Overview of Stats. for Comp. Biol.

Sayan Mukherjee

Preliminaries

The course is taught mainly in a case studies framework. The exceptions will be a few breaks to go over important methods or tools used in inference for a variety of statistical models in biology, namely expectation-maximization (EM) and Markov chain Monte Carlo (MCMC).

For a variety of biological problems we will proceed through the following steps:

1. Formalization of the biological problem – Statisticians call this modeling, the goal of this part is to come up with a likelihood
   a. What are the relevant variables to model the biological question ?
   b. Do we have the correct type of data to correspond to these variables ?
   c. Which variables are (un)observable ?
   d. What is our model of noise in the system ?
   e. What is our model of uncertainty in the system ?
   f. What are the parameters in our statistical model ?
   g. Give our model what biological assumptions have we made ?

2. Estimating parameters from data – Statisticians call this inference (more specifically parameter estimation)
   a. What is the uncertainty in my parameter estimates ?
   b. Do these parameter estimates make any biological sense ?
   c. How well does my data fit my modeling assumptions ?
   d. If I fit many models how should I think about my results, which model should I pick ?
   e. Does the model and parameters I chose fit future data or just explain what I have already seen ?

3. Answering the original biological questions using the inferred parameters and model(s).

Grading

There will be weekly homeworks which are really for practice and to help you get a better feel of the material.

The course grade is based on a take home midterm (40%) and a final project (60%). Ideally both of these will be based on data analysis.

The case studies

The following is a list of statistical concepts and biological problems that the case studies provide:

1. Inference of population structure – We start with a classical problem in population biology/genetics. The biological problem is provides us with some understanding of basic genetics and asks the question of given genetic data from several individuals what are the ancestral populations that make-up each individual.
   The statistical ideas covered are (1) modeling – mixture models, the multinomial model, conjugate priors, Bayesian modeling – (2) inference – MCMC and Gibbs sampling – (3) what kind of dogs does Sayan have ?

2. Inference regulatory motifs – We will find that the same model we used for a population biology problem can be used in a basic problem in molecular biology. Finding the motifs of targets of transcription factors. We use a different method for parameter inference.
   The statistical ideas covered are (1) modeling – mixture models, the multinomial model, conjugate priors, Bayesian modeling – (2) inference – we use expectation-maximization (EM).
3. Predicting phenotypes from gene expression microarray data – We look at problems like classifying cancer versus non-cancer from gene expression data.

The statistical ideas covered are (1) modeling – how to address many/many variables using shrinkage or regularization – (2) inference – maximum a posteriori (MAP) estimation.

4. Genome-wide association studies – Another problem with a massive number of variables. How to association single nucleotide polymorphisms (SNPs) with a phenotype?

There are many statistical challenges (1) modeling – missing heritability, univariate/multivariate models/tests – (2) multiple hypothesis testing.

5. Association mapping correcting for relatedness – How to associate SNPs to phenotypes when the samples are related or there is population structure.

There are many statistical challenges (1) modeling – mixed effects models – (2) inference – restricted maximum likelihood estimation (REML).

An example

In 1894 Karl Pearson was part of a debate on whether inheritance of traits is continuous – the position of the Darwinian group – or particulate – the position of the Neo-Mendelian. The resolution of this debate resulted in what is called the “modern synthesis.”

In the paper “Contributions to the Mathematical Theory of Evolution” Pearson developed Gaussian mixture models to address this question. The trait examined is the breadth of the carapace (forehead) of crabs. We will reexamine this work.