Physica A 585 (2022) 126427

Contents lists available at ScienceDirect

Physica A

journal homepage: www.elsevier.com/locate/physa

# Generalized Poisson ensemble

Rongrong Xie [a,*], Shengfeng Deng [a], Weibing Deng [a], Mauricio P. Pato [b]

[a] Key Laboratory of Quark and Lepton Physics (MOE) and Institute of Particle Physics, Central China Normal University, Wuhan 430079, China
[b] Instituto de Física, Universidade de São Paulo, Caixa Postal 66318, 05314-970 São Paulo, S.P., Brazil

ARTICLE INFO

*Article history:*
Received 25 January 2021
Received in revised form 27 April 2021
Available online 15 September 2021

*Keywords:*
Random matrix theory
Generalized Poisson ensemble
Nearest-neighbor distribution
Number variance

ABSTRACT

A generalized Poisson ensemble is constructed using the maximum entropy principle based on the non-extensive entropy. It is found that the correlations which are introduced among the eigenvalues lead to statistical distributions with heavy tails. As a consequence, long-range statistics, measured by the number variance, show super-Poissonian behavior and the short-range ones, measured by the nearest-neighbor-distribution show, with respect to Poisson, enhancement at small and large separations. Potential applications were found for the sequence data of protein and DNA, which display good agreement with the model. In particular, the ensuing parameter $\lambda$ of the generalized Poisson ensemble can be utilized to facilitate protein classification.



## 1. Introduction

Few years after Wigner's proposal of the now-famous ensemble of Gaussian random matrices [1,2], it was shown by Balian [3] that it can be obtained from a maximum entropy principle (MEP) based on the Shannon entropy. Several decades later, the so-called Tsallis entropy [4] became popular and, despite the various applications it has found, also controversial [5,6]. It has then become natural that, at some point, the generalized maximum entropy principle (GMEP) based on the new entropy would be used to generate random matrices. In fact, this was done independently by two groups [7,8] and several developments have followed the generation of the new ensemble via the GMEP [9–13]. As it could have been expected, the statistical measures of the new ensemble are distributions with heavy tails [8]. Another important feature of the generalized ensemble is its preservation of the symmetry under the unitary transformation of the Gaussian ensemble that leads to independence between the eigenvalue and the eigenvector distributions.

The dominant spectral characteristic of the Wigner Gaussian ensemble is the strong correlations among eigenvalues. On the other hand, at the opposite extreme, stays the Poisson ensemble, in which levels behave independently. At the beginning of the 80s, the Bohigas–Giannoni–Schmit conjecture stated that chaotic systems have spectra similar to those of the Gaussian ensemble while regular systems follow Poisson [14]. This conjecture has been corroborated by a large amount of evidence and, besides, with an analytic proof [15]. This shows that, in random matrix theory (RMT), the Poisson ensemble appears on an equal footing with the Gaussian ensemble. It is therefore entirely justified to investigate the correlations that are introduced among eigenvalues when a generalized Poisson ensemble is constructed by using the GMEP. As a matter of fact, this is something that is due in the GMEP program to generate random matrices.

As will be seen in Section 2, the behavior of the family of ensembles generated depends on the domain of variation of the Tsallis entropic parameter $q$. Thus, if $q < 1$, the matrix elements occupy a compact support, while for $q > 1$, their

* Corresponding author.
*E-mail addresses:* emilyxierr@gmail.com (R. Xie), gitsteven@gmail.com (S. Deng), wdeng@mail.ccnu.edu.cn (W. Deng), mpato@if.usp.br (M.P. Pato).

https://doi.org/10.1016/j.physa.2021.126427
0378-4371/

domain of variation is the whole real axis. Considering the limit of large matrix size $N$, for $q < 1$, the standard Poisson case is recovered when $N \to \infty$, but, for $q > 1$, this same limit can be combined with the limit $q \to 1$ in such a way that a new family of ensembles is defined.

The statistics and correlations, as well as the statistical physics of DNA and protein sequences, have been studied over the past few decades [16–18]. In recent years, RMT methods have also led to great success in the understanding of modular organizations of cell-constituent biological networks [19]. However, the distributions of spacings [16] in Protein and DNA sequences are not yet well characterized. It is gratifying to show that the results obtained in this paper can be applied to the sequence data of protein and DNA, which once again demonstrate the efficacy of RMT methods in distinguishing system-specific, nonrandom features from random noise in complex systems. This suggests that RMT methods provide an alternative avenue for the comparisons of biological sequence data.

The remainder of this paper is organized as follows: In the following section, we derive the generalized Poisson ensemble via the GMEP, in which the cases for $q < 1$ and $q > 1$ are treated separately. Section 3 features the applications of the generalized Poisson distribution to the sequence data of protein and DNA. Finally, Section 4 summarizes this work and provides a brief outlook.

## 2. The generalized Poisson ensemble

An ensemble of diagonal matrices $H$ whose elements are sorted from a normal distribution is a representation of the Poisson ensemble. The joint distribution of matrix elements can be written as

$$P_{GN}(H) = \frac{1}{(2\pi)^{N/2}} \exp\left(-\frac{1}{2}\mathrm{tr}H^2\right), \tag{1}$$

where the subscript "$G$" reflects the Gaussian nature of the matrix elements and "$N$" denotes the size of the matrix. From Eq. (1) we immediately deduce the density

$$\rho_G(x) = \frac{N}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}x^2\right), \tag{2}$$

and hence the unfold variable reads $s = \mathcal{N}_G(x) = \frac{N}{2}\mathrm{erf}\left(\frac{x}{\sqrt{2}}\right)$. Now we can state the main results of the Poisson ensemble. First, the probability that the interval $(-\frac{\theta}{2}, \frac{\theta}{2})$ is empty in the unfolded spectrum $s = 2\mathcal{N}_G(\frac{\theta}{2})$ is given by

$$E_G(s) = \exp(-s) = \exp\left[-2\mathcal{N}_G(\frac{\theta}{2})\right]. \tag{3}$$

Second, the fluctuations in the number of eigenvalues in the interval $(-\frac{\theta}{2}, \frac{\theta}{2})$ is characterized by the number variance $\langle n^2\rangle - \langle n\rangle^2$ which in terms of the unfolded interval length $L = 2\mathcal{N}_G(\frac{\theta}{2})$ is given by

$$\langle n^2\rangle = 2\mathcal{N}_G(\frac{\theta}{2}) + 4\mathcal{N}_G^2(\frac{\theta}{2}) = L + L^2 \quad \text{and} \quad \langle n\rangle = 2\mathcal{N}_G(\frac{\theta}{2}) = L\,, \tag{4}$$

such that the linear number variance of Poisson follows.

The above joint distribution can be easily derived by maximizing the Shannon entropy subjected to normalization and to the moment constraint $\langle \mathrm{tr}H^2\rangle = N$. The Shannon entropy can be generalized by the expression

$$S_q = \frac{1 - \int dH P^q(H)}{q-1} \tag{5}$$

such that, when the real parameter $q$ goes to one, the Shannon entropy is recovered. This is the Tsallis entropy and we want to maximize it subjected to the normalization constraint and to a moment constraint that in the GMEP formalism, is performed with respect to the escort distribution $P^q(H)$ as [20]

$$\int dH P^q(H)\mathrm{tr}H^2 - \mu \int dH P^q(H) = 0\,. \tag{6}$$

Constructing then the functional

$$S\{P(H)\} = S_q - \alpha_0 \int dH P(H) - \alpha\left[\int dH P^q(H)\mathrm{tr}H^2 - \mu \int dH P^q(H)\right], \tag{7}$$

where $\alpha_0$ and $\alpha$ are Lagrangian multipliers, and by imposing $\delta S\{P(H)\} = 0$, it is obtained that

$$P(H) \sim \left(\frac{1}{1-q} + \alpha\mu - \alpha\mathrm{tr}H^2\right)^{\frac{1}{1-q}}. \tag{8}$$

Comparing to Eq. (1), the immediate consequence of this expression is that the matrix elements, i.e. the eigenvalues, are not any more independent variables such that the correlation among them must be taken into account. Before proceeding

to show how this can be done, it is convenient to consider separately the cases $q \leq 1$ and $q \geq 1$. We emphasize that $H$ remains diagonal throughtout the above derivation. Consequently there is no Vandermonde term in the joint eigenvalue distribution, giving rise to a great simplification to the problem, as will be seen in what follows.

### 2.1. The case $q \leq 1$

Using the moment constraint, we find that $\alpha = \frac{N}{2\mu}$, and with $\mu = N$, we have

$$P_N(H) = \frac{\Gamma(\frac{1}{1-q} + \frac{N}{2} + 1)}{\left[2(\frac{1}{1-q} + \frac{N}{2})\pi\right]^{N/2} \Gamma(\frac{1}{1-q} + 1)} \left[1 - \frac{\mathrm{tr}H^2}{2(\frac{1}{1-q} + \frac{N}{2})}\right]^{\frac{1}{1-q}}, \tag{9}$$

which satisfies the condition that when $q \to 1$, $P_N(H) \to P_{GN}(H)$. Defining the quantities

$$\lambda_k = \frac{1}{1-q} + \frac{k}{2} \tag{10}$$

and

$$Q_k = 1 - \frac{1}{2\lambda_N} \sum_{i=1}^{k} x_i^2 , \tag{11}$$

the above distribution can be written as

$$P_N(H) = \frac{\Gamma(\lambda_N + 1) Q_N^{\lambda_0}}{(2\lambda_N \pi)^{N/2} \Gamma(\lambda_0 + 1)} . \tag{12}$$

By integrating out $N - k$ variables, the $k$-point correlation [2] that gives the probability of having $k$ eigenvalues at the positions $x_1, x_2, \ldots, x_k$ is found to be expressed as

$$R_k(x_1, \ldots, x_k) = \frac{N!}{(N-k)!} \frac{\Gamma(\lambda_N + 1) Q_k^{\lambda_{N-k}}}{(2\lambda_N \pi)^{k/2} \Gamma(\lambda_{N-k} + 1)} = \frac{N!}{(N-k)!} P_k(x_1, \ldots, x_k) . \tag{13}$$

These $n$-point functions can be used to generate a correlated sequence of the $N$ eigenvalues by writing the identity

$$P(x_1, \ldots, x_N) = P(x_1) \left[\frac{P(x_1, x_2)}{P(x_1)}\right] \left[\frac{P(x_1, x_2, x_3)}{P(x_1, x_2)}\right] \ldots \left[\frac{P(x_1, \ldots, x_N)}{P(x_1, \ldots, x_{N-1})}\right] , \tag{14}$$

in which each term can be interpreted as the conditional probability of sorting a new value once the previous ones have been sorted. We remark that all these conditional probabilities are univariate density distributions. Explicitly, once the set of $k - 1$ variables have been determined, the next one, $x_k$ with $k > 1$, is given by

$$x_k = \pm \sqrt{2\lambda_N Q_{k-1} t} , \tag{15}$$

where $t$ is sorted from the beta distribution

$$f(t; \frac{1}{2}, \lambda_{N-k} + 1) \tag{16}$$

and the signs $\pm$ are chosen with equal probability. With $Q_0 = 1$, these expressions are also valid for $k = 1$.

In the limit of large $N$, the density $R_1(x)$ rapidly approaches the Gaussian distribution of the original Poisson ensemble as $N$ increases, which is also followed by the two-point correlation that becomes the product of two independent Gaussians. As a consequence, statistical measures like nearest-neighbor-distribution (NND) and number variance (NV) that depends on the two-point correlation function become Poissonian, that is, exponential and linear, respectively (a result corroborated by numerical simulations). Notwithstanding, higher-order correlations that involve a number $k$ of points of order $N$ are preserved, since then the quantities $\frac{1}{2\lambda_N} \sum_{i=1}^{k} x_i^2$ cannot be treated as small. In particular, the joint distribution of matrix elements, Eq. (9), does not factorize as a product of the same individual functions.

### 2.2. The case $q \geq 1$

Considering now $q \geq 1$, the joint distribution of the matrix elements becomes the correlated distribution

$$P_N(H) = \frac{\Gamma(\frac{1}{q-1})}{(2\lambda\pi)^{N/2} \Gamma(\lambda)} \left(1 + \frac{\mathrm{tr}H^2}{2\lambda}\right)^{\frac{1}{1-q}} , \tag{17}$$

where the parameter

$$\lambda = \frac{1}{q-1} - \frac{N}{2} \tag{18}$$

was introduced. It is easily verified that the condition $q \to 1$, $P_N(H) \to P_{GN}(H)$ is again satisfied. In contrast to the case $q < 1$, in which the limit $N \to \infty$ leads back to the original Gaussian situation, there is the possibility of taking that limit by concomitantly making $q \to 1$ in such a way that the parameter $\lambda$ is kept fixed. This parameter therefore defines in the limit of large $N$ the generalized Poisson ensemble.

As above, defining

$$\hat{Q}_k = 1 + \frac{1}{2\lambda}\sum_{i=1}^{k} x_i^2 \,, \tag{19}$$

with $\hat{Q}_0 = 1$, all the eigenvalues, $x_k$, sequentially can be obtained as

$$\frac{x_k}{\sqrt{2\hat{Q}_{k-1}\lambda}} = \pm\sqrt{\frac{t}{1-t}} \,, \tag{20}$$

where $t$ is sorted from the beta distribution

$$f\left(t; \lambda + \frac{k-1}{2}, \frac{1}{2}\right) . \tag{21}$$

Alternatively, using the integral representation of the gamma function, we have the more manageable expression

$$P_N(H) = \frac{1}{(2\lambda\pi)^{N/2}\,\Gamma(\lambda)} \int_0^\infty d\xi \exp\left(-\xi - \frac{\xi \,\mathrm{tr} H^2}{2\lambda}\right) \xi^{\lambda+N/2-1} \,, \tag{22}$$

which shows that the statistics measures of the generalized ensemble are obtained by averaging the Gaussian measures with the distribution

$$w(\xi) = \frac{1}{\Gamma(\lambda)} \exp(-\xi)\xi^{\lambda-1} \,, \tag{23}$$

where $\lambda = \langle \xi \rangle = \bar{\xi}$. Also note that $\langle \sqrt{\xi} \rangle = \Gamma(\lambda + 1/2)/\Gamma(\lambda)$. Using then the above expressions, we obtain for the generalized Poisson ensemble the corresponding quantities as discussed at the beginning of this section for the Poisson ensemble. Starting from the density, the expression

$$\rho(x) = \frac{N}{\sqrt{2\pi}} \int_0^\infty d\xi\, w(\xi) \sqrt{\frac{\xi}{\lambda}} \exp\left(-\frac{\xi x^2}{2\lambda}\right) = \frac{N\Gamma(\lambda+1/2)}{\sqrt{2\pi\lambda}\Gamma(\lambda)} \left(1 + \frac{x^2}{2\lambda}\right)^{-\lambda-1/2} \tag{24}$$

is derived and the cumulative function reads

$$\mathcal{N}(x) = \int_0^x dy \rho(y) = \frac{N\Gamma(\lambda+1/2)\sqrt{x}}{2\sqrt{\pi\lambda}\Gamma(\lambda)} B\left(\frac{x^2}{2\lambda}; \frac{1}{2}, \lambda + \frac{1}{2}\right) , \tag{25}$$

where $B(t; a, b)$ is the incomplete beta function. For the probability that the interval $(-\frac{\theta}{2}, \frac{\theta}{2})$ in the middle of the spectrum has no levels, that is, the so-called gap probability, it is found that it is given by

$$E(s) = \int_0^\infty d\xi\, w(\xi) E_G\left[2\mathcal{N}_G\left(\sqrt{\frac{\xi}{\lambda}}\frac{\theta}{2}\right)\right] = \int_0^\infty \frac{d\xi}{\Gamma(\lambda)} \exp\left[-\xi - 2\mathcal{N}_G\left(\sqrt{\frac{\xi}{\lambda}}\frac{\theta}{2}\right)\right] \xi^{\lambda-1}, \tag{26}$$

where $s = 2\mathcal{N}(\theta/2)$ is the interval in the unfolded spectrum.

From Eqs. (2), (4) and (24), the variance $\langle n^2 \rangle - \langle n \rangle^2$ of the number of eigenvalues in the interval $(-\frac{\theta}{2}, \frac{\theta}{2})$ is given by

$$\begin{aligned} \Sigma^2(L) &= \int_0^\infty d\xi\, w(\xi) \left[2\mathcal{N}_G\left(\sqrt{\frac{\xi}{\lambda}}\frac{\theta}{2}\right) + 4\mathcal{N}_G^2\left(\sqrt{\frac{\xi}{\lambda}}\frac{\theta}{2}\right)\right] - L^2 \\ &= L + 4\int_0^\infty d\xi\, w(\xi)\, \mathcal{N}_G^2\left(\sqrt{\frac{\xi}{\lambda}}\frac{\theta}{2}\right) - L^2, \end{aligned} \tag{27}$$

where $L = 2\mathcal{N}(\frac{\theta}{2})$ is the unfolded interval.

As is usual in random matrix theory, the interest is in the asymptotic limit of large spectra, that is, in the expressions obtained when $N$ goes to infinity while keeping however the product $N\theta$ finite. Practically, this means to replace the cumulative functions by linear approximations that consist of multiplying their argument by the value of density at the origin, namely, $\mathcal{N}(x) \sim \rho(0)x = \frac{N\langle\sqrt{\xi}\rangle}{\sqrt{2\pi\lambda}}x$ and $\mathcal{N}_G(x) \sim \rho_G(0)x = \frac{N}{\sqrt{2\pi}}x$. Starting with the number variance, the parabolic expression

$$\Sigma^2(L) \simeq L + \left[\left(\frac{\sqrt{\lambda}}{\langle\sqrt{\xi}\rangle}\right)^2 - 1\right] L^2 \tag{28}$$

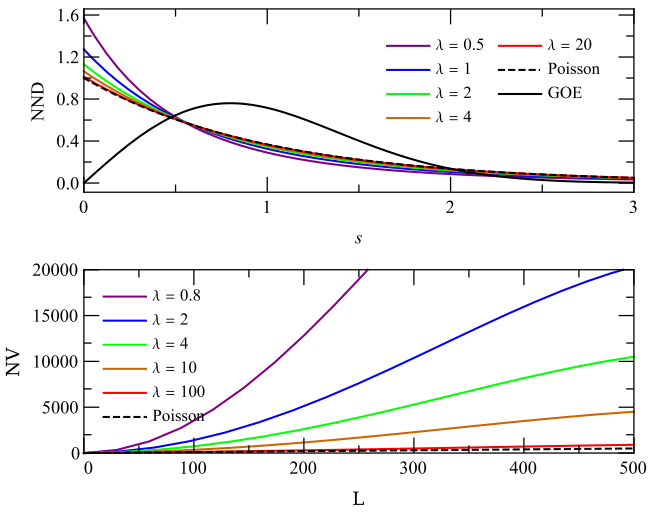


**Fig. 1.** The effect of the parameter λ on the generalized Poisson nearest-neighbor distribution and on its number variance. One can observe that when the parameter λ approaches to zero, a strong disorder regime is reached in which the local statistics show a power-law decay.

is derived in which the linear Poisson term competes with the square super-Poisson term.

Turning to the gap probability, using the approximation $s \simeq \frac{N\langle\sqrt{\xi}\rangle}{\sqrt{2\pi\lambda}}\theta$, Eq. (26) becomes

$$E(s) \simeq \frac{1}{\Gamma(\lambda)} \int_0^\infty d\xi \exp\left[-\xi - \frac{\sqrt{\xi}}{\langle\sqrt{\xi}\rangle} s\right] \xi^{\lambda-1}, \tag{29}$$

and from $p(s) = \frac{d^2E}{ds^2}$ the nearest-neighbor-distribution (NND) is given by

$$\begin{aligned} p(s) &\simeq \frac{1}{\Gamma(\lambda)\langle\sqrt{\xi}\rangle^2} \int_0^\infty d\xi \exp\left[-\xi - \frac{\sqrt{\xi}}{\langle\sqrt{\xi}\rangle} s\right] \xi^{\lambda} \\ &= \frac{k^2\Gamma(2\bar{\xi}+2)}{2^{\bar{\xi}}\Gamma(\bar{\xi})} \exp\left(\frac{k^2s^2}{8}\right) U\left(2\bar{\xi}+\frac{3}{2}, \frac{ks}{\sqrt{2}}\right), \end{aligned} \tag{30}$$

where $k = (\langle\sqrt{\xi}\rangle)^{-1}$ and $U(a,x)$ is the parabolic cylinder function [21]. Using the asymptotic form of $U(a,x)$ for $a \gg x$, that is, for $\lambda \gg s$, we find that $p(s) \simeq \exp(-s)$, which is the Poisson limit. We observe that, for the moments of $p(s)$, we have

$$\langle s^\gamma \rangle = \frac{\langle\sqrt{\xi}\rangle^{\gamma-1}\Gamma(1+\gamma)}{\Gamma(\lambda)} \Gamma\left(\frac{2\lambda+1-\gamma}{2}\right), \tag{31}$$

with which the usual NND normalizations, $\langle s^0\rangle = 1$ and $\langle s\rangle = 1$, are satisfied. Moreover, it shows that the distribution does not have moments for $\gamma > 2\lambda + 1$, a signature of a power-law decay for large $s$. Consistent with this result, at the origin we have the inequality

$$p(0) = \frac{\lambda}{\langle\sqrt{\xi}\rangle^2} = \frac{\Gamma(\lambda)\Gamma(\lambda+1)}{\Gamma^2(\lambda+1/2)} \geq 1 \tag{32}$$

that follows from the logarithmic convexity property of the gamma function. The above shows that, with respect to Poisson, there is enhancement at small and large separations.

To assess the effect of the parameter λ on the generalized Poisson ensemble, we plot the NND and the NV with respect to five λ values in Fig. 1. For comparison, the curves for the Poisson ensemble and the GOE are also displayed. As can be seen from the curves for both the NNDs and the NVs, the generalized Poisson approaches the Poisson when λ takes a large value. The NND is rather insensitive to the changes of λ for $\lambda > 1$. In stark contrast, the generalized Poisson drastically deviates from the Poisson and becomes super-Poissonian when λ approaches zero, indicating that smaller λ leads to a stronger disorder in the ensemble. In this respect, the parameter λ, originating from Eq. (23), describes the disorder intensity of the generalized Poisson ensemble.

## 3. Applications

RMT is a statistical theory of spectra [1]. As such, it can be applied to any sequence of points on a line that are not necessarily levels of a physical system. One of the first impressive success of RMT was its application to the statistics of

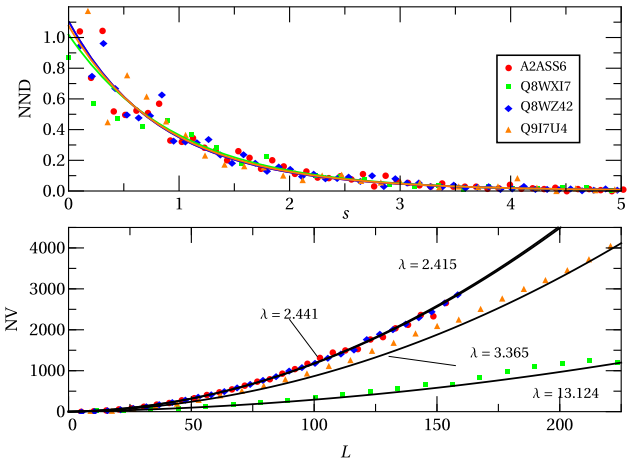


**Fig. 2.** The NNDs (top) and the NVs (bottom) of protein sequences. The NV curves are more separated and are hence utilized for fitting the parameter λ. Proteins with closer biological origins assume closer λ values.

the zeros of Riemann zeta function [2], which subsequently led to attempts of finding a Hamiltonian whose eigenvalues would be the zeros [22,23]. By removing punctuations from a text, it becomes a sequence of words separated by blanks that can be regarded as levels. Using RMT, two of us analyzed this spectrum of blanks as a tool to study the distributions of the lengths of words in literary texts of several languages [24]. Protein and DNA are also sequences of letters, and here we use the generalized Poisson model to investigate the spectra extracted from them. Different from what occurs with texts, Protein and DNA are sequences of letters without blanks. The spectra are then defined by choosing one given letter, as will be explained below, to play the role of blanks.

### 3.1. Protein

Protein sequences consist of 20 amino acids that are coded with corresponding letters. As an illustration, in Table 1 and Fig. 2, we present the analyses for four proteins from the Swiss-Prot protein database [25], the primary accession numbers of which are A2ASS6, Q8WZ42, Q8WXI7, and Q9I7U4 respectively. More results for a total number of 57 protein sequences are available in [26]. In order to obtain meaningful statistics, the lengths of the sequences being analyzed should be long enough. For example, the Titin protein with the accession numbers A2ASS6, Q8WZ42, and Q9I7U4, corresponding to the protein homologues in mouse, human and fruit fly, is the largest known protein [27]. All these three variants play similar protein functions: both A2ASS6 and Q8WZ42 are key components in the assembly and functioning of vertebrate striated muscles, while Q9I7U4 constitute the key component in the assembly and functioning of adult and embryonic striated muscles and muscle tendons. The Mucin-16 protein (also known as CA-125) with the accession number Q8WXI7 is another large protein, which plays a role in advancing tumorigenesis and tumor proliferation [28].

By exploiting the analogy of amino acids as words [29] in texts, it is natural to denote the most frequent amino acid of a protein as the blank. It is then straightforward to analyze the NNDs and the NVs of protein sequences as in the text sequence analyses [24]. In this respect, the NND of a sequence is just computed as the distribution of the reduced spacing $s = L_s/\langle L_s\rangle$ where $L_s$ denotes the length of spacing between two consecutive "blanks", while the NV simply measures the variance of the number of "blanks" contained in the interval of length $L$, averaged over all non-overlaping intervals taken from the sequence.

Fig. 2 suggests that the NNDs and the NVs of the protein sequences can be well fitted with the NND and NV [cf. Eqs. (28) and (30)] of the generalized Poisson ensemble. Note that both the NND and the NV only depend on the fitting parameter λ. The NND curves almost overlap entirely with each other regardless of their distinct λ values, thus the NNDs of the data can be considered as fluctuations around an average distribution for the generalized Poisson ensemble that is greater than one in the vicinity of the origin and displays a power-law decay in the tail. Therefore, as indicated at the end of Section 2.2, there is enhancement at small $s$ and large $s$. Furthermore, the parabolic behavior of the NV indicates that the NV of the generalized Poisson distribution is consistent with Taylor's law [30], which is caused by the fluctuation scaling mechanism [31–33]. These observations suggest that the NV is visually a better classifier. We can also observe that the NV of the generalized Poisson distribution has a super-Poissonian behavior that is characterized by a larger variance than the Poisson distribution; cf. Fig. 1.

Table 1 lists the average spacing length $\langle L_s\rangle$, the fitting parameter λ and the goodness-of-fit measure $R^2$, i.e. the coefficient of determination, for the corresponding sequence data. From Table 1 and Fig. 2, on the one hand, it can be inferred that the sequences corresponding to the Titin protein, i.e. A2ASS6, Q8WZ42, and Q9I7U4, give rise to similar values for the parameter λ. What is more, the sequences A2ASS6 and Q8WZ42 assume almost identical values for λ. This

**Table 1**
Parameter values of the protein and DNA sequence data. $\langle L_s \rangle$ is the average spacing length. $\lambda$ is the parameter for the NV of the generalized Poisson ensemble; see Eq. (28). $R^2_{\mathrm{NND}}$ and $R^2_{\mathrm{NV}}$ are measures for the fitting qualities of the NNDs and the NVs.

| Area | Sequence | $\langle L_s \rangle$ | $\lambda$ | $R^2_{\mathrm{NND}}$ | $R^2_{\mathrm{NV}}$ |
|---|---|---|---|---|---|
| Protein | A2ASS6 | 9.764 | 2.441 | 0.946 | 0.998 |
| | Q8WZ42 | 9.468 | 2.415 | 0.956 | 0.9997 |
| | Q9I7U4 | 5.663 | 3.365 | 0.901 | 0.996 |
| | Q8WXI7 | 4.478 | 13.124 | 0.978 | 0.961 |
| DNA | A2ASS6 | 17.063 | 2.45 | 0.953 | 0.997 |
| | Q8WZ42 | 13.937 | 2.03 | 0.945 | 0.9996 |
| | Q9I7U4 | 11.139 | 1.328 | 0.839 | 0.997 |
| | Q8WXI7 | 2.628 | 23.435 | 0.994 | 0.994 |

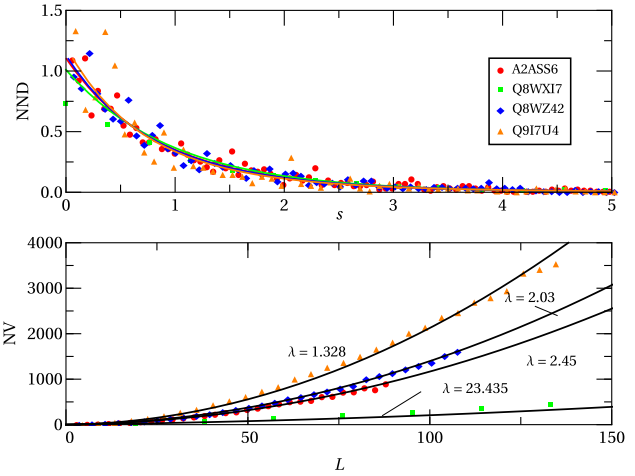


**Fig. 3.** The NNDs (top) and the NVs (bottom) for the DNA sequences corresponding to the proteins in Section 3.1.

remarkable distinction can be further ascribed to the fact that the sequences A2ASS6 and Q8WZ42 reside in mammals, while the sequence Q9I7U4 is the protein homologue in fruit fly, which should be more distant from those in mammals from an evolution point of view. On the other hand, the parameter $\lambda$ of Q8WXI7 is quite different from those of others since Q8WXI7 corresponds to a completely different kind of protein. We should remark that, in contrast to the case for text sequences [24], the average spacing length $\langle L_s \rangle$ shows a less discernible power in characterizing different groups of proteins.

### 3.2. DNA

DNA contains all the information that organisms need to live and reproduce themselves. This is realized through the so-called transcription–translation process for protein synthesis, in which DNA is first used as a template to produce mRNA, and then the genetic code in mRNA is translated to make a protein. Therefore we study the corresponding DNA sequences of the above proteins. The primary accession numbers for these DNA sequences are again A2ASS6, Q8WZ42, Q9I7U4, and Q8WXI7, which can be downloaded from [34]. There are four nucleotides (A,T,C,G) in DNA and thus we can consider the sequences are composed of four letters. We again selected the most frequent letter of a DNA sequence to form a spectrum of blanks and the corresponding NND and NV are analyzed in a similar manner as the previous subsection. More results for a total number of 57 DNA sequences, corresponding to the proteins in the previous subsection, are also available in [26].

Fig. 3 shows that the chosen DNA sequences display similar behaviors with respect to the corresponding protein sequences. Especially, from Table 1, we observe that the fitting parameter $\lambda$ of the DNA sequences led to the same grouping behavior as in the protein sequences. This is not at all surprising due to the maps between the DNA and the protein sequences. However, since these maps from DNAs to proteins are injective, the $\lambda$ values are distinct for each DNA and its corresponding protein.

These results indicate that this generalized Poisson spectral analysis works quite well in the case of protein classification, either directly through the protein sequences or through the corresponding DNA sequences, and therefore it could be served as a new method for protein analysis. In this classification scheme, the parameter $\lambda$ for the generalized Poisson

distribution is the signature characteristic for each protein. This is further corroborated by the analyses of more sequence data for protein and DNA in [26], which not only show that the NNDs and the NVs are well dictated by the generalized Poisson, but also demonstrate very diverse values for $\lambda$ from sequence to sequence. Furthermore, since all amino acids (or nucleotides) are statistically equal for most proteins (DNAs), one should expect that any given amino acid (or nucleotide) should distribute along the sequence as a generalized Poisson distribution.

## 4. Conclusion

In this work, a family of correlated ensembles was constructed by using the Tsallis non-extensive entropy. For values of the entropic parameter less than one, it was found that for large matrices, the NND and the NV are Poissonian, though higher-order correlations are not destroyed. Following Ref. [35], the case $q \geq 1$ can be interpreted as a situation in which an external source of randomness is superimposed to the Gaussian ones. This puts ensembles generated by GMEP in the context of disordered systems and sheds some light on the parabolic behavior of the number variance as a manifestation of the Taylor law caused by a fluctuation scaling mechanism. Regarding the NND, we remark that heavy tails already have been reported in studies of symbol frequency distributions [36]. We found that this generalized Poisson distribution has satisfactory applications to the distributions of protein and DNA sequences, with which the NNDs and the NVs of the data are adequately accounted for by the theoretical forms. Especially, the parameter $\lambda$ of the generalized Poisson ensemble provides a quite good characteristic for the classification of proteins and hence can be exploited to devise new methods for protein analysis. For instance, $\lambda$ can be considered as one of the reduced dimensions for certain clustering algorithms. This study also hints that disorders may prevail in many other systems that can be described by RMT and the generalized Poisson distribution may be applied to scenarios where the Poisson distribution fails.

## CRediT authorship contribution statement

**Rongrong Xie:** Conceptualization, Methodology, Software, Writing original draft. **Shengfeng Deng:** Software, Writing original draft, Formal analysis. **Weibing Deng:** Writing original draft, Supervision, Funding acquisition. **Mauricio P. Pato:** Conceptualization, Formal analysis, Writing original draft.

## Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

## Acknowledgments

M.P.P. was supported by grant 307807/2017-7 of the Brazilian agency Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq), Brazil and is a member of the Brazilian National Institute of Science and Technology-Quantum Information (INCT-IQ). This work was supported in part by the National Natural Science Foundation of China (Grant No. 11505071, 11747135, 11905163), the Programme of Introducing Talents of Discipline to Universities, China under Grant No. B08033, the self-determined research funds of CCNU, China from the colleges' basic research and operation of MOE, and the numerical simulations have been performed on the GPU cluster in the Nuclear Science Computing Center at Central China Normal University (NSC3).

## References

[1] C.S. Porter, Statistical Theories of Spectra, Academic Press, New York, 1965.
[2] M.L. Mehta, Random Matrices, third ed., Elsevier Academic Press, 2004.
[3] R. Balian, Random matrices and information theory, Nuovo Cimento B 57 (1968) 183.
[4] C. Tsallis, Possible generalization of Boltzmann–Gibbs statistics, J. Stat. Phys. 52 (1998) 479.
[5] H.J. Hilhorst, G. Schehr, A note on $q$-Gaussians and non-Gaussians in statistical mechanics, J. Stat. Mech. (2007) P06003.
[6] Petr Jizba, Jan Korbel, Maximum entropy principle in statistical inference: Case for non-Shannonian entropies, Phys. Rev. Lett. 122 (2019) 120601.
[7] A.C. Bertuola, O. Bohigas, M.P. Pato, Family of generalized random matrix ensembles, Phys. Rev. E 70 (2004) 065102.
[8] F. Toscano, R.O. Vallejos, C. Tsallis, Random matrix ensembles from nonextensive entropy, Phys. Rev. E 69 (2004) 066131.
[9] A.Y. Abul-Magd, Nonextensive random matrix theory approach to mixed regular-chaotic dynamics, Phys. Rev. E 71 (2005) 066207.
[10] K.A. Muttalib, J.R. Klauder, Family of solvable generalized random-matrix ensembles with unitary symmetry, Phys. Rev. E 71 (2005) 055101.
[11] G. Akemann, P. Vivo, Power law deformation of Wishart–Laguerre ensembles of random matrices, J. Stat. Mech. (2008) P09002.
[12] A.Y. Abul-Magd, G. Akemann, P. Vivo, Superstatistical generalizations of Wishart–Laguerre ensembles of random matrices, J. Phys. A 42 (2009) 175207.
[13] G. Akemann, J. Fischmann, P. Vivo, Universal correlations and power-law tails in financial covariance matrices, Physica A 389 (2010) 2566.
[14] O. Bohigas, M.J. Giannoni, C. Schmit, Characterization of chaotic quantum spectra and universality of level fluctuation laws, Phys. Rev. Lett. 52 (1984) 1.
[15] S. Heusler, S. Müller, A. Altland, P. Braun, F. Haake, Periodic-orbit theory of level correlations, Phys. Rev. Lett. 98 (2007) 044103.
[16] V. Brendel, P. Bucher, I.R. Nourbakhsh, B.E. Blaisdell, S. Karlin, Methods and algorithms for statistical analysis of protein sequences, Proc. Natl. Acad. Sci. USA 89 (1992) 2002–2006.

[17] William R. Pearson, Effective protein sequence comparison, Methods Enzymol. 266 (1996) 227–258.
[18] S.V. Buldyrev, N.V. Dokholyan, A.L. Goldberger, S. Havlin, C.-K. Peng, H.E. Stanley, G.M. Viswanathan, Analysis of DNA sequences using methods of statistical physics, Physica A 249 (1998) 430–438.
[19] F. Luo, P.K. Pradip, J. Zhou, Handbook of Data Intensive Computing: Application of Random Matrix Theory To Analyze Biological Data, Springer, New York, 2011.
[20] C. Tsallis, R.S. Mendes, A.R. Plastino, The role of constraints within generalized nonextensive statistics, Physica A 261 (1998) 534.
[21] M. Abramowitz, I. Stegun, Handbook of Mathematical Functions, Dover, New York, 1965.
[22] M.V. Berry, J.P. Keating, Supersymmetry and Trace Formulae, Springer, New York, 1999, pp. 355–367.
[23] C.M. Bender, D.C. Brody, M.P. M¨ uller, Hamiltonian for the zeros of the Riemann zeta function, Phys. Rev. Lett. 118 (2017) 130201.
[24] Weibing Deng, Mauricio P. Pato, Approaching word length via level spectra, Physica A 481 (2017) 167.
[25] https://www.uniprot.org.
[26] https://github.com/xiephysics/Generalized_Poisson_ensemble.
[27] https://en.wikipedia.org/wiki/Titin.
[28] https://en.wikipedia.org/wiki/CA-125.
[29] Nils Strodthoff, Patrick Wagner, Markus Wenzel, Wojciech Samek, UDSMProt: universal deep sequence models for protein classification, Bioinformatics 36 (8) (2020) 2401–2409.
[30] L.R. Taylor, Aggregation, variance and the mean, Nature 189 (1961) 732.
[31] Z. Eisler, J. Kertész, Scaling theory of temporal correlations and size dependent fluctuations in the traded value of stocks, Phys. Rev. E 73 (2006) 046109.
[32] Z. Eisler, I. Bartos, J. Kertész, Fluctuation scaling in complex systems: Taylor's law and beyond, Adv. Phys. 57 (2008) 89.
[33] M. de Menezes, A.-L. Barabási, Phys. Rev. Lett. 92 (2004) 28701.
[34] https://www.ncbi.nlm.nih.gov/gene.
[35] O. Bohigas, J.X. de Carvalho, M.P. Pato, Disordered ensembles of random matrices, Phys. Rev. E 77 (2008) 011122.
[36] Martin Gerlach, Francesc Font-Clos, Eduardo G. Altmann, Similarity of symbol frequency distributions with heavy tails, Phys. Rev. X 6 (2016) 021009.