Model selection for time-reversible evolutionary models via linear invariants

Roser Homs Pons

Phylogenetic trees are used to understand and reconstruct evolutionary processes at a molecular level. One of the main problems in phylogenetics is to estimate which tree better describes the evolutionary history of some given data in the form of DNA or amino acid sequences. To reconstruct the phylogenetic tree, one needs to assume a substitution model that explains how characters are substituted at each site of the sequence according to biochemical properties.

Misspecification of the substitution model can lead to severe incongruences in the retrieved phylogenetic tree. Classically the selection of a suitable evolutionary model is based on heuristics or relies on the choice of an approximate input tree. Felsenstein suggested that certain linear equations satisfied by the expected probabilities of patterns observed at the leaves of a phylogenetic tree could be used for model selection. It remained an open question, however, whether these equations were sufficient to fully characterize the evolutionary model under consideration.

Using techniques from algebraic geometry and group theory, Casanellas et al. (2012) proved that the space of phylogenetic mixtures under equivariant models (such as JC69, K80, K81, SSM and GMM) is a linear space that fully characterizes the evolutionary model. In Kedzierska et al. (2011), they successfully implemented a method for model Selection in Phylogenetics based on linear INvariants (SPIn) which outperformed approaches existing at the time for simulated data under a variety of single-tree and mixture settings, though only available for JC69, K80, K81, SSM.

Extending this model selection approach to other substitution models (also for protein data) requires computing linear invariants beyond equivariant models. A first step in this direction was done in Casanellas and Steel (2016), where a they describe the structure and dimension of the vector spaces of phylogenetic mixtures and of linear invariants for phylogenetic trees on any number of leaves under the Equal-Input model.

In this talk, we will provide a general framework to study any (algebraic) time-reversible model with any number of states and how to extend linear invariants from small to larger trees. We will focus on the model TN93, show how to compute their linear invariants and provide a full description of the linear space of mixtures for trees on 4 leaves.

This is a joint work with Marta Casanellas and Angélica Torres.