

5th Annual Joint Bioinformatics Workshop

July 19, 2005
2229 Seamans Center
University of Iowa

by

Raul Piaggio-Talice
Iowa State University

Evolutionary History Model Selection via Improved Phylogenetic Compression

ABSTRACT

A central problem in phylogenetics is to select a hypothesis that best describes the evolutionary history leading to the sequences in a given alignment. Such a hypothesis can consist of a single tree for the whole alignment or multiple trees for different sections of it (as in the case of horizontal transfer or gene duplication). Ane and Sanderson recently proposed a method of approaching this model selection problem by using the minimum description length principle from algorithmic information theory. In this approach, the alignment is described by a two-part encoding composed of a code for a candidate hypothesis plus a code to recover the alignment given such hypothesis. The hypothesis assumed to be correct will be the one that minimizes the length of such encoding. Note that a minimum length encoding is also the best compression possible of the sequence alignment. We present a modification to the Ane and Sanderson method that results in provably shorter codes, closer to the minimum description length. The improvement is achieved by using ranking (and unranking, if the code is to be uncompressed) techniques. This is applied to code the hypothesis (tree or trees) as well as to code the sequence alignment given the hypothesis. The shorter code provides a better compression mechanism and is expected to sharpen the hypothesis decision criterion. The new method still produces (efficiently) computable codes despite the fact that finding the hypothesis that minimizes a two-part code is akin to computing the Kolmogorov complexity of the alignment, known to be uncomputable [LV93]. When tested in real-world datasets from [SDEL03], the new method shows hypothesis distinction capabilities similar to that of the original version while the compression improved on average by 26.18%, with gains that range from 8.79% to 49.02%.