Maximum Entropy Principle for Composition Vector Method in Phylogenetics

Raymond H. Chan
Department of Mathematics, The Chinese University of Hong Kong, Hong Kong

Molecular Phylogenetics is the study of evolutionary relatedness among species through molecular sequencing data. The composition vector (CV) method is an alignment-free method for phylogenetics. Since biological sequences are often obscured by noise and bias, denoising is necessary when using the CV method. By using the maximum entropy principle for denoising and utilizing the special structure of the constraint matrix to simplify the optimization, we derive several new denoising formulas. By comparing with existing formulas on ten different data sets, we found that one of our formulas gives more accurate phylogenetic trees. An example is the tree for the tetrapod data set where we can correctly group birds and reptiles together, a result that cannot be obtained previously by either alignment method or other denoising formulas.

Joint work with Wei Wang (CAS-MPG Partner Institute and Key Lab for Computational Biology).