IMPERIAL COLLEGE LONDON

MSc COURSE IN COMPUTATIONAL METHODS IN ECOLOGY AND EVOLUTION

EXAM 1

For Internal Students of Imperial College of Science, Technology and Medicine

Exam Date: Tuesday, 13th Jan 2015, 1300 – 1600

Length of Exam: 3 HOURS

Instructions: All sections are weighted equally. It is a three-hour exam, and there are 5 sections, so it is a reasonable guideline to spend about 35 minutes on each section. All sections allow you to choose between two questions, answering one. Read instructions carefully at the head of each section.

PLEASE PUT ANSWERS TO EACH SECTION IN A SEPARATE EXAM BOOK.

WE REALLY MEAN IT. PLEASE PUT ANSWERS TO EACH SECTION IN A SEPARATE EXAM BOOK. THE REASON FOR THIS IS THEN WE CAN PARALLELIZE MARKING AMONG THE DIFFERENT LECTURERS AND YOU GET THE MARKS BACK SOONER.

Continues on next page

Section 1: GIS and Genomics

Please select exactly **one question** and answer it.

A. You have been provided with a dataset showing the locations of the insectivorous Venus flytrap (*Dionaea muscipula*) plants across a wetland in North Carolina. A sticky insect trap was positioned near the location of each plant located in the wetland and used to quantify the number of potential food insects for the plant over a one month period. You also have a shapefile showing the boundary of the wetland.

Using this data (for each plant, GPS location and a count of food insects), how might you use GIS to assess the hypotheses that the density of plants is predicted by local variation in insect densities. Specifically (each question equally weighted):

(i) How might you use GIS buffers to establish the variation in plant density around individual point locations?

(ii) What options might be appropriate for quantifying insect density for each plant density estimate?

(iii) What effect does the spatial scale at which density is calculated have on your analysis? How might you assess the magnitude of this effect and how might behaviour of typical prey insects help resolve this question?

(iv) What effect might the edge of the wetland have on your analysis and how might you counteract this?

(v) What analysis might you use on your extracted data to test the hypothesis that plant density is predicted by local insect density? What assumptions of the analysis might you be particularly concerned about?

(vi) What assumptions does this analysis make about the original sample data?

B. Answer the following:

(i) Describe how you could generate the data required for a genome assembly project using Sanger sequencing and Next Generation Sequencing techniques (NGS). Briefly compare the differences between both methodologies. (30%)

(ii) Briefly outline the steps that you would follow to obtain a eukaryote genome assembly once you got your NGS data from a sequencing facility. How would you deal with the presence of repetitive regions in the genome and its effect on the assembly? What statistic could you use to assess the quality of the resulting assembly? (70%)

Continues on next page

Section 2: Stats and model fitting

Please select exactly **one question** and answer it.

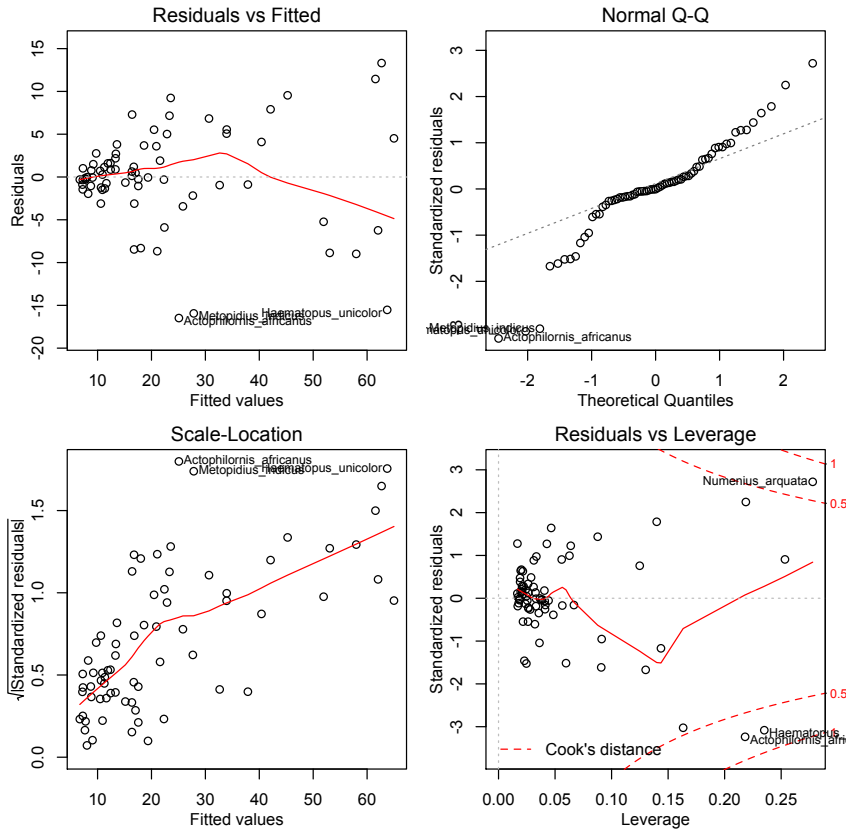
A. A student you are supervising has presented you with an initial linear model from his project looking at predicting variation in egg mass between species of shorebirds. The model summary and diagnostic plots are as follows. `M.Mass` and `F.Mass` are male and female body mass, respectively.

```
Call:
lm(formula = Egg.Mass ~ M.Mass * F.Mass, data = shorebird.data)

Residuals:
     Min       1Q   Median       3Q      Max
-16.4843  -1.3953  -0.0291   2.7322  13.3129

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)    4.212e+00  1.445e+00   2.915  0.00483 **
M.Mass         7.080e-02  3.402e-02   2.081  0.04126 *
F.Mass         4.847e-02  2.686e-02   1.805  0.07560 .
M.Mass:F.Mass -5.097e-05  1.913e-05  -2.664  0.00965 **
---
Signif. codes:  0    ***    0.001    **    0.01    *    0.05    .    0.1          1

Residual standard error: 5.761 on 67 degrees of freedom
Multiple R-squared:  0.8915,    Adjusted R-squared:  0.8866
F-statistic: 183.5 on 3 and 67 DF,  p-value: < 2.2e-16
```



Why does your student need to go back the drawing board and what suggestions do you have for improvement?

B. In nature, predators and prey body sizes are typically correlated, and a reasonable model of the size of predator expected to take a prey item of a given size seems to be a power law relationship:

Continues on next page

$$M_{pred} = M_0 M_{prey}^a$$

where M_{pred} is predator mass (in g), M_{prey} is prey mass (in g), M_0 is a constant, and the power-law exponent a is another constant. One argument is that this power-law relationship arises from the fact that the effects of body size on the underlying traits responsible for consumer-resource interactions (e.g., body velocity, jaw crushing power) can be captured by power-laws.

Somebody has given you a large dataset on the predator-prey body size relationship in marine organisms. When you plot predator body size against prey body size, you get the result shown in Figure 1a. The other two plots show the body size distributions of predators (Figure 1b) and prey (Figure 1c).

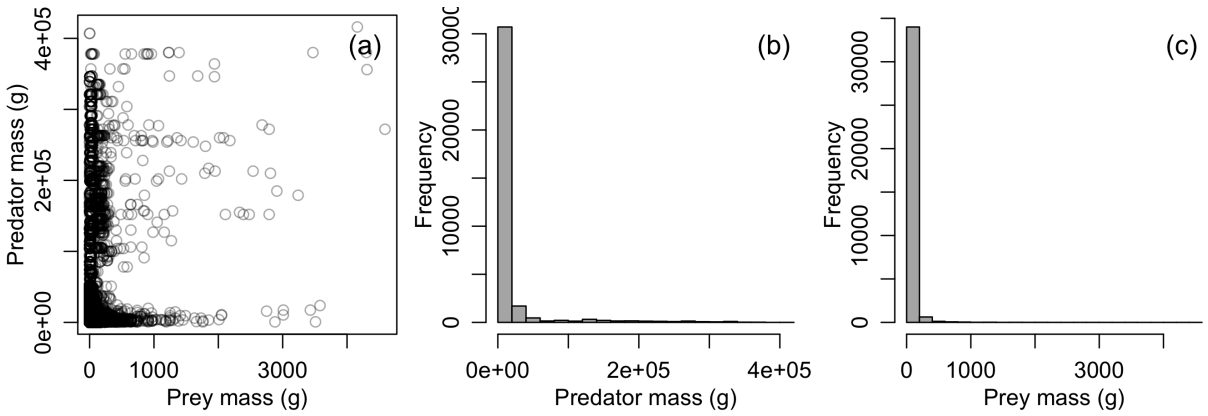


Figure 1: Predator-prey body sizes in marine organisms: (a) predator body mass as a function of prey body mass; (b) a histogram of predator body mass; and (c) a histogram of prey body mass. All masses in grams.

How would you go about testing whether the above equation is a good fit to these data? In particular, please answer the following (each question equally weighted):

(i) Why you might expect a positive relationship between predator and prey size — what biological constraints would apply, and why do you think such a relationship is not so evident in the above scatter plot?

(ii) Write out the equation of a model that should allow you to fit a linear regression model to the data and state what assumptions you would make about the data to allow you to fit the model to the data. Do you foresee any problem with using a linear model with these data?)

(iii) Sketch a graph that would best express how you think the fitted model would look like.

(iv) How you would estimate uncertainty around the parameters a and M_0?

(v) State what M_0 and a mean, in words. In particular, what physiological attribute should M_0 naturally include, especially if the organism is an ectotherm?

(vi) Would you consider this model to be mechanistic? Why, or why not?

Continues on next page

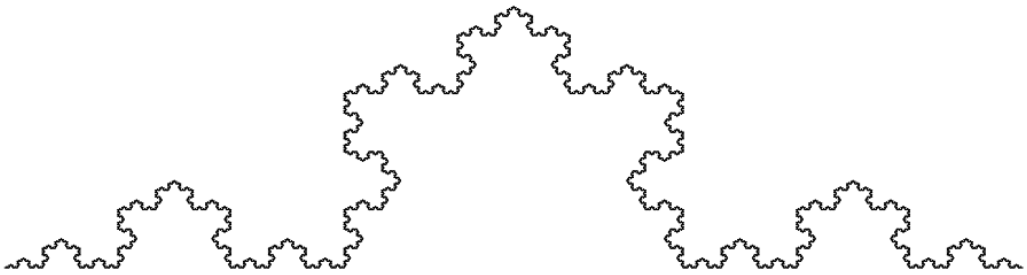
Section 3: HPC & fractals

Please select exactly **one question** and answer it.

A. You only need give brief bullet point style answers to these questions:

(i) Give three reasons why fractals occur in the natural world. (30%)

(ii) Calculate the dimension of the following fractal; show your workings. (20%)



(iii) Consider the function `Draw_fract` written below in pseudo code:

```
Draw_fract(x, y, r, n){
   if (n>0) {
      Draw_fract( x         ,y          ,r*1/3   ,n-1  )
      Draw_fract( x         ,y+r*2/3    ,r*1/3   ,n-1  )
      Draw_fract( x+r*2/3   ,y          ,r*1/3   ,n-1  )
      Draw_fract( x+r*2/3   ,y+r*2/3    ,r*1/3   ,n-1  )
   } else {
      Draw a square of width r with it's bottom left corner at (x,y)
  }
}
```

a. Show what `Draw_fract(0,0,27,0)` would draw. (10%)
Hint: this means x = 0 , y = 0 , r = 27 and n = 0, read through the code carefully trying to think like a computer.

b. Show what `Draw_fract(0,0,27,1)` would draw so n = 1 now (10%).
Hint: you will use your answer to i) above.

c. Show what `Draw_fract(0,0,27,2)` would draw so n = 2 . (20%)

d. As `n` becomes large the function `Draw_fract` described above generates a fractal, what is its fractal dimension; explain your answer? (10%)

B. You only need give brief bullet point style answers to these questions.

Consider a simple individual based neutral model containing a fixed number (J) of individual organisms.

In each time step, an individual is chosen at random to die and be replaced with the offspring of another individual in the system. With probability v the new-born individual is of an entirely new species in the system (speciation) otherwise it is of the same species as its parent.

(i) For the special case where $v = 0$. Given an initial condition where there are many species in the system, describe what will happen to the diversity of the system as the simulation progresses and why. (20%)

Continues on next page

(ii) 2.) What happens to the system in the more general case where $v > 0$ starting from different initial conditions of both high and low diversity. (20%)

(iii) Consider two simulation tasks for this neutral model: A) simulating the system many times for a wide range of different values of v and B) simulating the system for a single value of v but with a very large value of J.

a. Can either of these simulation tasks run in parallel on multiple CPUs and why? (10%)

b. For running these tasks on a high performance computing facility you would need to write a shell script file. Your shell script contains the code `#PBS -l mem=800mb` what does this mean and how might you need to change this to perform the required simulations? (10%)

c. Your shell script also contains the code `#PBS -l walltime=12:00:00` what does this mean and how might you need to change this to perform the required simulations? (10%)

(iv) Give three advantages or disadvantages to the use of simple models such as the individual based neutral model described here. (30%)

Continues on next page

Section 4: Maths I

Please select exactly **one question** and answer it.

A. Gompertz Growth

The Gompertz growth curve is given by

$$N(t) = K \exp(-a \exp(-bt)),$$

where K and b are strictly positive constants.

Now answer the following (equal weighting for all questions):

(i) Find a formula for a in terms of $N(0)$.

(ii) What is the behaviour of $N(t)$ as t goes to infinity?

(iii) Show that $N(t)$ is strictly increasing if $N(0) < K$ and strictly decreasing if $N(0) > K$.

(iv) Under what conditions on the constants b, K and $N(0)$ does $N(t)$ have an inflection point? If so, for what time t?

B. Taylor series expansion of the ascsine transform

The transform

$f(x) = \arcsin(x)$

is often used in biology as a data preprocessing step for normalising proportions, and consists of taking the arcsine of the square root of a data point. Calculate the Taylor series approximation of the arcsine function at $x = 0$, up to fourth order. This will be significantly easier to calculate than the arcsine function when you don't have a computer at hand (Hint: recall that $sin^{-1}x = \frac{1}{\sqrt{1-x^2}}$)

Continues on next page

Section 5: Maths II

Please select exactly **one question** and answer it.

A. Tidal Currents in Fjords

In a simple mathematical model of the tidal currents in a fjord we model the fjord as a canal of length L and with a constant parabolic cross-section:

$$y_{can}(x) = d\left(\left(\frac{2x}{w}\right)^2 - 1\right).$$

Here, x is a horizontal coordinate perpendicular to the flow of water, $y_{can}(x)$ is the height (or depth, if negative) of the canal floor as a function of x (with respect to average sea level), and d and w are given parameters (describing the depth and width of the canal, respectively).

The diurnal change of the sea level due to the tides is modeled as

$$y_{sea}(t) = y_0 \sin(4\pi t).$$

Here, t is time measured in days, $y_{sea}(t)$ is the sea level at time t, again with respect to average sea level, and y_0 is a given parameter describing the maximal height of the tides.

You may assume that the fjord never runs dry, so $y_0 \leq d$.

Now answer the following (equal weighting for all questions):

(i) Calculate the volume of water $V(t)$ in the fjord as a function of time. You may ignore the fact that it takes time for water to enter or leave the fjord.

(ii) Calculate the strength of the tidal current by finding the instantaneous change of $V(t)$.

(iii) Calculate the sea level y_{sea} at the time when the tidal current is the strongest (I am not asking for this time t). You can use the fact that the solution is positive.

To check that your solution of part iii is correct, you can verify that for a value of y_0 that is very small compared to d, the solution is $y_{sea} \approx 0$. For $y_0 = d$ the solution is $y_{sea} = y_0/3$.

B. A Recurrence Relation

Find a formula for x_k as a function of k from the recurrence $x_k = x_{k-1} + (3/4)x_{k-2}$, with $x_1 = 0$ and $x_2 = 8/3$.

End of paper