High Performance Computing Programming Exercises

28th November - 2nd December 2016

James Rosindell (j.rosindell@imperial.ac.uk)

On each question, it will be indicated [in brackets] how many marks are available. The marks add up to 100, but together they only count for 60% of your final score for the assignment. The remaining 40% of your final score will be discretionary and based on the overall quality of your text answers and the extent to which they demonstrate your understanding of the topics, the style and quality of the computer code that you handed in and your answers to any of the challenge questions. 'Challenge questions' are therefore not essential but attempting some of them will always improve your result. It is advisable that you only work on 'challenge questions' after you have answered the main questions and checked your answers thoroughly. The reason for this mark scheme is to bring the mark distribution more in line with that which you would expect from an essay question.

You should hand in three files:

- *A typed document giving your written answers to the questions marked ⋆.*
- *A single file containing all the commented R code that you ran on your own computer to complete the worksheet.*
- *A zip file containing all your results files from the cluster along with the shell script and R code that was run on the cluster to produce them.*

Many of the functions in your R code file will be marked automatically – so be careful to:

- *Name the files by your username e.g. jrosinde.R*
- *Name all your functions exactly as in the instructions.*
- *Put "rm(list=ls())" and "graphics.off()" at the top of your file.*
- *Do not use packages, they should not be needed for this.*
- *If I run the source command on your file it should run without error and load all the functions into memory so that they can be tested, it should not actually run the functions or do anything else.*

Neutral Theory Simulations

These questions build on one another step by step so by the end you will have produced your own individual based simulation code in R.

You will store the state of your simulated system as a vector of individuals called ***'community'***. Each entry in the vector is a number that tells you the species of the individual in that position.

1.) You will need to know the species richness of your system so write a function ***'species_richness'*** to measure the species richness in the vector of individuals in your system. For example, ***species_richness(c(1,4,4,5,1,6,1,2))*** should return 5. (Hint: use the 'unique' command) [2 marks]

2.) Write a function ***'initialise_max'*** to generate an initial state for your simulation community with the maximum possible number of species for the community of size J individuals. For example ***initialise_max(7)*** should return a vector { 1 2 3 4 5 6 7 } (Hint: use the 'seq' command) [1 mark]

3.) In this type of simulation, it's important to consider the effect of the initial condition so write another function ***'initialise_min'*** to generate an alternative initial state for your simulation of a certain size with the minimum possible number of species (that's mono-dominance of one species). For example ***initialise_min(4)*** should return a vector { 1 1 1 1 }. [1 mark]

You will need to pick an individual to die and another to reproduce to fill the gap left by the death - they should not be the same individual (though they could be of the same species). It is important not to confuse the index of your ***community*** vector (representing the ID of an individual organism) with the number stored in that position of the vector (representing the ID of the species that individual belongs to).

4.) Write a function ***'choose_two(x)'***. This function should first choose a random number according to a uniform distribution between 1 and x inclusive of the endpoints. Next it should choose a second random number also between 1 and x but not equal to the first number. The numbers should be returned as a vector of length 2. So ***'choose_two(4)'*** should return one of the following vectors with equal probability:
{1 2} , {1 3} , {1 4} , {2 1} , {2 3} , {2 4} , {3 1} , {3 2} , {3 4} , {4 1} , {4 2} , {4 3}
(Hint: use the 'sample' command) [2 marks]

5.) Write a function ***'neutral_step'*** to perform a single step of a simple neutral model simulation, without speciation, on a community vector. That is, use ***choose_two*** to pick an individual to die and another to reproduce - then perform the death and reproduction on your community vector and return the result. For example ***neutral_step(c(10,5,13))*** should return one of the following six community states with equal probability:
{ 5 5 13 } when the first individual dies and is replaced by the second's offspring
{ 13 5 13 } when the first individual dies and is replaced by the third's offspring
{ 10 10 13 } when the second individual dies and is replaced by the first's offspring
{ 10 13 13 } when the second individual dies and is replaced by the third's offspring
{ 10 5 10 } when the third individual dies and is replaced by the first's offspring
{ 10 5 5 } when the third individual dies and is replaced by the second's offspring
[1 mark]

6.) Write a function ***'neutral_time_series'*** that will do a neutral theory simulation and return a time series of species richness in the system. The function should have three inputs: ***initial*** (the initial condition, which also determines the simulation size), ***duration*** (the number of time steps), and ***interval*** (the interval between time steps where you record the species richness of the system). The first value in the time series should be the species richness of the initial condition at time 0. The function should return a list – which enables you to return two objects. The first element in the list will be the time series as a vector and the second element will be the state of the community at the end of the simulation (to allow for further analysis to be carried out on the resulting community, or to allow further simulation). For example ***neutral_time_series (initial = initialise_max(7) , duration = 20 , interval = 2)*** should return a list containing firstly a time series vector of length 11 with the first value being 7 and secondly a community vector of length 7 with values giving species identities. Running the simulation again with a duration of 21 will make no difference to the size of the output time series though the time series itself is stochastic so will probably be different (Hint: use %% and list) [4 marks].

7.) ⋆ Plot a time series graph of your neutral model simulation from an initial condition of maximal diversity in a system size of 100 individuals. Run the simulation for 10,000 time steps and record the species richness every 10 time steps. Make sure that the x-axis on

your plot shows the number of model steps not the number of readings - you are recording richness only every 10 time steps so there will be 10 model steps between any pair of readings that get returned to you by ***neutral_time_series***. (Hint: use 'seq'). Include the code you wrote for this question in a function called ***'question_7'*** which should require no inputs to run. What state will the system always converge to if you wait long enough? Why is this? [3 marks]

8.) Write a new function ***'neutral_step_speciation'*** which will perform a step of a neutral model with speciation. In each time step, speciation will replace a dead individual with a new species (with probability ν) otherwise the dead individual is replaced with the offspring of another individual as before in ***'neutral_step'***. You should leave speciation rate, ν , as a parameter in your function. For example, ***'neutral_step_speciation(c(10,5,13),v = 0.2)'*** should behave like ***'neutral_step(c(10,5,13))'*** with probability 0.8, and with probability 0.2 it should instead be equally likely to return any of the following three vectors { 1 5 13 } , { 10 1 13 } , { 10 5 1 } with the number 1 representing the new species. (Hint: use the 'runif' command, also be careful to make sure that any new species really have a unique number assigned to them that has not been used before - it needn't be 1 in the above example but it cannot be 5, 10 or 13) [3 marks]

9.) Make a new function ***'neutral_time_series_speciation'*** which uses a neutral simulation with speciation, but otherwise performs in the same way as ***'neutral_time_series'***. The new function should have four inputs, the same three as ***'neutral_time_series'***, and an additional input ***v*** for the speciation rate. The return should be in the same format as before, a list containing a time series vector and a community vector [2 marks].

10.) ⋆ Perform a neutral theory simulation with speciation and plot species richness against time as you above. Use a speciation rate of ν = 0.1, a community size of J = 100 and run your simulation for 10,000 time steps taking readings every 10 time steps. Plot two time series on the same axes in different colours showing how the simulation progresses from two different initial states given by ***initialise_max*** and ***initialise_min***. Include the code you wrote for this question in a function called ***'question_10'*** which should require no inputs to run. Explain what you found from this plot about the effect of initial conditions. Why does the neutral model simulation give you those particular results? [4 marks]

11.) You are going to study the species abundance distribution of these neutral simulations. First you need to write a function ***'species_abundance'*** to tell you what the abundances of all the species are in the system from an input of your community vector. For example ***species_abundance(c(1,5,3,6,5,6,1,1))*** should return 3 2 2 1 (in that order - decreasing). This is because there are 3 of species '1', 2 of species '6', 2 of species '5' and 1 of species '3'. (Hint: use table and sort) [3 marks]

12.) Write a function ***'octaves'*** to bin the abundances of species (e.g. the output of the ***species_abundance*** function) into what would be called 'octave classes'. The first value of the returned vector should tell you how many species have an abundance of only 1, the second value of the returned vector should tell you how many species have an abundance of either 2 or 3 and in general the n^{th} value of the returned vector should tell you have many species have an abundance greater than or equal to 2^{n-1} whilst strictly less than 2^n. For example, ***octaves(c(100,64,63,5,4,3,2,2,1,1,1,1))*** is asking us to sort 12 species into bins, the first species has an abundances of 100, the second 64 and the 4 rarest species

are all represented by one individual only. ***octaves(c(100,64,63,5,4,3,2,2,1,1,1,1))*** should return 4 3 2 0 0 1 2 in that order. (Hint: use the log and floor functions) [3 marks]

The simulations are stochastic you will therefore need to average the result from a number of independent readings to get an idea of the overall behaviour of the system. You will find that the octave vectors that are not always the same length, so R will not allow you to simply add them, or worse will sum them in a way that you do not intend so will give the wrong answer.

13.) Write a function ***'sum_vect(x, y)'*** which accepts two vectors as inputs, x and y, and returns their sum, after filling whichever of the vectors that is shorter with zeros to bring it up to the correct length. For example ***sum_vect(c(1,3),c(1,0,5,2))*** should return (2,3,5,2). (Hint: use length and if) [2 marks]

14.) ⋆ Run a neutral model simulation using the same parameters as in question 10 for a 'burn in' period of 10,000 time steps. Next record the species abundance octave vector. Then repeatedly continue the simulation for a further 1000 time steps, and record the species abundance octave vector again. You should repeat the further simulation and recording process 100 times. Produce a bar chart plot of the average species abundance distribution (as octaves). Include the code you wrote for this question in a function called ***'question_14'*** which should require no inputs to run. (Hint: it's OK to use a for loop here, it will also be helpful to use the ***sum_vect*** , ***octaves*** and ***species_abundance*** and ***neutral_time_series_speciation*** functions that you already wrote). [4 marks]

Challenge Question A: ⋆ Plot the mean species richness as a function of time (measured in simulation steps) across a large number of repeat simulations using the same parameters as in question 14. Add a 97.2% confidence interval on the species richness at each point in time. Repeat this for both initial conditions (high initial diversity and low initial diversity). Estimate the number of time steps needed for the system to reach dynamic equilibrium. Include the code you wrote for this question in a function called ***'challenge_A'***.

Challenge Question B: ⋆ Plot a graph showing many averaged time series for a whole range of different initial species richnesses. In each initial community state, each individual should be equally likely to take any species identity. Include the code you wrote for this question in a function called ***'challenge_B'***. (Hint: it's OK both here and elsewhere to make additional functions of your own to help make your code neater)

Simulations using HPC

15.) You are going to be running a much larger simulation of the same type that you conducted for your answer to question 14 and with more repeat readings. To do this requires use of high performance computing (HPC) and some adaptation of your R code. (Hint: I did not call the neutral_time_series_speciation function or the question_14 function for this and instead copied the code down and created a new function for running on the cluster.)

First, create a function ***'cluster_run'*** which accepts seven input parameters: speciation_rate , size , wall_time, rand_seed, interval_rich , interval_oct and burn_in_time.

- Your code will need to run for a predefined amount of time – given in minutes by the variable wall_time. Don't keep checking the time as this will waste a lot of

calculations. The simplest solution is to check the time every generation – where one generation involves a birth or death for every individual in the system (if size = 1000, one generation is 500 time steps). (Hint: use the proc.time command)

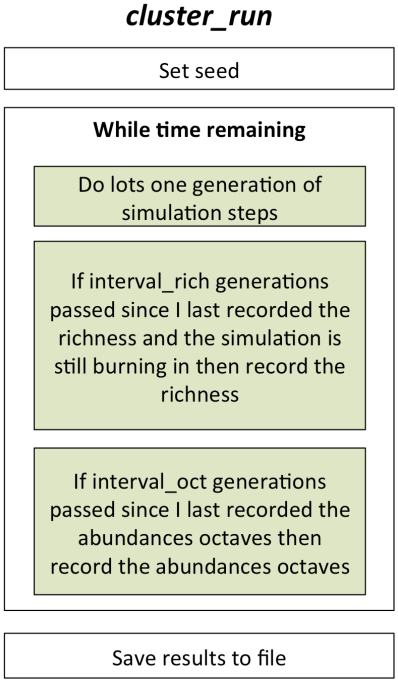
- You need to control the random number seeds so that each parallel simulation takes place with a different seed. If you run two simulations with the same seed, you will get the same answer regardless of the fact that it's a stochastic simulation. So your function should set the random number seed as rand_seed

- Your code should run for the number of generations specified by burn_in_time and store the species richness at intervals of interval_rich during this period. After the number of generations exceeds the burn in time, stop recording the species richness. Remember that generations are different from time steps that you used before.

- For the entire simulation, until the simulation runs out of time, you should record the species abundances as octaves every interval_oct generations. (Hint: use list)

- You should save your simulation results in a file including the following data: the time_series recorded during the burn_in_time, the list of species abundance octaves, the state of the community at the end of the simulation, the total amount of time actually consumed on the simulation and all seven of the input parameters for the function. Record all this in a file where the end of the file name is the number of the random seed. This way simulation files will not overwrite one another on the cluster where there will be no nice warning message asking if you want to replace the file or save under a different name! Also this gives you have a sanity check that no pair of simulations were conducted with the same random seed. (Hint: use the save and paste commands)

- Test your code locally before proceeding further using the same parameters from question 14 and a short time limit of 5-10 minutes.



Simple example

size = 200 , wall_time = 15 , interval_rich = 2
interval_oct = 3 , burn_in_time = 5

Action	Generation	Time (mins)
START (and setup)	0	0.0
Record richness	0	
Record octaves	0	
200 steps, check time	1	1.9
200 steps, check time	2	2.4
Record richness	2	
200 steps, check time	3	4.0
Record octaves	3	
200 steps, check time	4	5.7
Record richness	4	
200 steps, check time	5	7.0
200 steps, check time	6	8.3
Record octaves	6	
200 steps, check time	7	10.0
200 steps, check time	8	11.6
200 steps, check time	9	13.3
Record octaves	9	
200 steps, check time	10	15.1
END (and write to file)	10	

Burning in because generations ≤ 5

Still running because time < 15

Now you're ready to write lines of code in your R file around the functions you have written so that when you run the file using the source command you will get the simulation you want. Create a new file for this to run on the cluster, keep your original for the remainder of the questions.

- Your code should include a new variable ***'iter'*** and should start with the line: ***iter <- as.numeric(Sys.getenv("PBS_ARRAY_INDEX"))***. Your code will be run 100 times in parallel on the cluster, it will be run with iter = 1,2,3, …, 100 so you should use the variable iter in your code to make sure you don't just repeat the identical simulation 100 times.

- Everyone will use the same values for community size J in their simulations (500,1000,2500,5000) but each person will have different speciation rates (written at the top of your worksheet). You will have to select the correct value for J in each parallel simulation based on the value of iter (e.g. J = 500 when iter = 1,5,9,13,….).

- I suggest a time limit of 12 hours for all your jobs (you will put 11.5 hours into your code and tell the cluster 12 hours just in case).

- Use iter to set the seed for your simulation

- I suggest interval_rich = 1, interval_oct = roughly size/10 and burn_in_time = 8*size

- Test your code locally before running it on the cluster

[11 marks for correct test code and correct code for running on the cluster]

16.) Write shell script for running your code on the cluster. Use sftp and ssh to set your jobs running on the cluster as instructed during the lecture (also see lecture notes). Run a small job first just to test, then run the full set of jobs to the cluster [11 marks for all your output files shell script code and R code in a zip]

17.) ⋆ While your job is running on the cluster you can write R code to read in and process the output files. Your code should provide a mean species abundance result for each set of parameter values and plot the species abundance distribution as octaves in a multi-panel graph. Only use data of the abundance octaves after the burn in time is up. (Hint: use the load function on your .rda files). Please hand in all your plots and the actual mean abundance octave numbers along with the exact speciation rate that you used. [11 marks for your graphs and results]

Challenge Question C: ⋆ Plot a graph of mean species richness against simulation generation and use it to inform you more precisely how long should have been allowed as a burn in period for different values of J. Include the code you wrote for this question in a function called ***'challenge_C'***.

Challenge Question D: ⋆ Conduct further simulations of the same system using coalescence (see the pseudo code below). Check that your results from the cluster agree with those from coalescence and compare the speed of the two approaches. How many CPU hours were used on the coalescence simulation and how many on the cluster to do an equivalent set of simulations? Why were the coalescence simulations so much faster? Include the code you wrote for this question in a function called ***'challenge_D'***. To get a coalescence simulation in R of the neutral model from question 14 as a function of community size J and speciation rate ν.

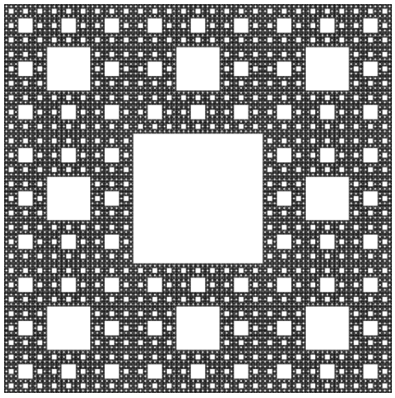
a. Initialise a vector ***lineages*** of length J with 1 as every entry.
b. Initialise an empty vector ***abundances***.
c. Initialise a number $N = J$.
d. Calculate θ, where $\theta = \nu \frac{J-1}{1-\nu}$.
e. Choose an index ***j*** of the vector ***lineages*** at random according to a uniform distribution.
f. Pick a random decimal number ***randnum*** between 0 and 1.

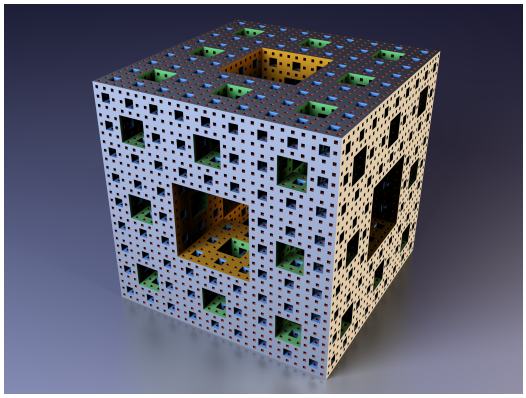
g. If ***randnum*** $< \frac{\theta}{\theta+N-1}$ append ***lineages[j]*** to the vector ***abundances***.
h. If ***randnum*** $\geq \frac{\theta}{\theta+N-1}$ choose another index ***i*** of the vector ***lineages*** at random, but not allowing ***i*** = ***j***. Then set ***lineages[i]*** = ***lineages[i]*** + ***lineages[j]***.
i. remove ***lineages[j]*** from ***lineages*** so that the ***lineages*** vector is now one shorter.
j. Decrease N by one so that N still gives the length of the ***lineages*** vector.
k. If N > 1 repeat the code again from e through to here.
l. Add the only element left in ***lineages*** to the end of ***abundances***.
m. END: a vector of simulated species abundances is stored in ***abundances***.

Fractals in nature

18.) ⋆ What are the fractal dimensions of these objects? Show and briefly explain your workings. [4 marks]

Hint: the object on the right looks the same from all six faces and is hollow in the very center; it should have a dimension somewhere between 2 and 3.





19.) ⋆ The chaos game

a. Store the following three points that correspond to coordinate on a graph: a=(0,0), b=(3,4) and c=(4,1).
b. Initialize the point vector X to indicate the point (0,0).
c. Plot a **very small** point on the graph at X. (hint: use cex)
d. Choose one of the three points (a, b or c) at random and move X half way towards whichever of the three points you chose.
e. Write a loop to repeat the code of c. and d. 100 times – what do you see? Now try increasing the number of repeats to 1000 or more. The function that does this should be called ***'chaos_game'*** [8 marks]

Challenge question E: ⋆ Try starting the chaos game from a completely different initial position X what happens now and why? Try plotting the first n steps in a different colour for various values of n to help you answer this. Try starting with the points of an equilateral triangle as a , b and c to produce a classic Sierpinski Gasket. If you're feeling super enthusiastic you could have more than 3 points and a distance of movement different from a half towards the next point.

20.) Create a function ***'turtle'*** in R to draw a line of a given length from a given point (defined as a vector) and in a given direction. So, ***'turtle'*** will have three inputs: start

position, direction and length. As well as drawing the line, turtle should return the endpoint of the line it just drew as a vector. Turtle should not open the plot it should just draw the line on an already open plot, this is because in a moment you are going to use successive calls of the function to draw things and you want all the lines to be on the same axes. (Hint: you need to use sin and cos). [2 marks]

21.) Now create another function ***'elbow'*** that calls ***'turtle'*** twice to draw a pair of lines that join together. ***'elbow'*** should accept as an input: the starting point, direction and length of the first line. The second line should start at the end point of the first line, have a direction that is 45 degrees ($\pi/4$ radians) to the right of that of the first line and a length that is 0.95 times the length of the first line. [2 marks]

22.) ⋆ Now copy and paste your 'elbow' function and rename it ***'spiral'***. Spiral will be an iterative function that draws a spiral. Instead of calling ***'turtle'*** twice to draw the first and second lines, spiral should call ***'turtle'*** to draw the first line and then call itself ***'spiral'*** instead of ***'turtle'*** to draw the second line. What happens and now and why? (Hint: if you get an error message that might be what's expected! Try and think about why you're getting it - think like a computer – run through the code you just wrote in your own head and see where it gets you) [2 marks]

23.) ⋆ Edit the 'spiral' function calling it ***'spiral_2'***. The edit should make it so that 'spiral 2' will only act if it's called with a line length that's above a certain size (e, a variable that you can experiment with). Now your code will draw a spiral shape on the graph without crashing or giving any error messages. [3 marks for a working spiral plotting function]

24.) ⋆ Now, copy and paste the ***'spiral_2'*** function and rename the copy ***'tree'***. Instead of having ***'tree'*** call itself only once (as ***'spiral_2'*** did), you should have it call itself twice: with directions that are 45 degrees to the right and 45 degrees to the left. Also, make the length of each subsequent call 0.65 times the length of the previous call (instead of 0.95 as it was for drawing the spiral). Don't forget that ***'tree'*** should still call ***'turtle'*** once as well as ***'tree'*** twice. You should get an attractive tree shape as your output plot. [4 marks for a working tree plotting function]

25.) Now copy and paste ***'tree'*** and rename it to ***'fern'***. Change your variables so that whilst one of the two branches goes 45 degrees to the left (as it did in f.) the other goes straight up (instead of to the right). Length multiples should now be 0.38 for the branch going to the left and 0.87 for the branch going straight up (instead of 0.65 for both as it was before). [3 marks]

26.) ⋆ Now copy and paste the function 'fern' and rename it to ***'fern_2'***. This should have an input parameter 'dir' which will decide whether the side branch of the fern goes to the left or right (it's easiest to do this with a variable that takes the value of either -1 or +1). When calling ***'fern_2'*** iteratively from within itself allow the direction of the side branch to alternate by passing on the 'dir' variable that has been multiplied by -1 to revers the direction. You should now get an attractive fern picture. [4 marks for a working fern plotting function]

Challenge question F: ⋆ What do you notice about the image produced and the time the program takes to run as you vary the value of e (the line size threshold)? experiment with the variables and colours to produce other types of fern and tree. Try using multiple colours – bonus points for being imaginative. Include the code you

wrote for this question in a function called ***'challenge_F'***, you can create a '***challenge_F2'*** and a ***'challenge_F3'*** if you need to.

Challenge question G: See how small you can make your code answer to question 26 without breaking it. To beat the record you would need to do it in less than 154 characters on one line of code (it would fit in a single text message). If you attempt this challenge, please make your shortened code a separate function named ***'challenge_G'*** from your main answer to question 26 because you'll have to remove all your comments to shorten the code. In the past there has been a fair amount of spirited debate about who had done this the best! To be clear, the rules are: your code should work fine even after the workspace has been cleared and should require no libraries, also, no marks should appear on your output axis apart from the lines that make up the fern. Finally, the fern should be of reasonable quality and at least very close in appearance to the fern produced in question 26 – if your code is shorter but doesn't produce the correct result then it doesn't count! This is what it should look like when displayed within in a normal window in R on a normal display, though there could be axes around it. (note: the definition of Challenge_G and open and close {} do not count towards the character count, you should define another function inside Challenge_G that calls itself)