1. Mars module expansion

Use simulated annealing to do the following:

- Major axis for Earth is 1AU
- Elliptical Orbit for Earth, with arbitrary focus (f), major axis direction (d), eccentricity e
- Line of intersection of Mars' plane with the ecliptic (l)
- Inclination of Mars' plane relative to the ecliptic (i)
- Elliptical orbit for Mars on its plane, with arbitrary focus (f'), major axis direction (d'), eccentricity e'
- Major Axis for Mars (m)
- The areal speed for Earth about its focus (s)
- The areal speed for Mars about its focus (s')
- Find values for f, d, e, l, i, f', d', e', m, s, s' so as to minimize the max Oppositions Long. Discrepancy, Triangulations Long. Discrepancy and Oppositions Lat. Discrepancy
- Use heuristics to cut down your search space for efficiency as discussed in class
- Remember all work on longitudes needs to be done by projecting Mars' orbit on the ecliptic place

2. Colour blindness (In fact, Cancer)

Given the following sequencing data, find the mutation. You will need to extend what we did in class to handle insertions and deletions.

- In the first file (combine all the STRAN-2020-21019-* files into one with the various reads to be aligned) and look for a 5-10 character insertion/deletion
- In the second file (STRAN-2021-22881-*), you have to look for an 100-200 length insertion/deletion

The reads are contained in fastq files. A fastq file contains both reads as well as quality indicators on how good each character in each read is. You can ignore the quality indicators in your project. Also note that file names occur in pairs labeled R1, R2. Each read in an R1 file has a corresponding read in the R2 file such that the two reads are obtained from the same DNA fragment, and therefore must align close to each other. However, for simplicity, you can treat the various reads as independent.

The hg19.fa file has the reference sequence. Unlike in the assignment, you have all the chromosomes now. Nevertheless, the total number of reads is a little lower than what you had in the assignment.

The hg19_last_col.txt and hg19_mapping.txt files contain the relevant BWT information as in the project, but for all the chromosomes together. Note the way this works is that the Ns in the hg19.fa file are skipped over and only the non-Ns are present in the last column. The mapping file contains accurate indices back to the hg19.fa reference though. The

chr_bwt_boundary.txt file can help you determine which chromosome a particular entry in the mapping file corresponds to.

Note that the files above come from real individuals and the above mutations cause these individuals to be at a very high risk for cancer. See this popular example to get a sense for the nature of this risk.

To find insertions and deletions, use the BWT algorithms discussed to class to come up with a shortlist of candidate matches and then use an edit distance algorithm to compute the min number of insertions and/or deletions needed at each candidate match position. You can restrict your search for reads that map to chr13 and chr17 for this project.

Remember that the insertion/deletion events we are looking for are contiguous. See if you can modify the above edit distance algorithm so that it allows only one contiguous insertion/deletion (of course, that one insertion/deletion could span multiple characters).

The output of your project should be a list of contiguous insertions/deletions on chr13/17 of the sizes indicated above that have evidence from at least 10 reads.

You need to absolutely ensure that these files are not used for any purposes beyond this project and please delete any material that might be residual on your systems once the project is completed.

3. More detailed modelling of COVID-19.

In the assignment, you started the simulator on 16 March 2021 (for the seven-day averaging, you may have started on 09 March 2021). Now you will start calibration on 11 October 2020 and will do two tunings.

- Start the simulator with susceptible, infected and recovered fractions matched to the round-1 seroprevalence data projected to 11 October 2020. Tune the contact rate parameters during 11 - 31 October 2020 to match the number of reported cases on 01 November 2020 to within 10%. Use the identity matrix for mobility during this period, so the disease evolves independently in each unit.
- Tune the unit-wise contact rate parameters (one scalar per unit) during the period 01 November 2021 - 28 February 2021 and the antibody waning parameter to bring the per day squared error to within 0.1 for each unit and to match the serosurvey data. Use the identity matrix for mobility during this period also. We update the contact rates in each iteration after evolving the time-series over the duration.
- Tune the unit-wise contact rate parameters during 01 - 15 March 2021 to match the number of model-reported cases on 15 March 2021 within 10% of the actual reported cases. Then tune the unit-wise contact rate parameters during 16 March - 07 April 2021 to minimise the per day squared error during this period. To account for the stochasticity due to a low number of reported cases in many units, we introduce moderate uniformly mixing mobility, i.e., M(t) =

(1-eps) I +eps J, where I denotes the identity matrix of size 38x38, J denotes the all-one matrix of size 38x38, and eps=0.01.

- Repeat the above for the duration 08-22 April 2021 to arrive at the corresponding calibrated contact rates for this period. From this period onward, since the infections have been seeded, we go back to diagonal mobility, i.e., M(t) = I.
- For the projections, we redo the calibration for the period 08 April - 01May 2021 and hold the resultant contact rate parameters constant into the future.
- To account for the delay in sample collection and test outcomes, we delay the trajectory of reported cases by one week.
- From 02 May 2021, use the vaccination data and effectiveness of vaccines, run your model forward assuming contact has been reduced by a factor ½ until 31 July 2021 and then we open up from 01 August 2021. Plot the estimated cases against the actual cases.

4. Setting targets based on multidimensional robust synthetic control method - an alternative to the D/L method

Implement the method that is given in this paper.

https://arxiv.org/abs/1905.06400

Compare the targets obtained from this implementation with those of your class assignment implementation and highlight differences. Give examples to show differences, and highlight situations where the other method will likely be better and where this method may be better.