Probability distribution of intersymbol distances in random symbolic sequences: Applications to improving detection of keywords in texts and of amino acid clustering in proteins

Pedro Carpena,[*] Pedro A. Bernaola-Galván, Concepción Carretero-Campos, and Ana V. Coronado
Departamento de Física Aplicada II, E.T.S.I. de Telecomunicación, Universidad de Málaga, 29071, Málaga, Spain

(Received 17 June 2016; revised manuscript received 4 October 2016; published 4 November 2016)

Symbolic sequences have been extensively investigated in the past few years within the framework of statistical physics. Paradigmatic examples of such sequences are written texts, and deoxyribonucleic acid (DNA) and protein sequences. In these examples, the spatial distribution of a given symbol (a word, a DNA motif, an amino acid) is a key property usually related to the symbol importance in the sequence: The more uneven and far from random the symbol distribution, the higher the relevance of the symbol to the sequence. Thus, many techniques of analysis measure in some way the deviation of the symbol spatial distribution with respect to the random expectation. The problem is then to know the spatial distribution corresponding to randomness, which is typically considered to be either the geometric or the exponential distribution. However, these distributions are only valid for very large symbolic sequences and for many occurrences of the analyzed symbol. Here, we obtain analytically the exact, randomly expected spatial distribution valid for any sequence length and any symbol frequency, and we study its main properties. The knowledge of the distribution allows us to define a measure able to properly quantify the deviation from randomness of the symbol distribution, especially for short sequences and low symbol frequency. We apply the measure to the problem of keyword detection in written texts and to study amino acid clustering in protein sequences. In texts, we show how the results improve with respect to previous methods when short texts are analyzed. In proteins, which are typically short, we show how the measure quantifies unambiguously the amino acid clustering and characterize its spatial distribution.

DOI: 10.1103/PhysRevE.94.052302

I. INTRODUCTION

A symbolic sequence can be defined as a series of N symbols $\{\mathcal{S}_1,\mathcal{S}_2,\dots,\mathcal{S}_N\}$ drawn from a given alphabet with m symbols, $\mathcal{S}_i \in \{\mathcal{A}_1,\mathcal{A}_2,\dots,\mathcal{A}_m\}$. Good examples of symbolic sequences can be written texts, DNA sequences, and proteins, which have been extensively analyzed within the statistical physics context over the past few years [1–21]. In a written text, the alphabet is formed by all the different words of the vocabulary of the text, and each word is a symbol. In deoxyribonucleic acid (DNA) sequences, depending on the property one wants to study, the alphabet can be composed of 4 symbols (A,T,C,G), 16 symbols (AA, ... ,TT), etc. In proteins, each symbol of the alphabet corresponds to one amino acid.

When analyzing symbolic sequences, one of the main properties of a given symbol is its spatial structure, i.e., how the symbol is distributed along the sequence. Actually, such spatial distribution is a key point when modeling the structure of a language [11,22] using written texts or when trying to automatically extract information from the symbolic sequence. Indeed, several successful techniques aimed at extracting keywords from written texts [23–30] or biologically relevant motifs from DNA sequences [21,31–34] are based directly or indirectly on quantifying the deviation of the spatial distribution of a given word or motif with respect to that expected for randomness. The reason is that relevant words (motifs) are unevenly distributed in texts (DNA sequences), and thus far from randomness, and form *clusters*, while nonrelevant words are used with the same likelihood everywhere in the text and then its spatial distribution is expected to be more homogeneous or closer to the random expectation [23] (see Sec. IV for a more detailed explanation).

As a consequence, the knowledge of the exact, randomly expected distribution of a given symbol in a symbolic sequence is crucial for all the analyses just mentioned. In many cases, the random expectation for the intersymbol distance distribution is assumed to be the geometric distribution [21,24,31–34] or even its continuous counterpart, the exponential distribution [11,13,23,29]. However, these two distributions are only correct in the asymptotic limit of very large symbolic sequences and many occurrences of the analyzed symbol. Thus, when analyzing short or medium-sized symbolic sequences, such as short texts or protein sequences, the geometric or exponential assumption could lead to misleading results.

In this paper, we obtain the exact, randomly expected interoccurrence distance distribution of a given symbol within a symbolic sequence, valid for any sequence size and any number of symbol occurrences (Sec. II), and we analyze its main properties (Sec. III). In particular, we show, first, that given a sequence length, there exists a number of occurrences of a symbol for which the variability of the distance distribution is maximized and, second, that the geometric distribution is the asymptotic limit of our result. Using the exact distribution, we are able to define a measure well suited to quantify the degree of *clustering* of a given symbol (Sec. IV) that can be used to improve keyword detection in texts, especially of short length (Sec. V), and also to detect unambiguously amino acid clustering in protein sequences (Sev. VI).

II. DISTRIBUTION OF WORD INTEROCCURRENCE DISTANCES IN SYMBOLIC SEQUENCES

For the sake of simplicity, in this section we refer to a generic symbolic sequence as the "text," and any symbol of the alphabet of that sequence as a "word." In this context, a simple way to characterize the spatial structure of a given word

[*] pcarpena@ctima.uma.es

along the text consists in using the probability distribution $p(d)$ of the interoccurrence distances d of the word: if we consider a text of length N, and a word which appears n times at positions $j_1, j_2, \dots, j_n$, then the interoccurrence distances d_i ($i = 1,2,\dots,n-1$) are given by $d_i = j_{i+1} - j_i$. If the word is distributed at random, then it is normally assumed [11,23–25,27,31–33] that the resulting $p(d)$ distribution is given by the geometric distribution:

$$p_{\rm geo}(d) = q(1-q)^{d-1}, \tag{1}$$

with $q \equiv n/N$ being the *probability* of finding that word in the text. In contrast, content-bearing (relevant) words are not randomly distributed along the text and then should present a distribution $p(d)$ that clearly differs from $p_{\rm geo}(d)$, and the larger the difference, the higher the relevance of the word to the text considered. Indeed, many techniques aimed at analyzing the statistical properties of symbolic sequences (written texts, DNA sequences, protein chains, etc.) and detecting keywords, relevant motifs, etc., use in some way the deviation of the observed $p(d)$ to the randomly expected $p_{\rm geo}(d)$ for that purpose.

However, the distribution $p_{\rm geo}(d)$ is the correct one only asymptotically, i.e., in the limits $N \to \infty$, $n \to \infty$, while keeping constant the ratio n/N. Our aim is the determination of the exact, randomly expected distribution for finite N and n, $p_{N,n}(d)$ from now on. Such distribution would describe better the statistical properties of low-frequency words in short texts (i.e., small N and n), corresponding to a more realistic situation where the use of the geometric distribution could lead to misleading results.

The exact formulation of the problem we want to solve is the following: Assuming that the distances are measured in words, our text can be considered as the interval $[1,N]$, where the positions of a given word must be integer numbers in that interval. Specifically, if we consider a word appearing randomly n times in the text ($n \leqslant N$), such a word would be placed at positions j_i ($i = 1,2,\dots,n$) such that $0 < j_1 < j_2 < \dots < j_n < N+1$. The set of the $n-1$ interoccurrence distances for that word is then given by $d_i = j_{i+1} - j_i$ ($i = 1,2,\dots,n-1$). However, we also include in the set two additional distances, $d_0 = j_1 - 0 = j_1$ and $d_n = N + 1 - j_n$, which can be understood as boundary conditions and is equivalent to considering that the word also appears at positions 0 and $N+1$. The inclusion of these two additional distances (then totaling $n+1$ distances) simplifies the calculation of $p_{N,n}(d)$ and, more importantly, does not modify the final result (see Appendix A): Since the word is assumed to be placed at random, all distances have the same statistical behavior, and the distribution $p_{N,n}(d)$ remains unchanged. We also note that, given N and n, the set of possible distances must be in the range $1,2,\dots,N+1-n$.

To calculate $p_{N,n}(d)$ we start by counting the number of ways in which a distance d can be obtained when all possible arrangements of n words within a text of length N are analyzed. Three different situations have to be taken into account in order to avoid counting the same distance more than once:

(i) A distance d is found at the beginning of the interval $[0,N+1]$, obtained by placing the first appearance of the word at $j_1 = d$. This leaves $N-d$ sites available for the remaining $n-1$ repetitions of the word, thus leading to $\binom{N-d}{n-1}$ different configurations where a distance d appears at the beginning of $[0,N+1]$.

(ii) A distance d is found at the end of the interval $[0,N+1]$, obtained by placing the last appearance of the word at $j_n = N + 1 - d$. By symmetry, this situation is equivalent to (i): there are $N-d$ sites available for the remaining $n-1$ words to place, and thus we have again $\binom{N-d}{n-1}$ different ways to get a distance d at the end of the interval $[0,N+1]$.

(iii) A distance d is found neither at the beginning nor the end of $[0,N+1]$. In such a case, d is a "real" distance between two repetitions of the word, and it must come from one of the following situations: $[1,1+d],[2,2+d],[3,3+d]\dots,[N-d,N]$. Thus, there are $N-d$ different cases to find a distance d bounded by two words, thus leaving $N-d-1$ sites to arrange the remaining $n-2$ words. As a result, we have $(N-d)\binom{N-d-1}{n-1}$ different ways to obtain a distance d in the inner part of $[0,N+1]$.

Adding up the three cases, the total number of ways of obtaining a word interoccurrence distance d when considering all possible arrangements of n words in a text of length N is:

$$2\binom{N-d}{n-1} + (N-d)\binom{N-d-1}{n-2} = (n+1)\binom{N-d}{n-1}. \tag{2}$$

Finally, the normalized probability distribution $p_{N,n}(d)$ can be obtained just dividing (2) by the total number of ways of obtaining any distance d:

$$p_{N,n}(d) = \frac{(n+1)\binom{N-d}{n-1}}{\sum_{k=1}^{N-n+1}(n+1)\binom{N-k}{n-1}} = \frac{\binom{N-d}{n-1}}{\binom{N}{n}}, \tag{3}$$

with $d = 1,2,\dots,N+1-n$. We show in Fig. 1 several examples of $p_{N,n}(d)$. Although for small N and n the distribution $p_{N,n}(d)$ could be obtained numerically in an exact way (i.e., checking all possible configurations of the n positions of the word), for larger N and n this cannot be done exactly in a reasonable time due to the huge number of configurations,

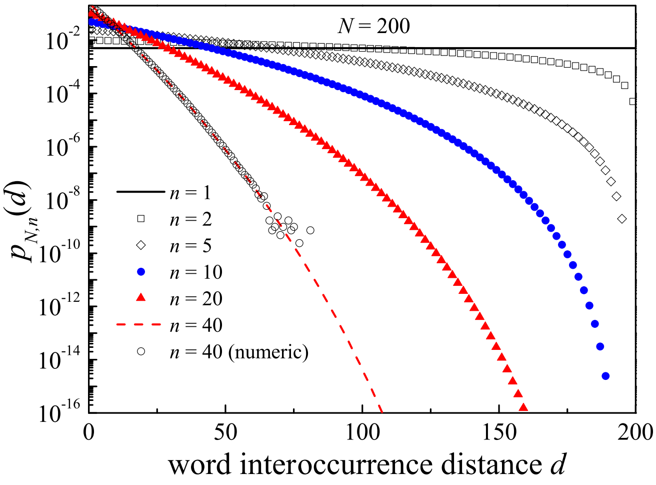

FIG. 1. Examples of distributions $p_{N,n}(d)$ obtained from Eq. (3) for different word counts n and for $N = 200$. For the $n = 40$ case we also show in circles the distribution $p_{N,n}(d)$ numerically obtained by generating 10^8 random configurations.

indicating the usefulness of the exact result in Eq. (3). We include an example in Fig. 1 with $N = 200$ and $n = 40$: The number of possible configurations is $\binom{200}{40} \simeq 2 \times 10^{42}$, impossible to check in a reasonable time. We show the exact result obtained from Eq. (3), and the numerical result obtained by generating 10^8 random configurations.

The probability of finding the smallest possible distance $d = 1$ can be easily calculated from (3):

$$p_{N,n}(1) = \frac{n}{N}, \tag{4}$$

i.e., the probability of finding a unit distance results to be the ratio between the word count n and the number of positions where the word can be located (the text length N) or, in other words, the probability of finding the word at a given position.

Conversely, the probability of finding the largest possible distance, $d = N + 1 - n$, is then

$$p_{N,n}(N + 1 - n) = \frac{1}{\binom{N}{n}}. \tag{5}$$

As for the dependence of $p_{N,n}(d)$ on n, some particular cases can be easily obtained from (3). For $n = 1$ we get

$$p_{N,1}(d) = \frac{1}{N}, \tag{6}$$

i.e., all distances d has the same likelihood when the word appears only once. For $n = 2$, from (3) we obtain

$$p_{N,2}(d) = \frac{2(N - d)}{N(N - 1)}. \tag{7}$$

In this case, the probability decreases linearly with the distance d. In general, given a word with n counts $p_{N,n}(d)$ is a $(n - 1)$-th degree polynomial defined in $d = 1,2, \ldots ,N + 1 - n$.

III. SOME PROPERTIES OF $p_{N,n}(d)$

Once the distribution $p_{N,n}(d)$ has been determined analytically, we present in this section some of its main properties.

Normalization. By construction, $p_{N,n}(d)$ is normalized:

$$\sum_{\ell=1}^{N+1-n} p_{N,n}(d) = \sum_{d=1}^{N+1-n} \frac{\binom{N-d}{n-1}}{\binom{N}{n}} = \frac{\binom{N}{n}}{\binom{N}{n}} = 1. \tag{8}$$

Mean. The inclusion of the boundary conditions $d_0 = j_1$ and $d_n = N + 1 - j_n$ implies that $\sum_{i=0}^{n} d_i = N + 1$. The mean distance $\langle d\rangle$ is then given exactly by $\langle d\rangle = (N + 1)/(n + 1)$ and obtained for any possible configuration of the n words. The same result can be obtained also from $p_{N,n}(d)$:

$$\langle d\rangle = \sum_{d=1}^{N+1-n} d\, p_{N,n}(d) = \sum_{d=1}^{N+1-n} d \frac{\binom{N-d}{n-1}}{\binom{N}{n}} = \frac{N + 1}{n + 1}. \tag{9}$$

Note that without including the boundary conditions (i.e., considering only the $n - 1$ inner distances), the expected mean is the same since the distribution for those distances is also $p_{N,n}(d)$ (see Appendix A). However, individual means obtained for different configurations of the n words in general differ from (9). This is one of the advantages of including boundary conditions: The sample mean is fixed *a priori* and coincides with the expected mean (9).

Second and higher moments: Variance. For the second moment, we have

$$\langle d^2\rangle = \sum_{\ell=1}^{N+1-n} d^2 \frac{\binom{N-d}{n-1}}{\binom{N}{n}} = \frac{(N + 1)(2N - n + 2)}{(n + 1)(n + 2)}, \tag{10}$$

while the third moment is given by

$$\langle d^3\rangle = \frac{(N + 1)(12N - 7n - 6Nn + 6N^2 + n^2 + 6)}{(n + 1)(n + 2)(n + 3)}. \tag{11}$$

Higher-order moments can be calculated straightforwardly using $p_{N,n}(d)$, although with increasing complexity in its explicit expression for increasing order.

Concerning the variance, using the mean (9) and the second moment (10) we get

$$\sigma^2 = \langle d^2\rangle - \langle d\rangle^2 = \frac{n(N + 1)(N - n)}{(n + 1)^2(n + 2)}. \tag{12}$$

For fixed N, σ^2 is maximal at $n = 1$ and decreases as n increases and tends to 0 as $n \to N$. Conversely, for fixed n, $\sigma^2 \sim N^2/n^2$ for large N.

Cumulative distribution. The cumulative distribution $P_{N,n}(k) \equiv \text{Prob}\{d \leqslant k\}$ can be obtained as:

$$P_{N,n}(k) = \sum_{d=1}^{k} \frac{\binom{N-d}{n-1}}{\binom{N}{n}} = 1 - \left(\frac{N - k}{n}\right) \frac{\binom{N-(k+1)}{n-1}}{\binom{N}{n}}, \tag{13}$$

which can be expressed in terms of the probability density (3):

$$P_{N,n}(k) = 1 - \left(\frac{N - k}{n}\right) p_{N,n}(k + 1) \tag{14}$$

or in a more compact way as

$$P_{N,n}(k) = 1 - \frac{\binom{N-k}{n}}{\binom{N}{n}}. \tag{15}$$

Equations (14) and (15) allow us to obtain straightforwardly the complementary cumulative distribution function $Q_{N,n}(k) \equiv \text{Prob}\{d > k\} = 1 - P_{N,n}(k)$:

$$Q_{N,n}(k) = \left(\frac{N - k}{n}\right) p_{N,n}(k + 1) = \frac{\binom{N-k}{n}}{\binom{N}{n}}. \tag{16}$$

A. Asymptotic properties

We address here the functional form of the interword distribution of distances in the limit of a very long text (or a very long symbolic sequence in general), i.e., for large N. Starting from (3), and using the definition of combinatorial numbers, we have:

$$p_{N,n}(d) = \frac{(N - d)!n!(N - n)!}{(n - 1)![N - d - (n - 1)]!N!}. \tag{17}$$

For large N we can write

$$\frac{(N - d)!}{N!} \sim \frac{1}{N^d}, \tag{18}$$

$$\frac{(N - n)!}{[N - n - (d - 1)]!} \sim (N - n)^{d-1}. \tag{19}$$

As $n!/(n-1)! = n$, from (17) we finally obtain

$$\hat{p}_{N,n}(d) = \frac{n}{N^d}(N-n)^{d-1} = \frac{n}{N}\left(1-\frac{n}{N}\right)^{d-1}, \tag{20}$$

where we have termed $\hat{p}_{N,n}(d)$ to the distribution for large N. Noting that the ratio n/N represents the word counts over the total number of words in the text, i.e., the *probability* of finding the word at a given position [see also (4)], then by defining $q \equiv n/N$ the distribution $\hat{p}_{N,n}(d)$ transforms straightforwardly into the *geometric distribution* $p_{\mathrm{geo}}(d)$ (1). As we stated above, $p_{\mathrm{geo}}(d)$ is the one usually assumed to be valid for a randomly distributed word (symbol) in a text (sequence). However, this is only correct in the limit of large N and n, and using it in a short text can produce misleading results (see Sec IV).

B. Maximum diversity of interoccurrence distances

The interoccurrence distance distribution $p_{N,n}(d)$ is obtained by assuming that the analyzed word is randomly distributed, which in many contexts is a synonym of "homogeneity." For this reason, we study here the expected variability of the distribution. A convenient measure to estimate the variability of a distribution is the coefficient of variation (c_v), defined as the ratio between the standard deviation and the mean,

$$c_v \equiv \frac{\sqrt{\sigma^2}}{\langle d\rangle} = \frac{\sigma}{\langle d\rangle}, \tag{21}$$

where c_v is a dimensionless quantity that shows the extent of variability of a distribution in relation to its mean, and that is commonly used to characterize deviation from randomness of time series in many scientific fields such as immunology [35], human dynamics [36], and complex systems [37] or the analysis of energy levels in disordered systems [38,39]. By using (9) and (12), for $p_{N,n}(d)$ we get

$$c_v(N,n) = \sqrt{\frac{n(N-n)}{(N+1)(n+2)}}. \tag{22}$$

On the one hand, the coefficient c_v is a monotonously increasing function of N, and tends to 1 as $N \to \infty$. On the other hand, and more interestingly, c_v presents a maximum as a function of the word count n. The behavior of c_v as a function of n for several values of the text length N is shown in Fig. 2, where the maxima of c_v can be clearly seen. Indeed, considering n a continuous variable and solving for n the equation $\partial c_v/\partial n = 0$, we obtain that the maximum variability of interoccurrence distances occurs when the word count is given by

$$n_{\max} = \sqrt{2N+4} - 2. \tag{23}$$

We also include in Fig. 2 the line of the maxima of c_v ($c_{v,\max}$) obtained as $c_{v,\max} = c_v(N,n_{\max})$ and plotted as a function of $n_{\max}$. According to (23), and noting that typically the text length N is at least of several hundreds, we can conclude that the spatial distribution of a given word presents a maximal diversity (or minimal homogeneity) for $n \simeq \sqrt{2N}$. Note that usually the random distribution is a reference for homogeneity, but our results imply that, depending on n, the homogeneity changes and even presents a minimum despite the fact of considering random distribution in all cases.

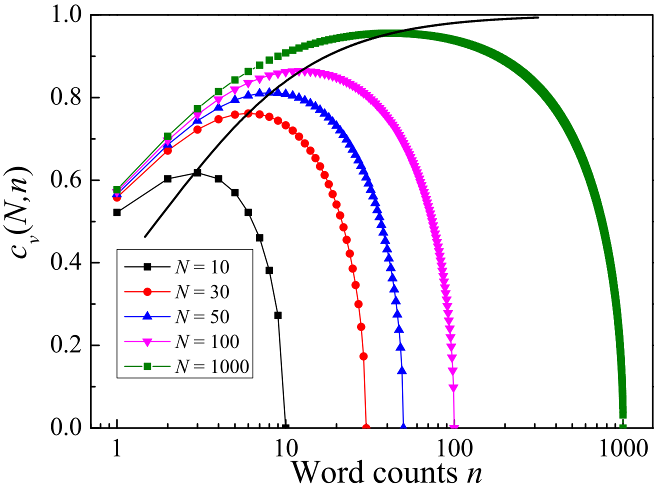

FIG. 2. Coefficient of variation $c_v(N,n)$ as a function of the word counts n for several values of the text length N. Note how in every case there is a clear c_v maximum $c_{v,\max}$ for a particular n value, $n_{\max}$. The solid line joins the $c_{v,\max}$ values as a function of $n_{\max}$.

IV. MEASURING CLUSTERING IN SYMBOLIC SEQUENCES

As we stated in the Introduction, many techniques aimed at detecting keywords in texts, or biologically relevant motifs in DNA sequences, are based on the deviation of the spatial distribution of a particular word or motif with respect to the random expectation. In texts, for example, the subjacent idea is that a relevant, content-bearing word is used more frequently in certain contexts (when the concept associated to the word is discussed) and less frequently or very rarely in other contexts. This results in large fluctuations in the spatial distribution of the word, which is concentrated in certain regions and (almost) does not appear in others. In this sense, using a physical analogy, we could say that a relevant word or motif attracts itself and tends to form *clusters*. In contrast, a nonrelevant word (prepositions, articles, etc.) is used with the same likelihood everywhere in the text, does not interact with itself, and should appear randomly. Thus, its spatial distribution is expected to be more homogeneous and closer to the random expectation. In this sense, the concepts of *clustering* and *relevance* are synonymous in written texts. A similar argument can be applied to biological sequences such as DNA or proteins. For example, in the case of DNA, a useless motif does not have any restriction in its location, whereas a functional one should appear more concentrated in regions associated to biological functions (genes, promoters, etc.) and therefore should also exhibit clustering.

As a consequence, it is convenient to define an appropriate *clustering measure*, able to capture such deviations from the random expectation for a given word or symbol in general. In principle, this could be done by comparing somehow the *observed* interword distance distribution with the *randomly expected* one. However, for practical purposes, this strategy is not convenient, especially in the case of words (or symbols in general) with low frequencies since in this case there will be only a small number of distances available, which is not enough to properly compare both distributions. This is the case of short or moderately large texts, where any word of the vocabulary appears a small number of times. This is also the case of large motifs (i.e., with low likelihood) in DNA sequences, as well as of amino acids in proteins, which are typically short.

Instead of using the distributions, we propose to use a simple and robust clustering measure defined as

$$C(N,n) = \frac{c_{v,\text{obs}}(N,n)}{c_{v,\text{exp}}(N,n)}, \tag{24}$$

where N is the text (symbolic sequence) length and n is the word (symbol) count, and $c_{v,\text{obs}}(N,n)$ $[c_{v,\text{exp}}(N,n)]$ is the *observed* (randomly *expected*) coefficient of variation (see Sec. III B). A similar clustering measure was originally proposed in Ref. [23], and later refined in Refs. [25,27]. However, in these previous works both $c_{v,\text{exp}}(N,n)$ and $c_{v,\text{obs}}(N,n)$ were calculated in a different way as we propose here (see Secs. IV A and IV B). In particular, $c_{v,\text{exp}}(N,n)$ was obtained using the geometric distribution (1) instead of the correct one (3), for which c_v is given in Eq. (22).

Some authors, especially for application in written texts, define a clustering measure only as the numerator of Eq. (24) $[c_{v,\text{obs}}(N,n)]$ with the name of intermittency index [11,13,29]. The reason is that they use as reference for randomness the exponential distribution (the continuous counterpart of the geometric distribution), for which $c_{v,\text{exp}}(N,n) = 1$ (see also Sec. IV A). However, note that the exact $c_{v,\text{exp}}$ (Fig. 2) is not a constant and, depending on N and n, with values far from unity.

With the definition in Eq. (24), C is a dimensionless quantity with a clear interpretation: (i) Values $C > 1$ mean that the fluctuations in the observed distance distribution are larger than those randomly expected, indicating that the word attracts itself and is clustered. (ii) Values of $C \simeq 1$ indicate that distance fluctuations are essentially random, and therefore the word is randomly distributed along the text. (iii) Values of $C < 1$ suggest low distance fluctuations, implying the existence of word self-repulsion.

As we mentioned above, the results shown in Secs. II and III suggest to change the traditional way in which both $c_{v,\text{obs}}(N,n)$ and $c_{v,\text{exp}}(N,n)$ are usually estimated, therefore changing substantially the clustering measure. In the following, we discuss both cases.

A. The value of $c_{v,\text{exp}}(N,n)$: Geometric vs. exact distribution

Typically, the random expectation for the interword distance distribution is assumed to be the *geometric distribution* (1), $p_{\text{geo}}(d)$ both in written texts and DNA sequences. For convenience, we reproduce its expression here:

$$p_{\text{geo}}(d) = q(1-q)^{d-1}, \tag{25}$$

with $q \equiv n/N$ being the probability of that word in the analyzed text. However, we have already obtained the exact distribution for finite N and n, $p_{N,n}(d)$ [Eq. (3)], which is also reproduced here for convenience:

$$p_{N,n}(d) = \frac{\binom{N-d}{n-1}}{\binom{N}{n}}. \tag{26}$$

The geometric distribution $p_{\text{geo}}(d)$ is only asymptotically correct, as we have shown in Sec. III A. Thus, its validity is limited to very large texts (symbolic sequences) ($N \to \infty$) and infinitely many instances of the word (symbol) ($n \to \infty$). In the case of written texts, for long books one may have N of the order of 10^5–10^6 words, and for relevant words n it is in the range from several tens to a few hundreds. With these numbers, the geometric approximation is usually enough since the differences between the asymptotic case and the exact distribution are small (see below). Then, automatic keyword detectors using clustering measures similar to (24) with the geometric distribution as the random expectation work fairly well when tested in long-enough books [23–25,27,30].

However, for short or moderately large texts (such as scientific articles, reports, web pages, etc.), $N \sim 10^3$–10^4 and n is of the order of a few tens at the very most. In this case, if $p_{\text{geo}}(d)$ is assumed to be the distance distribution expected for the random case instead of the correct one $p_{N,n}(d)$, then important errors can be introduced when estimating the probability of a given distance.

In Fig. 3, we illustrate the differences between the geometric distribution and the exact result in different cases. Figure 3(a) shows the exact distributions (symbols) and the corresponding geometric distributions (solid lines) for different values of the text length N and the word frequency n, in all cases as a function of the natural spatial variable, the normalized distance $\bar{d} = d/\langle d\rangle$. For each N and n combination, the probability q of the corresponding geometric distribution is obtained as $q = n/N$. For large values of n [see the $n = 200$ cases in Fig. 5(a)], as the ones expected in a long book, the geometric distribution and the exact result are very similar, independently of N. In contrast, for small n values [see the $n = 5$ cases in Fig. 3(a)], the geometric distribution is similar to the exact result only in the range $\bar{d} \leqslant 2$, while for $\bar{d} > 2$ the exact result is orders of magnitude smaller than the geometric case. Thus, if the geometric distribution is used to estimate the probability of a distance in this range, then such a probability would be severely overestimated.

To better compare both cases for different n values, in Fig 3(b) we represent the ratio $r(d) \equiv p_{N,n}(d)/p_{\text{geo}}(d)$ as a function of the normalized distance $\bar{d}$. First, we notice that the curves with the same n value collapse, independently of N, thus indicating that n is the natural variable to measure the deviation from the asymptotic case. Second, we observe that for large n, $r(d) \sim 1$ in the whole studied range of $\bar{d}$, as expected. In contrast, for small n, we observe two regimes: (i) In the range $0 < \bar{d} \leqslant 2$ the ratio $r(d)$ is larger than 1 with a maximum in between, and then the probability of a distance in that range is *underestimated* when using $p_{\text{geo}}(d)$. Actually, with a more detailed calculation using (25) and (26) one can prove that for large N the maximum of $r(d)$ occurs at $d_{\max} = N/n$ and has a value $r(d_{\max}) \simeq 1 + 1/2n$. (ii) In the range $\bar{d} > 2$, $r(d)$ falls abruptly for increasing $\bar{d}$, indicating the strong overestimation when using $p_{\text{geo}}(d)$ in this range.

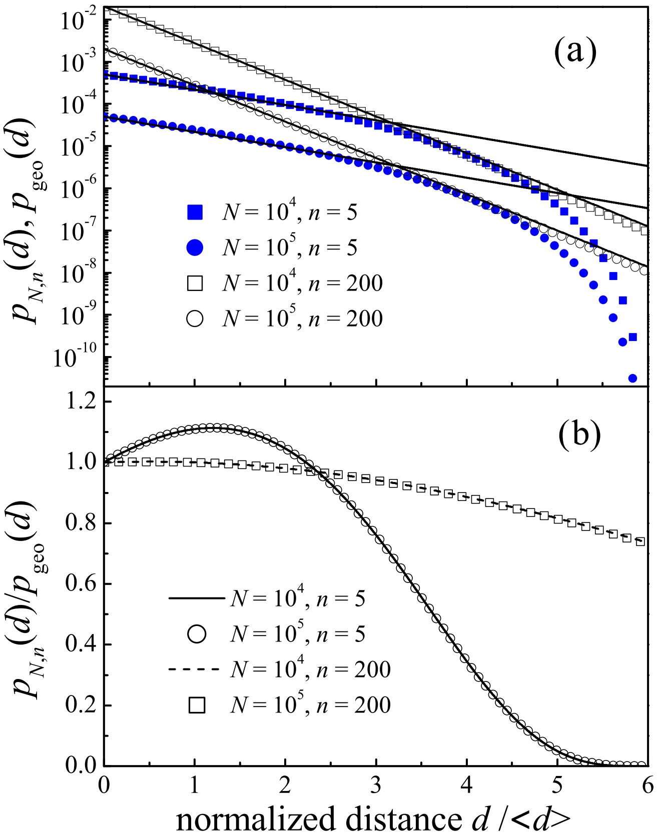

FIG. 3. (a) Exact interoccurrence distance distribution $p_{N,n}(d)$ (symbols) and the corresponding geometric distributions $p_{\rm geo}(d)$ (solid lines) as a function of the normalized distance $d/\langle d\rangle$ for different combinations of N and n values. For the geometric distributions, $q = n/N$. (b) Ratio $p(d)/p_{\rm geo}(d)$ for the four cases shown in (a). Note the collapse of the curves for words with the same frequency n, thus indicating that n is the natural variable to measure the deviation from the asymptotic case.

These differences between the exact and the geometric distributions are also reflected in the corresponding coefficients of variation. The denominator of the clustering measure (24), $c_{v,\rm exp}$, corresponds to the randomly expected coefficient of variation. We note that the geometric distribution is commonly chosen as the random expectation, so using the definition of c_v (21), it is easy to obtain for the geometric case that:

$$c_{v,\rm geo}(N,n) = \sqrt{1 - n/N} \equiv \sqrt{1-q}. \tag{27}$$

This value is the one widely used as the denominator in the clustering measure (24). In some cases, instead of the geometric distribution, its continuous counterpart (the exponential distribution) is considered as the reference for randomness [11,13,23,29]. In this latter case, one considers that N and n are both large, but $q = n/N$ is very small, and then $c_{v,\rm exp} \simeq 1$. Obviously, this assumption can lead to misleading results (see below) when either N or n or both are small, or when q is not negligible, as it happens very often in protein chains.

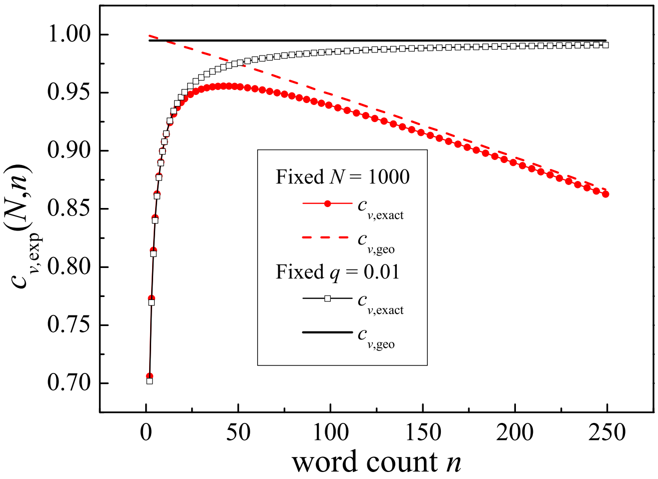

FIG. 4. Expected coefficients of variation obtained from the geometric distribution ($c_{v,\rm geo}$) and the exact result ($c_{v,\rm exact}$) as a function of the word count n. We distinguish between a text with a fixed length of $N = 1000$ words, and a word with a fixed probability $q = 0.01$, and then the text length is $N = n/q$.

However, the exact distribution is $p_{N,n}(d)$ and the corresponding coefficient of variation is [see Eq. (22)]:

$$c_{v,\rm exact}(N,n) = \sqrt{\frac{n(N-n)}{(N+1)(n+2)}}. \tag{28}$$

Both $c_{v,\rm geo}$ and $c_{v,\rm exact}$ are shown in Fig. 4 as a function of the word count n. We consider two cases: (i) We fix the text length N, and then the geometric probability q varies with n as $q = n/N$ (dashed line and circles). (ii) We fix the geometric probability q, and then the text length N varies as $N = qn$ (solid line and squares). In both cases we find that for increasing n, $c_{v,\rm exact}$ tends asymptotically to $c_{v,\rm geo}$, as expected. Then, for large n values, as it happens in long books, there should be no differences when using $c_{v,\rm exact}$ or $c_{v,\rm geo}$ in the relevance measure (24). However, for small n values, which is the normal situation found in short texts, the differences are substantial. Indeed, $c_{v,\rm geo}/c_{v,\rm exact} > 1$ and the smaller the n, the larger the ratio, thus indicating that the use of $c_{v,\rm exact}$ instead of $c_{v,\rm geo}$ in (24) is more sensitive to the detection of clustering in low-frequency words. We will show some examples of this property in Secs. V and VI.

B. The value of $c_{v,\rm obs}(N,n)$: The need for boundary conditions

Traditionally, the observed coefficient of variation $c_{v,\rm obs}(N,n)$ is calculated as follows: given a word appearing n times in a text of length N at positions $j_1, j_2, \ldots, j_n$, the set of the $n-1$ interword distances is obtained as $d_i = j_{i+1} - j_i$. Then,

$$\langle d_{\rm obs}\rangle = \frac{\sum_{i=i}^{n-1} d_i}{n-1}, \tag{29}$$

$$\langle d^2_{\rm obs}\rangle = \frac{\sum_{i=i}^{n-1} d_i^2}{n-1}, \tag{30}$$

from where $\sigma^2_{\rm obs} = \langle d^2_{\rm obs}\rangle - \langle d_{\rm obs}\rangle^2$ and $c_{v,\rm obs}(N,n) = \sigma_{\rm obs}/\langle d_{\rm obs}\rangle$. Note that when obtaining these expressions, boundary conditions are not imposed. In this case, the observed mean $\langle d_{\rm obs}\rangle$ differs in general from the randomly expected mean $\langle d\rangle = (N+1)/(n+1)$, since, depending on the particular spatial distribution of the word, $\sum_{i=i}^{n-1} d_i$ varies in the range $n-1 \leqslant \sum_{i=i}^{n-1} d_i \leqslant N-1$. Recall that the expected mean $\langle d\rangle$ is the same no matter whether boundary conditions are considered (9) or not (see Appendix A).

However, when considering boundary conditions, the observed mean is always identical to the expected mean, since in this case there are always $n+1$ distances, and $\sum_{i=1}^{n+1} d_i = N+1$. Obviously, $\langle d_{\rm obs}^2\rangle$ will also differ considerably if evaluated as in (30) with $n-1$ distances or considering boundary conditions with $n+1$ distances. As a consequence, the observed coefficient of variation $c_{v,{\rm obs}}(N,n)$ will change substantially if boundary conditions are imposed or not, as it is usually done.

Indeed, we propose here to calculate $c_{v,{\rm obs}}(N,n)$ using boundary conditions. Otherwise, when using $c_{v,{\rm obs}}$ in the clustering measure (24), the resulting $C(N,n)$ value could be misinterpreted. We recall that values of $C>1$ should indicate clustering, values of $C\simeq 1$ should indicate random distribution, and values of $C<1$ should indicate self-repulsion. However, let us consider an example of extreme clustering which would be interpreted as strong self-repulsion if boundary conditions are not applied. Suppose we have a text of length N, and a word appearing n times in the text as clustered as possible, for example, all separated the unit distance from position m, i.e., in $m, m+1, \ldots, m+n-1$. Applying Eqs. (29) and (30), we get $\langle d_{\rm obs}\rangle = \langle d_{\rm obs}^2\rangle = 1$, from where $\sigma_{\rm obs}^2 = c_{v,{\rm obs}}(N,n) = 0$, then leading (24) to $C=0$ indicating extreme repulsion. In contrast, using boundary conditions, there are two additional distances in the set, $d_0 = m$ and $d_n = N+1-(m+n-1)$. When these distances are considered, we obtain the highest possible clustering, as it should be (see below).

C. Extreme clustering values

As we have shown in the two preceding subsections how to properly calculate $c_{v,{\rm exp}}$ and $c_{v,{\rm obs}}$ to be used in the clustering measure C (24), we study in this section which are the extreme values of C. Let us consider a text of length N and a particular word of the vocabulary appearing n times in the text. Once N and n are fixed, the value of $c_{v,{\rm exp}}(N,n)$ is given by Eq. (28).

Concerning $c_{v,{\rm obs}}$, obviously it will be maximum when $\sigma_{\rm obs}^2$ reaches also its maximum value. Assuming boundary conditions, this happens when all the n instances of the word appear concentrated at the beginning, or at the end, or split between the beginning and the end of the text. In all these cases, the set of $n+1$ interword distances consists of n distances of value 1 and a distance of value $N+1-n$. Then, noting that $\langle d_{\rm obs}\rangle = (N+1)/(n+1)$, we get:

$$\sigma_{\rm obs,max}^2 = \frac{n+(N+1-n)^2}{n+1} - \left(\frac{N+1}{n+1}\right)^2. \qquad (31)$$

Using this expression in (21) we can obtain the maximum value of $c_{v,{\rm obs}}$, from which, by using (24), we finally obtain the maximum clustering value:

$$C_{\max}(N,n) = \sqrt{\frac{(N-n)(n+2)}{N+1}} \simeq \sqrt{n+2}, \qquad (32)$$

where the approximation usually works because typically $n \ll N$.

A situation of extreme clustering close to the maximum value appears when there is a single cluster neither at the beginning nor the end of the text. Let us consider a single cluster starting at position j and then occupying positions $j, j+1, \ldots, j+n-1$. The set of interword distances contains $n-1$ distances of value 1 and two additional distances of values j and $N+1-(j+n-1)$. Proceeding as before, we can obtain in this case $\sigma_{\rm obs}^2$ and the corresponding $C(N,n;j)$, which is this case is also a function of j. Obviously, for $j=1$ and $j=N-n+1$, $C(N,n;j) = C_{\max}(N,n)$ since the cluster is located at the beginning or at the end of the text, respectively. For j in the range $1<j<N-n+1$, the function $C(N,n;j)$ has a minimum at $j_{\min} = (N-n+2)/2$, where the clustering results are

$$C(N,n;j_{\min}) = \sqrt{\frac{(N-n)(n+2)(n-1)}{2n(N+1)}} = \sqrt{\frac{n-1}{2n}}\,C_{\max}(N,n). \qquad (33)$$

From now on, we consider this last value as the boundary for extreme clustering and define $C_{\rm b}(N,n) \equiv \sqrt{(n-1)/2n}\,C_{\max}(N,n)$. Similar extreme clustering values are obtained for situations such that a word is located almost entirely at a single cluster with a few isolated instances out of the cluster or a word is located at two very distant clusters. Thus, this analysis leads us to conclude that words with extreme clustering can be identified whenever their corresponding clustering value $C(N,n)$ falls in the range $C_{\rm b}(N,n) \leqslant C(N,n) \leqslant C_{\max}(N,n)$. This property will be used in the next section.

To end this analysis, we want to remark that the extreme clustering values shown in Eqs. (32) and (33) have been obtained under the hypothesis of having clusters with many interword distances of value 1. Obviously, this does not occur in a real text (although it can happen in other symbolic sequences as DNA or proteins), where the typical interword distances are always larger than 1 even within a cluster. However, noting that within a cluster these typical distances are much smaller than the expected mean (9), and that the coefficient of variation measures the distance fluctuations as compared to the mean, the extreme clustering values in Eqs. (32) and (33) are also applicable in this case.

In this section, we have calculated the extreme values of C via calculating the maximum values of $c_{v,{\rm obs}}(N,n)$. Note that if we could determine the probability distribution of *all* the observable $c_{v,{\rm obs}}(N,n)$ values, we could develop a standard hypothesis test since we could associate a p value to any experimental result. However, it seems that such distribution cannot be obtained analytically and should be obtained with extensive numerical simulations for any combination of N and n, which is out of the scope of this work.

V. APPLICATION I: IMPROVING KEYWORD DETECTION IN TEXTS

We have shown in the preceding section how to properly calculate $c_{v,\mathrm{exp}}(N,n)$ and $c_{v,\mathrm{obs}}(N,n)$ as a consequence of knowing the exact distribution of interword distances randomly expected (26) and of applying boundary conditions. Then, the values of the clustering measure (24) can differ substantially if calculated as we suggest here [i.e., using that $c_{v,\mathrm{exp}}(N,n) = c_{v,\mathrm{exact}}(N,n)$ and applying boundary conditions to obtain $c_{v,\mathrm{obs}}(N,n)$] or if calculated as it is usually done [i.e., using that $c_{v,\mathrm{exp}}(N,n) = c_{v,\mathrm{geo}}(N,n)$ and without applying boundary conditions to get $c_{v,\mathrm{obs}}(N,n)$]. To distinguish between both cases, from now on we name $C(N,n)$ to the clustering measure calculated as we propose here and $C^*(N,n)$ to the clustering measure calculated traditionally.

Our hypothesis is then that using $C(N,n)$ instead of $C^*(N,n)$ should improve the results of word ranking in two aspects: (i) $C(N,n)$ should detect better the clustering associated to relevant words, especially in the case of short texts and/or low-frequency words. The reason is the use of $c_{v,\mathrm{exact}}(N,n)$ in $C(N,n)$ instead of the asymptotic $c_{v,\mathrm{geo}}(N,n)$ in C^*, which overestimates the randomly expected clustering for low-frequency words; and (ii) only when boundary conditions are considered (i.e., in C) can the different regimes of word spatial distribution be properly associated to clustering values: $C > 1$ implies clustering, $C \simeq 1$ indicates random distribution, and $C < 1$ points to repulsion. This is not always true when C^* is used, as we have shown above with an extreme clustering example interpreted as repulsion.

To show that this is the case, we choose as our benchmark the book *On the Origin of the Species* [40] by Charles Darwin, as it has become the standard reference for many word-ranking algorithms [24,27,30], partly because the book contains its own glossary, and then deciding the relevance of a word to the book is easier and less subjective than in other cases.

When the whole book is analyzed ($N \simeq 1.56 \times 10^5$, and a vocabulary of 6866 different words), all the relevant words are relatively frequent with n of several tens, and therefore there should not be great differences related to frequency since $c_{v,\mathrm{exact}}$ and $c_{v,\mathrm{geo}}$ are almost identical (see Fig. 4). Indeed, other keyword extractors should work well in this case: The text is large and there are no statistical problems related to low-frequency keywords. To illustrate this, we show in Table I the ranking of the top-10 most relevant words obtained using C and C^*. For comparison, we also include the ranking obtained using the entropic measure E_{nor} [24]. As expected, the three measures work nicely: The three rankings are quite similar and share many words, and the extracted keywords reflect very well the topic of the book.

Concerning the C and C^* comparison, both rankings are similar and contain almost the same words with small changes in the order. Indeed, they share eight words, and the two words in the C^* ranking which are absent from the C one ("workers" and "diagram") appear also in the top-35 positions of the C ranking. Conversely, the word "fertility," absent from the top-10 C^* ranking, is actually ranked 11th. However, the word "wax," absent from the C^* ranking, deserves special attention. Its frequency is relatively high (and therefore with little influence), but it is ranked 8th using C and ranked 1406th using C^*. This dramatic difference is originated by the effect of using boundary conditions in C and not in C^*, and we discuss it in detail in Sec. V A.

TABLE I. Ranking of the 10 most relevant words extracted from the book *On the Origin of the Species*, by Charles Darwin. Words are ordered in descending value of C (first column), C^* (second column), and the entropic measure E_{nor} [24] (third column).

C ranking	C^* ranking	E_{nor} ranking
formations	formations	hybrids
sterility	cells	sterility
hybrids	sterility	i
bees	hybrids	species
instincts	bees	islands
instinct	instincts	forms
cells	workers	instincts
wax	slaves	varieties
fertility	instinct	breeds
slaves	diagram	fertility

We also want to mention that typical nonrelevant words (such as "the," "to," "and," "or," etc.) are not necessarily the last ones in the rankings. Note that C measures deviation from randomness, and such deviation can be attractive ($C > 1$, more frequent) or repulsive ($C < 1$, rare). These kinds of words are essentially distributed at random, and for them we observe $C \simeq 1$.

In conclusion, when analyzing long texts with C and C^*, the relevance rankings are quite similar with very few exceptions. However, when a shorter text is considered, the differences should be remarkable since the low frequency of the words accentuates the differences between C and C^* (and may strongly affect other keyword detectors). To show this, instead of the whole book, we choose to analyze its shortest chapter (Chapter 3), entitled *The Struggle for Existence*, with $N = 6239$ words. The results are shown in Table II, where we observe that, indeed, the C and C^* rankings differ markedly, with the C ranking being better since the C^* ranking includes words such as "had" or "said." We note also that the ranking obtained with E_{nor} includes some nonrelevant words as a consequence of the low word frequency.

To better understand the differences between C and C^* in short texts, we show in Table III some keywords of

TABLE II. Ranking of the seven most relevant words extracted from chapter 3 of the book *On the Origin of the Species*, by Charles Darwin. Words are ordered in descending value of C (first column), C^* (second column), and E_{nor} [24] (third column).

C ranking	C^* ranking	E_{nor} ranking
varieties	varieties	or
selection	had	been
bees	cattle	we
advantage	trees	the
heath	said	heath
individual	climate	i
competition	species	bees

TABLE III. Ranking of relevant words extracted from the shortest chapter (chapter 3) of the book *On the Origin of the Species*, by Charles Darwin. We include the word frequency in the chapter (second column) and the position of the word in the relevance ranking obtained using either C (third column) or C^* (fourth column).

Word	Word count n	rank C	rank C^*
varieties	16	1	1
selection	6	2	36
advantage	5	4	197
individual	6	6	219
competition	7	7	11
natural	7	16	207

the analyzed chapter, as well as their frequencies and their positions in the C and C^* rankings. In general, for (relatively) large word frequencies, both rankings are similar. That is the case of “varieties,” which is the first ranked word in both cases. In contrast, for relevant words with low frequencies (< 10), the general behavior is that the use of C improves the results considerably: The most relevant words to the chapter (“selection,” “advantage,” “individual,” “competition,” etc.) are boosted to the top of the rank, whereas the same words are relegated to rearward ranking positions when the usual C^* is used.

These results indicate that the use of the exact, finite-case distribution of distances between consecutive instances of a randomly distributed word together with the use of appropriate boundary conditions have potential benefits in word-ranking algorithms, especially for short texts and low-frequency words, where the statistical methods are usually less efficient.

A. Generic vs. specific keywords

The previous results show that C is a convenient keyword detector, especially in the case of short texts, and the algorithm is simple: Calculate the C value for any word in the vocabulary and rank all the words in descending order of C value, and the words at the top of the ranking would be the keywords of the text with high reliability. However, within the set of keywords (words with large C value), our results of extreme clustering in Sec. IV C allow us to go a step further and suggest a way to classify keywords and separate them into two classes, generic and specific keywords.

Generic keywords are words with large C values, but which are used all along the text. Such words can be identified by two main characteristics: First, the word should be relatively frequent and, second, its C value, though large, must be smaller than the extreme clustering boundary we determined in Sec. IV C, given by $C_b(N,n)$. Note that this boundary was obtained by assuming that the word was located in a single cluster.

Conversely, specific keywords are described by two main features: First, its frequency cannot be large and, second, its C value should be close to or larger than the extreme clustering boundary. Note that this can happen only when the word is concentrated in a single cluster or a similar situation, implying that the word is used solely in a very specific context of the text.

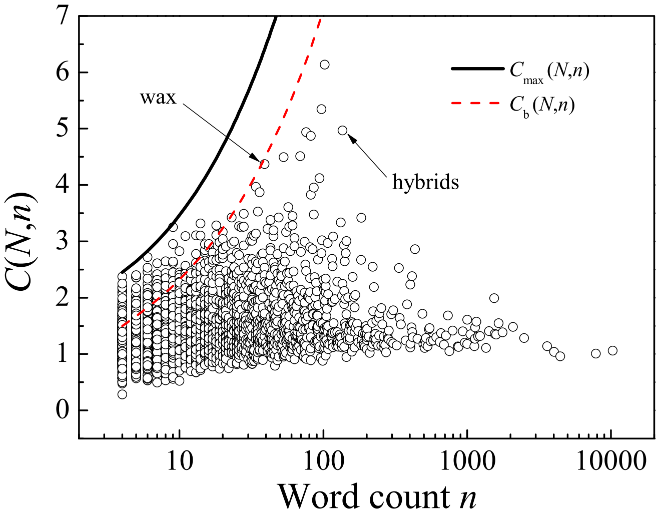

FIG. 5. Clustering values $C(N,n)$ for the words of the vocabulary of the book *On the Origin of the Species* as a function of the word count n. We include only words with $n > 3$. The lines correspond to the maximum clustering value $C_{\max}(N,n)$ and the extreme clustering boundary, $C_b(N,n)$.

In Fig. 5, we show the $C(N,n)$ value for all the words of the vocabulary of *On the Origin of the Species* as a function of the word count n. We also show with lines the maximum clustering value $C_{\max}(N,n)$ and the boundary for extreme clustering, $C_b(N,n)$. In the figure, we choose two typical examples corresponding to generic and specific keywords.

The case of “wax,” already mentioned above when discussing the results of Table I, corresponds to a specific keyword since its C value lies practically on top of the extreme clustering line. This word is only detected as a keyword if C is used (instead of C^*), as a consequence of applying boundary conditions. The specificity of “wax” can be better appreciated in Fig. 6 (top panel) where the positions of the 39 instances of the word in the whole text are indicated with

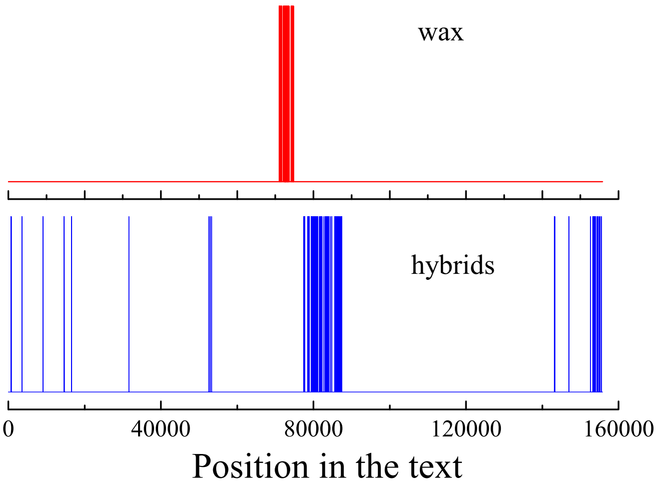

FIG. 6. Positions of the words “wax” ($n = 39$, top panel) and “hybrids” ($n = 136$, bottom panel) in the book *On the Origin of the Species*, for which $N \simeq 1.56 \times 10^5$ words.

vertical lines. Indeed, "wax" appears in the book only in the interval 71 066–74 741 words, entirely within chapter 7 of the book, where the behavior of bees is studied.

Concerning "hybrids," it is a word with larger frequency ($n = 136$) than "wax" and also with higher C value than "wax." However, the C value of "hybrids" lies well below the extreme clustering line (see Fig. 5), and this situation corresponds to a generic keyword: highly clustered but not used in a single context. Indeed, this is confirmed in Fig. 6 (bottom panel), where we show the positions of the 136 instances of "hybrids" along the whole text, and where it is shown how the word is clustered but also utilized in different contexts.

VI. APPLICATION II: AMINO ACIDS CLUSTERING IN PROTEINS

Proteins can be considered as another paradigmatic example of symbolic sequences: In general, a protein is a sequence of 20 different amino acids that can be read as a "text" where any amino acid is a "word" of the vocabulary. In addition, proteins are moderately short, with an average length of around 560 amino acids [41] for *Homo sapiens*, and the frequencies of individual amino acids are typically small. Then, when analyzing the spatial distribution of amino acids in a typical (short) protein searching for clustering, the use of C^* instead of C as will produce misleading results for two main reasons: (i) the use of the geometric distribution instead of the exact one as the random expectation is less justified than in texts (both N and n are small). In particular, the small frequency n of all the amino acids results in $c_{v,\mathrm{exact}}$ being clearly smaller than $c_{v,\mathrm{geo}}$ (see Fig. 4) and then contributing to rising C as compared to C^*. (ii) When applying boundary conditions, adding the two additional distances (then passing from $n-1$ to $n+1$ distances) can also strongly modify the value of $c_{v,\mathrm{obs}}$ when n is small. But in this latter case, the effect of considering boundary conditions does not always modify $c_{v,\mathrm{obs}}$ (and then C) in the same way, since, depending on the spatial configuration of the amino acid, the inclusion of boundary conditions can raise, lower, or leave unaffected $c_{v,\mathrm{obs}}$. As a consequence, these two effects combine to produce in general quite different C and C^* values. In the few cases of large proteins with large amino acid frequency n, we expect C and C^* to have similar values except for the cases where boundary conditions have an influence, typically for amino acids located in single clusters.

In order to have a global view of the different behavior of C and C^*, we calculate the C and C^* values for all the amino acids in the whole data set [41], and for any amino acid we obtain the ratio $r \equiv C/C^*$. The results are shown in Fig. 7, where we plot r as a function of the amino acid frequency n. As expected, for amino acids with large n, C and C^* are very similar [as the example in Fig. 9 (top panel)] and then $r \simeq 1$ in this case. However, for decreasing n we obtain a broad range of r values (broader for smaller n), but where the ratio r tends to be on average clearly greater than 1, implying that in general the clustering measure C has larger values than C^*, and then it is most sensitive to clustering detection. This does not mean that *all* the individual r values are larger than 1. Indeed, for some cases we observe that $r < 1$ and then $C < C^*$, but this situation is not frequent, as can be better appreciated in the inset of Fig. 7, where we show the probability density $p(r)$ of the r values: Despite having some r values smaller than 1, the majority of the values are in the $r > 1$ region, where $p(r)$ exhibits a long tail.

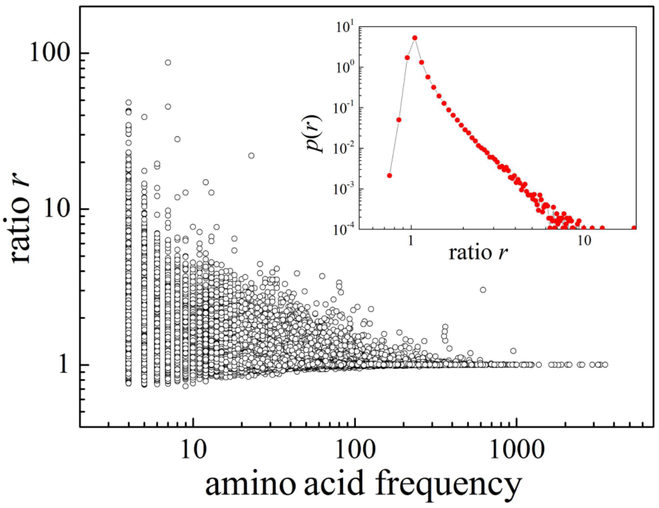

FIG. 7. Ratio $r \equiv C/C^*$ for all the amino acids in all the proteins of the Swiss-Prot database (20 204 proteins). Each ○ represents the r value obtained for any single amino acid (with frequency >3) in a single protein. Inset: Probability density of the r values.

Once we have shown that C is more appropriated for measuring amino acid clustering, we analyze the clustering behavior of the amino acids in the human proteins, which we show in Fig. 8. We observe a great diversity of behaviors, from extreme clustering ($C \gg 1$) to strong repulsion $C \ll 1$ passing through intermediate almost-randomly distributed cases ($C \simeq 1$), depending on the amino acid and the protein

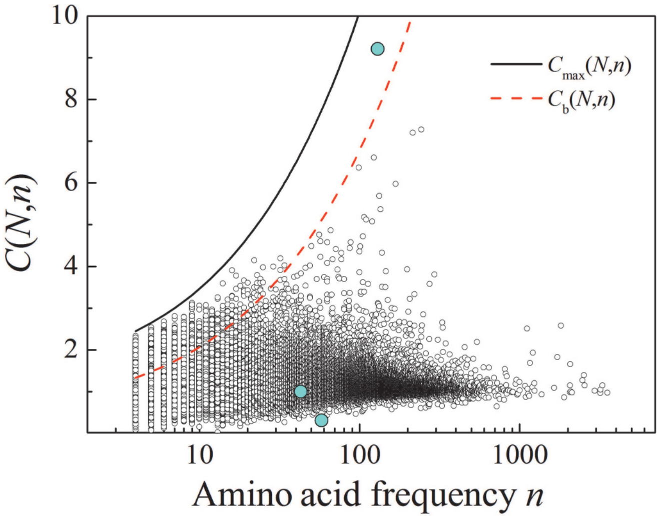

FIG. 8. Clustering measure $C(N,n)$ for all the amino acids in all the proteins of the Swiss-Prot database (20 201 proteins). Each ○ represents the $C(N,n)$ value obtained for any single amino acid (with frequency >3) in a single protein. The lines correspond to the maximum clustering value C_{max} and the boundary for extreme clustering $C_{\mathrm{b}}(N,n)$, obtained assuming that $N \gg n$ and then the approximation in Eq. (32) is valid. The three large circles correspond to the amino acids shown in Fig. 9.

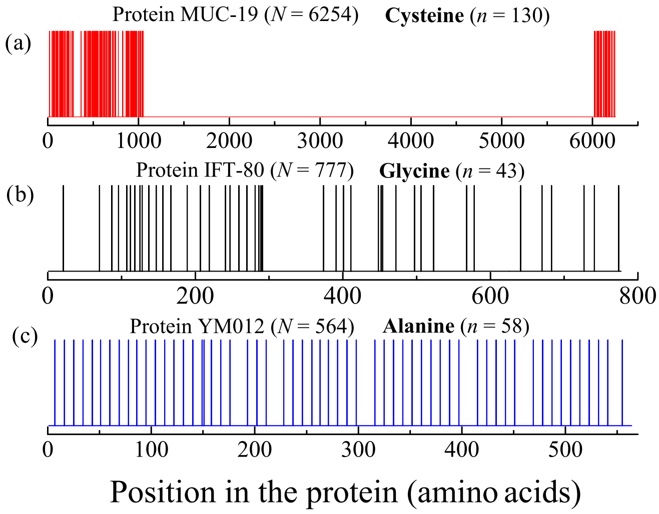

FIG. 9. Positions of three different amino acids in three different proteins. The three cases correspond to extreme clustering (a), random distribution (b), and strong repulsion (c).

considered. The advantage of using C is that its numerical values are always correctly interpreted in terms of the spatial distribution.

We end this section by showing in Fig. 9 some examples of amino acids with different behavior in the spatial distribution and how C captures this behavior. Figure 9(a) shows as vertical lines the positions of the amino acid cysteine in the human protein MUC-19. With a value of $C = 9.21$, it corresponds to the highest clustering value C for any amino acid in the whole data set (as can be also seen in Fig. 8). Indeed, for this amino acid, according to Eq. (33), we have that the boundary for extreme clustering is $C_b(6254,130) = 8.00$, clearly smaller than the observed clustering. Figure 9(b) depicts the positions of amino acid glycine in the protein IFT-80. In this case, we obtain $C = 1.00$, corresponding to a paradigmatic random distribution. Finally, Fig. 9(c) shows the positions of amino acid alanine in the protein YM102. In this case $C = 0.31$, indicating extreme self-repulsion, as can be appreciated in the figure where the positions of the amino acid follows an almost equidistant pattern.

VII. CONCLUSIONS

In this paper, we have obtained analytically the exact probability distribution of the interword distances of given words in a text (or, in general, of a symbol within a symbolic sequence), assuming that it appears at random. Traditionally, this distribution is assumed to be the geometric distribution, but this is only the case in the asymptotic limit. We analyze the main properties of the exact distribution, and in particular we show that there exists a certain frequency for the analyzed word for which the variability of the interword distances is a maximum. The knowledge of the distribution, together with the application of boundary conditions, allows us to improve the clustering detection in symbolic sequences. For written texts, our newly defined clustering measure improves considerably the keyword detection, especially in short texts. Also, an analysis of extreme clustering values allows us to classify keywords and distinguish between specific and generic keywords. For protein sequences, which are typically short, our clustering measure detects unambiguously the clustering of amino acids as compared to previous measures, properly characterizing their spatial distribution.

ACKNOWLEDGMENTS

We thank the Spanish Government (Grant No. FIS2012-36282) and the Spanish Junta de Andalucía (Grants No. FQM-7964 and No. FQM-362) for financial support.

APPENDIX A: DISTRIBUTION WITHOUT BOUNDARY CONDITIONS

Let us consider that we have a text of length N, and a given word appearing n times in that text. When no boundary conditions are considered, the interword distance distribution comes only from the "inner" $n-1$ distances (see Sec. II). When an inter-word distance d is found, such a distance appears because there are two neighbor instances of the word separated d words, and therefore their positions must correspond to one of the following situations: $[1,1+d],[2,2+d],[3,3+d]\dots,[N-d,N]$. Then, there are $N-d$ different cases to find a distance d bounded by two words, thus leaving $N-d-1$ sites to arrange the remaining $n-2$ words. As a result we have $(N-d)\binom{N-d-1}{n-2}$ different ways to obtain a distance d. If we notice that in this case the possible distances are again in the range $(1,N+1-n)$, then the probability to find a distance d can be obtained as

$$p_{N,n}(d) = \frac{(N-d)\binom{N-d-1}{n-2}}{\sum_{j=1}^{N+1-n}(N-j)\binom{N-j-1}{n-2}}. \tag{A1}$$

After performing the summation and simplifying, we obtain

$$p_{N,n}(d) = \frac{\binom{N-d}{n-1}}{\binom{N}{n}}, \tag{A2}$$

which coincides with Eq. (3), obtained using boundary conditions. Since the distributions are the same, all the statistical properties are also identical in both cases and in particular the expected mean (9).

APPENDIX B: RELATION TO GEOMETRIC FRAGMENTATION

Fragmentation problems have been an intense focus of research since the pioneering works of Lienau [42] and Mott [43–45]. Good historical reviews of the topic can be seen in Refs. [46,47]. The main aim of fragmentation research is to obtain the fragment size distribution of one-dimensional (1D), 2D, and 3D solid bodies after catastrophic events causing multiple fractures of the material. Usually the problem is tackled by using two different approaches. The first, simplest one, is called geometric fragmentation, and it consists essentially of statistically determining the fragment size distribution of a given topology when it is randomly partitioned into a number of discrete entities (see chapter 2 in Ref. [47]).

The second, more complex and realistic, approach is usually termed dynamic fragmentation, and this is a more physical approach in which loading conditions, material properties, cracks growth, fracture sites nucleation, etc., are included in the theoretical analysis, giving rise to different models [48–51].

If we restrict ourselves to the simplest case of geometric fragmentation, and in particular to the discrete fragmentation of 1D bodies, the result we have found for $p_{N,n}(d)$ can be extrapolated straightforwardly to the fragment size distribution in this case. Let us consider a segment of length L that can be broken at N evenlyspaced potential fracture sites:

$$\frac{L}{N+1},\frac{2L}{N+1},\ldots,\frac{NL}{N+1}. \tag{B1}$$

This problem is equivalent to considering the interval $[0,N+1]$, assuming that it can be fractured only at the integer positions $1,\ldots,N$ and then adopting $L/(N+1)$ as the unit of length. In this sense, the set of possible fracture points $1,\ldots,N$ coincides with the possible positions of a given symbol in a sequence of length N. If we randomly introduce n $(n \leqslant N)$ equally probable fracture points ("cuts") placed at positions j_i $(i=1,2,\ldots,n)$ such that $0<j_1<j_2<\ldots<j_n<N+1$, then we finally obtain $n+1$ fragments of lengths $\ell_0=j_1, \ell_i=j_{i+1}-j_i$ $(i=1,2,\ldots,n-1)$, and $\ell_n=N+1-j_n$. This is exactly the same situation as the one we considered before for word interoccurrence distances, and, in addition, the boundary conditions we introduced in that case appear here naturally as the lengths of the first and last fragments. Thus, the distribution of fragments lengths ℓ is then

$$p_{N,n}(\ell)=\frac{\binom{N-\ell}{n-1}}{\binom{N}{n}}. \tag{B2}$$

This same 1D discrete fragmentation problem was discussed in Ref. [48], where a recursive solution for the fragment size distribution was obtained instead of an explicit expression as (B2).

[1] I. Kanter and D. A. Kessler, Phys. Rev. Lett. **74**, 4559 (1995).

[2] R. Ferrer-i-Cancho and R. V. Solé, Proc. Natl. Acad. Sci. USA **100**, 788 (2002).

[3] C. Cattuto, V. Loretto, and L. Pietronero, Proc. Natl. Acad. Sci. USA **104**, 1461 (2007).

[4] Z. K. Zhang *et al.*, Eur. Phys. J. B **66**, 557 (2008).

[5] C. Cattuto *et al.*, Proc. Natl. Acad. Sci. USA **106**, 10511 (2009).

[6] M. A. Serrano, A. Flammini, and F. Menczer, PLoS ONE **4**, e5372 (2009).

[7] L. Lu, Z.-K. Zhang, and T. Zhou, Sci. Rep. **3**, 1082 (2013).

[8] X. Yan and P. Minnhagen, PLoS ONE **10**, e0125592 (2015).

[9] X. Yan and P. Minnhagen, Physica A **444**, 828 (2016).

[10] R. Ferrer-i-Cancho and R. V. Solé, Proc. R. Soc. B-Biol. Sci. **268**, 2261 (2001).

[11] E. G. Altmann, J. B. Pierrehumbert, and A. E. Motter, PLoS ONE **4**, e7678 (2009).

[12] L. Lu, Z.-K. Zhang, and T. Zhou, PLoS ONE **5**, e14139 (2010).

[13] D. R. Amancio, J. Stat. Mech. (2015) P03005.

[14] C. Bian *et al.*, Europhys. Lett. **113**, 18002 (2016).

[15] W. Li and K. Kaneko, Europhys. Lett **17**, 655 (1992).

[16] C.-K. Peng, S. V. Buldyrev, S. Havlin, M. Simons, H. E. Stanley, and A. L. Goldberger, Phys. Rev. E **49**, 1685 (1994).

[17] P. Bernaola-Galván, I. Grosse, P. Carpena, J. L. Oliver, R. Román-Roldán, and H. E. Stanley, Phys. Rev. Lett. **85**, 1342 (2000).

[18] I. Grosse, P. Bernaola-Galván, P. Carpena, R. Román-Roldán, J. Oliver, and H. E. Stanley, Phys. Rev. E **65**, 041905 (2002).

[19] P. Carpena, P. Bernaola-Galván, A. V. Coronado, M. Hackenberg, and J. L. Oliver, Phys. Rev. E **75**, 032903 (2007).

[20] P. Carpena, J. L. Oliver, M. Hackenberg, A. V. Coronado, G. Barturen, and P. Bernaola-Galván, Phys. Rev. E **83**, 031908 (2011).

[21] F. Dios *et al.*, Comput. Biol. Chem. **53**, 71 (2014).

[22] M. A. Montemurro and D. H. Zanette, PLoS ONE **8**, e66344 (2013).

[23] M. Ortuño *et al.*, Europhys. Lett. **57**, 759 (2002).

[24] J. P. Herrera and P. A. Pury, Eur. Phys. J. B **63**, 135 (2008); In this paper, the entropic measure E_{nor} is proposed as keyword detector. E_{nor} utilizes a *partition* of the text to calculate word entropies. The E_{nor} results in Table I (for the whole book) use the partition based on the chapters of the book, which seems to be the "natural" partition. The E_{nor} results in Table II, obtained for a single chapter of the book, use the partition based on the paragraphs of the chapter.

[25] P. Carpena, P. Bernaola-Galván, M. Hackenberg, A. V. Coronado, and J. L. Oliver, Phys. Rev. E **79**, 035102(R) (2009).

[26] A. Mehri and A. H. Darooneh, Phys. Rev. E **83**, 056106 (2011).

[27] C. Carretero-Campos *et al.*, Physica A **392**, 1481 (2013).

[28] Z. Yang *et al.*, Physica A **392**, 4523 (2013).

[29] D. R. Amancio *et al.*, PLoS ONE **8**, e67310 (2013).

[30] A. Mehri, M. Jamaati, and H. Mehri, Phys. Lett. A **379**, 1627 (2015).

[31] P. Carpena *et al.*, Gene **300**, 97 (2002).

[32] M. Hackenberg *et al.*, J. Theor. Biol. **297**, 127 (2011).

[33] M. Hackenberg *et al.*, BMC Genom. **11**, 327 (2010).

[34] M. Hackenberg *et al.*, Algorithm Mol. Biol. **6**, 2 (2011).

[35] G. F. Reed, F. Lynn, and B. D. Meade, Clin. Diagn. Lab. Immunol. **9**, 1235 (2002).

[36] K. I. Goh and A. L. Barabasi, Europhys. Lett. **81**, 48002 (2008).

[37] F. Guo *et al.*, arXiv:1506.09096.

[38] P. Carpena, P. Bernaola-Galván, and P. Ch. Ivanov, Phys. Rev. Lett. **93**, 176804 (2004).

[39] A. M. García-García and E. Cuevas, Phys. Rev. B **82**, 033412 (2010).

[40] Book downloaded from www.gutenberg.org.

[41] We have used the Swiss-Prot protein database. In particular, we have considered the set formed by the 20 204 reviewed proteins from *H. sapiens* downloaded from www.uniprot.org. For this set, the protein lengths are approximately log-normally distributed with an average length of 560 amino acids (SD = 605 amino acids).

[42] C. C. Lienau, J. Franklin Inst. **221**, 485 (1936).

[43] N. F. Mott and E. H. Linfoot, Ministry of Supply, AC 3348, January (1943).

[44] N. F. Mott, Ministry of Supply, AC 3642, March (1943).

[45] N. F. Mott, Ministry of Supply, AC 4035, May (1943).

[46] D. E. Grady and M. E. Kipp, J. Appl. Phys. **58**, 1210 (1985).

[47] D. Grady, *Fragmentation of Rings and Shells: The Legacy of N. F. Mott* (Springer, Heidelberg, 2006).

[48] D. E. Grady, J. Appl. Phys. **68**, 6099 (1990).

[49] B. L. Holian and D. E. Grady, Phys. Rev. Lett. **60**, 1355 (1988).

[50] F. Zhou, J. F. Molinari, and K. T. Ramesh, Appl. Phys. Lett. **88**, 261918 (2006).

[51] K. T. Ramesh *et al.*, Planet. Space Sci. **107**, 10 (2015).