In recent years, there have been many efforts to understand statistical properties of evolutionary processes on growing trees/sparse networks (i.e., when the number of nodes of the graph that represents the underlying evolutionary relationship increases, in contrast to the standard setting when the graph is fixed and the length of the genetic sequences increases). In most biologically realistic models of branching processes, correlations among the sampled data do not decay. This invalidates the standards assumptions for statistical analyses and makes many well-known statistical estimators inconsistent. The problem is further compounded by the fact that only the present’s individuals can be observed while characteristics of their ancestors are uncertain, and there’s no decomposable independent structures among the correlations of the variables.

In this talk, I will outline the main challenges for learning with evolutionary-related random variables. Using the problem of estimating the transition rate of a simple two-state symmetric model on phylogenetic trees as the guiding example, I will discuss some new approaches in our recent works and describe the main open questions in deriving concentration inequalities in this particular setting. (Received July 15, 2019)