VMMC – training – based homogeneity testing

Mikhail Malyutov, Northeastern University, Boston

Abstract

Testing homogeneity between finite Variable Memory Markov Chains (VMMC) (modeling protein families) was considered using very complicated statistical test BB on probabilistic suffix trees (Balding, Bush et al, 2008). If a universal compressor such as LZ - 78 compresses efficiently the long training string (such as literary text), then the homogeneity CCC - test with its theory developed by us two years ago is a computationally simple efficient substitute for the test BB.

The CCC-test cut the query into VERY small slices as compared to the length of the training string to avoid the UC self-adaptation to the query. Our test VMMCLR is the Studentized sum of empirical loglikelihood ratios between the query slices and simulated training string continuation of the same length. We prove exponential tails optimality and asymptotic normality of our test similarly to our study of the CCC-test. This approach is shown to be robust in recent papers by Ziv w.r.t. departures from the VMMC assumptions.