Summary of:

"Genetic Code, Hamming distance and Stochastic Matrices"

Matthew He

Division of Math, Science and Technology

Nova Southeastern University

The paper explains the usefulness of representing genetic code using Gray's code. It first explains the general form of this genetic code: mRNA, which is composed of codons, which are a series of three nucleotides, represented by the letters (C, A, G, U). Then it defines the notion of “Gray code”: a sequence of bits similar to binary, but whose increase of one unit only results from a single bit change. It is then possible to represent each nucleotide by a series of 2 bits {00, 01, 11, and 10} and therefore an entire codon per 6 bits.

The interest of this representation lies in the fact that the amino acids are formed from a sequence of codons attached to each other (CCC, CCU, CCA, and CCG); where only one nucleonic changes each time what happens naturally lends itself to the use of Gray's code. In fact, this allows us to get all the time an adjacent codon when a bit of the current codon is modified, reducing the risk of obtaining “mutations”, i.e. the Gray code is a safer and more intuitive way to manipulate genetic code.

The author then presents three different 8x8 matrices composed of codons, which are known in biology for their biochemical properties. The codons of the matrices are transformed into their representation under Gray's code, and then the Hamming distance of these codons is calculated. The hamming distance is simply the sum of the points where two strings of characters compared differ. Three new matrices are obtained composed of integers resulting from the Hamming calculation of the 6 bits of a codon. The author studies the properties of these matrices and notices that two of them are doubly stochastic, while the latter is not.

The paper also mentions that the Levenshtein distance is the most sophisticated since it can also compare two strings even if one has characters and the other does not.

Summerized by Chadha Sakka