

Title:

Vector Quantization of Amino Acids: Analysis of the HIV V3 Loop Region

Author(s):

A. B. Olshen, P. C. Cosman, A. G. Rodrigo, P. J. Bickel, and R. A. Olshen

Technical Report number (Dept. of Statistics, Stanford Univ.):

2003-29B/226

Date:

October 2003

Abstract:

This paper is about techniques for clustering sequences such as nucleic or amino acids. Our application is to defining viral subtypes of HIV on the basis of similarities of V3 loop region amino acids of the envelope (*env*) gene. The techniques introduced here could apply with virtually no change to other HIV genes as well as to other problems and data not necessarily of viral origin. These algorithms as they apply to quantitative data have found much application in engineering contexts to compressing images and speech. They are called *vector quantization*, and involve a mapping from a large number of possible inputs into a much smaller number of outputs. Many implementations, in particular those that go by the name generalized Lloyd or *k*-means, exist for choosing sets of possible outputs and mappings. With each there is an attempt to maximize similarities among inputs that map to any single output, or, alternatively, to minimize some measure of *distortion* between input and output. Here, two different types of vector quantization are brought to bear upon the cited problem of clustering V3 loop amino acid sequences. Results of this clustering are compared to those of the well known UPGMA algorithms, the unweighted pair group method in which arithmetic averages are employed.