Statistical analysis of adaptation in *Salmonella*

Background

Salmonella is a bacteria that causes grave diseases in humans resulting in around 300,000 deaths annually. A large collection of over 2000 isolates has been typed over the past five years using Multi-Locus Sequence Typing (MLST). MLST consists in sequencing 7 short fragments of genes for each isolate (Maiden *et al*., 1998; Torpdahl *et al*., 2005). The similarities and differences observed between the sequences reveal the relationships between the isolates. Approximately half of the typed isolates are from human sources and half from veterinary origins. It is clear from this data that certain types are more likely to infect humans than animals, although the level of adaptation to the human host has never been properly quantified, and no evolutionary explanation has been given.

Objectives

The aim of this project is to propose and apply a model for the evolution of adaptation in *Salmonella*. In particular, we would like to determine how often adaptation occurred during the evolution of the species, and to determine points where it happened. This is valuable in itself (the frequency of occurrence of these changepoints is unknown at the moment, as well as how profound the changes are), but this analysis should also be useful as a first step in a two-step association mapping study (Falush and Bowden, 2006). In such a study, closely related isolates that differ in adaptation are first identified, and then fully sequenced to try and find the genetic elements responsible for the difference.

Research plan

The first task will be to create a phylogenetic tree showing how the isolates are related, and to look at how the isolates from human origin are distributed on this tree. We will then propose a stochastic model for how adaptation to the human host evolved in *Salmonella*. This model will finally be applied to the data using a Monte-Carlo Markov Chain (Gilks *et al*., 1997). Full training in MCMC methodology will be given, but candidates should be able to program in the computing language of their choice.

References

Falush and Bowden, 2006 Genome-wide association mapping in bacteria? Trends in Microbiology 14:353-355
Gilks *et al*., 1997 Markov Chain Monte Carlo in Practice. Chapman & Hall
Maiden *et al*., 1998 Multilocus sequence typing: a portable approach to the identification of clones within populations of pathogenic microorganisms. PNAS 95:3140-3145
Torpdahl *et al*., 2005 Genotypic characterization of *Salmonella* by multilocus sequence typing, pulsed-field gel electrophoresis and amplified fragment length polymorphism Journal of Microbiological Methods 63:173-184