Gene Teams are on the Field: Evaluation of Variants in Gene-Networks Using High Dimensional Modelling

Suha Tuna
Istanbul Technical University
Cagri Gulec
Istanbul University
Emrah Yucesan
Bezmialem Vakif University
Ayse Cirakoglu
Istanbul University-Cerrahpasa
Yelda Tarkan Arguden (yeldata@iuc.edu.tr)
Istanbul University-Cerrahpasa

Research Article

Keywords:

Posted Date: March 15th, 2022

DOI: https://doi.org/10.21203/rs.3.rs-1343509/v1

Gene Teams are on the Field: Evaluation of Variants in Gene-Networks Using High Dimensional Modelling

Suha Tuna[1], Cagri Gulec[2], Emrah Yucesan[3], Ayse Cirakoglu[4], and Yelda Tarkan Arguden[4,*]

[1]Istanbul Technical University, Informatics Institute, Department of Computational Science and Engineering, Istanbul, 34469, Turkey
[2]Istanbul University, Istanbul Faculty of Medicine, Department of Medical Genetics, Istanbul, 34093, Turkey
[3]Bezmialem Vakif University, Faculty of Medicine, Department of Medical Biology, Istanbul, 34093, Turkey
[4]Istanbul University-Cerrahpasa, Cerrahpasa Faculty of Medicine, Department of Medical Biology, Istanbul, 34098, Turkey
[*]yeldata@iuc.edu.tr

ABSTRACT

Variation is a key concept in every biological aspect, particularly in medical genetics. In this field, each genetic variant is evaluated mostly as an independent entity in respect to its clinical importance. This approach may be sufficient to detect the pathogenic variants in single-gene disorders. However, in most of the complex diseases, the combination of the variants in specific gene networks, rather than the presence of a certain single variant predominates. Therefore, while considering a complex disease, the disease status can be evaluated as the success of a team composed of certain variants. To apply this approach, we tested the efficiency of high-dimensional modelling of gene-network restricted variants in distinguishing a disease status. To evaluate the proposed method, we select two gene networks, mTOR and TGF-β. For each pathway, we generate 400 control and 400 patient group samples. The considered mTOR and TGF-β pathways contain 31 and 93 genes of varying sizes, respectively. We performed Chaos Game Representation to each gene sequence to obtain 2-D binary patterns. Produced patterns are ordered successively, and a 3-D tensor structure is achieved for each gene network. Features for each data sample are acquired by exploiting Enhanced Multivariance Products Representation to 3-D data. The features are split as training and testing vectors. The training vectors are employed to train a Support Vector Machines classification model. We manage to achieve more than 96% and 99% classification accuracies for mTOR and TGF-β networks, respectively, using a limited amount of training samples.

Introduction

Recently, in parallel with the development of new technologies in genetics, the human genome can be studied holistically. Previously genes were being evaluated as single entities -we can call those times as "analysis era" of genetics- now the "synthesis era" is born, in which genes are examined as parts of a network consisting of the entire genome. Albert Lazslo Barabasi accounted this situation "disease phenotype is rarely a consequence of an abnormality in a single effector gene product, but reflects various pathobiological processes that interact in a complex network."[1]. In this remarkable concept, genes that encode proteins involved in a pathway or known to be associated with a particular disease are considered a "gene sub-network". Therefore, gene network/s analysis is now more reasonable and comprehensible than examining only single genes or pathways. The importance of this approach is evident in understanding the biogenesis of polygenic-multifactorial diseases that are commonly observed in the population and in which the cumulative effect of many mildly acting genes is determinative. Unlike single gene disorders, in polygenic/multifactorial diseases, there is not a single genetic change (mutation) in a single underlying gene. In addition to environmental factors, the combination of genetic changes called polymorphism or variant plays a role in the emergence of such diseases.

As an analogy, a gene network may be considered as a "team". The success of the team, relies on the efficiency of the metabolic pathway that contain proteins encoded by genes which make up the gene network. "Team success" is directly related to all players, not the one. The performance of any team depends on the harmonious working of its individual players. Individual players of a "gene team" are the specific variants of each one of the genes in the network a person carries. Depending on the efficiency of the variant combination, that individual is either healthy or affected in terms of a specific trait. This combinatorial effect of the genes is the mechanism of penetrance and expressivity. If there is a "marvelous" variant combination

-like a "dream team" of genes- then that person will be superior for that trait. When there are compensative genes in the gene network for a disease-causing mutation, then the mutant gene's deleterious effect can be suppressed, and phenotype appears normal. On the contrary, when many "weak" variants come together in the network, phenotype could be worse than expected from each of these variants. This is already known as one of the mechanisms of emergence of polygenic multifactorial traits.

Therefore, when a gene sub-network is determined, it is desirable to be able to identify the combination of variants in that module. If the differences between the gene network variant combinations among individuals can be determined, to foresee the susceptibility of that individual to the related diseases can be possible. The problem with this approach is that there are not practical techniques to examine a gene sub-network as a team.

Genome-wide association studies (GWAS) techniques are used to detect genomic variants that may be responsible for the predisposition to complex diseases. These studies enable the determination of the most significant variants in terms of the related trait/disease coexistence among the variants commonly found in people with a particular trait or disease. Using GWAS and bioinformatics methods, defining the gene sub-networks (or disease modules) underlying certain traits/diseases is possible. In this early days of "holistic genetics" era, a lot of research focused on this task[1–7].

One of the many application areas of the results obtained from GWAS studies is the prediction of an individual's susceptibility to a certain physical or mental illness based on their genetic profile. Polygenic Risk Score is the standard method used for this purpose, and it relies on the SNPs (Single Nucleotide Polymorphisms) that were determined as risky for that particular illness/trait by GWAS studies. By using the effect sizes determined in the GWAS study as the weights of the SNPs, the weighted total scores of all risk SNPs are calculated. Thus, a person-specific Polygenic Risk Score is determined. Although PRS is a method that can be used as a biomarker to determine individual susceptibility to diseases, there are currently some limitations that make its clinical application difficult. One of these is that GWAS studies are still limited to certain ethnic groups, and sometimes there are groups with different characteristics even within the same population. Another limitation is that many phenotypic traits are affected by too many genes (polygenicity). Besides, there is no consensus on which of the various methods used to calculate PRS is the most appropriate. In particular, the necessity of finding new methods to overcome the polygenicity problem is emphasized[6–14].

Methods such as GWAS are highly effective in identifying variants in genes in a particular disease-associated pathway that are common to most people with the disease. However, these methods are insufficient in determining patient-specific combinations of other variants in pathway genes. Regardless of whether they carry risky variants, clinical differences between individuals with complex diseases are considered to be the result of patient-specific combinations of variants. Papadimitriu et al. report a machine learning approach to identify digenic or bilocus variant combinations[15]. Nevertheless, it is emphasized that "the large number of known variants gives rise to an immense number of combinations, presenting mathematical, statistical, and computational challenges"[16]. Therefore, with the current techniques it is not possible to study the combinatorial effects of more than a few variants, let alone all of them. It is obvious that new approaches are required to overcome the problem.

Here, we propose a high dimensional modelling based method to analyse all the variants in a gene sub-network together, applying Chaos Game Representation (CGR)[17–20] as a pre-processing tool and statistical based high dimensional feature extraction technique named Enhanced Multivariance Products Representation (EMPR)[21–24]. Then, Support Vector Machines (SVM) which is a flexible and efficient classification algorithm[25] was performed in order to sort the gene network of an individual based on her/his sequence variants into control or patient groups. We created exemplary mTOR and TGF-β sub-networks consisting of 31 and 93 genes, respectively, to test our approach.

Approach

The biggest problem in processing variant combinations in gene sub-networks is the amount of sequence data. Therefore, to facilitate analysis, we considered to apply CGR, a technique to convert 1-D sequence data into 2-D pattern form[17–19]. The rationale was that the variants in each sequence data would result in slightly different CGR patterns, and computationally sorting out these pattern differences would be easier than comparing sequences. Afterwards, we had a 2-D pattern in hand for each gene in the sub-network that needed to be examined together as a team. To do that, we aligned each of the CGR patterns in succession to create a cube as a 3-D tensor, which would represent an individual's gene sub-network as a single entity. Then, we adopted EMPR to decompose this 3-D array and represent it in terms of less dimensional features with the aim of distinguishing control and patient groups according to their variant combinations[24].

In order to examine the efficacy and the distinguishing capability of our approach, we generated a data set for two gene sub-networks. These are the mTOR and TGF-β pathways, each of which contains 800 individual 3-D tensors after the application of CGR and aligning the images as a CGR cube. Half of these tensors stand for the control, while the other half denotes the patient groups. We split both groups as training and testing parts. Then, we fed SVM binary classification algorithm with three EMPR vector components of the training data and generated the learning model. Finally, we calculated the overall accuracy by predicting the class (control/patient) of each testing feature according to the constructed SVM model[25].

Methods

Data Source and Recruitment

The mTOR and TGF-β pathway genes were selected based on KEGG database (https://www.genome.jp/kegg/)[26]. Genomic sequences of the pathway genes were fetched from GRCh37 human genome database based on their genomic coordinates recorded in NCBI database (https://www.ncbi.nlm.nih.gov/projects/genome/guide/human/index.shtml).

As represented in Fig. 1, reference sequences composed of each gene sequence were used as a template to generate 400 control and 400 patient sequences for each pathway. At the first step, we created two lists of integers for both groups, that represent the positions of polymorphic and pathogenic variants ('polymorphic positions list'and 'pathogenic positions list'). Each integer in these lists has been randomly chosen to be within certain consecutive intervals and exclusive to the other list. This interval has been set to 100 and 200 for polymorphic and pathogenic variants, respectively (Any integer within the range 1-100, 100-200, 200-300, and so on, for 'polymorphic positions list', and any integer within the range 1-200, 200-400, 400-600, and so on, for 'pathogenic positions list'). In the second step, the reference base at each position represented in 'polymorphic positions list' was replaced by the variant base in 40% of both control and patient sequences. The alterations in these positions were accepted as non-pathogenic and/or common variants with 0.40 minor allele frequency in both groups. In the next step, reference base at each position represented in 'pathogenic positions list'was replaced by the variant base in 25% of control sequences and 30% of patient sequences. The alterations in these positions were accepted as disease-associated/pathogenic variants with 0.25 allele frequency in the control group and 0.30 allele frequency in the patient group. In all these steps, we set minor allele frequency (MAF) higher, because, in contrary to single-gene disorders where rare variants (with MAF$<$ 0.01) are causative, complex disorders are the consequences of the combination of the variants with higher allele frequency (MAF$>$ 0.01). All variant sequences were in haploid state. The properties of the datasets are summarised in Supp. Table 1 and Supp. Table 2.

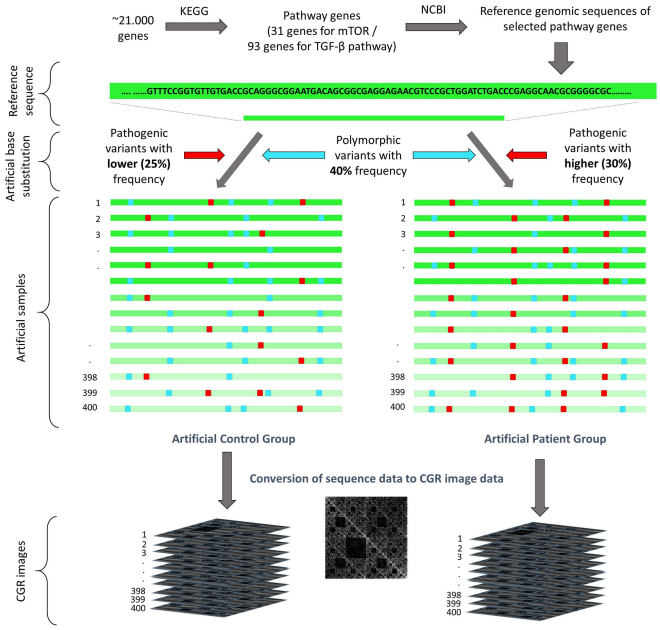

Figure 1. Fetching and pre-processing the genomic sequence data

The known available datasets, e.g., 1000 Genomes, GENESIS, Solve-RD, Munich Exome (EVAdB), Baylor-Hopkins Center for Mendelian Genomics (BH-CMG), 100KGP, GeneDx, and NHLBI-GO Exome Sequencing Project (ESP) databases, have not used preferably to avoid any bias (it is difficult to distinguish patient from control dataset). Therefore, we created datasets that we arranged according to the percentage of the allele frequency. Since real human samples or data were not used in the study, ethics committee approval was not considered necessary.

In order to evaluate the efficiency of the proposed method, both control and patient groups belonging to each pathway dataset were split into two independent and non-intersecting parts. The first part was considered as the training while the latter

was called the testing data. These separate subsets for each pathway dataset were symbolised as D_{train} and D_{test}, respectively. D_{train} were collected by generating randomly selected pathways among 400 control and 400 patient networks at a certain amount. For the classification phase, number of the elements in D_{train} was assumed to be less than the number of networks in D_{test}. D_{train} was utilised to train the classification algorithm, while D_{test} is employed to verify the efficacy of the training model. To provide a convenient learning model and determine whether a given network in D_{test} belongs to the control or the patient class, we applied a new feature extraction approach based on CGR and EMPR.

Chaos Game Representation

CGR is an efficient technique which converts long 1-D genomic sequences into 2-D images (see Fig. 2), say patterns[17–19]. In this manner, CGR enables to pull significant data parts out from the corresponding gene sequence using a convenient feature extraction method suitable for images.

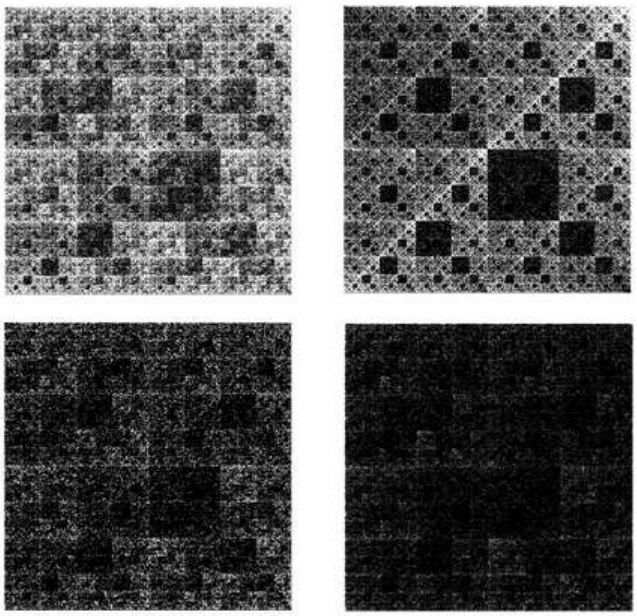

Figure 2. 700×700 CGR images corresponding to four genes in mTOR and TGF-β pathways: RPTOR of mTOR (top-left), GSK3B of mTOR (top-right), SMAD6 of TGF-β (bottom-left), SMAD7 of TGF-β (bottom-right)

In DNA sequence case, the corresponding CGR of a sequence is nothing but a square-shaped binary image whose bottom-left corner overlaps with the origin of 2-D Cartesian space. If Adenine is assumed to be depicted with the origin, which is the point $(0,0)$, Cytosine is placed at the point $(0,1)$, Guanine is located at $(1,0)$ while Thymine stands at the final corner, that is $(1,1)$. The pattern is initialized with a point on the centre of the image, that is $(0.5,0.5)$. The first point of the pattern is settled in the half way between the centre and the corner corresponding to the first nucleotide of the sequence. In general, the i-th point of the image is then placed just in the middle of the $(i-1)$-th point and the vertex corresponding to the i-th nucleotide.

Formally, if the horizontal and the vertical coordinates of the i-th nucleotide of a given sequence are defined as X_i and Y_i, respectively, these entities are determined using the following linear equations

$$\begin{aligned} X_i &= \frac{1}{2}\left(X_{i-1} + C_i^{(x)}\right) \\ Y_i &= \frac{1}{2}\left(Y_{i-1} + C_i^{(y)}\right) \end{aligned} \tag{1}$$

where $X_0 = Y_0 = 0.5$. In eqn. (1), $C_i^{(x)}$ and $C_i^{(y)}$ stand for the coordinates of the pre-defined corners of the unit-square, that is $[0,1]^2$, related to the corresponding nucleotide mentioned above.

The resolution of the CGR image is adjustable and may affect the representation quality of the gene sequence under consideration. For instance, if the size of the CGR image is selected too small, then some of the points can overlap and this fact can prevent the contribution of the overlapping points to the whole pattern. On the other hand, in case the size of the image is selected too large, some unnecessary gaps between the points may occur and the representation eligibility of the CGR pattern is influenced negatively. Thus, fixing the optimal resolution for a CGR image is also crucial to improve the representation quality.

In order to process the pathways under consideration as a whole and extract meaningful features using Enhanced Multivariance Products Representation, all CGR images of the genes in the pathways are aligned in succession. Thus, a 3-D representation for any individual mTOR or TGF-β gene network is constructed. The emerged 3-D data is named as *CGR cube* of a gene network and is suitable for processing by the proposed high dimensional modelling method.

Enhanced Multivariance Products Representation

Enhanced Multivariance Products Representation (EMPR) is a high dimensional data decomposition method[21–24]. It enables a representation of multidimensional data in terms of lower-dimensional entities. Accordingly, EMPR can be considered as a finite series of lower dimensional components. This aspect of EMPR enables to reduce the dimensionality of multidimensional data and simplifies further analysis.

In scientific experiments and applications, one of the crucial challenges in analysing data is the "curse of dimensionality"[27]. Therefore, governing this issue by reducing the number of dimensions becomes critical. Thus, EMPR can be regarded as an adequate technique for addressing multidimensional problems.

EMPR is an extension of a well-known statistical method called High Dimensional Model Representation (HDMR)[28,29]. HDMR was invented for decomposing and decorrelating the inputs in multidimensional input-output systems[28]. In a general multidimensional system, each input, say dimension, contributes the behaviour of the output individually or cooperatively with other inputs[29–31]. However, determining these contributions is significant in order to evaluate the corresponding model for meta-modelling[32,33], sensitivity analysis[34] and reduction[35], etc.

As HDMR, EMPR is capable of dealing with N-D data. But in this study, the 3-D case is considered without loss of generality. However, all formulations which will be presented here can be generalised to N-D case without any difficulty. Further in this section, EMPR for Gene Network Analysis (GNA) will be introduced and discussed.

Let $\mathbf{G}$ denote the 3-D CGR cube and assume that its size is $n_1 \times n_2 \times n_3$. This means the network $\mathbf{G}$ has n_3 gene sequences each of which has various size and is represented through $n_1 \times n_2$ binary images, thanks to CGR method. Then, the EMPR expansion of the CGR cube can be explicitly given as follows

$$\mathbf{G} = g^{(0)} \left[\bigotimes_{r=1}^{3} \mathbf{s}^{(r)} \right] + \sum_{i=1}^{3} \mathbf{g}^{(i)} \otimes \left[\bigotimes_{\substack{r=1 \\ r \neq i}}^{3} \mathbf{s}^{(r)} \right] + \sum_{\substack{i,j=1 \\ i<j}}^{3} \mathbf{g}^{(i,j)} \otimes \left[\bigotimes_{\substack{r=1 \\ r \neq i,j}}^{3} \mathbf{s}^{(r)} \right] + \mathbf{g}^{(1,2,3)}. \tag{2}$$

In formula (2), $g^{(0)}$, $\mathbf{g}^{(i)}$ and $\mathbf{g}^{(i,j)}$ denote the zero-way, the one-way and the two-way *EMPR components*, respectively and $\otimes$ stands for the outer product operation[36]. The 3-D EMPR expansion is a finite sum. Thus, it involves exactly 2^3 EMPR

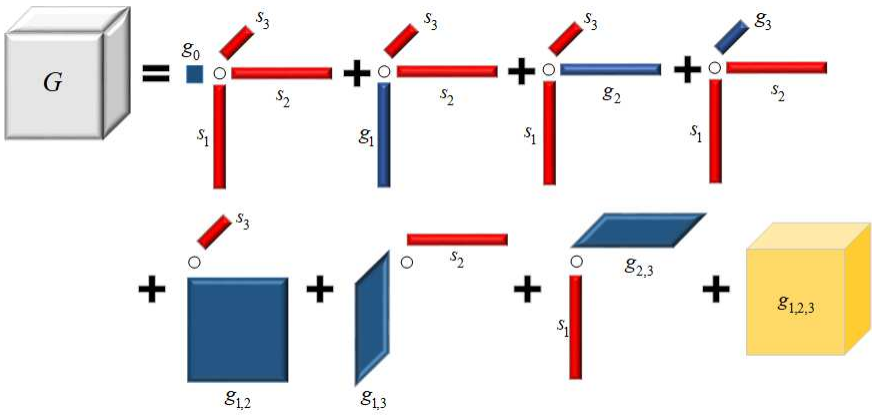

Figure 3. Graphical demonstration of EMPR expansion for 3-D case.

components[21–24]. The graphical expression of the EMPR decomposition is given through Fig. 3.

In (2), $g^{(0)}$ is a scalar that can be considered as a 0-D entity. $\mathbf{g}^{(i)}$ stands for 1-D structures which are the vectors, and $\mathbf{g}^{(i,j)}$ denotes the 2-D entities which can be acknowledged as the matrices. Additionally, other entities involved in (2) and denoted by $\mathbf{s}^{(r)}$ are 1-D elements and called the *support vectors*[24]. In this sense, $\mathbf{s}^{(r)}$ is the r-th support vector which resides on the r-th axis of the 3-D CGR cube where $r = 1, 2, 3$. Thus, one can easily verify that the r-th support vector is an entity composed of n_r elements. Support vectors are multiplied with the corresponding EMPR components in outer product manner and enhance its

dimensionality. Besides, they provide flexibility for EMPR expansion and must be selected rationally. This choice is crucial since it affects the representation eligibility of the EMPR expansion.

Since EMPR has an additive nature, $\mathbf{G}$ should be expressed in terms of 3-D structures. As a consequence of outer products between EMPR components and support vectors, new 3-D but less complicated entities are established. These new elements are called *EMPR terms*[21–24]. Each EMPR term is named regarding its EMPR component. Thus, the term constructed with $g^{(0)}$ and all three supports are called *zeroth EMPR term*. The term composed of $\mathbf{g}^{(i)}$ and the remaining two support vectors (except the i-th one) is called *i-th EMPR term*. Similarly, the term including $\mathbf{g}^{(i,j)}$ and the corresponding support vector are called (i,j)*-th EMPR term*. It is clear that all EMPR terms are of size $n_1 \times n_2 \times n_3$, just as the original data, $\mathbf{G}$.

Additionally, during the EMPR process, three weight vectors can be exploited to adjust the contributions of each CGR pixel in $\mathbf{G}$. The weight vectors are consisted of non-negative real values and must satisfy the following conditions

$$\left\|\boldsymbol{\omega}^{(1)}\right\|_1 = 1, \quad \left\|\boldsymbol{\omega}^{(2)}\right\|_1 = 1, \quad \left\|\boldsymbol{\omega}^{(3)}\right\|_1 = 1. \tag{3}$$

In (3), it is clear that the sum of all elements for each weight vector should be equal to 1. These equations hold due to the statistical necessities and they facilitate the computations in the evaluation process of EMPR components.

However, the EMPR components should satisfy the following constraints

$$\sum_{i_p=1}^{n_p} \boldsymbol{\omega}_{i_p}^{(p)} \mathbf{s}_{i_p}^{(p)} \mathbf{g}_{i_1,\ldots,i_m}^{(1,\ldots,m)} = 0; \quad 1 \leq p \leq m \in \{1,2,3\} \tag{4}$$

where $\mathbf{s}_{i_p}^{(p)}$ and $\boldsymbol{\omega}_{i_p}^{(p)}$ are the i_p-th elements of the p-th support vector and p-th weight vector, respectively. However, $\mathbf{g}_{i_1,\ldots,i_m}^{(1,\ldots,m)}$ stands for the $(i_1,\ldots,i_m)$-th entry of the corresponding EMPR component $\mathbf{g}^{(1,\ldots,m)}$. The equalities in (4) are called *vanishing conditions*. They lead to two essential properties of EMPR components, which are the uniqueness under a certain set of support vectors and the mutual orthogonality.

By employing the vanishing conditions in (4) and adopting the weight vectors given in (3) with the pre-selected support vectors, the scalar EMPR component, i.e. $g^{(0)}$, can be determined uniquely as follows

$$g^{(0)} = \sum_{i=1}^{n_1}\sum_{j=1}^{n_2}\sum_{k=1}^{n_3} \boldsymbol{\omega}_i^{(1)} \boldsymbol{\omega}_j^{(2)} \boldsymbol{\omega}_k^{(3)} \mathbf{s}_i^{(1)} \mathbf{s}_j^{(2)} \mathbf{s}_k^{(3)} \mathbf{G}_{ijk}. \tag{5}$$

It is possible to mark that the right-hand side of the equation (5) denotes a weighted sum of $\mathbf{G}$ multiplied by the relevant support vector elements over all axes. Thus, the zero-way EMPR component associates with a specific weighted average value of the CGR cube, $\mathbf{G}$.

If the conditions in (3) and constraints (4) are exploited again, the elements of three one-way EMPR components are calculated uniquely as follows

$$\begin{aligned}
\mathbf{g}_i^{(1)} &= \sum_{j=1}^{n_2}\sum_{k=1}^{n_3} \boldsymbol{\omega}_j^{(2)} \boldsymbol{\omega}_k^{(3)} \mathbf{s}_j^{(2)} \mathbf{s}_k^{(3)} \mathbf{G}_{ijk} - g^{(0)} \mathbf{s}_i^{(1)}, \\
\mathbf{g}_j^{(2)} &= \sum_{i=1}^{n_1}\sum_{k=1}^{n_3} \boldsymbol{\omega}_i^{(1)} \boldsymbol{\omega}_k^{(3)} \mathbf{s}_i^{(1)} \mathbf{s}_k^{(3)} \mathbf{G}_{ijk} - g^{(0)} \mathbf{s}_j^{(2)}, \\
\mathbf{g}_k^{(3)} &= \sum_{i=1}^{n_1}\sum_{j=1}^{n_2} \boldsymbol{\omega}_i^{(1)} \boldsymbol{\omega}_j^{(2)} \mathbf{s}_i^{(1)} \mathbf{s}_j^{(2)} \mathbf{G}_{ijk} - g^{(0)} \mathbf{s}_k^{(3)}.
\end{aligned} \tag{6}$$

while the rest of the components can be computed in a similar manner.

As addressed, the components $\mathbf{g}^{(1)}$, $\mathbf{g}^{(2)}$ and $\mathbf{g}^{(3)}$ are one-way entities. Therefore, each forms a vector lying on its corresponding axis. According to (5) and (6), $\mathbf{g}^{(1)}$ is obtained by squeezing the CGR cube through its front and upper sides, respectively. $\mathbf{g}^{(2)}$ is obtained by suppressing the CGR cube through its front and right sides. The last vector, that is $\mathbf{g}^{(3)}$ is evaluated by compressing the cube through its upper and right sides. After these suppression steps, the means associating with certain dimensions are procured. Then, the relevant support vector weighted with $g^{(0)}$ is subtracted from the calculated mean. Thus, each one-way EMPR term defines the attitude and individual contribution of the corresponding dimension (axis) to the whole network $\mathbf{G}$. In this sense, $\mathbf{g}^{(1)}$ and $\mathbf{g}^{(2)}$ terms specify both dimensions of the surrogate CGR pattern emerged from $\mathbf{G}$. This CGR pattern is a weighted average of CGR images belonging to all genes in the corresponding sub-network. However, the third one-way EMPR term, $\mathbf{g}^{(3)}$, interprets the interrelation among the CGR images of the genes of the network. Thus, each one-way EMPR term characterizes the $\mathbf{G}$ in its own way and can be exploited as low dimensional features for the 3-D gene network data on the focus.

Finally in this section, we will provide the details about the properties and selection process of the EMPR support vectors. As a beginning, the support vectors should satisfy the following normalization conditions

$$\sum_{i_p=1}^{n_p} \omega_{i_p}^{(p)} \left[\mathbf{s}_{i_p}^{(p)}\right]^2 = 1; \qquad p = 1,2,3. \tag{7}$$

under the given weight vectors. With the help of the conditions in (7), the support vectors can be selected independently from the magnitude. Thus, each support vector indicates the relevant direction where it acts as a weight vector to the contributions which are stored as the elements of EMPR components.

Any suitable set of vectors can be employed as the support vector team for EMPR, as long as they are in harmony with the conditions in (4) and (7). For this reason, the vectors whose elements are given explicitly as

$$\begin{aligned}
\mathscr{S}_i^{(1)} &= \sum_{j=1}^{n_2}\sum_{k=1}^{n_3} \omega_j^{(2)}\,\omega_k^{(3)}\,\mathbf{G}_{ijk},\\
\mathscr{S}_j^{(2)} &= \sum_{i=1}^{n_1}\sum_{k=1}^{n_3} \omega_i^{(1)}\,\omega_k^{(3)}\,\mathbf{G}_{ijk},\\
\mathscr{S}_k^{(3)} &= \sum_{i=1}^{n_1}\sum_{j=1}^{n_2} \omega_i^{(1)}\,\omega_j^{(2)}\,\mathbf{G}_{ijk}.
\end{aligned} \tag{8}$$

can be adapted as the support vectors of an EMPR expansion, after performing normalisation according to (7).

The support vectors in (8) can be calculated in a straightforward manner and exploited in EMPR expansion as long as they do not vanish[21,24]. From (8), it is obvious that each formula denotes a weighted average of the CGR cube $\mathbf{G}$ over all axes but the one direction (axis). Thereby, the equations in (8) indicate averaged directions for the CGR cube. To this end, these support vectors in (8) are called *Averaged Directional Supports (ADS)*[24] and can be encountered in several EMPR applications existing in the scientific literature[21–24]. In this study, the ADS are employed in order to extract features using EMPR. However, the constant weight vectors whose elements are as follows

$$\omega_i^{(1)} = \frac{1}{n_1}, \quad \omega_j^{(2)} = \frac{1}{n_2}, \quad \omega_k^{(3)} = \frac{1}{n_3} \tag{9}$$

will be exploited as the weights in EMPR processes.

In summary, EMPR enables to extract features from 3-D CGR cubes. These features are the vector EMPR components given in (6). The vectors are ensembled to form a long feature vector. Each of these vectors spans all dimensions of the CGR cube under consideration with one accord. Therefore, Support Vector Machines algorithm can be fed with the ensembled feature vectors and an efficient learning model can be constructed.

Support Vector Machines

Determining whether a given gene network belongs to the patient or the control group is the main aim of the present work. Thus, extracting practical and meaningful features and selecting an appropriate classifier which is in harmony with these features are crucial. Since data classification is one of the major challenges in machine learning, a vast of techniques are proposed both for supervised and unsupervised cases. Support Vector Machines (SVM), a flexible supervised classification algorithm, is considered as an effective technique in grouping pre-labeled data[25]. The aim of SVM is to construct a hyperplane whose margins with each cumulated point set (class) are the widest possible. If the collected data points are overlayed separate enough, then it becomes possible to distinguish them into homogeneous groups using a linear hyperplane (or linear kernel). Otherwise, a non-linear kernel should be exploited to obtain a satisfactory classification accuracy. This approach is called *the kernel trick*[37].

The main aim of this study is to determine whether a given gene network belongs to the control or patient group. Thus, we formulate this problem as a binary classification task. In order to classify the data in D_{test}, first, the SVM model should be trained using D_{train}. The elements of D_{train} and D_{test} are CGR cubes defined in subsection 3.2 are 3-D. Thus, it is hard to train the model by feeding SVM with the CGR cubes. To overcome this fact, the SVM algorithm is trained with the vector EMPR components of each CGR cube whose explicit formulae are given in (6). Therefore, a feature vector for each CGR cube is constructed by ensembling the one-way EMPR components corresponding to the CGR cube as follows

$$\mathbf{f} = \left[\mathbf{g}^{(1)^T} \quad \mathbf{g}^{(2)^T} \quad \mathbf{g}^{(3)^T}\right]^T. \tag{10}$$

If the CGR cubes are generated as size of $n_1 \times n_2 \times n_3$, then the length of each feature vector $\mathbf{f}$ becomes $n_1 + n_2 + n_3$. This means the hypersurface created by SVM algorithm lays in $n_1 + n_2 + n_3$ dimensional space. Though, this number may seem

quite large, the features whose distinguishing capabilities are satisfactory may reduce the computation complexity of SVM significantly.

To train the SVM model, **f** features of the CGR cubes in D_{train} are evaluated. Then, the SVM model is trained using these feature vectors. After the training phase, **f** features of the CGR cubes in D_{test} is given to the trained model and the class of each feature which belongs to D_{test} is predicted. Consequently, the statistics for the objective evaluation of the proposed estimator are calculated using the elements of the corresponding confusion matrix obtained in each independent run.

Results

In this section, we will provide the results obtained by assembling CGR, EMPR and SVM for the mTOR and TGF-β gene network datasets. To this end, we performed several computational efforts to emphasise the efficiency of the proposed method. Since the aim of this study is to present an efficient classification method for the gene pathways, the overall accuracy (OA) is considered as the fundamental objective assessment metric. The OA value for each experiment is calculated as follows

$$\text{OA} = \frac{\text{Number of correct predictions}}{\text{Number of testing samples}} \times 100. \tag{11}$$

However, since OA could yield limited information about the classifier performance, we also reported the true negative rate, true positive rate (precision), recall (sensitivity), specificity, and Matthew's Correlation Coefficient (MCC) metrics[38,39]. The reported statistics are the average of 100 independent SVM runs. Before the training stage, all features belonging to the training and the testing set were normalised. In SVM phase, we adopted Radial Basis Function (RBF) kernel as the SVM kernel. To determine the best classifier parameters c and γ which controls the behaviour of the RBF kernel, we performed 5-fold cross-validation and grid search on 9×9 grid $[10^{-4}, 10^{-3}, \ldots, 1, \ldots, 10^{3}, 10^{4}] \times [10^{-4}, 10^{-3}, \ldots, 1, \ldots, 10^{3}, 10^{4}]$. Finally, the model was trained using an SVM algorithm implemented by the LIBSVM package[40]. In Fig. 4, we provided the classification and cross-validation

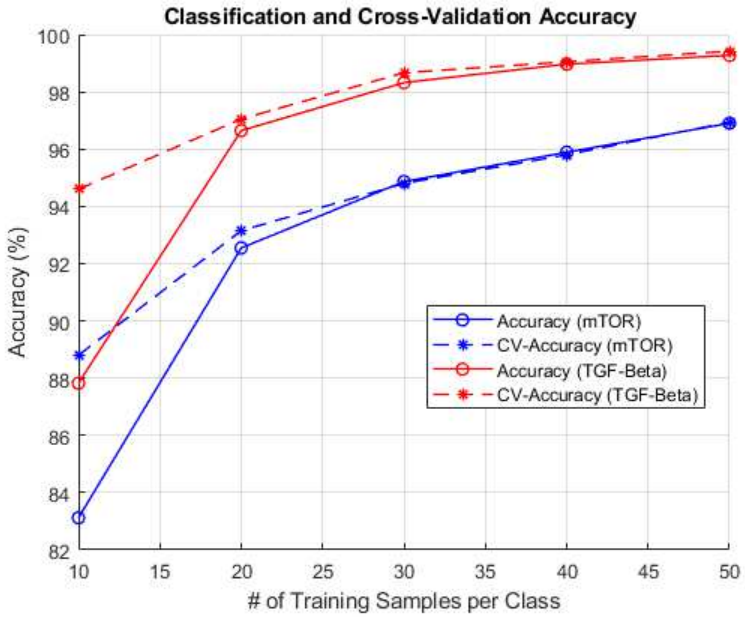

Figure 4. Average overall and cross-validation accuracies for varying training sample counts.

accuracies for both gene pathway datasets. We performed the trials for various training sample amounts both for control and patient groups. These amounts vary between 10 and 50 with an increment of 10. After resolving the number of training samples for each class, the fixed number of training samples were selected randomly. Then, the rest of the networks in each dataset were reserved for testing.

It is clear from Fig. 4 that the proposed method yields higher than 90% classification accuracy using only 20 training samples for both mTOR and TGF-β datasets. Initially, the OA values for mTOR and TGF-β networks are calculated as about 88% and 83% for 10 training samples from both classes, respectively. Then, these values increase to about 97% and 93% rapidly. The increments for both datasets are consistent as the number of training samples from control and patient classes

grows. Furthermore, the cross-validation (CV) accuracies for both datasets tend to escalate while the number of training samples increases and are in harmony with the observed OA results. It is evident that the gap between the corresponding OA and CV accuracy tends to decrease consistently both for mTOR and TGF-β while the training sample count grows, especially after dealing with 20 training samples.

Table 1. Classifier performance metrics for mTOR and TGF-β datasets.

	Metric / S	10	20	30	40	50
mTOR	True Neg. Rate	0.9087	0.9265	0.9398	0.9463	0.9579
	True Pos. Rate	0.8127	0.9315	0.9596	0.9738	0.9813
	Recall	0.9011	0.9227	0.9375	0.9438	0.9565
	Specificity	0.7613	0.9282	0.9598	0.9740	0.9816
	MCC	0.6894	0.8544	0.8984	0.9189	0.9386
TGF-β	True Neg. Rate	0.9608	0.9854	0.9902	0.9949	0.9949
	True Pos. Rate	0.8407	0.9506	0.9770	0.9846	0.9907
	Recall	0.9589	0.9853	0.9902	0.9949	0.9949
	Specificity	0.7945	0.9476	0.9763	0.9844	0.9906
	MCC	0.7765	0.9344	0.9668	0.9794	0.9855

After discussing the classification accuracy of the suggested method, we also need to evaluate the performance and stability of the proposed estimator based on CGR, EMPR, and SVM. To this end, the widely used machine learning metrics for the estimator assessment, such as true negative rate, true positive rate (precision), recall (sensitivity), specificity and MCC were provided in Table 1. In Table 1, the specified metrics are tabulated for increasing training sample counts from control and patient classes for both mTOR and TGF-β datasets.

It is obvious from Table 1 that each metric approaches to value 1 consistently as the number of training samples grows. However, the True Positive Rate, specificity, and MCC values may be considered a bit low at 10 training samples from control and patient classes for both datasets. Nevertheless, these values increase rapidly both for mTOR and TGF-β as 20 or more training samples were employed. We can easily verify from Table 1 that all stability metrics are calculated above 0.93 and 0.98 for mTOR and TGF-β datasets, respectively, by exploiting 50 training samples from both control and patient groups. The reported values address that the proposed estimator achieves significant success in accurately classifying the networks belonging to control and patient samples for the considered mTOR and TGF-β datasets. As the further assessment of the proposed CGR, EMPR, and SVM assemble, receiver operating characteristic (ROC) curves for both datasets are presented in Fig. 5 and 6, where the corresponding area under curve (AUC) values are provided therein. In Fig. 5 and 6, the dashed line demonstrates the random classifier, which can be evaluated as the worst case. In Fig. 5, five ROC curves for 10, 20, 30, 40 and 50 mTOR training samples were presented. On the other hand, for TGF-β dataset in Fig. 6, the ROC curves were plotted for only 10, 20 and 30 training samples, since the improvements in the results for higher training sample counts are not significant. One can easily observe from Fig. 5 and 6 that the AUC values increase consistently while the number of training samples grows for both datasets.

Discussion

In this new age of "holistic genetics", most efforts so far have been devoted to identifying specific gene sub-networks[2–5,41,42]. The attempts to study the behaviour of these variants in their context are still few and timid because of the technical difficulties of handling the vast amount of variants between individuals[15,16]. To the best of our knowledge, our proposed approach to decipher the outcomes of gene networks based on specific combinations of all the variants in the module is original and unique.

Today, one of the biggest challenges in genetics is to handle the enormous amount of data acquired. It seems almost impossible to interpret the impact and importance of the millions of variants obtained in a single Next Generation Sequencing study. Therefore, focusing on data in terms of patterns and corresponding less dimensional entities seems more rational than working on the particular variants. For this purpose, all gene sequences of the selected gene networks were converted into 2-D binary image patterns using CGR. Then, these patterns were aligned (as three-dimensional tensor) in succession, creating a cube. Afterwards, these tensors were decomposed and represented in terms of their less dimensional features using EMPR. Finally, SVM, which is a multi-class classification algorithm, was fed with three EMPR vector features for each network. Our findings revealed an accuracy higher than 96% employing only 50 training features out of 400 data samples from both control and patient groups. The AUC results indicate that the proposed classifier performance in distinguishing between two

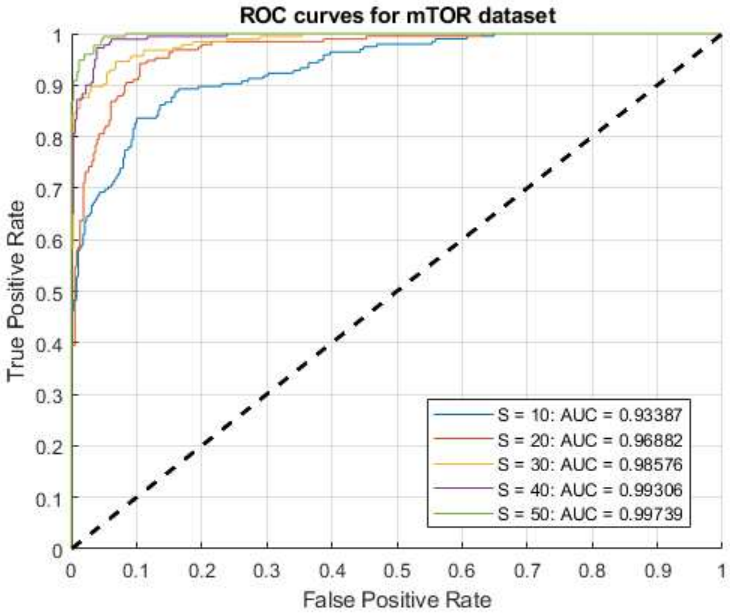

Figure 5. ROC curves and AUC values for mTOR dataset with varying training sample counts.

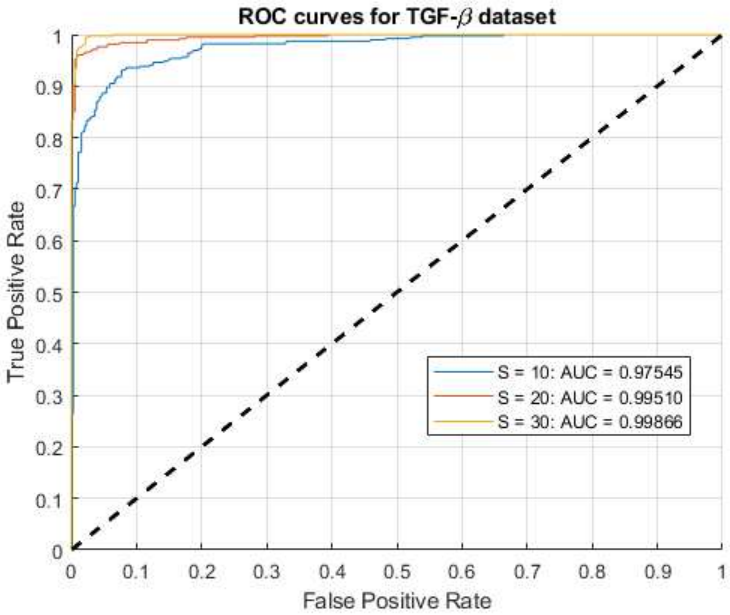

Figure 6. ROC curves and AUC values for TGF-β dataset with varying training sample counts.

classes is admirable. Consequently, our results indicate that the proposed CGR, EMPR and SVM ensemble provides efficient classification performance.

One of the strengths of our approach is its capability of handling data of various sizes. It is independent from the length of the sequences and the number of genes in the networks. It can be easily applied to all gene networks and is an easy to implement algorithm as well. To demonstrate these capabilities, we conducted our experiments on two different gene sub-network datasets, mTOR and TGF-β. These sub-networks contain 31 and 93 genes, respectively, and each gene is of different size. According to our observations, the proposed approach provides accurate and stable results adopting limited training samples.

Since human has a diploid genome and each variant in human genome has a zygosity (homozygous, heterozygous or

hemizygous) state, this method could be considered as challenging for human variant data. In addition, there are two positional possibilities (cis and trans) regarding to any two variants at heterozygous state. However, for a gene network, positional state of the variants in the network-genes is not important, because each gene works as a separate unit of the network. Therefore, positional state of the variants between different genes is not a limitation for our method. On the other hand, positional state of the variants within same gene may become a limitation, because each one of two heterozygous variants in a gene may be located in same or different protein molecule. To overcome this limitation, our method may require some modifications to be applied in diploid case. These modifications may include the representation of each base substitution with IUPAC codes (R for A/G, S for C/G, W for A/T, M for A/C, for instance) as additional features or properties to CGR rules. Considering these additional features, CGR process may be updated. Thus, a 4-D sample space may occur and the orthogonality of the sample space is preserved. Finally, EMPR, which is suitable for N-D structures may be implemented to extract the features of the network under consideration.

In this study, we examined our approach on two datasets prepared on reference sequences of the mTOR and TGF-β gene sub-networks. The observations and findings in this study encourage us that our approach has the potential to be a diagnostic tool as well as determine individual disposition to polygenic multifactorial conditions. In addition, comparative studies may be conducted as an evolutionary perspective. This study may also be adapted to different scientific fields e.g. population genetics, phylogenetics, advanced genomics studies etc. Furthermore, provided that the necessary fieldwork is done, this method can also be used in talent determination, thus providing the opportunity to receive appropriate training from an early age.

Conclusion

According to our results and observations, using high dimensional computational modelling for gene network and network-specific gene variant analyses in a holistic manner seem rational and reliable. Our promising results encourage us to perform the proposed approach on diploid sequence data for more comprehensive future studies.

Data availability

The datasets generated and/or analysed during the current study are available on figshare platform with DOI: 10.6084/-m9.figshare.19312214

References

1. Barabási, A.-L., Gulbahce, N. & Loscalzo, J. Network medicine: a network-based approach to human disease. *Nat. reviews genetics* **12**, 56–68 (2011).
2. Choobdar, S. *et al.* Assessment of network module identification across complex diseases. *Nat. methods* **16**, 843–852 (2019).
3. Hawe, J. S., Theis, F. J. & Heinig, M. Inferring interaction networks from multi-omics data. *Front. genetics* **10**, 535 (2019).
4. Maiorino, E. *et al.* Discovering the genes mediating the interactions between chronic respiratory diseases in the human interactome. *Nat. communications* **11**, 1–14 (2020).
5. Menche, J. *et al.* Uncovering disease-disease relationships through the incomplete interactome. *Science* **347** (2015).
6. Fang, G. *et al.* Discovering genetic interactions bridging pathways in genome-wide association studies. *Nat. communications* **10**, 1–18 (2019).
7. Wang, Y. R. & Huang, H. Review on statistical methods for gene network reconstruction using expression data. *J. theoretical biology* **362**, 53–61 (2014).
8. Uffelmann, E. *et al.* Genome-wide association studies. *Nat. Rev. Methods Primers* **1**, 1–21 (2021).
9. Konuma, T. & Okada, Y. Statistical genetics and polygenic risk score for precision medicine. *Inflamm. regeneration* **41**, 1–5 (2021).
10. Khera, A. V. *et al.* Genome-wide polygenic scores for common diseases identify individuals with risk equivalent to monogenic mutations. *Nat. genetics* **50**, 1219–1224 (2018).
11. Silberstein, M., Nesbit, N., Cai, J. & Lee, P. H. Pathway analysis for genome-wide genetic variation data: Analytic principles, latest developments, and new opportunities. *J. Genet. Genomics* **48**, 173–183 (2021).
12. Weng, L. *et al.* Snp-based pathway enrichment analysis for genome-wide association studies. *BMC bioinformatics* **12**, 1–9 (2011).

13. Xie, X., Kendzior, M. C., Ge, X., Mainzer, L. S. & Sinha, S. Varsan: associating pathways with a set of genomic variants using network analysis. *Nucleic acids research* **49**, 8471–8487 (2021).

14. Zhu, X. & Stephens, M. Large-scale genome-wide enrichment analyses identify new trait-associated genes and pathways across 31 human phenotypes. *Nat. communications* **9**, 1–14 (2018).

15. Papadimitriou, S. *et al.* Predicting disease-causing variant combinations. *Proc. Natl. Acad. Sci.* **116**, 11878–11887 (2019).

16. Mellerup, E. & Møller, G. L. Combinations of genetic variants occurring exclusively in patients. *Comput. Struct. Biotechnol. J.* **15**, 286–289 (2017).

17. Jeffrey, H. J. Chaos game representation of gene structure. *Nucleic acids research* **18**, 2163–2170 (1990).

18. Hoang, T., Yin, C. & Yau, S. S.-T. Numerical encoding of dna sequences by chaos game representation with application in similarity comparison. *Genomics* **108**, 134–142 (2016).

19. Kania, A. & Sarapata, K. The robustness of the chaos game representation to mutations and its application in free-alignment methods. *Genomics* (2021).

20. Yang, W.-F., Yu, Z.-G. & Anh, V. Whole genome/proteome based phylogeny reconstruction for prokaryotes using higher order markov model and chaos game representation. *Mol. Phylogenetics Evol.* **96**, 102–111 (2016).

21. Tunga, B. & Demiralp, M. The influence of the support functions on the quality of enhanced multivariance product representation. *J. Math. Chem.* **48**, 827–840, DOI: 10.1007/s10910-010-9714-2 (2010).

22. Tunga, M. A. & Demiralp, M. A novel method for multivariate data modelling: Piecewise generalized EMPR. *J. Math. Chem.* **51**, 2654–2667, DOI: 10.1007/s10910-013-0228-6 (2013).

23. Tuna, S. & Tunga, B. A novel piecewise multivariate function approximation method via universal matrix representation. *J. Math. Chem.* **51**, 1784–1801, DOI: 10.1007/s10910-013-0179-y (2013).

24. Tuna, S. *et al.* Iterative enhanced multivariance products representation for effective compression of hyperspectral images. *IEEE Transactions on Geosci. Remote. Sens.* DOI: 10.1109/TGRS.2020.3031016 (2020).

25. Cristianini, N. & Ricci, E. *Support Vector Machines*, 928–932 (Springer US, Boston, MA, 2008).

26. KEGG. KEGG website (2021). Accessed: 2021-02-20 https://www.kegg.jp/dbget-bin/www_bget?hsa04150.

27. Gualberto, E. S., De Sousa, R. T., Thiago, P. D. B., Da Costa, J. P. C. & Duque, C. G. From feature engineering and topics models to enhanced prediction rates in phishing detection. *Ieee Access* **8**, 76368–76385 (2020).

28. Sobol, I. Sensitivity estimates for nonlinear mathematical models. *Math. Model. Comput. Exp* **1**, 407–414 (1993).

29. Rabitz, H. & Aliş, Ö. F. General foundations of high-dimensional model representations. *J. Math. Chem.* **25**, 197–233 (1999).

30. Alış, Ö. F. & Rabitz, H. Efficient implementation of high dimensional model representations. *J. Math. Chem.* **29**, 127–142 (2001).

31. Rabitz, H., Aliş, Ö. F., Shorter, J. & Shim, K. Efficient input—output model representations. *Comput. physics communications* **117**, 11–20 (1999).

32. Ayres, D. & Eaton, M. Uncertainty quantification in nuclear criticality modelling using a high dimensional model representation. *Annals Nucl. Energy* **80**, 379–402 (2015).

33. Kubicek, M., Minisci, E. & Cisternino, M. High dimensional sensitivity analysis using surrogate modeling and high dimensional model representation. *Int. J. for Uncertain. Quantification* **5** (2015).

34. Liu, Y., Hussaini, M. Y. & Ökten, G. Global sensitivity analysis for the rothermel model based on high-dimensional model representation. *Can. J. For. Res.* **45**, 1474–1479 (2015).

35. Chowdhury, R., Rao, B. & Prasad, A. M. High-dimensional model representation for structural reliability analysis. *Commun. Numer. Methods Eng.* **25**, 301–337 (2009).

36. Kolda, T. G. & Bader, B. W. Tensor decompositions and applications. *SIAM review* **51**, 455–500 (2009).

37. Dagher, I. Quadratic kernel-free non-linear support vector machine. *J. Glob. Optim.* **41**, 15–30 (2008).

38. Carvalho, D. V., Pereira, E. M. & Cardoso, J. S. Machine learning interpretability: A survey on methods and metrics. *Electronics* **8**, 832 (2019).

39. Chicco, D. & Jurman, G. The advantages of the matthews correlation coefficient (mcc) over f1 score and accuracy in binary classification evaluation. *BMC genomics* **21**, 1–13 (2020).

40. Chang, C.-C. & Lin, C.-J. LIBSVM: A library for support vector machines. *ACM Transactions on Intell. Syst. Technol.* **2**, 27:1–27:27 (2011). Software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm.

41. Almeida, J. S., Carrico, J. A., Maretzek, A., Noble, P. A. & Fletcher, M. Analysis of genomic sequences by chaos game representation. *Bioinformatics* **17**, 429–437 (2001).

42. Cui, H., Srinivasan, S. & Korkin, D. Enriching human interactome with functional mutations to detect high-impact network modules underlying complex diseases. *Genes* **10**, 933 (2019).

Author contributions statement

The study was designed by all authors. S.T. performed the computational tasks and wrote the method, implementation and results sections. C.G. generated the data and wrote the data source and recruitment part. C.G., E.Y., A.C. and Y.T.A. wrote the rest of the manuscript, provided the bibliography and added the comments in discussion. All authors read and approved the final version of the manuscript.

Competing interests statement

Authors declare that there is no competing interest.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- SupplementaryTable1.xlsx
- SupplementaryTable2.xlsx