

BPPS: An Algorithm for Analyzing Protein Sequence Alignments

Jun Liu
Department of Statistics
Harvard University
jliu@stat.harvard.edu

Abstract

Aligning multiple biopolymer sequences has been recognized as a central activity in bioinformatics research. But the analysis of the resulting alignments has not been rigorously formulated and mathematically tackled. We have developed statistical procedures to decompose the multiple alignments into distinct categories and to pinpoint critical structural features within each category. A central part of our statistical procedures is a novel algorithm called the Bayesian partitioning with pattern

selection (BPPS), which is based on a two-way mixture model and can simultaneously classify protein sequences into distinct subfamilies and select conserved positions that are characteristic of these subfamilies. When applied to P-loop GTPases, this revealed within Rab, Rho, Ras, and Ran a canonical network of molecular interactions centered on bound nucleotide. This network presumably performs a crucial structural and/or mechanistic role considering that it has persisted for more than a billion years after the divergence of these families.