บทความวิจัย

การตรวจสอบการกลายพันธุ์ของโปรตีนหนามในสายพันธุ์น่ากังวลของเชื้อโคโรน่าไวรัสสายพันธุ์ใหม่โดยวิธีการจำแนกโดยใช้ซัพพอร์ตเวกเตอร์บนระนาบของระบบพิกัดดีอาร์พี

กบิล กาญจมาภรณ์กุล[1] ธัญญดา รุ่งโรจน์มงคล[2,3] และสุพจน์ หารหนองบัว[1,*]

[1]ศูนย์เชี่ยวชาญเฉพาะทางด้านเคมีเชิงคำนวณ ภาควิชาเคมี คณะวิทยาศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย กรุงเทพมหานคร 10330

[2]ศูนย์เชี่ยวชาญเฉพาะทางด้านตัวเร่งชีวภาพและเทคโนโลยีชีวภาพยั่งยืน ภาควิชาชีวเคมี คณะวิทยาศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย กรุงเทพมหานคร 10330

[3]หลักสูตรชีวคอมพิวเตอร์และชีวโมเลกุลเชิงคำนวณ บัณฑิตวิทยาลัย จุฬาลงกรณ์มหาวิทยาลัย กรุงเทพมหานคร 10330

Email:supot.h@chula.ac.th

รับบทความ: 28 พฤศจิกายน 2564 แก้ไขบทความ: 21 มกราคม 2565 ยอมรับตีพิมพ์: 14 กุมภาพันธ์ 2565

บทคัดย่อ

การวัดเชิงปริมาณของสภาวะควอนตัมในจีโนไทป์ของกรดอะมิโนที่กลายพันธุ์ในโปรตีนหนามของสายพันธุ์น่ากังวลต่างๆของโคโรน่าไวรัสสายพันธุ์ใหม่ได้ถูกตรวจวัดโดยกระแสของเชอร์นไซม่อนส์ ในงานวิจัยนี้เราได้เสนอวิธีใหม่โดยใช้กระแสของเชอร์นไซม่อนส์ในตัวอักขระของรหัสพันธุกรรมในการวาดกราฟเพื่อใช้แสดงค่าความโค้งงอของการบิดตัวของโครงสร้างทุติยภูมิของโปรตีนหนามซึ่งวิธีนี้ตั้งอยู่บนทฤษฎีภาพสะท้อนของกระจกสมมาตรและเพิ่มคุณสมบัติพิเศษของความเชื่อหลักของชีววิทยาเข้าไป ในที่นี้สนามยังมิลส์ในรหัสพันธุกรรมคือสนามที่เกิดจากการเหนี่ยวนำโดยใช้ตัวดำเนินการเคอร์บนการขนส่งแบบขนานของพันธะไฮโดรเจนในชีวโมเลกุลทั้งสามแบบในระบบลากรานของบนระนาบดีอาร์พีในเทนเจ้นของแมนนิโฟวน์สี่มิติที่ถูกนิยามโดยพื้นผิวของสนามพฤติกรรมในการบิดตัวบนทอรัสที่ถูกสแปนด้วยเบซีสโดยเบสสี่ตัวคือเอ ที ซีและจี โดยวิธีนี้ผลของการวิเคราะห์ข้อมูลถูกแสดงผลในรูปแบบของฮิสโตรแกรมของค่าตัวแปรของสภาวะควอนตัมในจีโนไทป์ของการกลายพันธุ์ของโปรตีนหนามโดยใช้ข้อมูลจากรหัสพันธุกรรมระหว่าง 17 มกราคม 2563 ถึง 30 มิถุนายน 2564 พบว่ากราฟเลื่อนตัวไปทางซ้าย การจำแนกกลุ่มและการทำนายผลของการกลายพันธุ์เพื่อใช้ตรวจสอบการกลายพันธุ์บนบริเวณรีเซ็พเตอร์โดเมนในโปรตีนหนามและทำนายผลในกรดอะมิโน 120 ตัวในบริเวณนี้ในสายพันธุ์น่ากังวลเก้าสายพันธุ์โดยการวิเคราะห์เชิงปริมาณโดยใช้อัลกอลิทึมของการเรียนรู้ของเครื่องที่ตั้งอยู่บนพื้นฐานการคำนวณแบบการจำแนกโดยใช้ซัพพอร์ตเวกเตอร์บนวิธีการฟิตค่าตัวแปรของสภาวะควอนตัมในจีโนไทป์ในสารพันธุกรรมให้ผลลัพธ์ของคำนวณหาค่าของตัวแปรสภาวะทางควอนตัมในจีโนไทป์ของกรดอะมิโนที่กลายพันธุ์ด้วยประสิทธิภาพที่ดี

คำสำคัญ: กระแสของเชอร์นไซม่อน จีโนไทป์ โคโรน่าไวรัสสายพันธุ์ใหม่ ซัพพอร์ตเวคเตอร์แมชชีน

อ้างอิงบทความนี้

กบิล กาญจมาภรณ์กุล, ธัญญดา รุ่งโรจน์มงคล และสุพจน์ หารหนองบัว. (2565). การตรวจสอบการกลายพันธุ์ของโปรตีนหนามในสายพันธุ์น่ากังวลของเชื้อโคโรน่าไวรัสสายพันธุ์ใหม่โดยวิธีการจำแนกโดยใช้ซัพพอร์ตเวกเตอร์บนระนาบของระบบพิกัดดีอาร์พี. **วารสารวิทยาศาสตร์และวิทยาศาสตร์ศึกษา**, 5(1), 58-71. http://doi.org/10.14456/jsse.2022.6

Research Article

Detection of spike protein mutation in SARS-CoV-2 variants of concern with support vector classification over (d,r,p)-layer coordinate system

Kabin Kanjamapornkul[1], Thanyada Rungrotmongkol[2,3] and Supot Hannongbua[1,*]

[1]Center of Excellence in Computational Chemistry (CECC), Department of Chemistry, Faculty of Science, Chulalongkorn University, Bangkok 10330, Thailand

[2]Center of Excellence in Biocatalyst and Sustainable Biotechnology , Department of Biochemistry, Faculty of Science, Chulalongkorn University, Bangkok 10330, Thailand

[3]Program in Bioinformatics and Computational Biology, Graduate School, Chulalongkorn University, Bangkok 10330, Thailand

**Email:supot.h@chula.ac.th*

Received <28 November 2021>; Revised <21 Janruary 2022>; Accepted <14 February 2022>

Abstract

Quantitative measurement of the quantum genotype of mutated amino acids in the S protein on different variants of SARS-CoV-2 was performed with Chern-Simons current. The Chern-Simons current in our proposed novel approach was used to plot the curvature of S protein folding based on mirror symmetry and add more extra properties of central dogma in biology. The Yang-Mills field in genetic code is an induced curl operator of parallel transport of unified field between hydrogen bonding between 3 types of biomolecules over the Lagrangian system in (d, r, p)-layer coordinate system on a tangent of the four-dimensional manifold defined by the surface of behavior field in twisted torus spanned by four nitrogen bases of A, T, C, and G. The empirical analysis of quantum genotype of S protein between 17 January 2020 and 30 June 2021 was reported in the histogram with left shift. The prediction of the genetic variation in S protein of various variants of interest and variants of concern in SARS-CoV-2 were computed by using a machine learning algorithm based on support vector classification over the fitting parameter of the quantum genotype with good performance.

Keywords: Chern-Simons Current, Genotype, SARS-CoV-2, Support Vector Machine

Cite this article:

Kanjamapornkul, K., Rungrotmongkol, T. and Hannongbua, S. (2022). Detection of spike protein mutation in SARS-CoV-2 variants of concern with support vector classification over (d,r,p)-layer coordinate system. **Journal of Science and Science Education**, 5(1), 58-71. http://doi.org/10.14456/jsse.2022.6

Introduction

The hidden relationship between the changing of curvature in the curve shape of Spike (S)-protein folding and mutations occurring in the Severe Acute Respiratory Syndrome Coronavirus 2 (SARS-CoV-2) genetic code was one of many open questions for scientists from different fields (Kanjamapornkul, Rongrotmongkol, and Hannongbua, 2021). In the quantum biology approach, this issue can be explained in the context of theoretical investigation. In the reproduction process of SARS-CoV-2, the ability to change the frame of reference in the adaptive behavior of the virus for survival in nature with parasitism state induced the quantum transition states by deleting some genetic codes of S protein inside the circular tube of viral RNA and producing the new mutated quantum genotype in the excited state. Delta variant contains a double mutation in 4 amino acids in the receptor-binding domain (RBD) of S1 and two deletions of residue numbers 157 and 158. This is in contrast to B.1.1.7 or Alpha variant with a single mutation and three deletions of amino acids in the S protein. Prediction of future mutation with a new location and new deletion position of amino acid in the S protein of the new variant is concerned (Li, F., 2016). The adaptive Yang-Mills field was defined in Chern-Simons current in genetic code of all living organisms, including viruses. If the S -protein curvature is not in an equilibrium state with ACE2, the virus will be blocked out and tried to change its curvature with an adaptive Yang-Mills field in the genetic code.

The source of the adaptive genetic fields related to the natural selection process induces the mutation in the genetic materials. The new definitions of these behavior fields in genetic code involved curvature in protein folding in the cohomology theory of central dogma (Capozziello et al., 2018). In this work, we redefine them by borrowing the mathematical structure of Yang-Mills fields in physics (Capozziello, S., Pincak, R. and Kanjamapornkul, K., 2017) and adding more properties in parallel transport of behavior fields in the four genetic codes. The new definitions of a mathematical model on four-dimensional manifolds in quantum biology were used to explain the source of adaptive curvature in S protein upon binding with the host cells in various living organisms (Pincak, R., Kanjamapornkul, K., and Bartos, E., 2020a). At the time of research, N501Y.V2 VUI2020 B.1.1.7.2 (Alpha), B.1.617.2 (Delta), Gamma, Iota, Eta, Lambda, Beta, Kappa, and Mu are SARS-CoV-2 variants of interest and concern. More mutations such as Delta plus and Alpha plus are under monitoring with an expected impact on business and society with lockdown phenomena (Mlcochova et al., 2021). The typical way to detect the mutation in the S protein of the SARS-CoV-2 variant is through an alignment method over the alphabet code of the genome (Ali et al., 2021). Many problems with this method exist, such as the time computation being too slow (Sathipati et al., 2022). Moreover, the sorting algorithm over the string alphabet gives all statistical parameters limited for quantitative measurement (Kumar et al., 2021). Therefore, the new approach of Chern-Simons current in quantum biology was introduced for measuring the curvature in mutated S proteins. Using this approach, all statistical parameters can be computed and input to support vector classification to predict the new possible mutations in S protein (Kanjamapornkul et al., 2017). The detection of the source of genetic variation in new types of nine variants of SARS-CoV-2 was an indirect application of algebraic topology and cryptography (Anshel et al., 1999). Another point of view is that the curvature in hydrogen bonging of [A]-[T] and [C]-[G] can be visualized with a two-form differential, which can be extended to three forms by using a Chern-Simons current (Capozziello et al., 2018). The curvature in three alphabet codes in codon was defined by the smallest subunit of amino acids in protein folding. On the other hand, the new coordinate system for plotting the curvature in the genetic code had a source from 20 amino acids with transitions in the (d, r, p) - layer (Figure 1). One question arises from the new theory of cell biology. What is the relationship between the smallest subunit used to represent the living organism and the curvature of the protein docking system? The answer to this question allows the new definition of the (d, r, p) - coordinate system used to construct the new algorithm to be applied. The Information in genetic code produces the curvature by rotating a biological clock like a Caesar key used to authenticate the correct folding in the protein-protein interaction system. In quantum cryptography for biology, this encrypting state is the analogy to the smallest subunit of ciphertext for encoding in the genetic code and transmission through the cell signal channel. The peptide bond in protein

folding can use the Minkowski cone with an embedded parabola equation for regression to define and detect the mutation in the configuration space of this undivided smallest subunit in the genetic code (Pincak, Kanjamapornkul and Bartos, 2020b). The transitions between DNA, RNA, and protein were used to redefine the old algorithm from public and private key pair in a qubit state in protein folding to the new definition of Hopf fibration in the Kolmogorov space of biological time-series data. (Kanjamapornkul, and Pincak, 2016).

The mutations on the SARS-CoV-2 S protein were highly considered (Andersen et al., 2020), because its composition is used for binding with receptor proteins. The S protein has a high degree of genetic variation (Schoeman and Fielding, 2019). To detect the genetic variation part of S protein means detecting the change in adaptive behavior field in the genetic code of SARS-CoV-2. When this component is replicated to the next generation through viral replication, some contaminated noises exist within the natural source of quantum genotype that gets mixed with the origin of the S clade into many SARS-CoV-2 variants. In this research, the nine variants of concern were selected as the mutated S proteins for quantitative measurement of the quantum genotype of SARS-CoV-2. The prediction of the new possible mutation was conducted by using the classification between mutated and non-mutated amino acids 401-520 that covers the area of the RBD domain (Kumar et al., 2021). The genetic variation detection of S protein can reveal the source and origin of changing curvature in different SARS-CoV-2variants.

Theory and Methodology

In quantum chemistry, the S protein folding state is visualized by using a coordinate of atoms in each amino acid and the electrostatic potential of the electrons in each atom that forms their molecule in the Euclidean space. For this reason, the computation is costly and takes a lot of time. To solve this problem, the new concept of quantum biology allows us to define the new coordinate system (d,r,p)-layer of three types of behavior fields in biomolecules, DNA, RNA, and protein (Pincak, Kanjamapornkul and Bartos, 2019). These new quantities of canonical spinor states are the foundation of unified hydrogen bonding in nitrogen bases [A]-[T] and [C]-[G] for the three undivided components of three types of biomolecules into behavior field in the twisted torus surface based on the central dogma (Kanjamapornkul, Rongrotmongkol, and Hannongbua, 2021). Moreover, the coordinate of hydrogen bonding in viral RNA, r_i has 1-1 maps to coordinate of S protein, p_i, and coordinate of host cell receptor protein $p_i^* = d_i$ in the chain sequence in the extended central dogma $r_i \to p_i \to d_i$.

Definition of (d,r,p)-Layer Coordinate System

Let $([A],[U],[C],[G]) := (x,y,z,t) \in X \subset \mathbb{R}_1^4 = H = C \times C$ be a coordinate of viral RNA in the open set X in the manifold of the genetic code of S protein. A chain sequence of transition state along with the central dogma from viral RNA state to S protein state in the replication process,

$$0 \to \mathrm{X} \to C_0(X) \overset{d}{\to} C_1(X) \overset{d}{\to} C_2(X) \overset{d}{\to} C_3(X) \to 0. \tag{1}$$

The expression above is equivalent with differential form in De Rham cohomology in differential calculus,

$$0 \to X \to \Omega^0(X) \overset{d}{\to} \Omega^1(X) \overset{d}{\to} \Omega^2(X) \overset{d}{\to} \Omega^3(X) \overset{d}{\to} \Omega^4(X) = 0, \tag{2}$$

$$0 \overset{\simeq}{\to} \mathrm{x} \overset{d}{\to} \mathrm{dx} \overset{d}{\to} d^2\mathrm{x} \overset{d}{\to} d^3\mathrm{x} \overset{d}{\to} d^4\mathrm{x} = 0. \tag{3}$$

In quantum biology, the new definitions of the coordinate system for four alphabet codes in SARS-CoV-2 are obtained using the chain sequence above with adding some extra properties of Lie transport of support vector of behavior field along the unit circle of the predefined plane wave function in Frechnet-Serret formulas for moving frame (Kanjamapornkul, Rongrotmongkol and Hannongbua, 2021).

Wave function of S protein in Quantum Biology

In this section, the plane wave function of four alphabet codes is defined with a Hamiltonian operator for the eigenvalue of their behavior field in fiber space of moving frame of (d, r, p) -layer.

Definition 1

Let ε be the evolutional field in genetic code, and $\theta(t)$ be the evolutional path along the biological clock $\mathbf{t}$. The $x_t = (\varepsilon\cos\theta(t), \varepsilon\sin\theta(t), t)$ is biological time-series data of genetic code. The (d, r, p) -coordinate of genetic code in three-dimensional coordinates of cylindrical coordinate over transformation from four alphabet codes in four-dimensional manifold to (d, r, p), and to (ε, θ, t). The evolutional field is defined by varying the distance in radius. For simplicity of the study, we normalized to the constant radius in the unit circle $\boldsymbol{\varepsilon}$=1. The eigenvectors of the transition state in four alphabet codes are defined by,

$$x_{[A]}(t) = (\cos 0, \sin 0, t),\ x_{[C]}(t) = \left(\cos\frac{\pi}{2}, \sin\frac{\pi}{2}, t\right), x_{[U]}(t) = \left(\cos\frac{2\pi}{2}, \sin\frac{2\pi}{2}, t\right), \tag{4}$$

and

$$x_{[G]}(t) = \left(\cos\frac{3\pi}{2}, \sin\frac{3\pi}{2}, t\right). \tag{5}$$

The spinor fields in genetic code are the projection from the quaternionic field into two complex planes with predefined wave functions in the first d-layer mirror of the left chiral plane in quantum biology.

Definition 2

The plane wave function of behavior fields in four alphabet codes of S protein is defined in general form over the unit circle by,

$$\Psi = e^{i\theta} = \cos\theta + i\sin\theta = (\cos\theta, \sin\theta) \epsilon S^1 \subset C, \tag{6}$$

with Hamiltonian operator for eigenvector of [A],[U],[C], and [G], in d-layer of fiber space,

$$H_{d=[C]}\Psi = \frac{d\Psi(\theta)}{d\theta} = ie^{i\theta} = e^{i\frac{\pi}{2}}e^{i\theta} = e^{i\left(\theta+\frac{\pi}{2}\right)}, \frac{d\Psi(\theta)}{d\theta}|_{\theta=0} = \cos\frac{\pi}{2} + i\sin\frac{\pi}{2} = x_{[C]}(t), \tag{7}$$

$$H_{d=[U]}\Psi = \frac{d^2\Psi(\theta)}{d\theta} = (i)(i)e^{i\theta} = e^{i\pi}e^{i\theta} = e^{i(\theta+\pi)}, \frac{d^2\Psi(\theta)}{d\theta}|_{\theta=0} = \cos\pi + i\sin\pi = x_{[U]}(t), \tag{8}$$

$$H_{d=[G]}\Psi = \frac{d^3\Psi(\theta)}{d\theta} = (i)(i)(i)e^{i\theta} = e^{i\frac{3\pi}{2}}e^{i\theta} = e^{i\left(\theta+\frac{3\pi}{2}\right)}, \frac{d^3\Psi(\theta)}{d\theta}|_{\theta=0} = \cos\frac{3\pi}{2} + i\sin\frac{3\pi}{2} = x_{[G]}(t), \tag{9}$$

$$\Psi = \frac{d^4\Psi(\theta)}{d\theta} = -(i)(i)e^{i\theta} = e^{i2\pi}e^{i\theta} = e^{i(\theta+2\pi)}, \frac{d^4\Psi(\theta)}{d\theta}|_{\theta=0} = \cos 2\pi + i\sin 2\pi = x_{[A]}(t). \tag{10}$$

The eigenvalue of the behavior field in genetic code is a projection from the quaternionic field to one side of the left chiral mirror complex plane of their hidden spinor field. It is a hidden eigenstate because these quantities are not actual values.

Definition 3

The(d, r, p)-layer coordinate system is the cyclic cylindrical coordinate with fiber space in the plane of cyclic polar coordinate in the projection of viral RNA state discontinuous transcript into S protein state. We have $((x, y), z) = ((r, p), d)$ with

$$\mathbf{x}(t) = (\cos\theta(t), \sin\theta(t), t), \tag{11}$$

$$\mathbf{r}(t) = \frac{d}{dt}(\cos\theta(t), \sin\theta(t), t) = (-\sin\theta(t), \cos\theta(t), 1). \tag{12}$$

The normal vector is $\mathbf{p}(t)$ state with

$$\mathbf{p}(t) = \frac{d\mathbf{r}}{dt} = (-\cos\theta(t), -\sin\theta(t), 0), \mathbf{d} = \mathbf{r} \times \mathbf{p}, \tag{13}$$

$$\frac{d\mathbf{d}(t)}{dt} = -\beta\mathbf{p}(t), \tag{14}$$

$$\frac{d\mathbf{r}(t)}{dt} = \kappa\mathbf{p}(t), \tag{15}$$

$$\frac{d\mathbf{p}(t)}{dt} = -\kappa\mathbf{r}(t) + \beta\mathbf{d}(t). \tag{16}$$

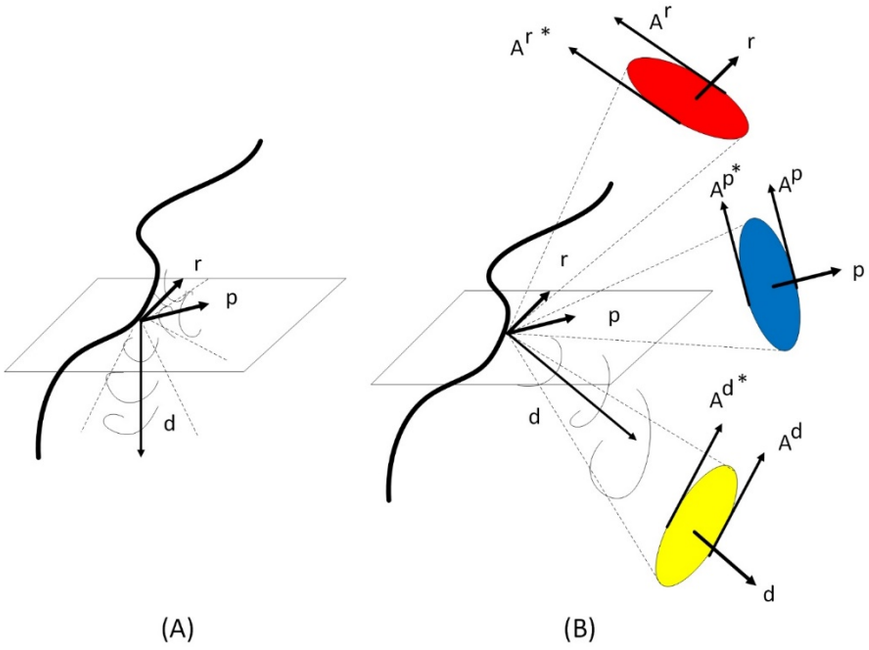

Figure 1. (A) The moving frame of (d,r,p)- coordinate system in the curve shape of protein folding with tangent and normal vector. (B) The induced behavior field in genetic code from moving frame in fiber space of tangent of the manifold of curve shape in protein folding with Minkowski cone in the ray of (d,r,p)-layer.

Relationship between Yang-Mills behavior field and the curvature in S protein folding

We define β with $A_{\mu=\beta}$, an inertial behavior in genetic code analogy with mass in quantum biology. It is an inertial behavior in the frame of reference in the Jacobian transform of (d,r,p)-coordinate system defined by quantum genotype g_{ij}. The curvature in viral RNA folding is as a source of Yang-Mills behavior field in SARS-CoV-2's viral RNA. In this research, we choose $-\Gamma_{ij}^{k}(g) = -A_{\mu} = f(\beta)$ for function f, implying that the quantum genotype is a source of curvature in the docking system in S protein binding to ACE2. This curvature change is the source of Chern-Simons current in the evolutional path of parasitism state between virus and host cell. The curvature in S proteins of SARS-CoV-2 can directly be computed from their genetic code of amino acid sequence appeared by using coordinate transform over the behavior field in genetic code $A^{r}(g_{ij}^{r}) = A_{\mu} = \Gamma_{ij}^{\mu}$. In general, $A_{\mu=k}(g)$ is an inertial behavior in the adaptive evolutional field in quantum genotype analogy with parallel transport of (d,r,p)- -moving frame in fiber space of tangent of the manifold. It is an inertial behavior in the frame of reference with the left translation of hydrogen bonding between [A]-[U] and [C]-[G] in viral quantum genotype g_{ij}^{r} of S protein. In this case, $\Gamma_{ij}^{k}(g) = A_{\mu} = V^{ext}(r)$ for van Der Waals force field $V^{ext}(r) = \frac{1}{r^{6}}$ was used. By the fact that the inversed expression contains six roots, the quantum genotype is a source of localized Lennard Jones potential (Jones, 1924) along the curvature of protein folding in the docking system of protein-protein interaction with six hidden transition states.

Let the coordinate transform g_{ij} be a cocycle of quantum genotype. The coordinate system is on the surface of twistor in a four-dimensional manifold represented the behavior field in genetic code as a source of evolution to exchange the hydrogen bonding between configuration spaces of the viral RNA state $r_i, r_j \in \{[A],[T],[C],[G]\}$ in two charts of open set in the tangent of the manifold of the co-living organism. The connection $\Gamma_{ij}^{k}(g)$,

$$\Gamma_{ij}^{\mu}(g) = \frac{1}{2} g^{\mu l}(\partial_j g_{il} + \partial_i g_{lj} - \partial_l g_{ji}). \quad (17)$$

The connection localized between three types of changing fields in each cocycle along with the fiber space of gauge group action $g_{ij} \in G$ in the tangent of the manifold as a source of Lie transport of support vector in Poisson bracket for parasitism state between SARS-CoV-2's (+) mRNA and host cell synthesis of (-)-mRNA prior

to the transition to (+) mRNA again as the next generation of new synthesis of viral mRNA, and S protein from host cell tRNA.

We use Lagrangian, $L_{Virus}^{Type-II}(r,r')$ for the system of viral (+)-mRNA- host cell (-) mRNA interaction with underlying genotype written in terms of cocycle g_{ij} of Jacobian transform of the coordinate-system in viral plus strand-RNA of SARS-CoV-2, $<r|$, with quantum genotype g^r and host cell minus strand of discontinuous transcription, $|r'>$ for this internal reproduction process between SAS-CoV-2 and host cell. It is a change of coordinate system underlying plus and minus-RNA sequence of quantum genotype g^r between $<r|r'>:=g^r<r,r'>$. Let $X=S^3$ be a manifold of a living organism with the localized coordinate system of the spinor field in genetic code along the fiber space of protein state $x=(r,p(\mathrm{r}),\mathrm{d}(r))\in X, L_{Virus}^{Type-II}: T_xX\times\mathbb{R}\to\mathbb{R}$.

The viral replication process in the host cell is the analogy with the flow of coordinate system by the transformation of alphabet code in four-dimensional space to Minkowski space with the extra dimension of the biological clock in each (d,r,p)-layer. The changing frame of inertial behavior field in genetic code of living organisms generate all behavior fields by six differential two forms in $\Omega^2(X)$. We denoted these six components by $(A^d,A^r,A^p,A^{d^*},A^{r^*},A^{p^*})$. These connections induce adjoint cofunctors of the interaction between six types of behavior field which formed an open set of sheave cocycles in Groethendieck cohomology in biology for a source of the canonical coordinate of quantum genotypes in all living organisms, $M=\{\mathcal{O}_{A^d},\mathcal{O}_{A^r},\mathcal{O}_{A^p},\mathcal{O}_{A^{d^*}},\mathcal{O}_{A^{r^*}},\mathcal{O}_{A^{p^*}}\}$. The Groethendieck topology for measuring the distance in the six coordinates system along the parallel transport in biomolecule is a discrete topology in Kolmokorov space with power set $P(M), n(P(M))=2^6$, as a source of 64 transition states in the codon. The elements of M are pairs of localized hydrogen boinding between two biomolecules with left and right chiral supersymmetry. According to mirror symmetry in the extra dimension, the observation can occur only in one direction. Therefore, we observe only by using the Hodge start operator between three differential forms over six-dimensional manifold M, $*:\Omega^3(M)\to\Omega^{6-3}(M)$. This star product is an origin of anti self-dual, AdS form in the 20 amino acids of protein state since the total number of differential three form over six-dimensional manifold is equal to $6_{C_3}=20$, the number of 20 amino acids in nature by biogenesis of 64 codons. The least action of the geodesic path over 64 genetic codes is a new definition of Chern-Simons current over behavior field,

$$A:=A_{\mu=k}, F=dA=curl(A),\ *F=-F,\ S_{CS}=\frac{k}{4\pi}\int tr(A\wedge dA+\frac{2}{3}A\wedge A\wedge A)\ , \qquad (18)$$

where $k=1,2,\cdots,64$. Let the least action defines the curvature of L-amino acids S_{CS} over the momentum space of protein state $p=\partial A_\mu$ and spatial space of RNA $r=A_\mu$, where A_μ is a behavior field in genetic code with inertial behavior $\beta_{i,i=1,\cdots,20}$, for 20 L-amino acids.

Definition 4

The Yang-Mills field in genetic code is an induced field by using curl operator over a connection,

$$\mathrm{F_{\mu\nu}}=\partial_\mu \mathrm{A_\nu}-\partial_\nu \mathrm{A_\mu}. \qquad (19)$$

The dual behavior field in AdS is $*\,\mathrm{F_{\mu\nu}}=-\mathrm{F^{\mu\nu}}$, and $<\mathrm{F}*\mathrm{F}>=-\mathrm{F}^2$, where $\mathrm{A_\mu}$ is a connection or intrinsic gravitational field from the spacetime curvature effect in the protein-protein interacting system.

Algorithm to plot the curvature in (d,r,p)-Layer Coordinate System

The Chern-Simons current in genetic code is different from Chern-Simons current in physics due to the underlying manifold of a living organism $X_t=S^3=\mathbb{R}_1^4=H=C\times C$.

Definition 5

The Chern-Simons current in genetic code $\mathrm{J^\mu}$ is defined by a change of the curvature of the interaction Yang-Mills fields between the SARS-CoV-2 S protein and host cell receptor protein or antibody,

$$\mathrm{J^\mu}=-\frac{1}{4}<\mathrm{F_{\mu\nu}}*\mathrm{F^{\mu\nu}}>=\frac{\partial \mathrm{S_{CS}}}{\partial \mathrm{A}}=\frac{\partial}{\partial \mathrm{A}}\frac{\mathrm{k}}{4\pi}\int \mathrm{tr}(\mathrm{A}\wedge \mathrm{dA}+\frac{2}{3}\mathrm{A}\wedge\mathrm{A}\wedge\mathrm{A}), \qquad (20)$$

where $\mathrm{F_{\mu\nu}}$ is the Yang-Mills behavior field curvature in proteins of the SARS-CoV-2 under the cocycle of quantum genotype $\mathrm{g_{ij}}$ of the viral RNA state. The quantity of $\mathrm{F^{\mu\nu}}$ measures the curvature change between

the group action of the cocycle of the quantum genotype $\mathbf{g^{ij}}$ of the behavior field in the underlying DNA sequence of the host cell receptor protein.

The explicit form with transition state k is written by

$$J^{\mu=k} = \sqrt{\frac{2}{k+2}} sin(\frac{\pi}{k+2}), \quad k = 1, \cdots, 64. \tag{21}$$

The resulting computation of all propagators in 64 transitions in the codon table for plotting the curvature directly from the average of those values into 20 amino acids is shown in Table 1 (Capozziello et al., 2018).

Dataset

All samples downloaded from GISAID were used to study the genetic variation between unmutated S protein in S clade of SARS-CoV-2 and mutated S protein in nine variants of concerns. In this research, the empirical analysis was employed on the two data groups with 774 selected samples. The first group was the biological time series of chosen genetic code randomly with only one sample of sequence of about 1273 amino acids in S protein per one submission day from 17 January 2021 to 30 June 2021 within 504 samples. The second group of samples was selected 30 samples per a variant of concerns. The list of nine variants is Delta, Gamma, Beta, Alpha, Mu, Iota, Lambda, Kappa, and Eta within 270 samples of selected S protein randomly. We have 774 samples of S protein from two different groups of the study.

Results

The new algorithm of the Chern-Simons current was performed over all the samples in the dataset yielded two main results. The resulted running with the first type of algorithm of Chern-Simons current for computing the time series of quantum genotype for detecting the mutations within the two-year samples. The second algorithm was Chen-Simons current in the physiology of time series in which performed over the second group of S protein with nine variants concerned for the classification of mutated amino acids. This result was used to forecast the mutation of variants in S-protein with a support vector machine between nine variants and S clade. The details of both two algorithms are shown in Figure 2. The new method of plotting the genetic code using the Chern-Simons current was applied to classify the similarity of the shape of the wave function of the S protein of samples. The similar wave function shapes imply that the quantum biological states in the genetic code have similar transition states. The Chern-Simons current in the genetic code of the SARS-CoV-2 S-protein was computed to measure quantum genotype by fitting the parameters in the parabola curve. The quantum genotype in the Chern-Simons current of 30 samples in nine variants of concern was caulculated. After that, we evaluated the fitting regression parameter of the plot in the trend of the Chern-Simons current with quadratic and cubic polynomials and measured the difference between the regression coefficients to allow a quantitative measurement of the transition state from the mutated and non-mutated amino acids in different variants. It was a complementary approach to the qualitative sequence alignment method with input parameters for the machine learning algorithm. The computed regression coefficients β_1, β_2 and β_3 from the equation $y_t = \beta_1 x_t^2 + \beta_2 x_t + \beta_3$ of β in the genetic code of the S proteins from 504 samples shown in the histogram of Figure 3. The fitting parameters are summarized in Tables S1-S6 in Supporting Information. The spectrum image in the mutation of amino acids 480-520 in January 2021 shows that amino acids 484, 485, and 486 are more yellow than others (Figure 4). These locations in RBD offer more evolution than other amino acids in periods of 1-31 January 2021. Moreover, we observed from the image pattern in each amino acid in RBD that Mu is similar to Lambda as seen from the pattern of blue color (Figure 5).

The performance of prediction of mutated amino acids in S protein

The calculations of the quantum genotype in mutated S protein of nine variants were performed by using the second algorithm of Chern-Simons current in physiology for 270 samples and fitting with parameters in quadratic equation $y_t = \beta_1 x_t^2 + \beta_2 x_t + \beta_3$.

After that, the classification between mutated and non-mutated amino acids was considered using support vector classification with a radial-based kernel and the input of two classes with two factors. These two factors are the coefficient $\beta_1$and β_3. The first class was assigned the value of 1 for the mutated amino acid. The second class was given -1 for non-mutated amino acid in S clade. The results of the classification of the whole sample of mutated S protein between Delta variants and other variants with 30 selected samples are the coefficients $\beta_2$and β_3 in Figure S1 in supporting information. The forecast of mutated and non-mutated amino acids in between nine variants and S clade is shown in the scattering plot with support vector, and the separated plane is illustrated in Figure S2 in supporting information. The machine learning algorithm in this research supports vector classification with radial base function as the chosen kernel for training with 80% of random amino acids 401-520 in S clade and variants of concern. Training the known samples is a support vector with the separated plane (Figure S2 in supporting information) for classifying the unknown sample.

The performance of forecasting of unknown test sample for 120 amino acids in residue numbers 401-520 are given in Tables S7-S9 in supporting Information with percentage of correction counting from 120 amino acids in unknown test samples. The results are as follows: 75% for S (non-mutated amino acids) 80% for Delta variant, 91.66% for Alpha variant, 79.16% for Beta, 79.16% for Gamma variant, 81.67% for Mu variant, 83.33% for Eta variant, 86.67% for Kappa variant, 80.83% for Iota variant, and 86.67% for Lambda variant. The quantitative measurement of quantum genotype in each mutated amino acid in nine variants is focused. In our 270 samples, 39 mutated amino acids are detected in the nine variants. The computation result was comparative with non-mutated amino acids in S clade (Table 1). This procedure was successfully applied to detect and classify the mutation between unknown and known variants. The difference between these two quantities implies the changing of curvature in S protein folding in each amino acid in quantum biology.

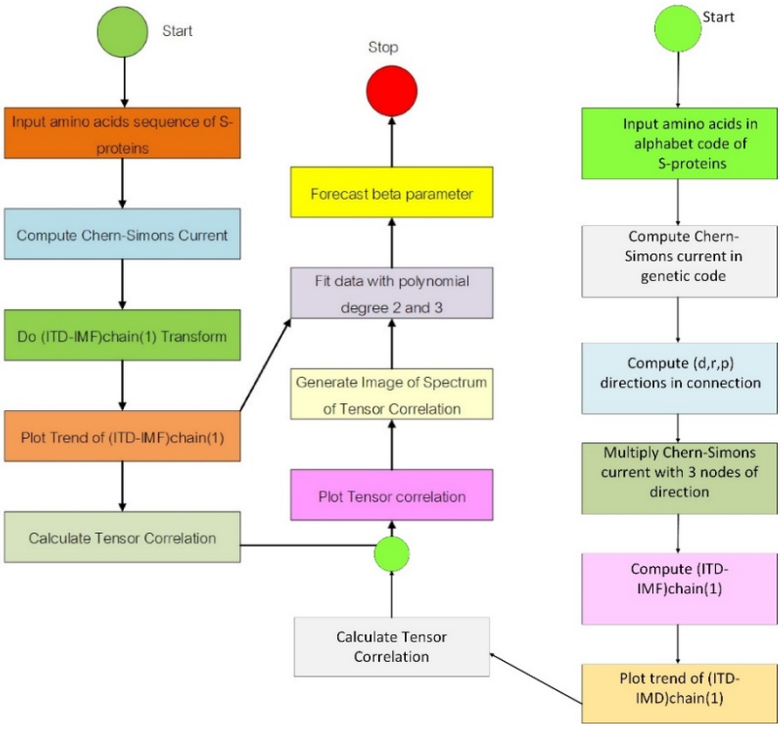

Figure 2. The flowchart of an algorithm for the plot in (d,r,p) coordinate system. (Left) The first type of algorithm of Chern-simons current in genetic code, and (right) the new algorithm of physiology in biological time series data for the measurement of quantum genotype

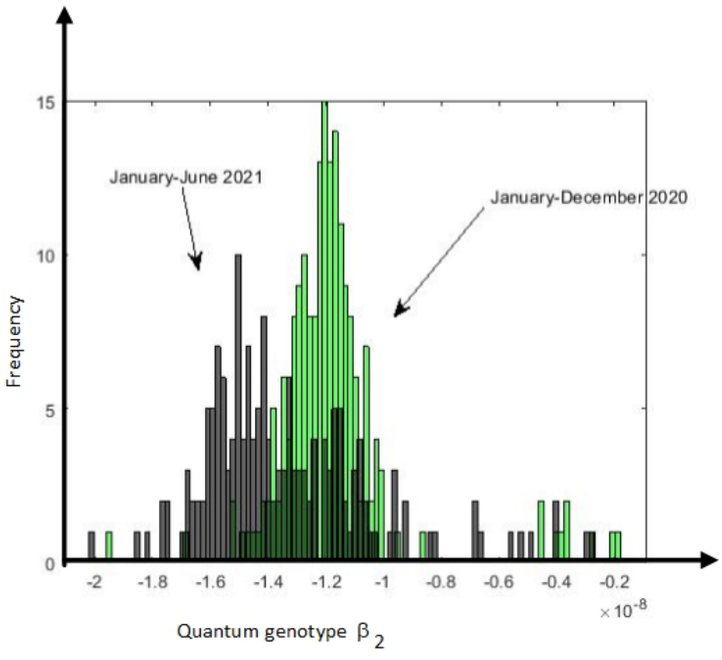

Figure 3. The histogram of quantum genotypes of S protein in 2020-2021. There is a value shift between January-June 2021 and January-December 2020 in the direction of decreased value.

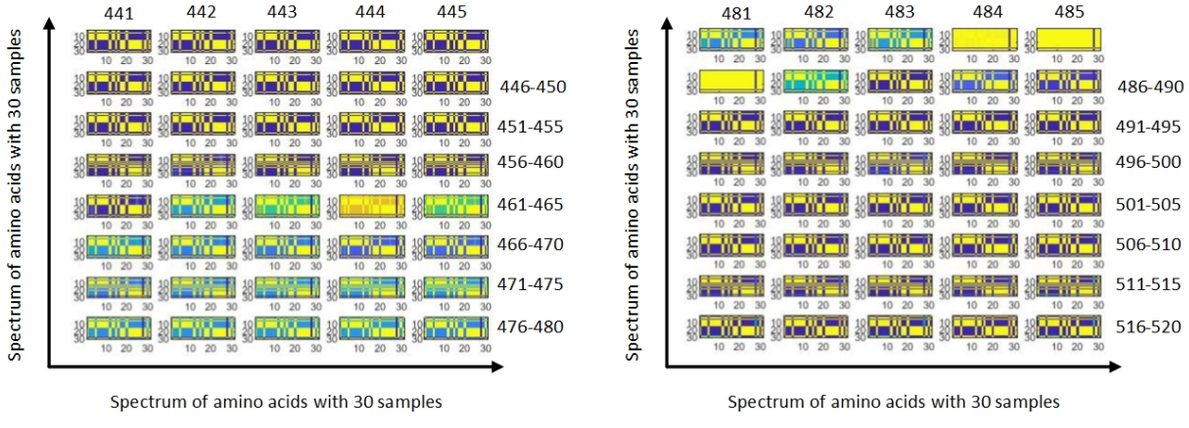

Figure 4. The image of the spectrum of tensor correlation of $(ITD - IMF)chain(1)$ of S protein in 31 samples of 1-31 January 2021. The image of each amino acid from left to right and up to down for amino acids 441-480 (A) and 481-520 (B)

Table 1. All mutated amino acids in S proteins from different variants of concern. The measurement of quantum genotype between mutated and non-mutated residue is also shown.

No	Amino Acids	Variant	β_S	$\beta_{variant}$	β_{diff}
1	L5F	Iota	-4.2572	4.2359	8.4931
2	L18F	Gamma	0.5784	3.6582	3.0798
3	T19R	Delta	0.0616	0.0568	0.0048
4	T20N	Delta	0.4536	0.0172	0.4382
5	P26S	Gamma	-0.1726	0.1193	0.2919
6	Q52R	Eta	-0.0112	0.0743	0.0855
7	Q75S, G75V	Iota, Lambda	0.0385	1.3523	1.3143
8	T76I	Lambda	0.2344	0.5273	0.2929
9	D80A	Beta	-0.1781	0.2618	0.4399
10	T95I	Iota, Mu	-3.8385	0.1659	4.0044
11	D138Y	Gamma	0.1707	0.0489	0.1218
12	Y144T	Mu	-0.2014	0.0587	0.2601
13	E154K	Kappa	0.1172	-0.0223	0.0949
14	E156G	Delta	0.0293	0.0309	0.0016
15	R190S	Gamma	0.4232	0.0502	0.3730
16	D215G	Beta	-0.0217	-0.0223	0.0006
17	D253N, D253G	Lambda, Iota	0.7827	-0.0195	0.8022
18	R346K	Mu	0.1344	0.2434	0.1090
19	K417T, K417N	Gamma, Beta	-2.4618	0.0567	2.5185
20	L452Q, L452R	Lambda, Delta, Kappa	-0.0412	0.1154	0.1566
21	T478K	Delta	0.2191	0.0015	0.2176
22	E484K, E484Q	Kappa, Eta, Beta, Iota, Mu	0.0948	0.1294	0.0346
23	F490S	Lambda	-0.0365	0.1161	0.1526
24	N501Y	Alpha, Gamma, Mu, Eta	-2.6878	0.0859	2.7737
25	A570D	Alpha	0.1553	0.0705	0.0848
26	D614G	All variants	-0.0400	0.0903	0.1303
27	H655Y	Gamma	0.1729	0.3125	0.1396
28	Q675H	Lambda	-0.1431	0.1101	0.2532
29	Q677H	Eta	-0.2581	-0.0584	0.1997
30	P681R	Mu, Kappa, Gamma, Alpha, Delta	0.9368	0.0658	0.8710
31	A701V	Iota, Beta	0.2154	1.1109	0.8955
32	T859N	Lambda	0.0224	-0.0509	0.0733
33	F888L	Eta	-0.2248	0.0049	0.2297
34	D950N	Mu, Delta	0.7321	-0.0727	0.8048
35	S982A	Alpha	0.3482	0.0296	0.3186
36	T1027I	Kappa	0.1537	0.0298	0.1239
37	H1101D	Delta	0.6623	0.0076	0.6547
38	V1176F	Gamma	1.8372	4.0715	2.2343
39	D1118H	Alpha	0.3570	-0.0625	0.4195

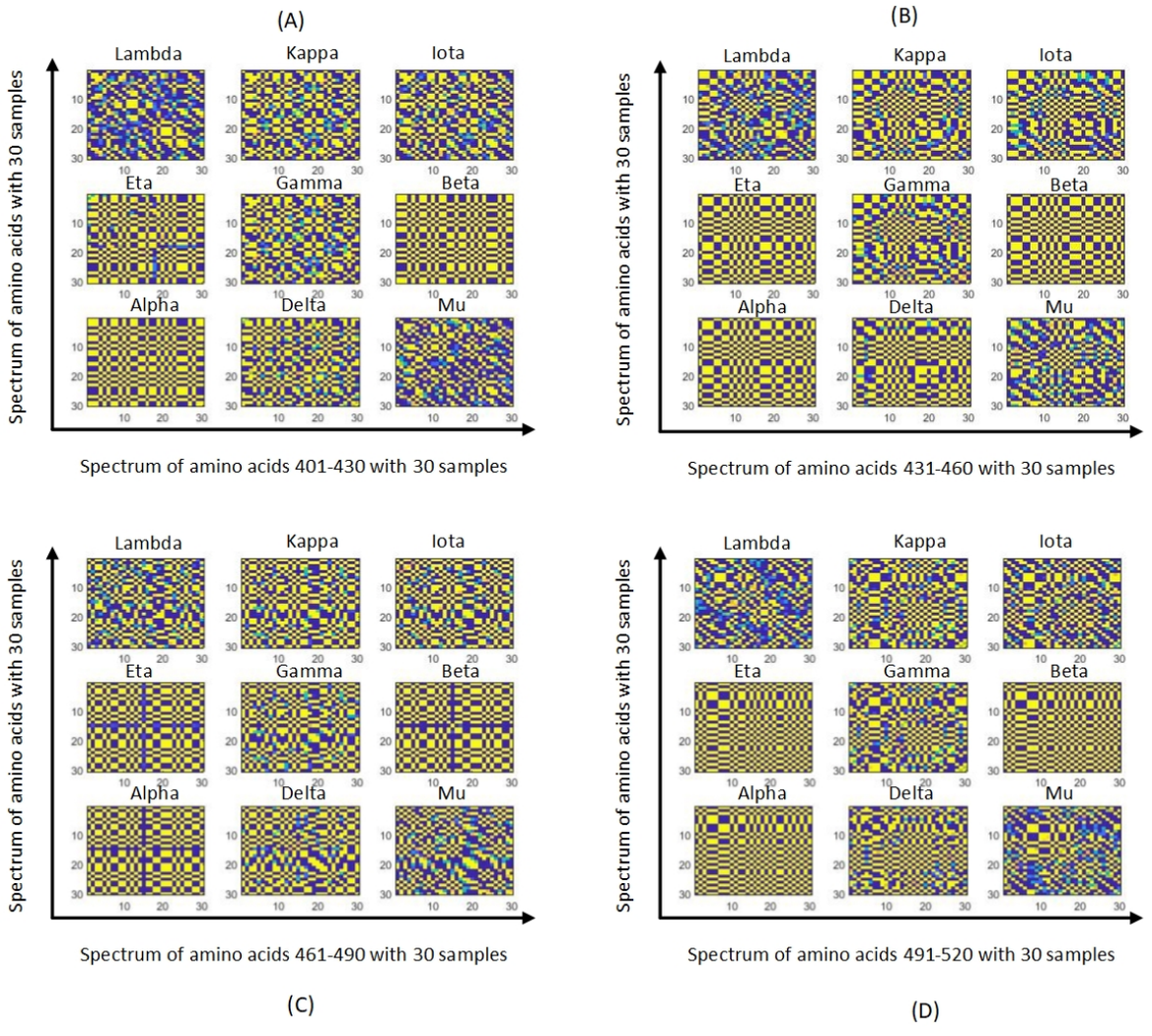

Figure 5. The image of the spectrum of tensor correlation of (ITD-IMF)chain(1) of S protein in nine variants with amino acids between 401 and 520 in panel (A)-(D). The nine spectrum images in each panel from left to right and up to down are Lambda, Kappa, Iota, Eta, Gamma, Beta, Alpha, Delta, and Mu. The blue color of the spectrum means few mutations, and the yellow color represents more mutations. The size of the image is 30 means 30 selected samples. The more irregular pattern in the image means more mutation exists among the 30 samples in each variant.

Discussion

First, we notice from the shape of the histogram of quantum genotype in Figure 3; there is a shift of mean value of the highest peak of the frequency to the more left side of the histogram within one year from 2020 to 2021. The evolutional path implies that the coming new variant of concern in the year 2022 will locate the left shift of the frequency of quantum genotype more than the location of the highest peak of the year 2021. Secondly, in Figure S1 in supporting information, we notice from the shape of scattering points that the evolution of Eta, Iota, and Lambda are closed to each other. Secondly, Mu is more closed to Kappa and Gamma. Finally, in the performance measurement of the prediction of unknown amino acids in nine variants, the average performance is about 80% which is less than the previous study with the alignment method (Ali et al., 2021). On the other hand, the prediction performances were calculated from unknown 120 amino acids in the S protein of nine variants that cover all mutated amino acids in the RBD domain. The shade areas with pink color in panel(A)-(I) on Figure S2 in supporting information indicate the boundary of unmutated amino acids with S-clade. Therefore, we use the result of the separated plane of these boundaries to classify the support vectors for predicting the unknown samples with the assigned value to 1 in the pink area and -1 for

outside. The -1 means the behavior field spin down with the adaptive behavior of mutated amino acids. The other work (Kumar et al., 2021) did not show the separated plane of support vector machine within the area of quantum genotype. Their performance was calculated from the classification of mutation from whole-genome in about 1273 amino acids in S protein between each variant counted from principle component of dimension reduction from 1273 dimensional vector space span by alphabets in 1273 amino acids. For this reason, the result of their prediction will be higher than our result, but no meaning is involved with quantum genotype like in our work.

Conclusion

This paper proposes a novel approach in quantitative measurement of quantum genotype in biology to plot the genetic code of various selected samples of S protein from SARS-CoV-2 variants of concern in new quantities, referred to as the (d, r, p)-layer coordinate system. The result of the quantum genotype computation in this work was performed in comparison with non-mutated S protein and mutated S-protein and observed the evolutional behavior field according to demonstrate the new theoretical investigation approach with empirical for quantitative result measurement of detecting the changing curvature in S protein folding in each amino acid in the new quantum biology approach based on support vector classification. The performance of forecasting of unknown test samples for 120 amino acids between amino acid numbers 401 and 520 was presented in this work with good results. The result of the computation was performed in comparison with non-mutated amino acids in the S clade. This result can be applied to detect and classify mutations between unknown and known variants. The application of this research work can be used to detect the new incoming novel variant of SARS-CoV-2 in the future. If we update the input data, we can expand the size of samples and input the new incoming future variants of concern into account more S protein.

Supporting Information

Lists of the following Tables and Figures are in the supporting Information:

-Tables S1-S6 show the result of the quantum genotype of Alpha and Delta variants.

-Tables S7-S9 show the result of forecasting.

-Figures S1-S2 show the results of scattering plot between 9 variants and results of support vector with hyperplane in support vector classification.

For more information on accession number of time series of genetic code in S protein in this research is available for download at http://sproteins.info

Acknowledgments

Kabin Kanjamapornkul (K.K.) thanks to Chulalongkorn University Graduate Scholarship to Commemorate the 72nd Anniversary of his majesty King Bhumibol Adulyadej, and the 90th Anniversary of Chulalongkorn University Fund (Ratchadaphiseksomphot). K.K., Thanyada Rungrotmongkol (T.R.), and Supot Hannongbua (S.H.) are grateful for computational resources supported by NSTDA Supercomputer Center (ThaiSC) for this work.

References

Ali, S., Sahoo, B., Ullah, N., Zelikovskiy A. and Patterson, M. (2021). A *k*-mer based approach for SARS-CoV-2 variant identification. Bioinformatics Research and Applications. **Lecture Notes in Computer Science**, 13064, 153-16**.**

Andersen, K.G. Rambaut, A., Lipkin W. I., Holmes E.C. and Garry R. F. (2020). The proximal origin of SARS-CoV-2. **Nature Medicine**, 26(4), 450-452.

Anshel, Anshel, M. and Goldfeld, D. (1999). An algebraic method for public-key cryptography. **Mathematical Research Letters**, 6(3), 287-291.

Capozziello, S., Pincak, R., Kanjamapornkul, K. and Saridakis, E.N. (2018). Chern-Simons current in systems of DNA-RNA transcriptions. **Annalen der Physik**, 530(4), 1700271.

Capozziello, S., Pincak, R. and Kanjamapornkul, K. (2017). Anomaly on superspace of time series data. **Zeitschrift fuer Naturforschung A**, 72(12), 1077-1091.

Jones J. E. (1924). On the determination of molecular fields. -II. From the equation of state of a gas. **Proceedings of the Royal Society A**, 106(738), 463-477.

Kanjamapornkul, K., Rongrotmongkol, T. and Hannongbua, S. (2021). Frenet-Serret formulas for moving frame with (d,r,p)-layer in protein folding. **Journal of Science and Science Education**, 4(1), 38-50.

Kanjamapornkul, K. , Pincak R., Chunithpaisan, S. and Bartos, E. (2017). Support spinor machine. **Digital Signal Processing**, 70, 59-72.

Kanjamapornkul, K. and Pincak, R. (2016). Kolmogorov space in time series data. **Mathematical Methods in the Applied Sciences**, 39(15), 4463-4483.

Kumar, S., Thambiraja, TS., Karuppanan, K. and Subramaniam, G. (2021). Omicron and Delta variant of SARS-CoV-2: A comparative computational study of spike protein. **Journal of Medical Virology**, 94(4),1641-1649.

Li, F. (2016). Structure, Function, and Evolution of Coronavirus Spike Proteins. **Annual Review of Virology**, 3, 237-261.

Mlcochova, S., Kemp, S.A., Dhar, M.S.et al. (2021). SARS-CoV-2 B.1.6.17.2 Delta variant replication and immune evasion. **Nature**, 599, 114-119.

Pincak, R., Kanjamapornkul, K. and Bartos, E. (2020a). Cohomology theory for biological time series **Mathematical Methods in the Applied Sciences**, 43(2), 552-579.

Pincak, R., Kanjamapornkul, K. and Bartos, E. (2020b). A theoretical investigation on the predictability of genetic patterns. **Chemical Physics**, 535, 110764.

Pincak, R., Kanjamapornkul, K. and Bartos, E (2019). Forecasting Laurent Polynomial in the Chern-Simons Current of V3 Loop Time Series. **Annalen der Physik**, 531(7),1800375.

Sathipati, S.Y. Shukla, S.K. and Ho, S.Y. (2022). Tracking the amino acid changes of spike proteins across diverse host species of severe acute respiratory syndrome coronavirus 2. **iScience**, 25, 103560.

Schoeman, D. and Fielding, B.C. (2019). Coronavirus envelope protein: Current knowledge. **Virology Journal,** 16(1), 69.