Question Bayesian integration,

Bayesian data integration to predict new ciliary genes. The site there are three files for today:

CiliaCarta_table_for_students_positive. A set of 100 genes that are known to be associated with the cilium.

CiliaCarta_table_for_students_negative. A set of 100 genes that are known not to be associated with the cilium.

CiliaCarta_table_for_students_unknown. A set of 7 genes for which we would like to know how likely it is that they are associated with the cilium based on the genomics data we have for them.

Each file contains a number of datasets: direct protein-protein interactions as measured by Masspec or by Yeast 2 hybrid, the presence of an evolutionary conserved RFX transcription factor binding site and the co-evolution of the gene with the presence of cilia (cilia co-occurrence). The cilia co-occurrence is expressed as Dollo Parsimony : the number of independent evolutionary events with respect to the presence of cilia and with respect to the presence of the gene. We have estimated that a score of 9 or lower (maximally 9 independent gains/losses of the cilium and of the gene along the evolutionary tree) is a good indicator for being ciliary.

a) Calculate for each type of data its predictive value: i.e. the fraction of ciliary genes that have that property divided by the fraction of non-ciliary genes that have that property. You will also need to calculate the fraction of ciliary genes that do not have that property divided by the fraction of non-ciliary genes that do not have that property. Which type of data has the highest predictive value? Which one has the lowest?

b) Now you will calculate the (relative) likelihoods that the 7 genes of the CiliaCarta_table_for_students_unknown are ciliary. You will do that by, for each candidate gene adding the logarithms of the relevant fractions from a) : i.e. if a gene has a property (e.g. the protein physically interacts with a known cilium protein) you will add the log. of the fraction of ciliary genes that physically interact with a known ciliary protein divided by the fraction of non-ciliary genes that interact with a known ciliary protein. If a protein does not interact with a known ciliary protein, you will add the fraction of ciliary proteins that do not interact with ciliary protein divided by the fraction of non-ciliary proteins that do not interact with a ciliary protein.

c) Thus far we have forgotten about the “prior” in our calculations. In this case the prior is the probability that any gene is ciliary. We estimate that there are about 1200 ciliary genes in the human genome. Add the prior to score (remember that you have to take the logarithm…) and obtain the logarithm of the probability that each candidate gene is ciliary.

d) With the current data (the positive and the negative sets of genes), what is the maximum likelihood that any gene could be predicted to be ciliary.

e) One of the nice things about Bayesian data integration is that adding new data does not affect the previous calculations, you can simply add the new scores to the existing ones. Suppose all the candidate genes are co-expressed with the cilium, and ciliary genes are 8 times more likely to be co-expressed with other cilium genes than non-cilium genes. Which genes are now more likely to be ciliary than non-ciliary?