Revisiting the Genomics and Genetic Codes Using Walsh-Hadamard Spectrum Analysis

Mayasar Ahmad Dar and Deepmala Sharma(✉)

Department of Mathematics, National Institute of Technology, Raipur, India
deepsha.maths@nitrr.ac.in

Abstract. Walsh-Hadamard spectrum is widely used in the field of science and technology like classification of cancer cells, image processing, speech processing, signal and image compression etc. In this paper, a genomic analysis using Walsh-Hadamard spectrum and cross-correlation has been done. Transformation of genetic code using Walsh-Hadamard spectrum has been given. We redefine the Walsh-Hadamard spectrum in genomics and analyse the origin of mRNA features by using this spectra. Finally, using Walsh-Hadamard spectrum the overall energy of the mRNA sequence has been evaluated.

Keywords: Walsh-Hadamard spectrum · DNA · mRNA

1 Introduction

Every living organism starts its life from a single cell that contains DNA, its double helix structure carries genetic instructions for the development, functioning, growth and reproduction. Almost every activity of a living organism is based on the gene expression, regulation and protein synthesis that is based on the DNA codons. The DNA and mRNA sequences are key factors that determined the proteins. The process of DNA transformation into mRNA and subsequently proteins synthesis is basically governed by genetic code. The transparency order and Walsh-Hadamard spectrum have been widely used as an important tool for research in cryptography from last so many years, especially in the construction of cryptographically important Boolean functions, used in various cryptosystems. The Walsh-Hadamard spectrum has also been used in biological science, like Zhao and Pompili [1] classified human cancer cells (normal and diseased cells) based on Walsh-Hadamard spectrum of DNA Methylation Profile and showed that the transform domain vector is unique for a particular tissue type. The various characteristics, reactions and theories of DNA are explored by the DNA cryptography, in which DNA is considered as an information carrier [2]. A five-stage algorithm based on DNA cryptography was used to encrypt information [3]. The central dogma of biology is used in inspired pseudo biotic DNA cryptography [4]. A DNA based symmetric key cryptography for secure data transfer over the communication channel was analysed [5]. The effect of Walsh-Hadamard spectrum on various generalized Boolean functions has been analyzed by different authors. The various results of Boolean functions based on spectral

V. Agarwal et al. (Eds.): BIOSANGAM 2022, ABSR 24, pp. 106–113, 2023.
https://doi.org/10.2991/978-94-6463-020-6_11

analysis were provided by Sarkar and Maitra [6], and Zhou et al. [7]. In this paper, we investigate a Walsh-Hadamard spectrum analysis of the genetic code in the bit stream of Boolean functions and evaluate some important properties of such codes. We take the DNA and mRNA sequences as the bit streams only.

2 Preliminaries

Let $\mathbb{F}_2^n$ be the n-dimensional vector space over the finite field $\mathbb{F}_2$. A function from $\mathbb{F}_2^n$ to $\mathbb{F}_2$ is known as Boolean function. We denote by $\mathcal{B}_n$ the set of all n-variable Boolean functions. Any Boolean function can be represented as a multivariate polynomial called Algebraic Normal Form (ANF). The support of a Boolean function f is given by $S_f = \{x \in \mathbb{F}_2^n : f(x) \neq 0\}$, whose cardinality $|S_f|$ is known as the Hamming weight of f. The Hamming distance $d(f, g)$ between $f, g \in \mathcal{B}_n$ is the number of elements $x \in \mathbb{F}_2^n$ where these functions differ.

The Walsh-Hadamard transform of a function $f \in \mathcal{B}_n$ is the integer-valued function over $\mathbb{F}_2^n$ defined by

$$W_f(a) = \sum_{x \in \mathbb{F}_2^n} (-1)^{f(x) + <a,x>}$$

where $a \in \mathbb{F}_2^n$ and $<a, x>$ is an inner product.

The Cross-Correlation between the function f and g at $a \in \mathbb{F}_2^n$ is given by

$$C_{f,g}(a) = \sum_{x \in \mathbb{F}_2^n} (-1)^{f(x) + g(x+a)}$$

Moreover for $f = g$, the sum $C_f(a)$ is called Autocorrelation of f at a.

The sum-of-squares indicator of the Cross-Correlation between $f(x), g(x) \in \mathcal{B}_n$ is defined by

$$\sigma_{f,g} = \sum_{a \in \mathbb{F}_2^n} C_{f,g}^2(a)$$

The transparency order of a Boolean function $f \in \mathcal{B}_n$ is defined by [8]

$$TO_f = 1 - \frac{1}{2^n(2^n - 1)} \sum_{a \in \mathbb{F}_2^{n*}} |C_f(a)| \tag{1}$$

The below given lemma is the consequence of Lemma 2.2 [9].

Lemma 2.1. Let $k \in \mathcal{B}_n$ then for any $v \in \mathbb{F}_2^n$

$$\sum_{y \in \mathbb{F}_2^{n*}} |C_k(y)| \geq \left| W_k^2(v) - 2^n \right|$$

From above lemma and (1) we compute

$$W_k^2(v) \leq (1 - TO_k)2^n\big(2^n - 1\big) + 2^n \tag{2}$$

The upper bound for the above relations is obtained in case of bent function.

A sequence of amino acids determines a particular protein that is coded by mRNA. The DNA contains the genetic information which is transferred from one generation to another. A DNA molecule consists of four bases $Adenine(A)$, $Gaunine(G)$, $Thymine(T)$ and $Cytocine(C)$ which follow complementary base pairing rule. The bases in DNA and mRNA are same except that T is replaced by U in mRNA. The mRNA acts as a template for protein synthesis. Some of the protein features of mRNA codons are given below

GCA	*GCC*	*GCG*	*GCU*	*Alanine*
UGC	*UGU*			*Cystine*
GAC	*GAU*			*Aspartic Acid*

3 Genomic Analysis

From above section, we notice that transparency order, cross-correlation and Walsh-Hadamard spectrum are interrelated by various types of equalities or inequalities. They may be applied in genomics, a genomic analysis on any of these, directly or indirectly has impact on the other. The genomic data helps in assessing the various interactions between the biological processes. The genomic data may be treated as a function of genomic positions to evaluate the cross-correlation between these functions, here each genomic feature is considered as a Boolean function of genomic position t. In biological systems, the genomic features are of special importance not only at the same genomic positions, but also at the proximal positions. The transparency order definition wholly depends upon the correlation between the functions, so what type of correlation exists in the genes, the transparency order may change accordingly. These correlations may result from various types of interactions between the genes. The cross-correlation between two genomic functions $h(t)$, $r(t)$ is defined as follow

$$C_{h,r}(x) = \sum_{t \in \mathbb{F}_2^n} (-1)^{h(t)+r(t+x)}$$

where n is the length of genome.

There are various types of correlations having different roles [10], this cross-correlation in a similar way may be relevant in determining the nature of interactions like the scale of interactions between the genes.

The Walsh-Hadamard spectrum for a genomic sequence of length n is defined as

$$W(x) = \frac{1}{n}\sum_{i=0}^{n-1} h(i)M(x, i)$$

where $h(i)$ is the i^{th} genomic position in the genomic sequence of length n before transformation. $M(x, i)$ is the x^{th} row and i^{th} column position of the Walsh matrix and

$W(x)$ is the x^{th} position of the sequence after transformation and $x = 0, 1, \ldots n-1$. The Walsh matrix is given by

$$M_n = \begin{bmatrix} M_{n-1} & M_{n-1} \\ M_{n-1} & -M_{n-1} \end{bmatrix}$$

for $n \geq 1$ and $M_0 = 1$. The Walsh-Hadamard spectrum decomposes the original genomic sequence into a series of basic functions of Walsh matrix. This Walsh-Hadamard spectra and cross-correlation may help in understanding some properties of genomics.

Since DNA and mRNA sequences are bit streams, the Walsh-Hadamard spectrum at $y \in \{0, 1\}^n$ may alternately be defined as

$$W(y) = \frac{1}{2^n} \sum_{x \in \mathbb{F}_2^n} (-1)^{f(x) + <x,y>}$$

where n is the length of bit stream.

The basis function of Walsh-Hadamard spectrum is given by

$$B_y(x) = (-1)^{<x,y>}$$

where x and y are bit stream of length n. Here y is called the partition. The Walsh basis are orthonormal, therefore

$$\sum_{x \in \mathbb{F}_2^n} B_y(x) B_z(x) = \sum_{x \in \mathbb{F}_2^n} (-1)^{\langle x, y+z \rangle} = \begin{cases} 2^n, & when\ y = z \\ 0, & otherwise \end{cases}$$

Any function can be represented by Walsh-Hadamard basis. A Walsh-Hadamard basis helps in the construction of a function $f : X^n \rightarrow R$, which is expressed as the linear sum of Walsh-Hadamard spectrum as given below

$$f(x) = \sum_{x \in \mathbb{F}_2^n} W(y) B_y(x)$$

Thus Walsh-Hadamard spectrum can be considered as the relative contribution of the partition y to the function value of $f(x)$. In other words the absolute value of $W(y)$ is considered as the significance of the corresponding partition y. When magnitude of $W(y)$ is small then the y^{th} partition is said to be insignificant and ignore its contribution.

Suppose a function $f : X^n \rightarrow Y$, from the data $\{(x_1, y_1), (x_2, y_2), \ldots (x_t, y_t)\}$ is generated by some function : $X^m \rightarrow Y$, such that $\hat{f}$ approximates f. To learn $\hat{f}$ can be understood as the problem of approximating the Walsh-Hadamard spectrum of f. We can estimate the significant Walsh-Hadamard spectrum of f and use it to define $\hat{f}$. The complexity of inducing a function in Walsh representation is directly proportional to the number of such spectra. The Walsh-Hadamard spectrum in the Boolean domain has 2^n values and estimating all of them will require exponential time.

3.1 Transformation of Genetic Code Using Walsh-Hadamard Spectrum

In this section we investigate the effect of the genetic code like representation transformation using Walsh-Hadamard spectrum. There is a correspondence between mRNA and the proteins defined by genetic code. The codon is trinucleotide sequence of DNA or RNA that corresponds to a specific amino acid. The size of codon is three, but we may treat it as a parameter. This analysis investigates the effect of such transformations in the bit stream. The transformations for these codons using bit stream may be analysed analogously using Boolean functions. Although the strings are from the bit stream, we will use mRNA, proteins and genetic code accordingly for maintaining the link between the bioscience and the Walsh-Hadamard spectrum.

Let us take a mapping from the mRNA sequence (r) to the corresponding protein sequence (p) using genetic code as

$$F : R_{n_r} \rightarrow P_{n_P}$$

where n_r and n_P be the respective lengths of mRNA and protein sequence. In case of Boolean functions $R = P = \{0, 1\}^n$. Since mRNA codons are in the form of triplets, we represent these triplets in the form of bit streams having value of 1 or 0. Let us take an example of Code Z as:

AUU	100	1
GUA	101	1
CGA	011	0
GCU	001	0

Thus a particular mRNA codon maps to a single bit protein feature. As there are a large number of Boolean functions having the same value despite the fact that the functions are having different combinations of the bit streams. In a similar way there may be some codons which code for the same protein feature and we denote that class by τ_p. Thus the Walsh-Hadamard spectrum in the mRNA space may be written as

$$\begin{aligned} W(y) &= \frac{1}{2^{n_r}} \sum_{r\in\mathbb{F}_2^{nr}} (-1)^{f(r)+<r,y>} \\ &= \frac{1}{2^{cn_p}} \sum_{p\in\mathbb{F}_2^{np}} (-1)^{f(p)} \sum_{r_i\in\tau_p} (-1)^{<r_i,y>} \end{aligned} \tag{3}$$

Let S_0 and S_1 be the total number of codons that map to a protein feature value of 0 and 1 respectively. The cardinality of τ_p is $|\tau_p| = S_0^{n_{p,0}} S_1^{n_{p,1}}$, where $n_{p,0}$ and $n_{p,1}$ is the number of 0's and 1's in n_p. The magnitude of $\sum_{r_i\in\tau_p}(-1)^{<r_i,y>}$ may take values between 0 and $S_0^{n_{p,0}} S_1^{n_{p,1}}$. This may be considered as the scaling factor of every protein sequence to the y^{th} Walsh-Hadamard spectra. If the value of $(-1)^{<r_i,y>}$ depends only on features of r corresponding to 1's in the partition y, then mRNA features may belong to the same mRNA codon, different codons, or a combination of both. In other words they may originate from the same protein feature, different protein features, or a combination of both. If there are equal number of 0's and 1's in the protein feature, then we call it balanced. The protein feature is said to be perfect if

$$\sum_{x\in\mathbb{F}_2^{np}} (-1)^{f(x)+f(a+x)} = \begin{cases} 2^n, & if\ a = 1 \\ 0, & otherwise \end{cases}$$

3.2 Energy of mRNA Sequence

Some proteins recognize specific bases, consider a signal to be a DNA or mRNA sequence pattern that is recognized by a protein. In thermodynamical sense "recognize" means binding. The energy of Walsh-Hadamard spectrum can be defined as

$$E = \sum_{y\in\mathbb{F}_2^n} W^2(y)$$

The energy of the spectrum in terms of genetic code like representation is described below. From Eq. (3), we can write

$$W^2(y) = \frac{1}{2^{2cn_p}} \sum_{p,q\in\mathbb{F}_2^{np}} (-1)^{f(p)}(-1)^{f(q)} \sum_{r_i,k_i\in\tau_p} (-1)^{<r_i,y>}(-1)^{<k_i,y>}$$

$$\sum_{p\in\mathbb{F}_2^{np}} W^2(y) = \frac{1}{2^{2cn_p}} \sum_{p\in\mathbb{F}_2^{np}} (-1)^{2f(p)} \sum_{r_i,k_i\in\tau_p} \sum_{y\in\mathbb{F}_2^{np}} (-1)^{<r_i+k_i,y>}$$

Using the orthonormality condition we can write

$$\sum_{p\in\mathbb{F}_2^{np}} W^2(y) = \frac{1}{2^{cn_p}} \sum_{p\in\mathbb{F}_2^{np}} (-1)^{2f(p)} S_0^{n_{p,0}} S_0^{n_{p,1}}$$

We now specialize this result for the Code Z, for this code $S_0 = 2$, $S_1 = 2$ and $c = 3$, therefore we have from above equation

$$\sum_{p\in\mathbb{F}_2^{np}} W^2(y) = \frac{1}{2^{3n_p}} \sum_{p\in\mathbb{F}_2^{np}} (-1)^{2f(p)} 2^{n_{p,0}+n_{p,1}} = \frac{1}{2^{2n_p}} \sum_{p\in\mathbb{F}_2^{np}} (-1)^{2f(p)}$$

In other words we can say that the energy of protein sequence is same as that of mRNA sequence and is given by

$$E_{p,r} = \frac{1}{2^{2n_p}} \sum_{p\in\mathbb{F}_2^{np}} (-1)^{2f(p)}$$

Thus overall energy remains invariant under the transformation Code Z. The overall energy is an important property, the number and location of Walsh-Hadamard spectrum constitute the critical properties that significantly contribute to the overall energy. The energy of protein sequence is crucial in growth and maintenance, providing structure, maintaining proper pH, improving immune system, causing biochemical reactions etc.

4 Conclusion

In this paper, we give genomic analysis through Walsh-Hadamard spectrum and cross-correlation. We established a genetic code type transformation in terms of Walsh-Hadamard spectrum using the concept of Boolean functions and evaluated the origin of mRNA sequence with reference to their codons. Also we find the energy of the mRNA sequence with the help of Walsh spectra. Future scope of the work is to find more applications of Walsh-Hadamard spectrum and transparency order in identifying gene promoter regions and in studying the structural properties (conformational and physiochemical) of DNA.

References

1. X. Zhao and D. Pompili. Walsh-hadamard transform of DNA methylation profile for the classification of human cancer cells. in: Proceedings of the 5th International Conference on Bioinformatics and Computational Biology. 2017, p. 26–29 DOI: https://doi.org/10.1145/3035012.3035026
2. G. Xiao, M. Lu, L. Qin, and X. Lai, New field of cryptography: DNA cryptography. Chinese Science Bulletin. 51(12), 2006, p. 1413–1420. DOI: 1420 (2006). https://doi.org/10.1007/s11434-006-2012-5
3. M. Najaftorkaman and N. S. Kazazi, A method to encrypt information with DNA-based cryptography. International Journal of Cyber-Security and Digital Forensics (IJCSDF). 4(3), 2015, p. 417–426. DOI: https://doi.org/10.5120/ijca2016907719
4. E. S. Babu, C. N. Raju, and M. H. K. Prasad, Inspired Pseudo Biotic DNA based Cryptographic Mechanism against Adaptive Cryptographic Attacks. DOI: https://doi.org/10.6633/IJNS.201603.18(2).11
5. B. B. Raj, J. F. Vijay, and T. Mahalakshmi, Secure data transfer through DNA cryptography using symmetric algorithm. International Journal of Computer Applications. 133(2), 19–23. DOI: https://doi.org/10.5120/ijca2016907719
6. P. Sarkar and S. Maitra, Cross-correlation analysis of cryptographically useful Boolean functions and S-boxes. Theory of Computing Systems. 35(1), 2002, p. 39–57. DOI: https://doi.org/10.1007/s00224-001-1019-1
7. Y. Zhou, M. Xie, and G. Xiao, On the global avalanche characteristics between two Boolean functions and the higher order nonlinearity. Information Sciences. 180(2), 2010, p. 256–265. DOI: https://doi.org/10.1016/j.ins.2009.09.012
8. K. Chakraborty, S. Sarkar, S. Maitra, B. Mazumdar, D. Mukhopadhyay, and E. Prouff, Redefining the transparency order. Designs, codes and cryptography. 82(1), 2017, p. 95–115.
9. Q. Wang and P. Stănică, Transparency order for Boolean functions: analysis and construction. Designs, Codes and Cryptography. 87(9), 2019, p. 2043–2059. DOI: https://doi.org/10.1007/s10623-019-00604-1
10. E. D. Stavrovskaya, T. Niranjan, E. J. Fertig, S. J. Wheelan, A. V. Favorov, and A. A. Mironov, StereoGene: rapid estimation of genome-wide correlation of continuous or interval feature data. Bioinformatics. 33(20), 2017, p. 3158–3165. DOI: https://doi.org/10.1093/bioinformatics/btx379