Distributions of Distances in Information Strings

Milan Kunz*

Jurkovicova 13, 63800 Brno, Czech Republic

Zdenek Rádl

Rennenská 2, 60200 Brno, Czech Republic

Received September 21, 1997

Distances between identical symbols in information strings (biological, language, computer programs (*.exe files) are described with a different precision with four distributions: exponential, Weibull, log-normal and negative binomial. The correlations are sometimes highly significant.

INTRODUCTION

Statistical properties of information distributions, especially their extreme skewness, raised the notion of their specificity.[1−4] Determining frequencies of symbols or words was a time-consuming task suitable for shortening unbearably long time periods.[5] These linguistic studies had some pragmatical value, too; that is, learning of languages starting with the most frequent words and phrases and attribution of texts to authors.

The inverse function to frequencies (or adjacencies) are distances between identical counted objects. The distance function, the Wiener index,[6] gained a significant role in mathematical chemistry.[7]

Distances between identical symbols exist in all information strings with any number of symbols or their *k*-tuples (words). Their manual counting was even more troublesome than counting words. Therefore, such studies were made only for neighbor symbols where the local transitivity (frequencies of 2-tuples, e. g. $a \rightarrow a$, $a \rightarrow$ b) was studied by Harary and Paper.[8] Time intervals between consecutive patent applications of patentees[9] and time intervals between consecutive publications[10] were determined for some small samples. Visits in a library were analyzed by Fourier spectral analysis.[11]

Another situation exists in studies of DNA fragments and proteins. The geometrical distances between individual nucleotides, codons, and amino acids were investigated very thoroughly because gaining information from the biological material is a laborious task, whereas an additional evaluation is comparatively cheap. The long-range statistical properties of nucleic acid sequences were investigated as planar trajectories[12] or by using point geometry analysis.[13] Even the methods of linguistics were applied.[14] The biological DNA strings were, for example, decomposed as texts into syllables, words, or group of words.[15]

A stochastical generation of a string of *m* repeatings of an alphabet of *n* symbols is conventionally modeled by tossing a dice with *n*-sides. A coin is the first nontrivial model of the dice with two sides. When a coin is tossed, different long sequences appear when one result prevails. If frequencies of heads and tails in long runs are not equal, the coin is considered to be biased. The distribution of sequences between successive events (head or tail) in all possible runs is known as the negative binomial distribution. The negative binomial distribution is the inverse of the binomial distribution, and it evaluates frequencies of distances between consecutive binary symbols in their all strings. This distribution was a statistical curiosity until some decades ago. Its evaluation was a rather difficult task[16,17] because its distribution function does not exist in a closed form. Now this distribution is included in standard statistical software program packages.[18]

The distances between counted symbols in a string can be transformed to distances d_{ij} in graphs. To use graph notions, familiar for representing molecules, a string of symbols can be treated in two ways: as a star with multiple m_j indexed arcs leading from one added root to *n* individual symbols forming *n* leaves of the star, or as a forest of *n* stars, each token being a leaf of the corresponding star. There must be added *n* individual roots for all symbols. Both models have different combinatorial properties,[19] with forests of *n* stars having more elements of symmetry.

Information strings looks different; that is, coded string of the four nucleotide bases (a gene fragment FRAXGE 52 seq, beginning[20] is used as the example) has the form:

```
GAATTCAGGT   AAGCTATCTT   GAAAGGGGAA   ATATCAAAAG

CTAGAGATCA   GAGTAAGGCT   GAGACTCAGA   GTCAAGTGGG

GAAGACTAAG   TTGCAGTATG   TACTGGCAGT   GAAGATAAGT...
```

The bases are written conventionally in the groups of 10 symbols. This string can be compared with a result of tossing a tetrahedron (maybe biased, because the frequencies of all nucleotide bases are not equal). We can ask if such a string is stochastical or if this sequence is unlikely to be produced by a mere chance.

When this string of four bases is transformed into the string of 64 3-tuples, known as codons (each codon was replaced by an arbitrary ASCII symbol; for example, = represents TTC, B represents TCA, and other symbols are given in the Table 2), a part of the string of this gene fragment looks like

S0095-2338(97)00067-X CCC: $15.00
Published on Web 04/17/1998

```
=RBiPkpPXyIKw?G
QvY=DLywgYNYAQEZmS[F
= >gUOUYQQqqIjYikIarTRy=jPmAMVf^<¡wM >LpaWUJ
DkUBz^VVJ
^jG
G
fd^ABF
```

The transcription is a quite unintelligible text, as if a secret code were used. Each symbol in this string corresponds to an amino acid, except the ending symbols (G, F, J) in the rows of the transcription. These endings are three cistronic codons that divide the sequences of amino acids as blanks divide words, points sentences, and enters paragraphs. The biological strings can be compared with, for example, this text or with a computer program in the form of an exe file (APPEND.EXE DOS 6.22, an arbitrary fragment; hexadecimal coding is used to facilite printing; compare with the original in your PC):

```
00000000000000000000000000000000
009700B1FF000000005B5D7C3C3E2B3D
3B2200AE000000BC000000000004EA00
D400F70004010000000001CA0004EA00
D400F700040100012001004401D30000
```

The first question is if these three different strings have some common statistical properties, even if we know that biological, language, and computer information strings are programs that are processed differently. A gene is a biological information string that is processed by the cell for the development of the qualities that have been passed on from its parents. We are attempting to understand all details of this process. A text is an information string that is processed by the reader. We know from our experience what a reader does but we do not know yet in detail how our brain works. A computer program is processed by the computer according to an algorithm. The computers were constructed by people. Thus, all details are understandable, at least for experts.

We see immediately that the exe file contains a long sequence of the same symbol 0. This feature can be compared with blank spaces in the printed texts (empty rows or their ends), or to the CGG repeat in the gene[20] FMR-1. Such a sequence has the analagous function as the graphics and punctuation marks have for reading, they form breakpoints, and they divide significant parts of the file and give time for processing them. The sequences of amino acids between the cistronic codons have lengths that can be compared with single words, which are divided in texts by single blank spaces, or with whole sentences, which are divided by points. We know from our own experience that we do not control distances between symbols when we write something, except in verses ending with rhymes, and avoiding repeating of words. Thus, the distances appear in our texts quite spontaneously. Similarly, computer programs are formed. Fifty years ago, the axiomatic theory of communication was published,[21] which had a profound effect on the basic concepts of thermodynamics as related to quantum mechanics and statistical approaches. The bridging between the classical thermodynamics and the information theory has not yet happened.[22] For finding the common roots, new ideas and new comparative data are necessary.

Table 1. Types of Distributions of Distances between Letters in the Paper (ref 23)

			significance of the χ^2 test[b]			
letter	frequency	range	E.	W.	L. N.	N. B.
a	1010	2−88	0.000	0.000	0.000	0.000
A	63	2−1940	0.360	0.059	0.000	0.005
b	227	1−566	0.892	0.796	0.080	0.896
c	576	1−191	0.820	0.927	0.001	0.802
d	368	1−297	0.634	0.458	0.046	0.692
e	1460	1−81	0.000	0.000	0.000	0.000
E	26	3−3750	0.001	*a*	*a*	*a*
f	318	1−357	0.925	0.888	0.042	0.928
g	168	1−657	0.867	0.830	0.086	0.870
h	454	1−248	0.577	0.654	0.000	0.000
i	1148	1−102	0.000	0.000	0.000	0.000
I	75	2−1263	0.154	0.689	0.027	0.145
j	34	24−2396	0.136	0.222	0.523	*a*
k	61	1−1613	0.126	0.350	0.581	0.125
l	480	1−228	0.168	0.121	0.000	0.280
m	286	1−471	0.291	0.235	0.000	0.265
n	903	1−108	0.000	0.000	0.056	0.000
o	943	1−116	0.000	0.000	0.001	0.000
p	320	1−428	0.043	0.036	0.000	0.041
r	831	1−129	0.000	0.001	0.000	0.005
s	914	1−129	0.000	0.056	0.000	0.002
q	17	90−2568	*a*	*a*	*a*	*a*
t	1077	1−131	0.000	0.000	0.003	0.000
u	361	3−355	0.645	0.644	0.086	0.000
v	115	4−620	0.245	0.104	0.022	0.000
w	171	2−551	0.804	0.957	0.287	0.000
x	33	7−2001	0.415	0.224	0.202	*a*
y	188	3−428	0.001	0.054	0.004	0.000
z	28	1−2382	0.619	*a*	*a*	0.614

[a] The program did not make the test, insufficient degrees of freedom. [b] Distributions: E., exponential; W., Weibull; L. N., log normal; and N. B., negative binomial.

EXPERIMENTAL SECTION

The information strings in ASCII form were at first indexed with the position index *i* (*i* going from 1 to *m*) of each individual symbol in the string, and then the differences between these d_j position indexes were determined. The differences are the topological distances between the same symbols. The sets of these values were evaluated by the commercial program (STATGRAPHICS,[18] Vers. 4.0) using different statistical tests. From all available tested distributions, only four distributions gave significant results, the exponential distribution, the Weibull distribution, the log-normal distribution, and the negative binomial distribution. The actual values (mean, standard deviation, skewness, kurtosis, distribution parameters, etc.) are of little interest in this preliminary report because they differ considerably between similar tested files. Evaluation of other parallel samples has shown too big variance of results. Therefore, the results are presented in the form of the significance of the χ^2 tests (Tables 1−3) to show the general behavior of the distance distribution.

The differences between experimental and calculated values were usually great at the shortest distances (1−10), therefore the range was always adjusted to 0. Adjusting the lowest possible value to greater distances by pooling these distances increased the significance of the χ^2 tests in some cases. The significance improved dramatically sometimes, but if the pooling continued over an optimum, the significance could decrease again, as the number of freedom degrees became too low for the test.

We start our discussion with texts (Table 1). The lower and

Table 2. Types of Distributions of Distances between Consecutive Codons and Amino Acids in FRAXCDNA (ref 24)

amino acid			significance of the χ^2 test[a]					amino acid			significance of the χ^2 test[a]				
codon	F.	range	E.	W.	L. N.	N. B.	note[b]	codon	F.	range	E.	W.	L. N.	N. B.	note[b]
TTT: < Phe	184	1−458	0.000	0.238	0.092	0.000		CGG: Ä Arg	54	1−552	0.761	0.854	0.026	0.730	
TTC: = Phe	92	1−708	0.109	0.255	0.055	0.118		AGA: j Arg	132	2−213	0.517	0.367	0.001	0.000	
phenylalanine	276	1−458	0.000	0.551	0.159	0.000		AGG: k Arg	184	1−188	0.247	0.182	0.001	0.233	
TCT: ’ Ser	125	1−303	0.623	0.615	0.018	0.623		arginine	486	1−92	0.072	0.042	0.000	0.073	
TCC: A Ser	114	1−258	0.623	0.622	0.006	0.657		ATT: Ö Ile	93	2−421	0.002	0.001	0.000	0.000	p
TCA: B Ser	100	2−367	0.581	0.877	0.084	0.039		ATC: Ü Ile	91	1−357	0.238	0.198	0.043	0.239	
TCG: C Ser	35	3−853	0.935	0.428	0.469	0.029		ATA: ˆ Ile	80	1−724	0.023	0.445	0.426	0.023	l
AGT: h Ser	85	2−526	0.025	0.021	0.008	0.001		Ileucine	264	1−199	0.004	0.007	0.031	0.032	
AGC: i Ser	115	1−286	0.273	0.170	0.273	0.271	l	ATG: _ Met	68	1−487	0.443	0.293	0.129	0.444	
serine	574	1−97	0.089	0.220	0.000	0.131		ACT: ‘ Thr	102	1−246	0.863	0.574	0.038	0.899	
TAT: D Tyr	65	2−568	0.105	0.071	0.043	0.000	l	ACC: a Thr	102	3−459	0.774	0.587	0.116	0.000	
TAC: E Tyr	62	2−555	0.257	0.324	0.026	0.164	p	ACA: b Thr	101	1−241	0.885	0.826	0.226	0.868	
tyrosine	127	1−330	0.090	0.061	0.002	0.106		ACG: c Thr	33	1−988	0.355	0.178	0.032	0.370	p
TAA: F ochre	78	1−545	0.229	0.069	0.069	0.254		threonine	338	1−129	0.284	0.286	0.001	0.103	
TAG: G amber	77	1−417	0.192	0.132	0.007	0.254	p	AAT: d Asn	108	1−345	0.277	0.205	0.007	0.278	c
TGT: H Cys	105	1−294	0.978	0.951	0.144	0.977		AAC: e Asn	78	1−456	0.001	0.001	0.013	0.001	p, l
TGC: I Cys	110	1−415	0.104	0.184	0.054	0.088	p	asparagine	186	1−250	0.948	0.942	0.048	0.759	
cysteine	215	1−150	0.929	0.936	0.034	0.854		AAA: f Lys	257	1−282	0.000	0.001	0.000	0.000	l, p
TGA: J opal	133	3−347	0.708	0.289	0.084	0.000		AAG: g Lys	109	1−338	0.012	0.168	0.004	0.012	l, p
TGG: K Try	152	1−279	0.937	0.991	0.299	0.912		lysine	366	1−206	0.000	0.000	0.000	0.000	
TTA: > Leu	81	1−611	0.797	0.925	0.187	0.640		GGT: l Val	74	1−446	0.311	0.246	0.062	0.346	
TTG: ? Leu	125	1−284	0.813	0.609	0.432	0.712		GTC: m Val	70	2−601	0.050	0.306	0.153	0.630	p
CTT: L Leu	124	1−324	0.343	0.233	0.002	0.341		GTA: n Val	62	4−636	0.021	0.018	0.007	0.008	
CTC: M Leu	155	1−347	0.389	0.666	0.015	0.282		GTG: o Val	118	1−272	0.199	0.182	0.002	0.196	
CTA: N Leu	88	1−381	0.771	0.815	0.009	0.764		valine	324	1−151	0.072	0.047	0.000	0.265	
CTG: O Leu	163	1−344	0.227	0.255	0.268	0.144		GCT: p Ala	126	1−377	0.544	0.411	0.202	0.570	
leucine	736	1−63	0.000	0.000	0.000	0.000		GCC: q Ala	132	1−221	0.417	0.346	0.003	0.437	
CCT: P Pro	160	1−213	0.644	0.497	0.035	0.634	p	GCA: r Ala	110	1−318	0.272	0.586	0.377	0.272	
CCC: Q Pro	133	1−439	0.247	0.842	0.247	0.664		GCG: s Ala	42	2−421	0.353	0.111	0.042	0.113	
CCA: R Pro	161	1−292	0.263	0.140	0.000	0.369	c	alanine	410	1−142	0.265	0.408	0.011	0.073	l
CCG: S Pro	57	2−569	0.437	0.602	0.109	0.000		GAT: t Asp	105	1−346	0.750	0.852	0.146	0.749	
proline	511	1−87	0.000	0.000	0.000	0.002		GAC: u Asp	61	2−634	0.464	0.301	0.003	0.062	p
CAT: T His	107	1−263	0.568	0.586	0.138	0.450	l	aspartic acid	166	1−207	0.312	0.312	0.009	0.598	p
CAC: U His	116	1−442	0.140	0.647	0.010	0.136		GAA: v Glu	100	1−395	0.230	0.353	0.141	0.234	p
histidine	223	1−150	0.602	0.538	0.006	0.382		GAG: w Glu	179	1−240	0.313	0.852	0.045	0.229	l
CAA: V Gln	112	1−301	0.434	0.432	0.126	0.395	p	glutamic acid	279	1−180	0.006	0.101	0.003	0.001	l
CAG: W Gln	166	1−242	0.213	0.103	0.002	0.177	p	GGT: x Gly	105	1−346	0.615	0.852	0.145	0.749	
glutamine	278	1−150	0.602	0.538	0.006	0.382		GGC: y Gly	158	2−209	0.192	0.145	0.006	0.000	p
CGT: X Arg	41	1−667	0.615	0.829	0.565	0.596		GGA: z Gly	123	1−278	0.022	0.015	0.000	0.021	
CGC: Y Arg	39	1−632	0.104	0.184	0.054	0.099	p	GGG: ä Gly	144	1−245	0.086	0.854	0.093	0.730	l, p
CGA: Z Arg	36	8−738	0.823	0.631	0.304	0.000		glycine	530	1−81	0.000	0.001	0.000	0.000	

[a] Distributions: E., exponential; W., Weibull; L. N., log normal; and N. B., negative binomial. [b] p= peak, single higher count representing about one-half of the χ^2 value; l = low, single lower count representing about one-half of the χ^2 value; c crater, the lower count with neighbor counts higher than expected.

upper case letters were counted together, except the vowels. The distributions of distances between the vowels are highly irregular, but the upper case vowels are dispersed more regularly, at least A and I in the given example. The frequency of other upper case vowels was too low to give the sufficient number of freedom degrees for the χ^2 test. Some distributions of distances between consonants are highly regular, especially their tails, if the low distances inside words are pooled. They are described with a different precision with four distributions: exponential, Weibull, log-normal, and negative binomial. Sometimes it is rather difficult to decide which distribution is the best one for fitting. Three distributions are closer: exponential, Weibull, and negative binomial. The log-normal distribution appears at other letters as the best one.

A similar picture is obtained with codons in the FRAXCDNA seq fragment (Table 2). Here some highly regular distributions of codons give irregular distributions of amino acids. These distributions are polymodal, some distances are significantly lower or higher than expected. There appear structures that are suggestive of impact craters, two peaks divided by one value lower than expected. The forbidden distances are either shortened or prolongated. This property of this string can be compared with long lists of chemical names of derivatives, when the basic parts with their distinct letters repeat at the end of each name, which can explain the irregularities of the vowels. The vowels are in all kind of words, as auxiliary and key words are, and these distributions are mixed together as codons in amino acids. The numerals from the exe file (Table 3) with the hexadecimal coding correspond to the nucleotide bases in the DNA. There appeared one outlier distance >1000 at all numerals. This distance was produced by one long string of the numeral 0, which together with the property of the hexadecimal code having zero in many 2-tuples, gave an unexpected high relative frequency 0.3998 of the distance 1 of this numeral. The eventual deletion of this long string somewhat improved the correlations. The odd frequencies of the numeral 0 are significantly lower than even ones due to the use of the zero in the 2-tuple hexadecimal code. The shorter CGG repeat coincident with a breakpoint cluster region in the FRAXGE 52, found by inspecting the gene fragment[23] and corresponding to strings of the numeral 0 in the exe file, gave the high value of the corresponding χ^2 test, too. Thus, such a feature can be detected by a statistical analysis of the distance distribution.

DISCUSSION

To understand the results, we at first construct the limiting cases of possible strings, comparing them with physical

Table 3. Types of Distributions of Distances between the Numerals Letters in the Exe File APPEND.EXE from DOS 6.22

			significance of the χ^2 test[b]			
numeral	frequency	range	E.	W.	L.N	N. B.
1	1049	1−974	0.174	0.116	0.002	0.199
2	1705	1−1005	0.000	*a*	0.000	0.000
3	1317	1−1017	0.058	0.000	0.000	0.093
4	1294	1−1054	0.032	*a*	0.005	0.000
5	1113	1−1040	0.000	0.007	0.073	0.000
6	1355	1−1167	0.000	*a*	0.026	*a*
7	1263	1−1028	0.000	*a*	0.131	0.000
8	1342	1−1143	0.000	*a*	*c*	0.000
9	554	1−1027	0.009	0.020	0.076	0.010
0	4966	1−40	*a*	*a*	*a*	*a*

[a] The program did not make the test, insufficient degrees of freedom. [b] Distribution: E., exponential; W., Weibull, L. N., log normal; N. B., negative binomial. [c] When the extremal value 1143 was excluded, the significance of the χ^2 test was 0.254.

bodies, a crystal and an unmixed blend of substances. The analogy is obscured by the different dimensionalities of physical bodies and information strings, respectively, by the size of the given examples. A three-dimensional crystal has the distances uniformly distributed in its axes. A two-dimensional section could be compared with a page of 24 repeats of ACGT, where all horizontal and vertical distances between the equal symbols are 4:

ACGTACGTACGTACGTACGTACGT
CGTACGTACGTACGTACGTACGTA
GTACGTACGTACGTACGTACGTAC
TACGTACGTACGTACGTACGTACG

The distribution of distances is monotone, and the standard deviation is zero.

An unmixed blend of substances could be compared with the string, where equal symbols are together in one place, all distances between them are 1, except the distances of the last symbol in the group to the first symbol in the following group and the distances from the beginning. These distances are different but they can be made equal by by cycling the string:

AAAAAACCCCCGGGGGGTTTTTT
AAAAAACCCCCGGGGGGTTTTTT
AAAAAACCCCCGGGGGGTTTTTT
AAAAAACCCCCGGGGGGTTTTTT

The distribution of distances is highly skewed, the standard deviation is maximal, if all equal symbols are lumped into one group. This distribution can be compared with the position of types in the printer's case before the typographer forms meaningfull strings. His work changes the position of symbols, which is equivalent with the mixing. Because the frequency of symbols is not changed by setting, the process is changing the symmetry of the string. All strings can be described by their cycle index.

The visual inspection shows, that the strings studied are not in these limiting states. As the tabulated results show, the distributions of distances between some identical symbols in the information strings can be fitted well by distributions of the probability theory, especially their tails, when the shortest distances are pooled. The distributions of distances are described with a different precision with four skewed distributions:[25] exponential, Weibull, log-normal, and negative binomial. The true form of the distribution lies between them.

The exponential distribution has only one adjustable parameter. The log-normal, the negative binomial, and the Weibull distributions have two adjustable parameters. For finding the best form of the distribution, it could be better to use a more parametrical distribution, maybe the three parametrical distribution that was proposed for the distributions of molecular weights of polymers by Kubín[26−27] could be suitable.

The negative binomial distribution appears spontaneously as their mean in long runs of tossing coins. The individual results do not depend on previous ones. This is a genuine stochastic process. In the information theory flipping coin is known as white noise.

Three distributions that also gave significant fits with the behavior of distances between some symbols, exponential, Weibull, and log-normal, demand other relations between symbols, but they can be considered to be stochastic, too. Any deviations from these distributions show that either a rare event occurred accidentally with a low expectation, or some efforts or forces effecting distances shaped the string, showing that the formation of the string was not a simple stochastic process. For example the unmixed zeroes in the exe file and some regularities in the other two strings.

Thus it is interesting that the highest χ^2 values were observed at groups of codons that are common to one amino acid. These codons, corresponding to synonyms, appear in the string quite randomly, with the highly significant χ^2 tests, but the corresponding amino acids have the unsignificant χ^2 tests. The lower significance of the χ^2 test at the amino acids coded by more codons is due to the modality of their distributions. There appear peaks, valleys, and craters. This phenomenon can perhaps explain the lower significance of the χ^2 tests of the codons corresponding to only one amino acid and of the vowels. They must be used as necessary for the given purpose.

This obvious interpretation of the results has one fault, that is, it needs an explanation of why this difference exists, why the codons are used differently, and why they do not simply copy the distributions of their common amino acids. It looks like some rules of a good style existed for the DNA design.

It can be concluded that the preliminary analysis of distances in information strings gave some interesting results. Their importance can be evaluated only after obtaining and comparing more data. The results with with other English samples and Czech texts were analogous to the given example. Unfortunately, the STATGRAPHICS program was not suitable for evaluating longer sequences and it will be necessary to elaborate techniques for comparing parallel results.

There appeared one unsolved problem connected with the entropy. The information entropy is calculated from frequencies of symbols and does not depend on their positions, there is no measure for the effect of mixing on the calculated value. Because the mixing in thermodynamics is a spontaneously going process, the entropy is growing by mixing. Information extent and information distance were studied by Ruch and co-workers.[28−31] They proposed these concepts to describe the quantitative aspect of statistical information and gain information. They tried to use the concept of mixing distance to develop an intuitive understanding of partially ordered properties. They believed that these concepts are superior to statistical entropy concepts, but they did not made any actual measurements of distances.

The determination of distances in information strings gives a possibility to calculate the distance entropies. They play perhaps some role till now unrecognized. Computer programs were developped within our lifetimes, languages within some millions of years, and genetic code within some billions years. The distributions of distances between identical symbols in information strings have similar properties. These similarities of information strings could help to understand mysteries of genetics.

REFERENCES AND NOTES

(1) Haitun, S. D. Stationary Scientometric Distributions I: Different Approximations. *Scientometrics* **1982**, *4*, 525.

(2) Haitun, S. D. Stationary Scientometric Distributions II: Non Gaussian Nature of Scientific Activities. *Scientometrics* **1982**, *4*, 89−101.

(3) Haitun, S. D. Stationary Scientometric Distributions III: The Role of the Zipf Distribution. *Scientometrics* **1982**, *5*, 375−395.

(4) Kunz, M. Plots against Information Laws, *Science and Science of Science* **1995**, *3* (1−2), 91−97.

(5) Yule, G. U. *The Statistical Study of Literary Vocabulary*; Cambridge University Press: Cambridge, 1944.

(6) Wiener, H. Structural Determination of Paraffin Boiling Points. *J. Am. Chem. Soc*. **69**, *17*, 1949.

(7) Mihalić, Z.; Veljan, D.; Amić, D.; Nikolić, S.; Plavsić, D.; Trinajstić, N. The Distance Matrix in Chemistry. *J. Math. Chem.* **1992**, *11*, 223−258.

(8) Harary, F.; Paper, H. H. Toward a General Calculus of Phonemic Distribution. *Language* **1957**, *33*, 143−169.

(9) Kunz, M. Time Spectra of Patent Information. *Scientometrics* **1987**, *11*, 163−173.

(10) Kunz, M. About Metrics of Bibliometrics. *J. Chem. Inf. Comput. Sci.* **1993**, *33*, 193−196.

(11) McGrath, W. E. Periodicity in Academic Library Circulation, a Spectral Analysis. In *Book of Abstracts*, *Part I*; Kretschmer, H., Ed., Fourth International Conference on Bibliometrics, Informetrics and Scientometrics, September 11−15, 1993, Berlin.

(12) Ninio, J.; Mizraji, E. Perceptible Features in Graphical Representations of Nucleic Acid Sequences. In *Visualizing Biological Information*; Pickover, C. A., Ed.; Word Scientific: Singapore, 1995; pp 33−42.

(13) Huen, Y. K. Representation of Biological Sequences Using Point Geometry Analysis. In *Visualizing Biological Information*; Pickover, C. A., Ed.; Word Scientific: Singapore, 1995; pp 165−182.

(14) Stanley, H. E.; Buldyrev, S. V.; Goldberger, A. L.; Havlin, S.; Mantegna, R. N.; Peng, C.-K.; Simons, M. *NATO ASI Ser., Ser. E*, **1996**, *322* (Physics of Biomaterials: Fluctuations, Selfassembly and Evolution), 219−234.

(15) Schmitt, A. O.; Ebeling, W.; Herzel, H. The Modular Structure of Informational Sequences. *Biosystems* **1996**, *37*, 199−210.

(16) Irwing, J. O. The Place of Mathematics in Medical and Biological Statistics. *J. Roy. Stat. Soc. A* **1963**, *126*, 1−45.

(17) Schilling, M. F. Long Run Predictions. *Math. Horizon* **1994**, *Spring*, 10−12.

(18) STATGRAPHICS, Statistical Graphics Corporation.

(19) Kunz, M. Entropies and Information Indices of Star Forests. *Collect. Czech. Chem. Commun.* **1986**, *51*, 1856−1863.

(20) Verkerk, A. J. M. H.; Pieretti, M.; Sutcliffe, J. S.; Fu, Y.-H.; Kuhl, D. P. A.; Pizzuti, A.; Reiner, O.; Richards S.; Victoria, M. F.; Zhang, F. Identification of a Gene (FMR-1) Containing a CGG Repeat Coincident with a Breakpoint Cluster Region Exhibiting Length Variation in Fragile X Syndrome. *Cell* **1991**, *65*, 905−914.

(21) Shannon, C. E. The Mathematical Theory of Communication. *Bell System Technical Journal* **1948**, *27*, 379 and 623.

(22) Tribus, M. *Information and Thermodynamics: Bridging the Two Cultures* **1986**, *11*, 347−359.

(23) Exner, O.; Kunz, M. Citation Histories of Related Papers in the Field of Chemical Correlation Analysis. *Scientometrics* **1995**, *32*, 3−10.

(24) Verkerk, A. J. M. H.; de Graaff, E.; De Boulle, K.; Eichler, E. E.; Konecki, D. S.; Reyniers, E.; Mance, A.; Poustka, A.; Willems, P. J.; Nelson, D. L.; Oostra, B. A. Alternative Splicing in Fragile X Gene FMR-1. *Human Mol. Gen.* **1993**, *2*, 399−404.

(25) Hastings, N. A. J.; Peacock, J. B. *Statistical Distributions*; Butterworths: London, 1975.

(26) Kubín, M. A Generalized Exponential Function as a Model of Polymer Distribution Curves. *Collect. Czech. Chem. Commun.*, **1967**, *32*, 1505−1517.

(27) Kubín, M. A Generalized Exponential Function as a Model Distribution Curves of Polymers. II. Determination of Parameters by Means of Number, Weight and z-Averages. *Collect. Czech. Chem. Commun.*, **1969**, *34*, 703−707.

(28) Ruch, E. The Diagram Lattice as Structural Principle, *Theor. Chim. Acta (Berl.)* **1975**, *38*, 167−183.

(29) Ruch, E.; Schranner, R.; Seligman, T. H. The Mixing Distance *J. Chem. Phys.* **1978**, *69*, 386−392.

(30) Ruch, E.; Lesche, B. Information Extent and Information Distance *J. Chem. Phys.*, **1978**, *69*, 393−401.

(31) Ruch, E.; Schranner, R.; Seligman, T. H. Generalization of a Theorem by Hardy, Littlewood, and Pólya. *J. Math. Anal. Appl.* **XXXX**, *76*, 222−229.

CI970067E