Modelling with MATLAB: Assignment 4 2019 - Getting to grips with current research

Background: This assignment is based on the manuscript "Selection-based model of prokaryote pangenomes" by Maria Rosa Domingo-Sananes and James O. McInerney, which is available here:

https://www.biorxiv.org/content/10.1101/782573v1

Please download a copy of this document as a PDF. This is a complicated and exhaustive piece of current research. The manuscript asks, "Can simple mechanisms explain the variation in genetic content between individuals from the same prokaryote (i.e. bacterial) species?" **You do not need to understand, or even to read, the entire paper in order to answer the questions in this assignment. Your task is to implement, to test, and possibly to extend, some of its mathematical ideas.**

Please note that both authors have given their permission for their work to be used for this assignment. You must **not** contact the authors in relation to this assignment, or in any other way, until the assignment deadline has passed. (The authors have agreed not to answer any such questions, and to refer any inappropriate communications to the University of York.)

Your solutions should be uploaded to Moodle as a single written document, which may contain graphics and mathematics (but which does not just list your code), and which makes clear links to well-labelled MATLAB files which should be uploaded to Moodle (as .m files) at the same time. Your solutions may be in PDF (e.g. generated via LaTeX), Word, or any other appropriate format, but they must NOT depend on the marker having access to any additional software beyond a PDF reader, Microsoft Word, and a copy of MATLAB. Credit will be given for providing working code, and for providing suitable comments within the code to allow it to be used accurately. **Simply submitting a collection of MATLAB files is not enough.**

You will need to understand some basic biology (prokaryote, genome, core genes, accessory genes), but not at an advanced level. Wikipedia is a good start.

There are 4 questions for the H-level version of this assignment, and 5 questions for the M-level version of this assignment. Q1 carries 40 marks, and each of the other questions carries 20 marks. Q1 and Q2 concern simple models for a single gene with a fixed fitness effect, *s*. Q3 and Q4 concern genomes made up of a large number of genes, where the value of *s* for each gene in a given environment is simulated from a defined probability distribution. **Q5 is for students taking the M-level module only.**

Your answers need to be uploaded on Moodle by **1200 on MONDAY 6th JANUARY 2019**, please.

Q1. (a) Write a MATLAB function which will give the time derivatives for the manuscript's "Eq. 1" (p. 3, explained in lines 114-136) as a function of appropriate input variables and parameters.

(b) Choose a single set of parameter values for Eq.1, using the manuscript to provide justification for your choices.

(c) For your chosen set of parameter values, use a numerical method to identify any biologically realistic equilibrium values (fixed points) for Eq. 1.

(d) Use an appropriate numerical method to investigate the stability of any fixed points identified in 1(c). Describe the time-dependent behaviour of the system for your choice of parameters.

Q2. The idea that each gene has its own fitness effect, s, is explained in lines 139-148. By adapting your solutions to question 1, or otherwise, verify that the results shown in Figure 2b (p. 6), which show how gene frequency x changes with s, are correct. (The simplest way to do this is to use your numerics to reproduce Figure 2b using the parameter values given in the manuscript.)

Q3. The manuscript shows how the dynamics of a large number of different genes can be simulated, where the parameters for each gene may be drawn from random distributions, as explained in lines 234-304. Show how the single gene code from Q1 and Q2 can be adapted to describe a genome of a large number of independent genes, each of which has a fixed value of s simulated from an exponential distribution with rate 1, mean shifted to -1. All genes are otherwise identical. Hence, verify that the results shown in Figure 3 III c and d (p. 8) are qualitatively correct. (The simplest way to do this is to use your numerics to qualitatively reproduce Figure 3 III c and d using the parameter values given in the manuscript.)

Q4. The manuscript assumes that the environment in which the organisms live is constant. Suppose, instead, that the environment changes between two states, "A" and "B", periodically every τ time units. When the environment is in state "A" each gene i has a fixed value of s_{Ai} simulated from an exponential distribution with rate 1, mean shifted to -1. Similarly, in environmental state "B" each gene i has a fixed value of s_{Bi} simulated from an exponential distribution with rate 1, mean shifted to -1. Parameters s_{Ai} and s_{Bi} are independent, and all genes are otherwise identical. Use computer simulations to assess to what extent the conclusions of Figure 3 III are changed by this incorporation of environmental variability. **Your answer for Q4 must not exceed 1 page of A4** (in addition to any code you may wish to submit).

Q5. For students taking the M-level module only. Suppose that the model in Eq. 1 is modified so that the rate of gene loss is described by the delay differential equation

$$dx/dt = r_g\,(1 - x(t)) + s\,x(t)\,(1 - x(t)) - r_l\,x(t-\delta)$$

where δ represents a fixed non-negative delay. For your choice of parameters in 1c., investigate whether such a delay can destabilise the system. **Your answer for Q5 must not exceed 1 page of A4** (in addition to any code you may wish to submit).