Identifying mRNA Sequences and Proteins by Use of BCH Codes

Mario Enrique Duarte González

Universidad Antonio Nariño
Facultad de Ingeniería Electrónica, Biomédica y Mecatrónica (FIBEM)
mardugo@gmail.com

Abstract

One of the great challenges of the scientific community on theories of genetic information, genetic communication, genetic coding and proteomics is to determine a mathematical structure related to mRNA sequences and proteins. In [1], a model of an intra-cellular transmission system of genetic information, similar to a model of digital communication system, has been proposed; and narrow sense BCH cyclic linear error correcting codes (LECC) over $\mathbb{Z}_4$ and $\mathbb{F}_4$ have been used for identifying a mathematical structure in DNA and mRNA sequences. In this presentation, for mRNA sequences, we use the proposed transmission system and extend its capability by considering all possible BCH cyclic LECCs, hence, we are able to identify a mathematical structure for an increased number of mRNA sequences. For proteins, we establish an analogy between properties of error correcting codes and proteins; and propose a methodology for establishing a mathematical structure and for representing proteins through BCH cyclic LECCs. The analogy between LECCs and proteins is based on the error detection capability of chaperone molecules and is defined by the following links: 1) sequences over an alphabet of cardinality 20 with amino acid chains, 2) code's codewords with biologically functional proteins, 3) a code with a set of functional proteins and 4) correctable sequences for codeword c with proteins similar to a specific functional protein. We use BCH cyclic LECCs over alphabets $\mathbb{Z}_{20}$ and $\mathbb{F}_4 \times \mathbb{F}_5$ (both with 20 elements) for representing proteins. The labelling is made by using the isometry between $\mathbb{Z}_{20}$ and $\mathbb{F}_4 \times \mathbb{F}_5$ and the similarity between the mathematical structure of $\mathbb{Z}_{20}$ and the amino acid representation introduced in [2]. The BCH codes over ring alphabets $\mathbb{Z}_{20}$ and $\mathbb{F}_4 \times \mathbb{F}_5$ are designed according to [3] by considering their decomposition on local rings. As results, we show the algebraic structure of the corresponding codes for some mRNA sequences and proteins: alphabets, labels, primitive polynomials ($p(x)$), code generator polynomials ($g(x)$) and minimum distance. The mRNA sequences and proteins have been obtained from RCSB Protein Data Bank (PDB) and National Center for Biotechnology Information (NCBI) databases. Considering BCH codes over $\mathbb{Z}_4$ and $\mathbb{F}_4$, we identified, with one nucleotide difference, the mRNA sequences Rhagoletis pomonella contig 18598 and Acropora millepora SeqIndex1763; among others. Considering BCH codes over $\mathbb{Z}_{20}$ and $\mathbb{F}_4 \times \mathbb{F}_5$, we reproduce the IAAL-E3/K3 heterodimer protein and identify the Hepatitis GB virus B with one difference in one position; among others. The characterization of mRNA sequences and proteins may contribute to the development of a methodology that can be applied in mutational analysis, production of new drugs and genetic improvement, among other things, resulting in

the reduction of time and laboratory costs.

References

[1] L.C.B. Faria, A.S.L. Rocha, and R. Palazzo Jr. Transmission of intra-cellular genetic information: A system proposal. *Journal of Theoretical Biology*, 358(0):208–231, 2014.

[2] William Ramsay Taylor. The classification of amino acid conservation. *Journal of Theoretical Biology*, 119(2):205–218, 1986.

[3] Steven T. Dougherty and K. Shiromoto. MDR codes over $\mathbb{Z}_k$. *Information Theory, IEEE Transactions on*, 46(1):265–269, Jan 2000.