers, I. M., Gray, S. J., Urwin, R., Russell, J. E., Bygraves, J. A., Kaczmarek, E. B., and Maiden, M. C. 1999. Multilocus sequence typing and antigen gene sequencing in the investigation of a meningococcal disease outbreak. *J. Clin. Microbiol.* 37: 3883-3887.
- Huang, X. 1994. On global sequence alignment. *Comput. Appl. Biosci.* 10, 227-235.
- Maiden, M. C., Bygraves, J. A., Feil, E., Morelli, G., Russell, J. E., Urwin, R., Zhang, Q., Zhou, J., Zurth, K., Caugant, D. A., Feavers, I. M., Achtman, M., and Spratt, B. G. 1998. Multilocus sequence typing: a portable approach to the identification of clones within populations of pathogenic microorganisms. *Proc. Natl. Acad. Sci. USA.* 95: 3140-3145.

