

AP
Résumé de publication scientifique
Genetic Code, Hamming Distance and Stochastic
Matrices

Matéo Pasquier

Janvier 2023

1 Résumé

Le papier explique l'utilité de représenter du code génétique en utilisant du code de Gray. Il explique d'abord la forme générale de ce code génétique : l'ARNm est composée de codons, les codons étant une série de trois nucléotides représentés par des lettres (C, A, G, U).

il définit ensuite la notion de "Gray code" : une séquence de bits similaire à du binaire, mais dont l'augmentation d'une unité ne résulte que d'un seul changement de bit. Il est alors possible de représenter chaque nucléotide par une série de 2 bits (4 états différents) et donc un codon entier par 6 bits.

L'intérêt de cette représentation réside dans le fait que les acides aminés sont formés à partir d'une séquence de codons "accolés" les uns aux autres (CCC, CCU, CCA, CCG); où seulement un nucléotide change à chaque fois ce qui se prête naturellement à une utilisation du code de Gray. En effet, cela permet tout le temps d'obtenir un codon adjacent lorsqu'un bit du codon actuel est modifié, réduisant le risque d'obtenir des "mutations", i.e le code de Gray est un moyen plus sûr et intuitif de manipuler du code génétique.

L'auteur présente alors trois différentes matrices 8x8 composées de codons, ces compositions étant connues en biologie pour leurs propriétés biochimiques. Les codons des matrices sont transformés dans leur représentation sous code de Gray, puis est calculé la distance de Hamming de ces codons. La distance de Hamming est tout simplement la somme des points où deux chaînes de caractères comparées diffèrent. Sont alors obtenues trois nouvelles matrices composées d'entiers résultants du calcul de Hamming des 6 bits d'un codon. L'auteur étudie alors les propriétés de ces matrices et note que deux d'entre elles sont doublement stochastiques, tandis que la dernière n'est pas stochastique.