CSCI2951 HW2 | Ben Siranosian | 2014-10-21

Problem 1: Tagging SNPs and the minimum informative set of SNPs.

In each of the given haplotype matrices, there are several possible minimum informative subsets. I will report one of the possible solutions. Let MIF(*M*) designate the minimum informative subset of haplotype matrix *M*.

$$\text{MIF}(M_1) = \{s_1, s_2\}$$
$$\text{MIF}(M_2) = \{s_1\}$$
$$\text{MIF}(M_3) = \{s_1, s_2, s_3\}$$

Problem 2: EM haplotype phasing – programming

Code for this problem can be found in em_haplotype.py. Instructions for use can be found by running the script with no arguments or in the README.txt file.

Problem 3: EM haplotype phasing – worked out solution

Genotypes	Possible haplotypes	Initial frequency	Rd 1 frequency
020	000	0.1667	0.1667
120	010	0.1667	0.1667
122	100	0.1667	0.25
	110	0.1667	0.25
	111	0.1667	0.0833
	101	0.1667	0.0833

I found the haplotype frequencies to converge after a single iterations of EM in this case.

Problem 4: Fermat urns

A classical problem in probability!

Piere de Fermat is best known for his last theorem, postulating in 1637:

$$x^n + y^n \neq z^n \quad for\, n > 2$$

The theorem remained unproven for hundreds of years until 1995 when Andrew Wiles nailed down the final part of the proof. Fermat's last theorem can be used to show the answer to the urn problem provided.

Assume there are *a* white and *b* black balls in the first urn, and *c* white and *d* black in the second. The total number of balls in each urn is constant, so *a*+*b*=*c*+*d*. *n* balls are drawn from each urn.

The probability of drawing all white balls from the first urn is:

$$\left(\frac{a}{a+b}\right)^n$$

The probability of drawing all white balls OR all black balls from the second urn is:

$$\left(\frac{c}{a+b}\right)^n + \left(\frac{d}{a+b}\right)^n$$

For the probabilities to be equal,

$$\left(\frac{a}{a+b}\right)^n = \left(\frac{c}{a+b}\right)^n + \left(\frac{d}{a+b}\right)^n$$

Which reduces to:

$$a^n = c^n + d^n$$

Which due to Fermat's last theorem, cannot be true for $n>2$.