Genetic diversity and the Ewens sampling Formula

July 17, 2017

Evolution theory relies on two key elements: *mutations* and *selection.* A basic question is then to assess the relative importance of those two elements.

- For Darwin, *selection* was the key factor: it is selection that shapes species, and that lead a given species to diversify into several different species. Mutation is necessary to create the diversity necessary for this evolution, but it doesn't play a central role in the becoming of species.
- For Motoo Kimura, Jack Lester King and Thomas Hughes Jukes (1968-1969) however, mutation may lay a more important role: In some situations, when selection is not very strong, random mutations can indeed have a decisive impact for the evolution of species.

An example of this is the $\beta-$globin, the most common type of Hemoglobin protein in humans. The specific $\beta-$globin molecule that we produce depends on a well defined part of our DNA genome. The precise DNA code is not the same for everyone: there are at least 6 different DNA sequences, leading to different $\beta-$globin molecules (the most common molecule is carried by 25% of the population only). The distribution of those various DNA codes is not the result of a complicated selection pattern: it seems on the contrary that all those codes are equivalent, and that this diversity is simply the result of randomness.

The mathematical question that quickly appear is the following: what is the impact of randomness on the diversity of a population. The central result here is the *Ewens sampling formula*, that is a relatively simple result on Wright-Fisher processes, in line with the lecture of Vincent Bansaye. This project will rely on the book *Probability Models for DNA Sequence Evolution*, by Rick Durett (p.1-45). There is also a beamer presentation, with many references, for the genetics aspects of this problem.

This theory has gained attention recently for two independent reasons:

- The progress of rapid sequencing: the cost of sequencing has dropped dramatically in the last decade. One can then sequence parts of the genome that do not play an important coding role. Neutral models are then good models for the evolution of those sequences. Going further than the *Ewens sampling formula*, it is possible to investigate the history of the population (dynamics of the population size, for instance) through its genotypic diversity.
- In 2001, Stephen P. Hubbel extended this neutral theory idea to understand the structure of biodiversity, under the name *Unified neutral theory of biodiversity.* Here also, the idea is that species would appear randomly, and that selection would have a limited impact. He presented some stricking examples from coral reef communities, that seem to fit very well with this theory.

The project will be based on the book of Rick Durett for the theoretical aspects, and on numerical simulations. This could be completed by a survey of the biology litterature on the subject.