


Article

The Resolved Mutual Information Function as a Structural Fingerprint of Biomolecular Sequences for Interpretable Machine Learning Classifiers

Katrin Sophie Bohnsack ^{1,*} , Marika Kaden ¹ , Julia Abel ¹ , Sascha Saralajew ²  and Thomas Villmann ^{1,*} 

¹ Saxon Institute for Computational Intelligence and Machine Learning, University of Applied Sciences Mittweida, 09648 Mittweida, Germany; kaden1@hs-mittweida.de (M.K.); abel@hs-mittweida.de (J.A.)

² Bosch Center for Artificial Intelligence, 71272 Renningen, Germany; sascha.saralajew@de.bosch.com

* Correspondence: bohnsack1@hs-mittweida.de (K.S.B.); villmann@hs-mittweida.de (T.V.)

Abstract: In the present article we propose the application of variants of the mutual information function as characteristic fingerprints of biomolecular sequences for classification analysis. In particular, we consider the resolved mutual information functions based on Shannon-, Rényi-, and Tsallis-entropy. In combination with interpretable machine learning classifier models based on generalized learning vector quantization, a powerful methodology for sequence classification is achieved which allows substantial knowledge extraction in addition to the high classification ability due to the model-inherent robustness. Any potential (slightly) inferior performance of the used classifier is compensated by the additional knowledge provided by interpretable models. This knowledge may assist the user in the analysis and understanding of the used data and considered task. After theoretical justification of the concepts, we demonstrate the approach for various example data sets covering different areas in biomolecular sequence analysis.

Keywords: mutual information; sequence analysis; classification; machine learning; interpretable models



Citation: Bohnsack, K.S.; Kaden, M.; Abel, J.; Saralajew, S.; Villmann, T. The Resolved Mutual Information Function as a Structural Fingerprint of Biomolecular Sequences for Interpretable Machine Learning Classifiers. *Entropy* **2021**, *23*, 1357. <https://doi.org/10.3390/e23101357>

Academic Editor: Ting Hu

Received: 19 September 2021

Accepted: 14 October 2021

Published: 17 October 2021

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Copyright: © 2021 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

The accumulation of information based on physical organization processes like structure generation and self-organization belongs to the key aspects of living systems [1–4]. Thus, information theoretic concepts play an important role in sequence analysis for understanding biomolecular entities like RNA, DNA, and proteins to explain biological systems [5–7]. In case of DNA/RNA, the biomolecular information is coded by the nucleotide sequence, particularly their sequence element's frequencies, correlations, and other topological features. The extensive influence of information theoretic concepts and applications in the fields of computational, molecular, and systems biology is captured in various reviews [8–11].

The study of sequences in consideration of their biological properties is still crucial for such diverse applications as drug design, phylogenetic analyses, prediction of molecular interactions, identification of polymorphisms or definition of pathogenic mutations [12–15]. With the availability of powerful machine learning methods like deep and convolutional networks [16,17], and support vector machines [18] as well as the supporting hardware (graphic processing units—GPU), self-learning procedures have entered (and revolutionized) many of these areas of biomolecular research [19–23]. Although these models provide promising performance by automated training and outperform many statistical approaches, the disadvantage is their general “black-box” behavior, i.e., the model decisions are usually hardly interpretable. Thus, explanations are at least difficult to give and usually require additional tools [24]. However, current focus is given to develop interpretable machine learning models instead [25,26]. According to [27], interpretable models are *designed* to be

interpretable in contrast to explainable methods, which can be comprehended post-hoc by experts in the field using additional tools and elaborate considerations. Generally, interpretability increases the trustworthiness of the machine learning method and hence contributes to making them transparent for the applicants. However, interpretable models require meaningful features describing the objects to be considered, ideally taking domain knowledge into account.

It is precisely this identification or rather the design of problem-adequate features that is the subject of research in the field of alignment-free sequence comparison in computational biology. By overcoming some of the major disadvantages of alignments, such as strong evolutionary assumptions [28], high computational costs [29] as well as non-numerical sequence representation, alignment-free methods evolved as a true alternative for quantifying sequence (dis-)similarity [30]. At present, respective methods are used in the domains of phylogenetics [31–33], (meta-)genomics [34,35], database similarity search [36], or next-generation sequencing data analyses [37–39].

In particular, information theoretic and statistical quantities provide a natural way to generate unique signatures or fingerprints of molecular sequence data by considering the distribution of nucleotides as elements of sequence as well as their statistical correlations [40–42]. Long-range correlations in sequences are well-known and intensively studied in alignment-free sequence comparison [43–46]. A promising statistical descriptor approach for sequences is the concept of natural vectors. It considers the moments of the distribution of the nucleotides in a sequence as the determining quantities to characterize the molecular sequence [47]. The equality of all statistical moments of two sequences implies the equality of statistical distributions and, hence, can be seen as an equivalence relation. Natural vectors were successfully applied for DNA-analysis, virus, and protein sequence classification [48,49].

The use of so-called mutual information functions (MIFs) as an alternative to correlation profiles as sequence descriptors was first investigated in [50]. This idea was reconsidered in [51] and renewed in [52,53]. In bioinformatics, this concept was established as average mutual information profile (AMI-profile) and proposed to serve as a genomic signature [54]. Molecular descriptors based on the mutual information are considered in [55]. Further applications of MIF in computational biology involve its use for species identification from DNA sequences [56], for finding (species-independent) patterns that differ in coding and non-coding DNA [57] as well as for investigating co-variation of mutations in protein sequences [58]. To make the idea accessible to a larger audience, a user interface program for MIF calculation is provided in [59] and more applications are reviewed in [60].

The mutual information is known as a similarity measure between distributions, which originally is based on the Shannon entropy [61]. It implicitly takes all correlation moments of the distributions for the comparison into account. Popular alternatives to the standard mutual information are the Rényi and the Tsallis mutual information, which are based on the Rényi and the Tsallis entropy, respectively [62–65]. The numerical estimations of these mutual information variants seem to be more robust than the estimations for the Shannon original [66].

These mutual information concepts can be used to generate information theoretic features for sequence analysis: Rényi entropic profiles were considered for DNA classification problems based on chaos game representation [67,68]. Molecular descriptors based on the Rényi entropy were investigated in [69], whereas long range correlation using Tsallis mutual information was considered in [70]. However, to our best knowledge, MIF for these variants are not known so far.

Furthermore, in [71], it is criticized that the (Shannon) AMI profile, i.e., the MIF, suffers from an average effect of 16 kinds of base correlations. Therefore, the authors of this study proposed, based on the earlier work in [40], a partial information correlation.

This critic together with the above mentioned robustness observations for the Rényi and the Tsallis entropy estimators motivated our investigations: first, we introduce a

resolved variant of the Shannon-based MIF as a more adequate information theoretic signature of molecular sequences reducing the average effects. Afterwards, we transfer this concept to both the Rényi and the Tsallis variants obtaining respective (resolved) mutual information functions. The resulting signature vectors serve as data descriptors for sequence classification problems to be tackled by machine learning methods. In this machine learning part, we focus on dissimilarity based and interpretable classifier models according to the above discussion about interpretability. Particularly, we apply a variant of learning vector quantization which delivers feature correlation information regarding the classification problem as an additional information beyond the classifier predicting performance [72]. Furthermore, this method is known to be robust and optimizing the class separating hypothesis margin [73].

The paper is structured as follows: first, we introduce variants of the mutual information functions for the Shannon-, the Rényi-, and the Tsallis-entropy and give theoretical justifications. Second, we describe the interpretable machine learning classifier based on learning vector quantization and show how knowledge about the decision process and the regarding data properties can be extracted. Thereafter, we apply this methodology for three biomolecular sequence data sets covering different application areas. For this purpose, we describe in detail the feature generation and the parameter setting. Furthermore, we show in the example for one data set how knowledge extraction from the trained classifier model is done to provide useful additional information. Concluding remarks and outlook for future work complete the paper.

2. Variants of Mutual Information Functions as Biomolecular Sequence Signatures

In the following section, we introduce the concept of variants of mutual information functions, which later serve as determining fingerprints of nucleotide sequences. These functions reflect structural characteristics and spatial relations within the sequences. For this purpose, we consider several types of mutual information regarding different entropy concepts. Thereby, we concentrate on those approaches, which are frequently used in machine learning. For a general overview of entropies, divergences, and mutual information, we refer to [74].

2.1. The Resolved Mutual Information Function Based on the Shannon Entropy

We consider the Shannon entropy

$$H(X) = \int_{\mathcal{X}} p(x) \cdot \log\left(\frac{1}{p(x)}\right) dx \quad (1)$$

of a random quantity $X \subseteq \mathcal{X}$ with the density measure $p(x)$ being the expectation value of the information $\log\left(\frac{1}{p(x)}\right)$. In the machine learning context here, we interpret X as a feature or object quantity. The maximum value of the entropy $H(X)$ is obtained for a uniform density $p(x)$ and, hence, $H(X)$ serves as a measure of uncertainty [75].

The corresponding divergence is the Kullback–Leibler-divergence

$$D_{KL}(p(x) \parallel p(y)) = \int_{\mathcal{X}} \int_{\mathcal{Y}} p(x) \cdot \log\left(\frac{p(x)}{p(y)}\right) dy dx \quad (2)$$

as dissimilarity measure between the densities $p(x)$ and $p(y)$ [61,76]. The corresponding mutual information is

$$I(X, Y) = D_{KL}(p(x, y) \parallel p(x) \cdot p(y)) \quad (3)$$

quantifying the joint information of $p(x)$ and $p(y)$. Here, $p(x, y)$ is the joint density. Alternatively, the mutual information can be written as

$$I(X, Y) = H(X) - H(X|Y) \quad (4)$$

using the conditional entropy $H(X|Y)$ which can be written as

$$\begin{aligned}
 H(X|Y) &= H(X, Y) - H(Y) \\
 &= - \int_{\mathcal{X}} \int_{\mathcal{Y}} p(x, y) \cdot \log \left(\frac{p(x, y)}{p(y)} \right) dy dx
 \end{aligned} \tag{5}$$

known as the chain rule of the entropies [61,76]. Equivalently, the mutual information can be formulated as the difference between the sum of the marginal entropies and the joint entropy, i.e.,

$$I(X, Y) = H(X) + H(Y) - H(X, Y) \tag{6}$$

is valid.

We can rewrite the divergence formulation of the mutual information $I(X, Y)$ from Equation (3) as

$$I(X, Y) = \int_{\mathcal{X}} F(x, Y) dx$$

where

$$F(x, Y) = \int_{\mathcal{Y}} p(x, y) \cdot \log \left(\frac{p(x, y)}{p(x) \cdot p(y)} \right) dy \tag{7}$$

describes a mutual information relation of a particular object (feature) x with respect to the random quantity Y . We denote $F(x, Y)$ as the (feature) resolved mutual information (rMI).

The mutual information for sequences $X(t)$ and $Y(t + \tau)$ at time (position) t with shift $\tau \geq 0$ is defined as

$$I(X(t), Y(t + \tau)) = \int_{\mathcal{X}} \int_{\mathcal{Y}} p(x(t), y(t + \tau)) \cdot \log \left(\frac{p(x(t), y(t + \tau))}{p(x(t)) \cdot p(y(t + \tau))} \right) dy dx \tag{8}$$

which yields by setting $Y(t + \tau) = X(t + \tau)$

$$I(X(t), X(t + \tau)) = \int_{\mathcal{X}} \int_{\mathcal{X}} p(x(t), x(t + \tau)) \cdot \log \left(\frac{p(x(t), x(t + \tau))}{p(x(t)) \cdot p(x(t + \tau))} \right) dx(t + \tau) dx$$

as the auto mutual information at time/position t with shift (delay) τ [77,78]. If $p(x(t))$ is independent from t , only the joint probability $p(x(t), x(t + \tau))$ remains t -dependent or, more precisely, it becomes dependent only on the shift τ such that we simply write $p(x, x(\tau))$ for this. Thus, the auto mutual information in dependence on the shift τ is obtained as

$$I(X, \tau) = \int_{\mathcal{X}} \int_{\mathcal{X}} p(x, x(\tau)) \cdot \log \left(\frac{p(x, x(\tau))}{p(x) \cdot p(x(\tau))} \right) dx(\tau) dx \tag{9}$$

as an information theoretic analogous to the auto-correlation function. In [50,79], this shift-dependent auto mutual information is denoted as the mutual information function (MIF) $F(X, \tau) = I(X, \tau)$. Adapting the rMI from Equation (7) to the auto mutual information $I(X, \tau)$ results in the function

$$\begin{aligned}
 F(x, \tau) &= \int_{\mathcal{X}} p(x, x(\tau)) \cdot \log \left(\frac{p(x, x(\tau))}{p(x) \cdot p(x(\tau))} \right) dx(\tau) \\
 &= \int_{\mathcal{X}} p(x, x(\tau)) \cdot \log \left(\frac{p(x, x(\tau))}{p(x(\tau))} \right) dx(\tau) - p(x) \cdot \log(p(x))
 \end{aligned} \tag{10}$$

which can be seen as a quantity characterizing the inherent correlations of the sequence values $x(t)$. We denote $F(x, \tau)$ as the (feature) resolved mutual information function (rMIF), which trivially fulfills

$$I(X, \tau) = \int_{\mathcal{X}} F(x, \tau) dx \quad (11)$$

according to its definition. Note, more precisely would be the notation $F(X, x, \tau)$. We drop the dependency on X for better readability. For (finite) discrete distributions, it becomes simply a matrix \mathbf{F} and $I(X, \tau)$ constitutes a vector. Hence, we can compare those vectors in terms of respective norms, e.g., by the Euclidean norm for vectors or the corresponding Frobenius norm for matrices [80,81].

2.2. Rényi α -Entropy and Related Mutual Information Functions

The Rényi-entropy

$$H_{\alpha}^{\text{R}}(X) = \frac{1}{1-\alpha} \log \left(\int_{\mathcal{X}} (p(x))^{\alpha} dx \right) \quad (12)$$

is a generalization of the Shannon-entropy, where $\alpha > 0$ and $\alpha \neq 1$ is a parameter [62]. Depending on the context, it is also denoted as α -entropy. In the limit $\alpha \rightarrow 1$, the Shannon entropy is obtained. The corresponding Rényi-divergence is

$$D_{\alpha}^{\text{R}}(p(x) \parallel p(y)) = \frac{1}{\alpha-1} \log \left(\int_{\mathcal{X}} \int_{\mathcal{Y}} \frac{(p(x))^{\alpha}}{(p(y))^{\alpha-1}} dy dx \right) \quad (13)$$

with the limit $\lim_{\alpha \rightarrow 1} D_{\alpha}^{\text{R}}(p(x) \parallel p(y)) = D_{\text{KL}}(p(x) \parallel p(y))$ being valid, such that the α -dependent Rényi-mutual-information (RMI) is defined as

$$I_{\alpha}^{\text{R}}(X, Y) = D_{\alpha}^{\text{R}}(p(x, y) \parallel p(x) \cdot p(y)) \quad (14)$$

analogous to the Shannon case (3). This mutual information is widely applied in data analysis and pattern recognition as well as in information theoretic machine learning [82–90]. Unfortunately, a relation comparable to (6) does not hold, i.e.,

$$I_{\alpha}^{\text{R}}(X, Y) \neq H_{\alpha}^{\text{R}}(X) + H_{\alpha}^{\text{R}}(Y) - H_{\alpha}^{\text{R}}(X, Y)$$

is generally valid. This problem arises from the difficulty to define a conditional Rényi entropy to be consistent with the setting in the Shannon case [91–93]. Several variants are known [94,95]. The Jizba–Arimitsu conditional Rényi-entropy $H_{\alpha}^{\text{JA}}(X|Y)$ defined as

$$H_{\alpha}^{\text{JA}}(X|Y) = H_{\alpha}(X, Y) - H_{\alpha}(Y) \quad (15)$$

fulfills the chain rule by definition [96]. Obviously, $H_{\alpha}^{\text{JA}}(X|Y)$ can be interpreted as an extension of the conditional Shannon entropy $H(X|Y)$ because the definition (15) precisely coincides with Shannons chain rule (5). The resulting mutual entropy

$$M_{\alpha}^{\text{R}}(X, Y) = H_{\alpha}(X) + H_{\alpha}^{\text{JA}}(X|Y) \quad (16)$$

is consistent with (4) and preserves the symmetry [97]. However, it may violate the non-negativity as well as $I_{\alpha}^{\text{R}}(X, Y) \neq M_{\alpha}^{\text{R}}(X, Y)$ is valid. For further variants, we refer to [95].

Analogous to the resolved mutual information $F(x, Y)$ in the Shannon case from Equation (10), we denote

$$F_{\alpha}^{\text{R}}(x, Y) = \int_{\mathcal{Y}} \frac{(p(x, y))^{\alpha}}{(p(x))^{\alpha-1} \cdot (p(y))^{\alpha-1}} dy \quad (17)$$

as the α -scaled (feature) resolved Rényi mutual information (rRMI). Obviously,

$$I_\alpha^R(X, Y) = \frac{1}{\alpha - 1} \log \left(\int_{\mathcal{X}} F_\alpha^R(x, Y) dx \right)$$

holds. The Rényi variant of the cross mutual information for sequences $X(t)$ and $Y(t + \tau)$ at time t with shift $\tau \geq 0$ is defined as

$$I_\alpha^R(X(t), Y(t + \tau)) = \frac{1}{\alpha - 1} \log \left(\int_{\mathcal{X}} \int_{\mathcal{Y}} \frac{(p(x(t), y(t + \tau)))^\alpha}{(p(x(t)))^{\alpha-1} \cdot (p(y(t + \tau)))^{\alpha-1}} dy(t + \tau) dx(t) \right) \tag{18}$$

which gives by setting $Y(t + \tau) = X(t + \tau)$

$$I_\alpha^R(X(t), X(t + \tau)) = \frac{1}{\alpha - 1} \log \left(\int_{\mathcal{X}} \int_{\mathcal{X}} \frac{(p(x(t), x(t + \tau)))^\alpha}{(p(x(t)))^{\alpha-1} \cdot (p(x(t + \tau)))^{\alpha-1}} dx(t + \tau) dx(t) \right)$$

as the Rényi variant of the auto mutual information at time t with shift (delay) τ . Again, if $p(x(t))$ is independent from t , only the joint probability $p(x(t), x(t + \tau))$ remains t -dependent such that it becomes dependent only on the shift τ and we simply write $p(x, x(\tau))$ for this. Hence, the Rényi auto mutual information in dependence on the shift τ is obtained as

$$I_\alpha^R(X, \tau) = \frac{1}{\alpha - 1} \log \left(F_\alpha^R(X, \tau) \right) \tag{19}$$

with

$$F_\alpha^R(X, \tau) = \int_{\mathcal{X}} \int_{\mathcal{X}} \frac{(p(x, x(\tau)))^\alpha}{(p(x))^{\alpha-1} \cdot (p(x(\tau)))^{\alpha-1}} dx(\tau) dx \tag{20}$$

denoted as the Rényi variant of, or α -scaled Rényi mutual information function (RMIF). Accordingly, the α -scaled resolved version of the RMIF is

$$F_\alpha^R(x, \tau) = \int_{\mathcal{X}} \frac{(p(x, x(\tau)))^\alpha}{(p(x))^{\alpha-1} \cdot (p(x(\tau)))^{\alpha-1}} dx(\tau) \tag{21}$$

describing again the inherent correlations of the sequence and, hence, can serve as a characterizing quantity of the sequence. Accordingly, we denote the function $F_\alpha^R(x, \tau)$ as the α -scaled resolved Rényi mutual information function (rRMIF) for Rényi entropies. Obviously,

$$I_\alpha^R(X, \tau) = \frac{1}{\alpha - 1} \log \left(\int_{\mathcal{X}} F_\alpha^R(x, \tau) dx \right) \tag{22}$$

is valid analogous to Equation (11).

2.3. Tsallis α -Entropy and Related Mutual Information Functions

Recently, the Tsallis mutual information came into the focus for studying long range correlations in symbol sequences [70]. It is related to the Tsallis α -entropy

$$H_\alpha^T(X) = \frac{1}{\alpha - 1} \left(1 - \int_{\mathcal{X}} (p(x))^\alpha dx \right) \tag{23}$$

which becomes in the limit $\alpha \rightarrow 1$ the Shannon entropy $H(X)$. It was first introduced by HAVRDA AND CHARVÁT in 1967 [98] and later rediscovered by TSALLIS [64]. It is related to the Rényi α -entropy $H_\alpha^R(X)$ by

$$H_\alpha^R(X) = \frac{\log(1 - (1 - \alpha)H_\alpha^T(X))}{1 - \alpha}$$

as stated in [65]. The Tsallis-divergence is given by

$$D_\alpha^T(p(x) \parallel p(y)) = \frac{1}{\alpha - 1} \left(1 - \int_{\mathcal{X}} \int_{\mathcal{Y}} \frac{(p(x))^\alpha}{(p(y))^{\alpha-1}} dy dx \right) \tag{24}$$

as explained in [64,74]. Using the same procedure as for the Shannon case (3), we obtain

$$I_\alpha^T(X, Y) = D_\alpha^T(p(x, y) \parallel p(x) \cdot p(y)) \tag{25}$$

for the α -dependent Tsallis mutual information (TMI) [99]. As for the Rényi mutual information, the inequality

$$I_\alpha^T(X, Y) \neq H_\alpha^T(X) + H_\alpha^T(Y) - H_\alpha^T(X, Y)$$

is generally valid except the case $\alpha = 1$ being the Shannon case. It is symmetric and always non-negative but not consistent with the conditional Tsallis entropy

$$H_\alpha^T(X|Y) = \frac{H_\alpha^T(X, Y) - H_\alpha^T(Y)}{1 + (1 - \alpha) \cdot H_\alpha^T(Y)} \tag{26}$$

as explained in [65]. To avoid these and other difficulties, finally, the Tsallis α -entropy based mutual entropy (information) is suggested to be

$$M_\alpha^T(X, Y) = \frac{H_\alpha^T(X) + H_\alpha^T(Y) - H_\alpha^T(X, Y) + (1 - \alpha)H_\alpha^T(X)H_\alpha^T(Y)}{1 + (1 - \alpha)H_\alpha^T(X)}$$

as proposed in [65]. However, the inequality $M_\alpha^T(X, Y) \neq I_\alpha^T(X, Y)$ holds.

As for the Shannon and the Rényi variants of the mutual information, we consider a resolved Tsallis mutual information (rTMI)

$$F_\alpha^T(x, Y) = \int_{\mathcal{Y}} \frac{(p(x, y))^\alpha}{(p(x) \cdot p(y))^{\alpha-1}} dy \tag{27}$$

such that

$$I_\alpha^T(X, Y) = \frac{1}{\alpha - 1} \left(1 - \int_{\mathcal{X}} F_\alpha^T(x, Y) dx \right) \tag{28}$$

holds. For the auto mutual information with shift τ , we get

$$I_\alpha^T(X, \tau) = \frac{1}{\alpha - 1} \left(1 - F_\alpha^T(X, \tau) \right) \tag{29}$$

with

$$F_\alpha^T(X, \tau) = \int_{\mathcal{X}} F_\alpha^T(x, \tau) dx \tag{30}$$

as the Tsallis mutual information function (TMIF) by the same arguments as before with

$$F_\alpha^T(x, \tau) = \int_{\mathcal{X}} \frac{(p(x, x(\tau)))^\alpha}{(p(x) \cdot p(x(\tau)))^{\alpha-1}} dx(\tau) \tag{31}$$

denoted as the α -scaled resolved Tsallis mutual information function (rTMIF) for Tsallis entropies. In bioinformatics context, it can be seen as a α -scaled object dependent average Tsallis mutual information profile.

Comparing TMIF and RMIF as well as rTMIF and rRMIF, we can obviously state the equalities $F_\alpha^R(X, \tau) = F_\alpha^T(X, \tau)$ and $F_\alpha^R(x, \tau) = F_\alpha^T(x, \tau)$.

All quantities relevant for the later data analysis are summarized and adapted for biomolecular sequences in Table 2.

3. Interpretable Classification Learning by Learning Vector Quantization

Learning vector quantization (LVQ) as introduced by T. KOHONEN is a neural network approach for classification trained by Hebbian competitive learning to achieve an approximation of a Bayes classifier model [100,101]. It is based on the intuitive nearest prototype principle, i.e., prototype vectors are distributed in the data space during the learning phase to detect the data class distribution. In the recall phase, a data point is assigned to that class the nearest prototype is referencing based on a given data dissimilarity. It is known as a robust variant of the nearest neighbor principle [102]. In this way, LVQ is easy to interpret [27].

Particularly, LVQ supposes data vectors $\mathbf{x} \in \mathbf{X} = \{\mathbf{x}_k\}_{k=1}^K \subseteq \mathbb{R}^n$ together with class labels $c(\mathbf{x}) \in \mathcal{C} = \{1, \dots, C\}$ for training [100]. Furthermore, the LVQ-model requires prototype vectors $\mathbf{w}_j \in \mathbf{W} = \{\mathbf{w}_k\}_{k=1}^N \subseteq \mathbb{R}^n$ with class labels $c(\mathbf{w}_j)$ such that each class of \mathcal{C} is represented by at least one prototype. As already mentioned, a new data vector is assigned to a class by means of the nearest prototype principle

$$\mathbf{x} \mapsto c(\mathbf{w}^*) \quad \text{with} \quad \mathbf{w}^* = \underset{\mathbf{w}_j \in \mathbf{W}}{\operatorname{argmin}}(d(\mathbf{x}, \mathbf{w}_j))$$

where \mathbf{w}^* is denoted as the winning prototype for the input \mathbf{x} with respect to \mathbf{W} . Here, d is a predefined dissimilarity measure in \mathbb{R}^n frequently chosen as the (squared) Euclidean distance. According to [103], prototype learning in GLVQ can be realized as a stochastic gradient descent learning (SGDL) for the prototype set \mathbf{W} . The respective cost function

$$E = \sum_{\mathbf{x} \in \mathbf{X}} E(\mathbf{x}, \mathbf{W})$$

approximates the overall classification error for the training set \mathbf{X} by local errors

$$E(\mathbf{x}, \mathbf{W}) = f(\mu(\mathbf{x}, \mathbf{W}))$$

taking into account the classifier function

$$\mu(\mathbf{x}, \mathbf{W}) = \frac{d(\mathbf{x}, \mathbf{w}^+) - d(\mathbf{x}, \mathbf{w}^-)}{d(\mathbf{x}, \mathbf{w}^+) + d(\mathbf{x}, \mathbf{w}^-)}$$

such that $\mu(\mathbf{x}, \mathbf{W}) \in [-1, 1]$ is valid and f is a monotonically increasing sigmoid squashing function. Here, \mathbf{w}^+ is the closest prototype to \mathbf{x} with a correct label, whereas \mathbf{w}^- is the closest prototype with incorrect label, i.e.,

$$\mathbf{w}^+ = \underset{\substack{\mathbf{w}_j \in \mathbf{W} \\ c(\mathbf{x}) = c(\mathbf{w}_j)}}{\operatorname{argmin}} d(\mathbf{x}, \mathbf{w}_j) \quad \text{and} \quad \mathbf{w}^- = \underset{\substack{\mathbf{w}_j \in \mathbf{W} \\ c(\mathbf{x}) \neq c(\mathbf{w}_j)}}{\operatorname{argmin}} d(\mathbf{x}, \mathbf{w}_j) \quad (32)$$

such that $\mu(\mathbf{x}, \mathbf{W}) < 0$ holds in case of a correct classification.

The SGDL-step for a given input \mathbf{x} is

$$\Delta \mathbf{w}^\pm \propto \frac{\partial E(\mathbf{x}, \mathbf{W})}{\partial \mathbf{w}^\pm}$$

realizing an attraction scheme (vector shift) for \mathbf{w}^+ towards \mathbf{x} in case of the (squared) Euclidean distance as dissimilarity d , whereas \mathbf{w}^- is repelled from \mathbf{x} . This variant of LVQ is known as generalized LVQ (GLVQ, [103]).

The interpretability and the power of the GLVQ can be improved taking the dissimilarity d as

$$d_\Omega(\mathbf{x}, \mathbf{w}) = (\Omega(\mathbf{x} - \mathbf{w}))^2 \quad (33)$$

where $\Omega \in \mathbb{R}^{m \times n}$ is a mapping matrix with $m \leq n$. This mapping matrix is also the subject of adaptation during learning with

$$\Delta\Omega_{ij} \propto \frac{\partial E(\mathbf{x}, W)}{\partial \Omega_{ij}}$$

realizing the SGDL-step for a given input \mathbf{x} . This approach is known as the generalized matrix LVQ (GMLVQ) [72]. In case of $m < n$, it is the limited rank GMLVQ (LiRaM-LVQ) [104].

The resulting matrix $\Lambda = \Omega^T \Omega$ is denoted as classification correlation matrix (CCM) [105]. The matrix entries Λ_{ij} reflect after training those correlations between the i^{th} and j^{th} data features, which contribute to a class discrimination. More specifically, if $|\Lambda_{ij}| \gg 0$ is valid, the respective correlation of the features is important to separate the classes, whereas $|\Lambda_{ij}| \approx 0$ indicates that either the correlation between the i^{th} and j^{th} data feature does not improve the classification or that this correlation information is already contained in another significant correlation. The vector $\lambda = (\lambda_1, \dots, \lambda_n)^T$ with

$$\lambda_i = \Lambda_{ii}$$

being the non-negative diagonal elements of the classification correlation matrix is denoted as classification relevance profile (CRP) of the features [106]. It describes the relevance of the features for class discrimination with an analog interpretation as for $|\Lambda_{ij}|$. The classification influence profile (CIP), defined as $\kappa = (\kappa_1, \dots, \kappa_n)^T$ with

$$\kappa_i = \sum_j |\Lambda_{ij}|$$

provides the importance of the i^{th} data feature in combination with all other features for the separation of the data set. Both profiles, as well as the classification correlation matrix, provide additional information beyond the pure classification performance and, hence, contribute to a high interpretability of the classification model [107].

Moreover, all mentioned GLVQ variants are robust classification learning models maximizing the hypothesis margin for most appropriate class separation [73,108].

4. Applications of Mutual Information Functions for Sequence Classification

In the following, we apply the described information theoretic quantities as characterizing features for biomolecular sequences. Particularly, we use the introduced variants of mutual information functions summarized in Table 2 and natural vectors as feature generators. Their performance is evaluated in combination with the LiRaM-LVQ for three biological classification tasks.

4.1. Data Sets

The chosen data sets summarized in Table 1 are representatives of biological applications facing the common challenges of varying sequence lengths and containing ambiguous characters (see Section 4.2.3).

Table 1. Overview of the used data sets.

Data Set	Classes	Sequences	Per Class *	Mean Length	Std. Length
Quadruplex detection	2	368	175/193	62.1	43.7
lncRNA vs. mRNA	2	20,000	10,000 each	1197.3	710.8
COVID types	3	156	44/90/22	29,862.9	34.1

* Sample size per class.

4.1.1. Quadruplex Detection

This data set consists of 368 nucleotide sequences that were experimentally validated to either build or not build a G-quadruplex during folding. Quadruplexes are structural (3D) motifs of one or more nucleic acid strands consisting of at least two stacked tetrads. These are characterized by the planar arrangement of four nucleotides, each of which forms non-canonical bonds (base pairing schemes other than Watson-Crick) with two of the other nucleotides. If all tetrad-forming nucleotides are guanine, it is also denoted as the G-quadruplex, or G4. The utilized data are equivalent to that published by [109] without the random sequences (background sequences assumed to be non-G4). The data source is the G4RNA database [110].

4.1.2. lncRNA vs. mRNA

For the next task, we used a data set containing 10,000 human long non-coding RNA (lncRNA) sequences and 10,000 protein-coding transcripts (mRNA). lncRNA are transcripts that do not encode proteins, i.e., they are not translated, but play a role in gene regulation. Their typical length of more than 200 nucleotides (nt) delineates them from small non-coding RNA such as miRNAs or snoRNAs and similarities in sequence structure compared to mRNA make their differentiation challenging [111]. The data set was generated analogous to [111]: The data were retrieved from the GENCODE database [112] in the latest version v.38 at the time of access (11 August 2021). Data preprocessing comprising filtering of sequences with length 250–3000 nt and random selection of 10,000 sequences per class was applied. In contrast to [111], we decided to use the same interval for both classes in order to not prone our classifier to use the sequence length as class discriminating property.

4.1.3. COVID Types

As a third data set, we took 156 coronavirus sequences from human hosts of the types A, B, and C, implicitly coding the evolution in time of the virus. The SARS-CoV-2 sequence data source is the GISAID (Global Initiative on Sharing Avian Influenza Data) coronavirus repository from 4 March 2020 with types derived from a phylogenetic network analysis in [113]. Type A is most similar to the bat virus, type B evolves from A by a non-synonymous and a synonymous mutation (evolutionary substitutions that do or do not modify the resulting amino acid sequence, respectively) and type C is characterized by a further non-synonymous mutation.

4.2. Feature Generation

In the following, we introduce the concept of natural vectors and provide a description of how to generate feature vectors from the information theoretical quantities MIF and rMIF introduced in Section 2 for machine learning applications. Both feature generators are sequence length independent and capable of handling ambiguous characters in biological data as covered in more detail in Section 4.2.3.

4.2.1. Natural Vectors

Natural vectors (NV) in biomolecular context accumulate statistical descriptors concerning the distribution of nucleotide positions within a sequence $s = [s_1 \dots s_n]$ over the alphabet $\mathcal{A} = \{A, C, G, T\}$. They were generalized in [47] from [114]. Natural vectors are known to be characteristic fingerprints for biomolecular sequences reflecting statistical and, hence, information theoretic properties. Therefore, we consider them as baseline for comparison with mutual information functions.

To define NV accurately, let $n_k = \sum_{i=1}^n w_k(s_i)$ be the absolute frequency of nucleotide k in s , given that $w_k(s_i) \in \{0, 1\}$ indicates the absence (0) or presence (1) of k at sequence position s_i . Furthermore, let $\mu_k = \sum_{i=1}^n i \cdot \frac{w_k(s_i)}{n_k}$ denote the mean sequence position of nucleotide k and $D_k^j = \sum_{i=1}^n \frac{(i - \mu_k)^j w_k(s_i)}{n_k^{j-1} n^{j-1}}$ be the normalized central moment of order j . Then,

the natural vector is defined as

$$\mathbf{x} = (n_A, \mu_A, D_A^2, \dots, D_A^{n_A}, n_C, \mu_C, D_C^2, \dots, D_C^{n_C}, n_G, \mu_G, D_G^2, \dots, D_G^{n_G}, n_T, \mu_T, D_T^2, \dots, D_T^{n_T}) \tag{34}$$

Obviously, $\mu_k = D_k^1$ is valid and one can take just $n_k = D_k^0$ in terms of the statistical moments. Furthermore, it was stated that this setting guarantees a unique coding of the molecular sequences [47]. In practical applications, the maximum order j_{\max} of moments to be calculated is fixed equally in dependence of the data set for all nucleotides to achieve equal-length vectors for all considered sequences. Hence, the data dimension becomes $n = 4 \cdot (j_{\max} + 1)$.

In the experiments, we determined an optimal setting of j_{\max} under consideration of the sequence length via grid search. Therefore, we evaluated $j_{\max} \in \{2, 3, 4\}, \{2, \dots, 15\}$ and $\{2, \dots, 15\}$ for the Quadruplex detection, lncRNA vs. mRNA and COVID types data set, respectively. We directly take \mathbf{x} from Equation (34) as input (feature vector) for the LVQ model. The maximum order 15 for j_{\max} was taken as an upper bound because higher moments were numerically vanishing for the used data sets.

4.2.2. Mutual Information Functions

In case of mutual information functions, the feature vector $\mathbf{x} = (x_1, \dots, x_{\tau_{\max}})^T$ is generated from a sequence X by setting $x_\tau = F(X, \tau)$ or $x_\tau = F_\alpha^R(X, \tau)$ for Shannon and Rényi, respectively. The maximum distance between pairs of nucleotides considered in the sequence is τ_{\max} .

For the resolved mutual information functions, we take

$$\mathbf{x} = \left(x_1^A, \dots, x_{\tau_{\max}}^A \quad x_1^C, \dots, x_{\tau_{\max}}^C \quad x_1^G, \dots, x_{\tau_{\max}}^G \quad x_1^T, \dots, x_{\tau_{\max}}^T \right)^T$$

with $x_\tau^k = F(k, \tau)$ or $x_\tau^k = F_\alpha^R(k, \tau)$, where $k \in \mathcal{A} = \{A, C, G, T\}$.

Table 2 summarizes the applied mutual information functions for the Shannon and Rényi case.

Table 2. Overview of the computed mutual information functions for biomolecular sequences X from Sections 2.1 and 2.2.

	MIF	rMIF
Shannon	$F(X, \tau) = \sum_{x \in \mathcal{X}} F(x, \tau)$	$F(x, \tau) = \sum_{x(\tau) \in \mathcal{X}} p(x, x(\tau)) \cdot \log \left(\frac{p(x, x(\tau))}{p(x) \cdot p(x(\tau))} \right)$
Rényi	$F_\alpha^R(X, \tau) = \sum_{x \in \mathcal{X}} F_\alpha^R(x, \tau)$	$F_\alpha^R(x, \tau) = \sum_{x(\tau) \in \mathcal{X}} \frac{p(x, x(\tau))^\alpha}{(p(x) \cdot p(x(\tau)))^{\alpha-1}}$

In the literature on MIF, there is disagreement on how to calculate the marginal probabilities of the nucleotides: one camp propagates a symmetric version, i.e., $p(x)$ denotes the relative frequency of a nucleotide x in a sequence [52,54,56], while the other distinguishes the frequencies of the nucleotides at the positions x depending on $x(\tau)$, i.e., $p(x) = \sum_{x(\tau)} p((x, x(\tau))|x)$ and $p(x(\tau)) = \sum_x p((x, x(\tau))|x(\tau))$ [57–59]. We used the latter (non-symmetric) version, since biological sequences have a chemically reasonable reading direction, such that a nucleotide’s neighbor is determined in the 3’ direction.

An optimal setting of the hyper-parameter τ_{\max} was obtained under consideration of the sequence length via grid search. We evaluated $\tau_{\max} \in \{2, \dots, 8\}, \{10, 25, 50, 100\}$ and $\{5, 10, 50, 100\}$ for the Quadruplex detection, lncRNA vs. mRNA and COVID types data set, respectively.

The α -value for the Rényi variants was set to $\alpha = 2$ as usual. This choice leads to low computational costs and provides numerical stability [82].

4.2.3. Handling of Ambiguous Characters

Ambiguous characters are introduced by the IUPAC (International Union of Pure and Applied Chemistry) degenerate base notation [115]. Thereby, the notation ambiguous refers to the concept that a single character from the alphabet extension $\mathcal{E} = \{R, Y, M, K, S, W, H, B, V, D, N\}$ represents more than one nucleotide, present in data to describe incompletely specified bases or uncertainty of them [115]. For instance R denotes either A or G , the ambiguous character H stands for either A , C , or T , whereas N codes for all four possible nucleotides.

In order to make the feature generators cope with these representations, the weights $0 \leq w_k(s_i) \leq 1$ now code the probability for, and not just the presence (1) or absence (0) of, a nucleotide at one specific sequence position, i.e.,

$$w_A(s_i) = \begin{cases} 1 & \text{if } s_i = A \\ 0 & \text{if } s_i = C, G, T, Y, K, S \text{ or } B \\ \frac{1}{2} & \text{if } s_i = R, M \text{ or } W \\ \frac{1}{3} & \text{if } s_i = H, V \text{ or } D \\ \frac{1}{4} & \text{if } s_i = N \end{cases} \quad (35)$$

In [116], natural vectors were expanded to handle this extended alphabet. We designed a solution for the MIF variants analogously.

4.3. Classification

Following all the feature extractors mentioned above, we have applied a Z-score normalization in order to make the individual features comparable. Classification was then done using the LiRaM-LVQ implementation from the Python toolbox prototorch in 3-fold cross validation. In all cases, the prototypes for learning were initialized as randomly selected data points and the learning rate was set to 0.01 in all cases. The mapping dimension m was set to 10 independent of data set or feature set. The choice of the number of prototypes was data set depending: for Quadruplex detection and COVID types, we took only one prototype per class. For the lncRNA vs. mRNA data set, the grid search for optimal setting resulted in 50 prototypes per class as balance between complexity of the model and performance.

5. Results and Discussion

5.1. Classification Performance

Table 3 displays the achieved test accuracies by LiRaM-LVQ in combination with the optimal parameter setting of j_{\max} and τ_{\max} for the NV and MIF variant feature extractors, respectively.

Table 3. Achieved test accuracies \pm standard deviation by LiRaM-LVQ in percent and the respective parameter setting.

Data Set	NV	MIF		rMIF	
		Shannon	Rényi	Shannon	Rényi
Quadruplex detection	78.8 ± 1.0	68.9 ± 1.2	68.2 ± 1.7	77.4 ± 1.3	82.0 ± 1.0
	$j_{\max} = 4$	$\tau_{\max} = 7$	$\tau_{\max} = 7$	$\tau_{\max} = 8$	$\tau_{\max} = 7$
lncRNA vs. mRNA	71.9 ± 0.1	75.4 ± 0.2	75.5 ± 0.3	81.4 ± 0.1	76.3 ± 0.6
	$j_{\max} = 7$	$\tau_{\max} = 100$	$\tau_{\max} = 100$	$\tau_{\max} = 100$	$\tau_{\max} = 100$
COVID types	86.0 ± 1.2	98.1 ± 0.6	97.4 ± 1.0	99.7 ± 0.3	99.3 ± 0.5
	$j_{\max} = 5$	$\tau_{\max} = 50$	$\tau_{\max} = 50$	$\tau_{\max} = 50$	$\tau_{\max} = 50$

Considering these results in Table 3, we see that rMIF outperforms the MIF variants as well as the models which use NV for feature generations for all three data sets. Furthermore, the developed Rényi variant shows in the Quadruplex detection example for rMIF

significantly better results compared to the Shannon counterpart. However, for the second data set, the performance of rMIF depends on the choice of τ_{\max} . In general, it can be said that for long sequences τ_{\max} need to be chosen adequately if long range correlations are to be considered as well.

For deeper investigation of these results and to show the capabilities of the applied LiRaM-LVQ classifier, we will consider the CCM and CIP. Furthermore, visualizations of the mean MIF and rMIF per class and data set are considered for deeper understanding of the generated features and their potential differences between classes. In order to not overload the reader, we will restrict a more in-depth interpretation and discussion to one of the data sets, the quadruplex detection challenge.

It should be noted that a feature generation procedure based on pure statistics might achieve comparable or even better results. For this reason, it is not surprising that the statistical feature extractor Bag of words [117,118] has been successful in related works on the data sets mentioned: 92.8% AUC were achieved for the quadruplex data in combination with a simple neural network [109], an accuracy of 98.7% was described in [111] for the lncRNA vs. mRNA data by use of a convolutional neural network and 97.4% accuracy were obtained in [119] for COVID type detection using GMLVQ. However, the focus of this paper is on the investigation and further development of information theoretical methods and their suitability for sequence analyses in computational biology.

5.2. Visualization of MIF Variants

A closer look at the class-wise averaged MIF variant profiles in Figure 1 allows for assessing the methods behavior on the quadruplex data set. The plotted means suggest a clear class delineation, while the standard deviation adds depth/difficulty to the problem. All profiles are plotted prior to Z-score normalization, but with a slight vertical shift between the classes for better visual perception.

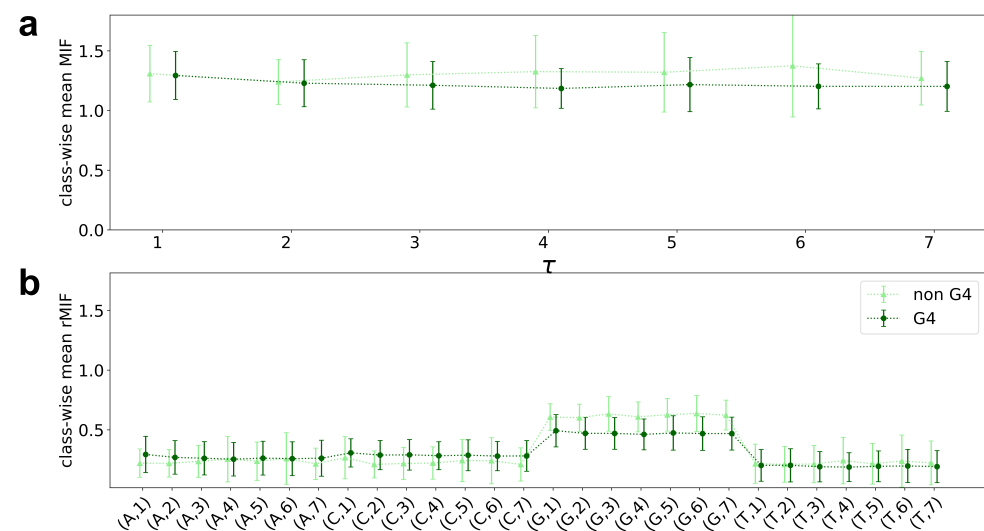


Figure 1. Data insights, i.e., class-wise mean and standard deviation of the MIF variants for the quadruplex data set (G4 and non-G4 forming sequences). (a) Rényi MIF; (b) Rényi rMIF.

Comparison of the MIF and rMIF clearly shows a more accurate resolution of the information for rMIF, not only in terms of inter-sequential distances, but also in terms of individual nucleotides. Obviously, the sum of the four $F(x, \tau)$ with $x \in \{A, C, G, T\}$ yields $F(X, \tau)$. The features with respect to G-nucleotides stand out in particular.

5.3. Interpretation of CCM and CIP of the Trained LiRaM-LVQ Model

The resulting CCM of the trained LiRaM-LVQ model gives domain experts, here biologists, immediate assistance to evaluate whether the classifier works reasonably. Furthermore, it allows statements to be made about whether the classification decision is

based on some data biases or artifacts that were not necessarily known during data generation [104]. This interpretation possibility of the LiRaM-LVQ model is a huge advantage in comparison to black box models [120] especially in biological issues: Together with meaningful data features, as given here, biologists can draw conclusions regarding expected biological and biochemical properties.

In the experiments, we verified that the CCM can serve as a basis for interpretation by repeating the classification process multiple times and analyzing whether the matrix is visually stable. If significant deviations had been seen, an interpretation would have been spurious. Each depicted CCM is the result of averaging the individual CCMs obtained from the three validation folds. Furthermore, we limit the visualization to the best hyperparameter setting according to our grid search.

For the quadruplex data set, the best choice $\tau_{\max} = 7$ gives a CCM with dimensionality 7×7 and 28×28 for the MIF and rMIF case, respectively. These quantities are visualized in Figure 2 giving insights into the classification decision of LiRaM-LVQ:

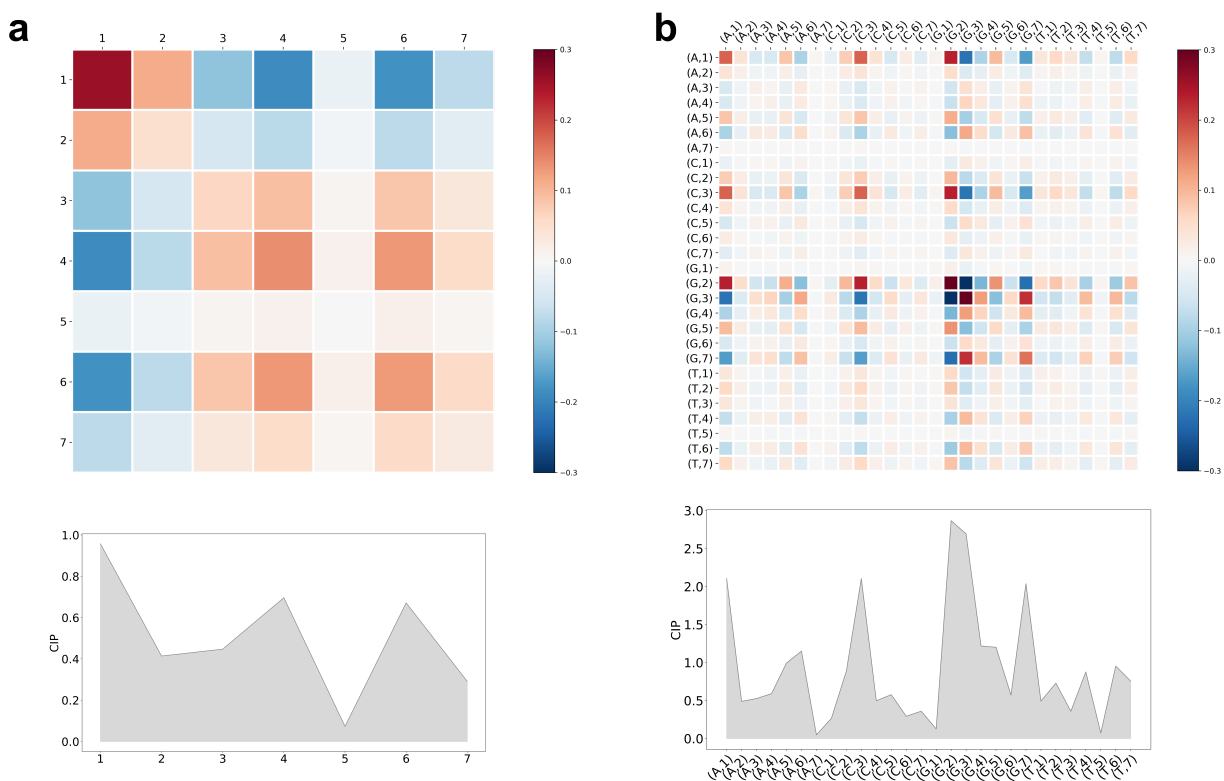


Figure 2. Classification insights, i.e., CCM and CIP of LiRaM-LVQ for the quadruplex data set. The color bars code the correlation values of the CCM. (a) Rényi MIF; (b) Rényi rMIF.

As can be seen from the CIP and from the CCMs' main diagonal, the CRP, in Figure 2a, the MIF values for τ equals 1, 4, and 6 mainly influence the classifier's decision to discriminate the classes of G-quadruplex (G4) and non-G4 forming sequences. Moreover, the CCM shows positive and negative correlations *between* the features. For example, features $\tau = 4$ and 6 are strongly positive and $\tau = 1$ and 4 are in strong negative correlation with each other. Thus, if $\tau = 1$ has a high value, it is important for the class discrimination, but only if $\tau = 4$ has a small value and vice versa. It is striking that the feature $\tau = 5$ does neither alone nor in combination with any other feature contribute to the differentiation for this learned model.

In Figure 2b, the CIP illustrates that eight features stand out with their influence on the class discrimination. Sorted by importance, these are: the information for (G, 2), (G, 3), (A, 1), (C, 3), (G, 7), (G, 4), (G, 5), and (A, 6). Taking the CCM into consideration, a high positive correlation between (A, 1) and (G, 2) as well as between (C, 3) and (G, 2) is

obvious. Examples for high negative correlations would be the pairs of (G, 2) and (G, 3) or between (A, 1) and (G, 3). The clearly recognizable significance of Gs at different distances is biologically sound due to the general characterization of a G-quadruplex by a pattern of recurring guanines in the sequence as described in [121]. Insights like this would not have been possible with the standard MIF but only with our introduced resolved variant rMIF.

At first glance, one might claim an inconstancy between the high influence values in the CIPs for MIF and rMIF features. However, in the MIF calculation, there is an averaging of the information over the alphabet, such that the classifier can make use of more detailed information with rMIF. This means that the summation of the classification influence values for all four (x, τ) does not necessarily result in the influence value for the MIF for a specific τ and vice versa. As the individual nucleotides play a key role in the bioinformatics domain, there might be an essential information loss if an averaging procedure takes place as it is done for the simpler MIF.

Beside biological interpretations, these insights offer the possibility to adjust or rather fine-tune the classification model. For example, by taking just the seven most important rMIF features into account, we still obtain a performance of $77.1 \pm 0.7\%$. Hence, we could reduce the model complexity with moderate performance decrease.

To sum up, the LiRaM-LVQ classifier is transparent in the decision process as well as in the Hebbian learning process. Now, the expert can start to evaluate the results and either extract knowledge from the classifier or question the quality of the data/model if the results seem peculiar.

For the sake of completeness, Figure 3 shows the CCM and CIP for the lncRNA vs. mRNA as well for the COVID type data set. The Shannon rMIF features were superior in these tasks. Our grid search resulted in high optimal values for τ_{\max} which is alright for pure performance evaluation but poses a problem in visually evaluating the CCMs and drawing conclusions. Therefore, we decided to take advantage of the same procedure described above: we identified the 30 most important/valuable features using the CIP, ran the classification procedure again using only these, and finally visualized both characteristics.

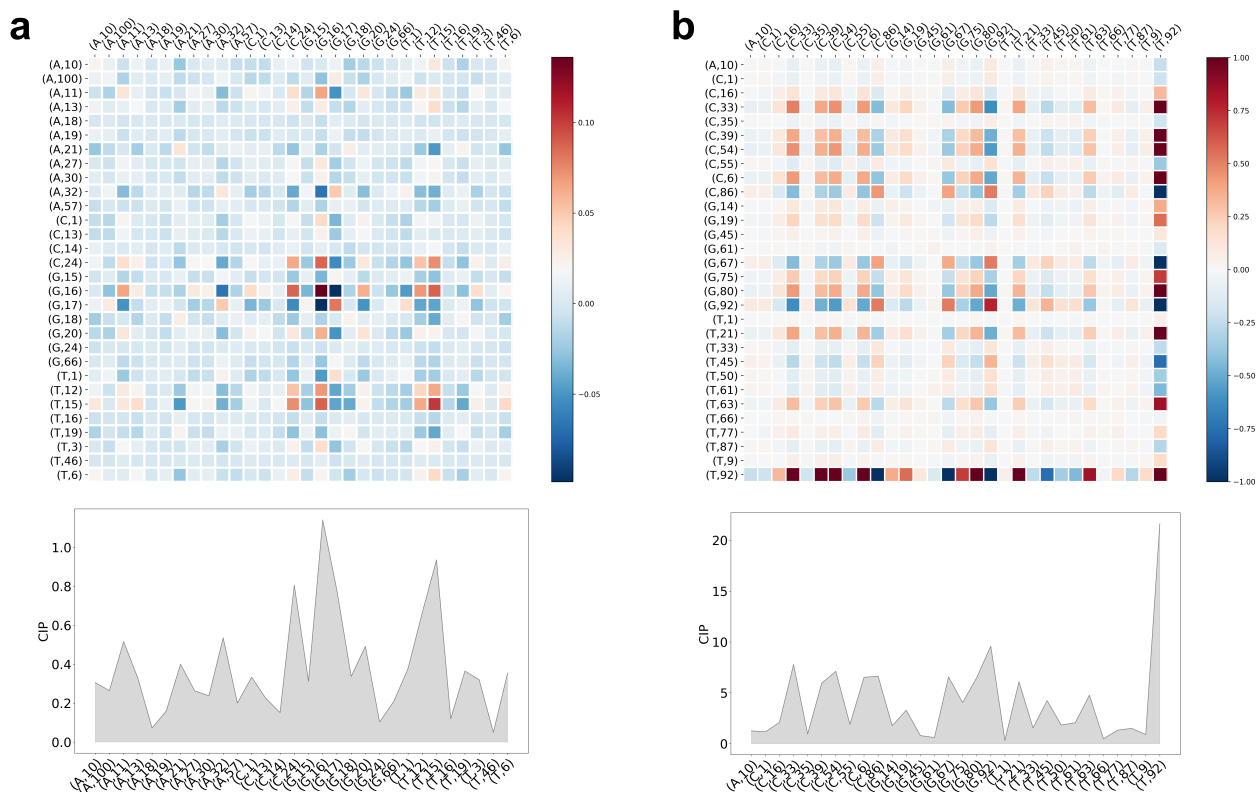


Figure 3. Classification insights, i.e., CCM and CIP of LiRaM-LVQ using the Shannon rMIF. The color bars code the correlation values of the CCM. (a) COVID data; (b) lncRNA vs. mRNA data.

An in-depth analysis of the results including biological interpretation is up to the well-disposed reader.

6. Conclusions, Remarks, and Future Work

In this contribution, we propose information theoretic concepts and quantities to characterize spatial correlations in sequences. In particular, we introduced several variants of mutual information functions for Shannon, Rényi, and Tsallis information theoretic approaches. In particular, the resolved mutual information functions provide subtle information regarding the internal spatial correlation of the sequences.

These functions/quantities can be used as sequence signatures/fingerprints and thus for comparison in machine learning approaches. In particular, interpretable machine learning models can make use of this resolved information to achieve insights about the sequence class differences. As we have shown using our favored LiRaM-LVQ, detailed information can be extracted as an add-on to the pure classification model.

We see applications for sequence analysis in bioinformatics, especially in the context of alignment-free sequence comparison. Additionally, we remark that this concept can be extended to the analysis of more general categorical sequential data such as natural language texts or sheet music.

In the future work, we will extend this approach to further mutual information concepts related to other widely considered entropy measures and information theoretic quantities, e.g., the Cauchy–Schwarz-divergence [85], or more general α -, β - and γ -divergences with related mutual information concepts [74,91,122]. Further considerations could be a generalization to higher than two-body correlations as suggested in [123] or performing the calculation for sequences not 1 by 1 residue (position), but multiple residues [59].

Furthermore, we want to compare these methods with other feature generators taking statistical (spatial) correlation into account such as the return time distribution [124] known from stochastic modeling, DMk method [125] incorporating the occurrence, location, and order relation of k -mers, compression based methods with the underlying concept of minimum description length [126], methods based on domain transform, i.e., Fourier/Wavelet [127,128], DNA walks [45,129] and iterated function systems, e.g., chaos game representation or universal sequence maps [42,130].

However, interpretability should be kept always as a key feature when considering alternative models [25,131,132]. Interpretability increases the trustworthiness and hence the acceptance of models for the potential users [27]. Further extensions improving transparency of the decision and already known for GLVQ approaches are the incorporation of reject options for ambiguous decisions or outliers as well as the use of interpretable probabilistic classifiers [133–135].

Author Contributions: Conceptualization, K.S.B., M.K. and T.V.; methodology, K.S.B., M.K., S.S. and T.V.; software and visualization, K.S.B. and M.K.; validation, resources and data curation, K.S.B. and J.A.; formal analysis, K.S.B., S.S. and T.V.; investigation, K.S.B., J.A. and M.K.; writing—original draft preparation, T.V. and K.S.B.; writing—review and editing, M.K., J.A. and S.S.; supervision and project administration, T.V.; funding acquisition, T.V. and M.K. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the European Social Fund Grant No. 100381749.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data set for quadruplex detection is publicly available at <https://academic.oup.com/bioinformatics/article/33/22/3532/4061281#supplementary-data>, that for lncRNA vs. mRNA at <https://www.encodegenes.org/human/> (version v.38), and the accession numbers for the COVID type detection at <https://www.springerprofessional.de/learning-vector-quantization-as-an-interpretable-classifier-for-/19111526?fulltextView=true> (all accessed on 11 August 2021). The toolbox prototorch is publicly available at <https://github.com/si-cim/prototorch>

and was used in version 0.2.0. The code for the NV and MIF variant calculation can be obtained from the authors upon request.

Acknowledgments: The authors would like to thank Mirko Weber, Daniel Staps, Alexander Engelberger and Jensun Ravichandran, all from the University of Applied Sciences Mittweida, for useful discussions and technical support.

Conflicts of Interest: The authors declare no conflict of interest.

Abbreviations

The following abbreviations are used in this manuscript:

AMI	Average Mutual Information
CCM	Classification Correlation Matrix
CIP	Classification Influence Profile
GISAID	Global Initiative on Sharing Avian Influenza Data
GLVQ	Generalized Matrix Learning Vector Quantization
IUPAC	International Union of Pure and Applied Chemistry
lncRNA	Long Non-Coding RNA
LiRaM-LVQ	Limited Rank Matrix Learning Vector Quantization
LVQ	Learning Vector Quantization
MIF	Mutual Information Function
mRNA	messenger RNA
NV	Natural Vectors
rMIF	resolved Mutual Information Function
RMIF	Rényi Mutual Information Function
rRMIF	resolved Rényi Mutual Information Function
rTMIF	resolved Tsallis Mutual Information Function
SGDL	Stochastic Gradient Descent Learning
TMIF	Tsallis Mutual Information Function

References

- Schrödinger, E. *What Is Life?* Cambridge University Press: Cambridge, UK, 1944.
- Eigen, M.; Schuster, P. Stages of emerging life—Five principles of early organization. *J. Mol. Evol.* **1982**, *19*, 47–61. [[CrossRef](#)]
- Haken, H. *Synergetics—An Introduction Nonequilibrium Phase Transitions and Self-Organization in Physics, Chemistry and Biology*; Springer: Berlin/Heidelberg, Germany, 1983.
- Haken, H. *Information and Self-Organization*; Springer: Berlin/Heidelberg, Germany, 1988.
- Baldi, P.; Brunak, S. *Bioinformatics*, 2nd ed.; MIT Press: Cambridge, MA, USA, 2001.
- Gatlin, L. The information content of DNA. *J. Theor. Biol.* **1966**, *10*, 281–300. [[CrossRef](#)]
- Gatlin, L. The information content of DNA. II. *J. Theor. Biol.* **1968**, *18*, 181–194. [[CrossRef](#)]
- Chanda, P.; Costa, E.; Hu, J.; Sukumar, S.; Hemert, J.V.; Walia, R. Information Theory in Computational Biology: Where We Stand Today. *Entropy* **2020**, *22*, 627. [[CrossRef](#)] [[PubMed](#)]
- Adami, C. Information Theory in Molecular Biology. *Phys. Life Rev.* **2004**, *1*, 3–22. [[CrossRef](#)]
- Vinga, S. Information Theory Applications for Biological Sequence Analysis. *Briefings Bioinform.* **2014**, *15*, 376–389. [[CrossRef](#)]
- Uda, S. Application of Information Theory in Systems Biology. *Biophys. Rev.* **2020**, *12*, 377–384. [[CrossRef](#)]
- Smith, M. DNA Sequence Analysis in Clinical Medicine, Proceeding Cautiously. *Front. Mol. Biosci.* **2017**, *4*, 24. [[CrossRef](#)] [[PubMed](#)]
- Mardis, E.R. DNA sequencing technologies: 2006–2016. *Nat. Protoc.* **2017**, *12*, 213–218. [[CrossRef](#)]
- Hall, B.G. Building Phylogenetic Trees from Molecular Data with MEGA. *Mol. Biol. Evol.* **2013**, *30*, 1229–1235. [[CrossRef](#)]
- Xia, X. Bioinformatics and Drug Discovery. *Curr. Top. Med. Chem.* **2017**, *17*, 1709–1726. [[CrossRef](#)] [[PubMed](#)]
- Goodfellow, I.; Bengio, Y.; Courville, A. *Deep Learning*; MIT Press: Cambridge, MA, USA, 2016.
- Krizhevsky, A.; Sutskever, I.; Hinton, G. ImageNet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems 25 (NIPS)*; Curran Associates, Inc.: San Diego, CA, USA, 2012; pp. 1097–1105.
- Schölkopf, B.; Smola, A. *Learning with Kernels*; MIT Press: Cambridge, MA, USA, 2002.
- Angermueller, C.; Pärnamaa, T.; Parts, L.; Stegle, O. Deep Learning for Computational Biology. *Mol. Sys. Biol.* **2016**, *12*, 878. [[CrossRef](#)] [[PubMed](#)]
- Min, S.; Lee, B.; Yoon, S. Deep learning in bioinformatics. *Briefings Bioinform.* **2016**, 1–16. [[CrossRef](#)]
- Nguyen, N.; Tran, V.; Ngo, D.; Phan, D.; Lumbanraja, F.; Faisal, M.; Abapihi, B.; Kubo, M.; Satou, K. DNA Sequence Classification by Convolutional Neural Network. *J. Biomed. Sci. Eng.* **2016**, *9*, 280–286. [[CrossRef](#)]

22. Jaakkola, T.; Diekhans, M.; Haussler, D. A discriminative framework for detecting remote protein homologies. *J. Comput. Biol.* **2000**, *7*, 95–114. [[CrossRef](#)]
23. Jumper, J.; Evans, R.; Pritzel, A.; Green, T.; Figurnov, M.; Ronneberger, O.; Tunyasuvunakool, K.; Bates, R.; Židek, A.; Potapenko, A.; et al. Highly accurate protein structure prediction with AlphaFold. *Nat.* **2021**, *596*, 583–596. [[CrossRef](#)] [[PubMed](#)]
24. Samek, W.; Montavon, G.; Vedaldi, A.; Hansen, L. *Explainable AI: Interpreting, Explaining and Visualizing Deep Learning*; Number 11700 in LNAI; Müller, K.R., Ed.; Springer: Berlin/Heidelberg, Germany, 2019.
25. Rudin, C. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nat. Mach. Intell.* **2019**, *1*, 206–215. [[CrossRef](#)]
26. Zeng, J.; Ustun, B.; Rudin, C. Interpretable classification models for recidivism prediction. *J. R. Stat. Soc. Ser. A* **2017**, *180*, 1–34. [[CrossRef](#)]
27. Lisboa, P.; Saralajew, S.; Vellido, A.; Villmann, T. The coming of age of interpretable and explainable machine learning models. In Proceedings of the 29th European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning (ESANN'2021), Bruges, Belgium, 6–8 October 2021; Verleysen, M., Ed.; i6doc.com: Louvain-La-Neuve, Belgium, 2021; pp. 547–556.
28. Zielezinski, A.; Vinga, S.; Almeida, J.; Karlowski, W.M. Alignment-Free Sequence Comparison: Benefits, Applications, and Tools. *Genome Biol.* **2017**, *18*, 186. [[CrossRef](#)]
29. Just, W. Computational Complexity of Multiple Sequence Alignment with SP-Score. *J. Comput. Biol.* **2001**, *8*, 615–623. [[CrossRef](#)]
30. Kucherov, G. Evolution of Biosequence Search Algorithms: A Brief Survey. *Bioinformatics* **2019**, *35*, 3547–3552. [[CrossRef](#)] [[PubMed](#)]
31. Haubold, B. Alignment-Free Phylogenetics and Population Genetics. *Briefings Bioinform.* **2014**, *15*, 407–418. [[CrossRef](#)] [[PubMed](#)]
32. Chan, C.X.; Bernard, G.; Poirion, O.; Hogan, J.M.; Ragan, M.A. Inferring Phylogenies of Evolving Sequences without Multiple Sequence Alignment. *Sci. Rep.* **2014**, *4*, 6504. [[CrossRef](#)]
33. Hatje, K.; Kollmar, M. A Phylogenetic Analysis of the Brassicales Clade Based on an Alignment-Free Sequence Comparison Method. *Front. Plant Sci.* **2012**, *3*, 192. [[CrossRef](#)]
34. Wu, Y.W.; Ye, Y. A Novel Abundance-Based Algorithm for Binning Metagenomic Sequences Using l-Tuples. *J. Comput. Biol. J. Comput. Mol. Cell Biol.* **2011**, *18*, 523–534. [[CrossRef](#)]
35. Leung, G.; Eisen, M.B. Identifying Cis-Regulatory Sequences by Word Profile Similarity. *PLoS ONE* **2009**, *4*, e6901. [[CrossRef](#)]
36. de Lima Nichio, B.T.; de Oliveira, A.M.R.; de Pierri, C.R.; Santos, L.G.C.; Lejambre, A.Q.; Vialle, R.A.; da Rocha Coimbra, N.A.; Guizelini, D.; Marchaukoski, J.N.; de Oliveira Pedrosa, F.; et al. RAFTS3G: An Efficient and Versatile Clustering Software to Analyses in Large Protein Datasets. *BMC Bioinform.* **2019**, *20*, 392. [[CrossRef](#)]
37. Bray, N.L.; Pimentel, H.; Melsted, P.; Pachter, L. Near-Optimal Probabilistic RNA-Seq Quantification. *Nat. Biotechnol.* **2016**, *34*, 525–527. [[CrossRef](#)]
38. Zerbino, D.R.; Birney, E. Velvet: Algorithms for de Novo Short Read Assembly Using de Bruijn Graphs. *Genome Res.* **2008**, *18*, 821–829. [[CrossRef](#)] [[PubMed](#)]
39. Pajuste, F.D.; Kaplinski, L.; Möls, M.; Puurand, T.; Lepamets, M.; Remm, M. FastGT: An Alignment-Free Method for Calling Common SNVs Directly from Raw Sequencing Reads. *Sci. Rep.* **2017**, *7*, 2537. [[CrossRef](#)] [[PubMed](#)]
40. Luo, L.; Lee, W.; Jia, L.; Ji, F.; Tsai, L. Statistical correlation of nucleotides in a DNA sequence. *Phys. Rev. E* **1998**, *58*, 861–871. [[CrossRef](#)]
41. Luo, L.; Li, H. The statistical correlation of nucleotides in protein-coding DNA sequences. *Bull. Math. Biol.* **1991**, *53*, 345–353. [[CrossRef](#)]
42. Jeffrey, H. Chaos Game Representation of Gene Structure. *Nucleic Acids Res.* **1990**, *18*, 2163–2170. [[CrossRef](#)] [[PubMed](#)]
43. Lin, J.; Adjeroh, D.; Jiang, B.H.; Jiang, Y. K_2 and K_2^* : Efficient alignment-free sequence similarity measurement based on Kendall statistics. *Bioinformatics* **2018**, *34*, 1682–1689. [[CrossRef](#)]
44. Li, W. The study of correlation structures of DNA sequences: A critical review. *Comput. Chem.* **1997**, *21*, 257–271. [[CrossRef](#)]
45. Peng, C.K.; Buldyrev, S.V.; Goldberger, A.L.; Havlin, S.; Sciortino, F.; Simons, M.; Stanley, H.E. Long-Range Correlations in Nucleotide Sequences. *Nature* **1992**, *356*, 168–170. [[CrossRef](#)]
46. Voss, R. Evolution of long-range fractal correlations and $1/f$ noise in DNA base sequences. *Phys. Rev. A* **1992**, *68*, 3805–3808. [[CrossRef](#)]
47. Deng, M.; Yu, C.; Liang, Q.; He, R.L.; Yau, S.S.T. A Novel Method of Characterizing Genetic Sequences: Genome Space with Biological Distance and Applications. *PLoS ONE* **2011**, *6*, e17293. [[CrossRef](#)]
48. Li, Y.; Tian, K.; Yin, C.; He, R.; Yau, S.T. Virus classification in 60-dimensional protein space. *Mol. Phylogenet. Evol.* **2016**, *99*, 53–62. [[CrossRef](#)] [[PubMed](#)]
49. Wang, Y.; Tian, K.; Yau, S. Proteine Sequence Classification using natural vectors and the convex hull method. *J. Comput. Biol.* **2019**, *26*, 315–321. [[CrossRef](#)]
50. Li, W. Mutual information functions versus correlation function. *J. Stat. Phys.* **1990**, *60*, 823–837. [[CrossRef](#)]
51. Herzel, H.; Grosse, I. Measuring correlations in symbol sequences. *Phys. A* **1995**, *216*, 518–542. [[CrossRef](#)]
52. Berryman, M.; Allison, A.; Abbott, D. Mutual information for examining correlations in DNA. *Fluct. Noise Lett.* **2004**, *4*, 237–246. [[CrossRef](#)]

53. Swati, D. Use of Mutual Information Function and Power Spectra for Analyzing the Structure of Some Prokaryotic Genomes. *Am. J. Math. Manag. Sci.* **2007**, *27*, 179–198. [[CrossRef](#)]
54. Bauer, M.; Schuster, S.; Sayood, K. The average mutual information profile as a genomic signature. *BMC Bioinform.* **2008**, *9*, 1–11. [[CrossRef](#)]
55. Gregori-Puigjané, E.; Mestres, J. SHED: Shannon Entropy Descriptors from Topological Feature Distributions. *J. Chem. Inf. Model.* **2006**, *46*, 1615–1622. [[CrossRef](#)] [[PubMed](#)]
56. Dehnert, M.; Helm, W.E.; Hütt, M.T. Information Theory Reveals Large-Scale Synchronisation of Statistical Correlations in Eukaryote Genomes. *Gene* **2005**, *345*, 81–90. [[CrossRef](#)]
57. Grosse, I.; Herzel, H.; Buldyrev, S.V.; Stanley, H.E. Species Independence of Mutual Information in Coding and Noncoding DNA. *Phys. Rev. E* **2000**, *61*, 5624–5629. [[CrossRef](#)]
58. Korber, B.T.; Farber, R.M.; Wolpert, D.H.; Lapedes, A.S. Covariation of Mutations in the V3 Loop of Human Immunodeficiency Virus Type 1 Envelope Protein: An Information Theoretic Analysis. *Proc. Natl. Acad. Sci. USA* **1993**, *90*, 7176–7180. [[CrossRef](#)]
59. Lichtenstein, F.; Antoneli, F.; Briones, M.R.S. MIA: Mutual Information Analyzer, a Graphic User Interface Program That Calculates Entropy, Vertical and Horizontal Mutual Information of Molecular Sequence Sets. *BMC Bioinform.* **2015**, *16*, 409. [[CrossRef](#)]
60. Nalbantoglu, Ö.U.; Russell, D.J.; Sayood, K. Data Compression Concepts and Algorithms and Their Applications to Bioinform. *Entropy* **2010**, *12*, 34–52. [[CrossRef](#)] [[PubMed](#)]
61. Shannon, C. A mathematical theory of communication. *Bell Syst. Tech. J.* **1948**, *27*, 379–432. [[CrossRef](#)]
62. Rényi, A. On measures of entropy and information. In Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability, Berkeley, CA, USA, 20–30 July 1960; Neyman, J., Ed.; University of California Press: Berkeley, CA, USA, 1961.
63. Rényi, A. *Probability Theory*; North-Holland Publishing Company: Amsterdam, The Netherlands, 1970.
64. Tsallis, C. Possible generalization of Boltzmann–Gibbs statistics. *J. Math. Phys.* **1988**, *52*, 479–487.
65. Sparavigna, A. Mutual Information and Nonadditive Entropies: The Case of Tsallis Entropy. *Int. J. Sci.* **2015**, *4*. [[CrossRef](#)]
66. Villmann, T.; Geweniger, T. Multi-class and Cluster Evaluation Measures Based on Rényi and Tsallis Entropies and Mutual Information. In *Proceedings of the 17th International Conference on Artificial Intelligence and Soft Computing-ICAISC, Zakopane*; Rutkowski, L., Scherer, R., Korytkowski, M., Pedrycz, W., Tadeusiewicz, R., Zurada, J., Eds.; Springer International Publishing: Cham, Switzerland, 2018; LNCS 10841, pp. 724–735. [[CrossRef](#)]
67. Vinga, S.; Almeida, J. Local Rényi entropic profiles of DNA sequences. *BMC Bioinform.* **2007**, *8*, 1–19. [[CrossRef](#)] [[PubMed](#)]
68. Vinga, S.; Almeida, J. Rényi continuous entropy of DNA sequences. *J. Theor. Biol.* **2004**, *231*, 377–388. [[CrossRef](#)]
69. Delgado-Soler, L.; Toral, R.; Tomás, M.; Robio-Martinez, J. RED: A Set of Molecular Descriptors Based on Rényi Entropy. *J. Chem. Inf. Model.* **2009**, *49*, 2457–2468. [[CrossRef](#)]
70. Papapetrou, M.; Kugiumtzis, D. Tsallis conditional mutual information in investigating long range correlation in symbol sequences. *Phys. A* **2020**, *540*, 1–13. [[CrossRef](#)]
71. Gao, Y.; Luo, L. Genome-based phylogeny of dsDNA viruses by a novel alignment-free method. *Gene* **2012**, *492*, 309–314. [[CrossRef](#)]
72. Schneider, P.; Biehl, M.; Hammer, B. Adaptive Relevance Matrices in Learning Vector Quantization. *Neural Comput.* **2009**, *21*, 3532–3561. [[CrossRef](#)]
73. Saralajew, S.; Holdijk, L.; Villmann, T. Fast Adversarial Robustness Certification of Nearest Prototype Classifiers for Arbitrary Seminorms. In Proceedings of the 34th Conference on Neural Information Processing Systems (NeurIPS 2020), Virtual-only Conference, 6–12 December 2020; Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M., Lin, H., Eds.; Curran Associates, Inc.: Red Hook, NY, USA, 2020; Volume 33, pp. 13635–13650.
74. Cichocki, A.; Amari, S. Families of Alpha- Beta- and Gamma-Divergences: Flexible and Robust Measures of Similarities. *Entropy* **2010**, *12*, 1532–1568. [[CrossRef](#)]
75. Mackay, D. *Inf. Theory, Inference Learn. Algorithms*; Cambridge University Press: Cambridge, UK, 2003.
76. Kullback, S.; Leibler, R. On information and sufficiency. *Ann. Math. Stat.* **1951**, *22*, 79–86. [[CrossRef](#)]
77. Kantz, H.; Schreiber, T. *Nonlinear Time Series Analysis*; Cambridge Nonlinear Science Series; Cambridge University Press: Cambridge, UK, 1997; Volume 7.
78. Fraser, A.; Swinney, H. Independent coordinates for strange attractors from mutual information. *Phys. Rev. A* **1986**, *33*, 1134–1140. [[CrossRef](#)]
79. Li, W. *Mutual Information Functions of Natural Language Texts*; Technical Report SFI-89-10-008; Santa Fe Institute: Santa Fe, NM, USA, 1989.
80. Golub, G.; Loan, C.V. *Matrix Computations*, 4th ed.; Johns Hopkins Studies in the Mathematical Sciences; John Hopkins University Press: Baltimore, MD, USA, 2013.
81. Horn, R.; Johnson, C. *Matrix Analysis*, 2nd ed.; Cambridge University Press: Cambridge, UK, 2013.
82. Erdogmus, D.; Principe, J.; II, K.H. Beyond second-order statistics for learning: A pairwise interaction model for entropy estimation. *Nat. Comput.* **2002**, *1*, 85–108. [[CrossRef](#)]
83. Hild, K.; Erdogmus, D.; Principe, J. Blind Source Separation Using Rényi’s Mutual Information. *IEEE Signal Process. Lett.* **2001**, *8*, 174–176. [[CrossRef](#)]

84. Jenssen, R.; Principe, J.; Erdogmus, D.; Eltoft, T. The Cauchy-Schwarz divergence and Parzen windowing: Connections to graph theory and Mercer kernels. *J. Frankl. Inst.* **2006**, *343*, 614–629. [[CrossRef](#)]
85. Lehn-Schiöler, T.; Hegde, A.; Erdogmus, D.; Principe, J. Vector quantization using information theoretic concepts. *Nat. Comput.* **2005**, *4*, 39–51. [[CrossRef](#)]
86. Principe, J. *Information Theoretic Learning*; Springer: Heidelberg, Germany, 2010.
87. Singh, A.; Principe, J. Information theoretic learning with adaptive kernels. *Signal Process.* **2011**, *91*, 203–213. [[CrossRef](#)]
88. Villmann, T.; Haase, S. Divergence based vector quantization. *Neural Comput.* **2011**, *23*, 1343–1392. [[CrossRef](#)] [[PubMed](#)]
89. Mwebaze, E.; Schneider, P.; Schleif, F.M.; Aduwo, J.; Quinn, J.; Haase, S.; Villmann, T.; Biehl, M. Divergence based classification in Learning Vector Quantization. *Neurocomputing* **2011**, *74*, 1429–1435. [[CrossRef](#)]
90. Bunte, K.; Haase, S.; Biehl, M.; Villmann, T. Stochastic Neighbor Embedding (SNE) for Dimension Reduction and Visualization Using Arbitrary Divergences. *Neurocomputing* **2012**, *90*, 23–45. [[CrossRef](#)]
91. Csiszár, I. Axiomatic Characterization of Information Measures. *Entropy* **2008**, *10*, 261–273. [[CrossRef](#)]
92. Fehr, S.; Berens, S. On the Conditional Rényi Entropy. *IEEE Trans. Inf. Theory* **2014**, *60*, 6801–6810. [[CrossRef](#)]
93. Teixeira, A.; Matos, A.; Antunes, L. Conditional Rényi Entropies. *IEEE Trans. Inf. Theory* **2012**, *58*, 4273–4277. [[CrossRef](#)]
94. Iwamoto, M.; Shikata, J. *Revisiting Conditional Rényi Entropies and Generalizing Shannons Bounds in Information Theoretically Secure Encryption*; Technical Report; Cryptology ePrint Archive 440/2013; International Association for Cryptologic Research (IACR): Lyon, France, 2013.
95. Ilić, V.; Djordjević, I.; Stanković, M. On a General Definition of Conditional Rényi Entropies. In Proceedings of the 4th International Electronic Conference on Entropy and Its Application (ECEA 2017), Online, 21 November–1 December 2017; Volume 2, pp. 1–6. [[CrossRef](#)]
96. Jizba, P.; Arimitsu, T. The world according to Rényi: Thermodynamics of multifractal systems. In *AIP Conference Proceedings*; American Institute of Physics (AIP): Ellipse College Park, MD, USA, 2001; Volume 597, pp. 341–348. [[CrossRef](#)]
97. Cai, C.; Verdú, S. Conditional Rényi divergence saddlepoint and the maximization of α -mutual information. *Entropy* **2020**, *21*, 316. [[CrossRef](#)]
98. Havrda, J.; Charvát, F. Quantification method of classification processes: Concept of structural α -entropy. *Kybernetika* **1967**, *3*, 30–35.
99. Vila, M.; Bardera, A.; Sbert, M.F.M. Tsallis Mutual Information for Document Classification. *Entropy* **2011**, *13*, 1694–1707. [[CrossRef](#)]
100. Kohonen, T. Learning Vector Quantization. *Neural Networks* **1988**, *1*, 303.
101. Kohonen, T. *Self-Organizing Maps*; Springer Series in Information Sciences; Springer: Berlin/Heidelberg, Germany, 1995; Volume 30.
102. Biehl, M.; Hammer, B.; Villmann, T. Prototype-based Models for the Supervised Learning of Classification Schemes. *Proc. Int. Astron. Union* **2017**, *12*, 129–138. [[CrossRef](#)]
103. Sato, A.; Yamada, K. Generalized learning vector quantization. In *Advances in Neural Information Processing Systems 8, Proceedings of the 1995 Conference*; Touretzky, D.S., Mozer, M.C., Hasselmo, M.E., Eds.; MIT Press: Cambridge, MA, USA, 1996; pp. 423–429.
104. Bunte, K.; Schneider, P.; Hammer, B.; Schleif, F.M.; Villmann, T.; Biehl, M. Limited Rank Matrix Learning, discriminative dimension reduction and visualization. *Neural Netw.* **2012**, *26*, 159–173. [[CrossRef](#)]
105. Villmann, T.; Bohnsack, A.; Kaden, M. Can Learning Vector Quantization be an Alternative to SVM and Deep Learning? *J. Artif. Intell. Soft Comput. Res.* **2017**, *7*, 65–81. [[CrossRef](#)]
106. Hammer, B.; Villmann, T. Generalized Relevance Learning Vector Quantization. *Neural Netw.* **2002**, *15*, 1059–1068. [[CrossRef](#)]
107. Biehl, M.; Hammer, B.; Villmann, T. Prototype-based models in machine learning. *Wiley Interdiscip. Rev. Cogn. Sci.* **2016**, *7*, 92–111. [[CrossRef](#)]
108. Cramer, K.; Gilad-Bachrach, R.; Navot, A.; Tishby, A. Margin analysis of the LVQ algorithm. In *Advances in Neural Information Processing (Proc. NIPS 2002)*; Becker, S., Thrun, S., Obermayer, K., Eds.; MIT Press: Cambridge, MA, USA, 2003; Volume 15, pp. 462–469.
109. Garant, J.M.; Perreault, J.P.; Scott, M.S. Motif Independent Identification of Potential RNA G-Quadruplexes by G4RNA Screener. *Bioinformatics* **2017**, *33*, 3532–3537. [[CrossRef](#)]
110. Garant, J.M.; Luce, M.J.; Scott, M.S.; Perreault, J.P. G4RNA: An RNA G-Quadruplex Database. *Database* **2015**, *2015*. [[CrossRef](#)] [[PubMed](#)]
111. Wen, J.; Liu, Y.; Shi, Y.; Huang, H.; Deng, B.; Xiao, X. A Classification Model for lncRNA and mRNA Based on K-Mers and a Convolutional Neural Network. *BMC Bioinform.* **2019**, *20*, 469. [[CrossRef](#)] [[PubMed](#)]
112. Frankish, A.; Diekhans, M.; Ferreira, A.M.; Johnson, R.; Jungreis, I.; Loveland, J.; Mudge, J.M.; Sisu, C.; Wright, J.; Armstrong, J.; et al. GENCODE Reference Annotation for the Human and Mouse Genomes. *Nucleic Acids Res.* **2019**, *47*, D766–D773. [[CrossRef](#)]
113. Forster, P.; Forster, L.; Renfrew, C.; Forster, M. Phylogenetic Network Analysis of SARS-CoV-2 Genomes. *Proc. Natl. Acad. Sci. USA* **2020**, *117*, 9241–9243. [[CrossRef](#)] [[PubMed](#)]
114. Liu, L.; Ho, Y.k.; Yau, S. Clustering DNA Sequences by Feature Vectors. *Mol. Phylogenet. Evol.* **2006**, *41*, 64–69. [[CrossRef](#)] [[PubMed](#)]
115. Cornish-Bowden, A. Nomenclature for Incompletely Specified Bases in Nucleic Acid Sequences: Recommendations 1984. *Nucleic Acids Res.* **1985**, *13*, 3021–3030. [[CrossRef](#)]

116. Yu, C.; Hernandez, T.; Zheng, H.; Yau, S.C.; Huang, H.H.; He, R.L.; Yang, J.; Yau, S.S.T. Real Time Classification of Viruses in 12 Dimensions. *PLoS ONE* **2013**, *8*, e64328. [[CrossRef](#)]
117. Blaisdell, B.E. Average Values of a Dissimilarity Measure Not Requiring Sequence Alignment Are Twice the Averages of Conventional Mismatch Counts Requiring Sequence Alignment for a Variety of Computer-Generated Model Systems. *J. Mol. Evol.* **1989**, *29*, 538–547. [[CrossRef](#)]
118. Goldberg, Y. Neural Network Methods for Natural Language Processing. *Synth. Lect. Hum. Lang. Technol.* **2017**, *10*, 1–309. [[CrossRef](#)]
119. Kaden, M.; Bohnsack, K.S.; Weber, M.; Kudła, M.; Gutowska, K.; Blazewicz, J.; Villmann, T. Learning Vector Quantization as an Interpretable Classifier for the Detection of SARS-CoV-2 Types Based on Their RNA Sequences. *Neural Comput. Appl.* **2021**, 1–12. [[CrossRef](#)]
120. Riley, P. Three pitfalls to avoid in machine learning. *Nature* **2019**, *572*, 27–29. [[CrossRef](#)] [[PubMed](#)]
121. Todd, A.K.; Johnston, M.; Neidle, S. Highly prevalent putative quadruplex sequence motifs in human DNA. *Nucleic Acids Res.* **2005**, *33*, 2901–2907. [[CrossRef](#)] [[PubMed](#)]
122. Csiszár, I. Information-type measures of differences of probability distributions and indirect observations. *Studia Sci. Math. Hungaria* **1967**, *2*, 299–318.
123. Hnizdo, V.; Tan, J.; Killian, B.J.; Gilson, M.K. Efficient Calculation of Configurational Entropy from Molecular Simulations by Combining the Mutual-Information Expansion and Nearest-Neighbor Methods. *J. Comput. Chem.* **2008**, *29*, 1605–1614. [[CrossRef](#)]
124. Kolekar, P.; Kale, M.; Kulkarni-Kale, U. Alignment-Free Distance Measure Based on Return Time Distribution for Sequence Analysis: Applications to Clustering, Molecular Phylogeny and Subtyping. *Mol. Phylogenet. Evol.* **2012**, *65*, 510–522. [[CrossRef](#)]
125. Wei, D.; Jiang, Q.; Wei, Y.; Wang, S. A Novel Hierarchical Clustering Algorithm for Gene Sequences. *BMC Bioinform.* **2012**, *13*, 174. [[CrossRef](#)]
126. Li, M.; Chen, X.; Li, X.; Ma, B.; Vitanyi, P. The Similarity Metric. *IEEE Trans. Inf. Theory* **2004**, *50*, 3250–3264. [[CrossRef](#)]
127. Yin, C.; Chen, Y.; Yau, S.S.T. A Measure of DNA Sequence Similarity by Fourier Transform with Applications on Hierarchical Clustering. *J. Theor. Biol.* **2014**, *359*, 18–28. [[CrossRef](#)]
128. Bao, J.; Yuan, R. A Wavelet-Based Feature Vector Model for DNA Clustering. *Genet. Mol. Res.* **2015**, *14*, 19163–19172. [[CrossRef](#)] [[PubMed](#)]
129. Berger, J.A.; Mitra, S.K.; Carli, M.; Neri, A. New Approaches to Genome Sequence Analysis Base Don Digital Signal Processing. In Proceedings of IEEE Workshop on Genomic Signal Processing and Statistics (GENSIPS), Raleigh, NC, USA, 12–13 October 2002; p. 4.
130. Almeida, J.S.; Vinga, S. Universal Sequence Map (USM) of Arbitrary Discrete Sequences. *BMC Bioinform.* **2002**, *3*, 1–11. [[CrossRef](#)] [[PubMed](#)]
131. Vellido, A. The importance of interpretability and visualization in machine learning for applications in medicine and health care. *Neural Netw. Appl.* **2020**, *32*, 18069–18083. [[CrossRef](#)]
132. Bittrich, S.; Kaden, M.; Leberecht, C.; Kaiser, F.; Villmann, T.; Labudde, D. Application of an Interpretable Classification Model on Early Folding Residues during Protein Folding. *Biodata Min.* **2019**, *12*, 1. [[CrossRef](#)] [[PubMed](#)]
133. Fischer, L.; Hammer, B.; Wersing, H. Efficient rejection strategies for prototype-based classification. *Neurocomputing* **2015**, *169*, 334–342. [[CrossRef](#)]
134. Villmann, A.; Kaden, M.; Saralajew, S.; Villmann, T. Probabilistic Learning Vector Quantization with Cross-Entropy for Probabilistic Class Assignments in Classification Learning. In *Proceedings of the 17th International Conference on Artificial Intelligence and Soft Computing-ICAISC, Zakopane, Zakopane, Poland, 3–7 June 2018*; Rutkowski, L., Scherer, R., Korytkowski, M., Pedrycz, W., Tadeusiewicz, R., Zurada, J., Eds.; Springer International Publishing: Cham, Switzerland, 2018; LNCS 10841, pp. 736–749. [[CrossRef](#)]
135. Saralajew, S.; Holdijk, L.; Rees, M.; Asan, E.; Villmann, T. Classification-by-Components: Probabilistic Modeling of Reasoning over a Set of Components. In Proceedings of the 33rd Conference on Neural Information Processing Systems (NeurIPS 2019), Vancouver, BC, Canada, 8–14 December 2019; MIT Press: Cambridge, MA, USA, 2019; pp. 2788–2799.