Algorithms for genetics (CSE280a) Assignment 1

January 14, 2019

Logistics

Submit using gradescope. All code must be zipped and submitted as a zip file.

Questions

1. Sign up on Piazza and Gradescope. Go through the Academic Integrity Policy (AIP) document on the course web-site, and send an email to the instructor and TA saying that you have read and agree with the AIP (2pt.).

2. The so called 'bread wheat' is hexaploid (6 copies of each chromosome). Consider a locus with 4 allelic values (a, b, c, d) with frequencies $0.5, 0.25, 0.15, 0.1$, respectively. (a) Compute the number of distinct possible genotypes. (b) Compute the expected number of occurrences of the genotype ab^3c^2 in a sample of $10,000$ individuals, assuming HW equilibrium holds (c) Generalize part (a) to compute the number of distinct genotypes given a ploidy of n (n copies of each chromosome) and m alleles.

3. Consider a diploid population with 4 alleles A_1, A_2, A_3, A_4 at a locus. An experiment on 100 individuals revealed the following counts of heterozygous individuals.

	A_1	A_2	A_3	A_4
A_1		18	21	12
A_2			7	3
A_3				5

   Design and implement a tool that estimates the most likely allele frequencies in the population, assuming HWE is satisfied. **Note.** Formulate this as an optimization, or a problem of maximizing the likelihood. The actual solution is less important, and a simple grid-search over the parameters will suffice.

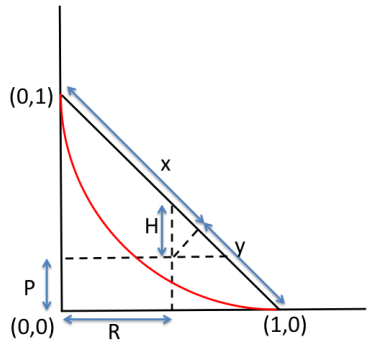

Figure 1: Triangle depicting genotype and haplotype frequencies.

4. (10 pts.) Fig. 1 is a convenient way to study HW equilibrium. Note that the hypotenuse joins coordinate $(0, 1)$ with $(1, 0)$. For any point (P, R) inside the triangle, let H denote the distance such that $(P + H, R)$ lies on the hypotenuse. Show the following:

(a) Show that $P+R+H=1$ and therefore, any point denoted by P, R, H can denote the genotype probabilities $(\Pr(AA), \Pr(aa), \Pr(Aa))$ of a diploid individual.

(b) Let the perpendicular drawn from (P, R) to the hypotenuse divide the hypotenuse into segments x and y. Show that

$$\frac{y}{x} = \frac{\Pr(A)}{\Pr(a)}$$

(c) Show that all genotypes that satisfy Hardy Weinberg equilibrium lie on the parabola defined by $P^2 + R^2 - 2PR - 2P - 2R + 1 = 0$.

(d) **non-credit question**. Can you use this figure to devise a test for computing deviation from HWE?

5. For a bi-allelic locus, (a) what is the maximum probability of seeing a heterozygote? (b) what is the maximum probability if the locus has n alleles? It is sufficient to answer this with a back-of-the-envelope calculation.

6. The occurrence of a rare event is considered to be *statistically significant*. Use HWE to argue that the observation of homozygotes at $k = 30$ consecutive bi-allelic loci (a 'loss-of-heteroygosity') is statistically significant, if you know that the population minor allele frequency is at each locus is at least 10%.

7. For each of the 6 data-sets (SNP matrices with n individuals/rows, and m sites/columns) provided

   (a) Determine if a perfect phylogeny exists or not. You have to write code to do this.

   (b) Plot the running time of your code as a function of $n \cdot m$. For full credit, it should scale linearly. Describe the key ideas using pseudo-code, and use a log-log plot so that all points are seen and show guide lines with slopes 1 and 2, to argue that you are scaling linearly.

8. Given the mtDNA genotypes of individuals sampled at the following locations on the globe (Figure 2), give your best guess for the migration routes.

	a	b	c	d	e	f	g	h	i	j	k	l	m	n	o
1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
2	0	0	0	0	0	0	1	0	1	0	1	0	0	1	0
3	0	1	0	0	1	0	1	0	1	0	1	0	0	1	0
4	0	0	0	0	0	1	0	1	0	0	1	0	0	1	0
5	0	0	1	1	0	1	0	1	0	0	1	0	0	1	0
6	1	0	0	1	0	1	0	1	0	0	1	0	0	1	0
7	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0
8	0	0	0	0	0	0	0	0	1	0	1	0	0	1	0
9	0	0	0	0	0	0	0	0	0	0	0	1	1	1	0
10	0	0	0	1	0	1	0	1	0	0	1	0	0	1	0
11	0	0	0	0	0	0	0	1	0	0	1	0	0	1	0
12	0	0	0	0	1	0	1	0	1	0	1	0	0	1	0
13	0	0	0	0	0	0	0	0	0	0	1	0	0	1	0
14	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
15	0	0	0	0	1	0	1	0	1	0	1	0	0	1	0
16	0	0	0	0	0	0	0	0	0	0	0	1	0	1	0

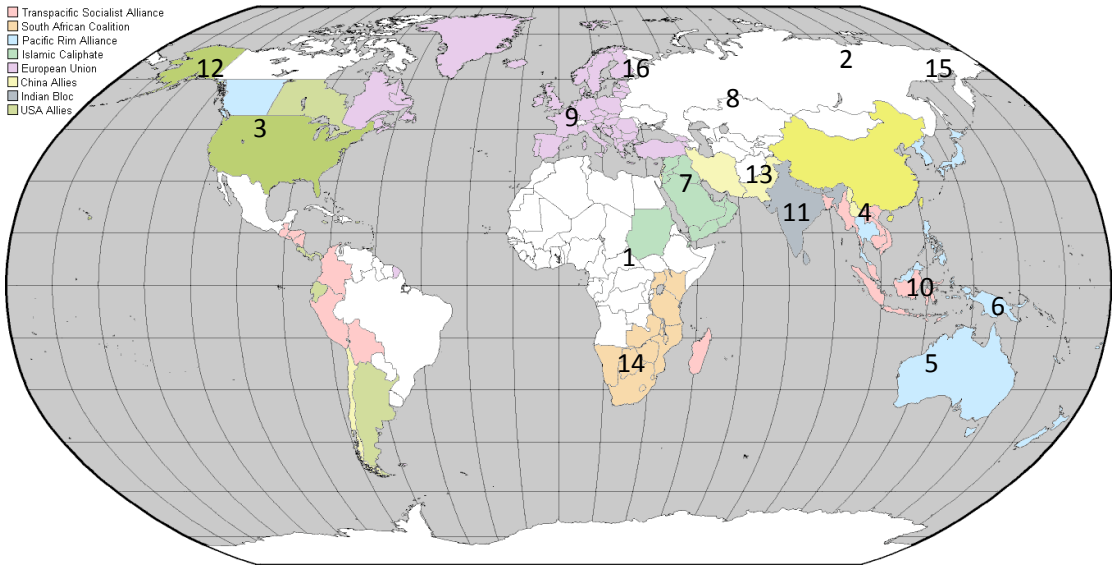

Figure 2: Locations sampled. Each individual is labeled with the location (s)he is sampled from.