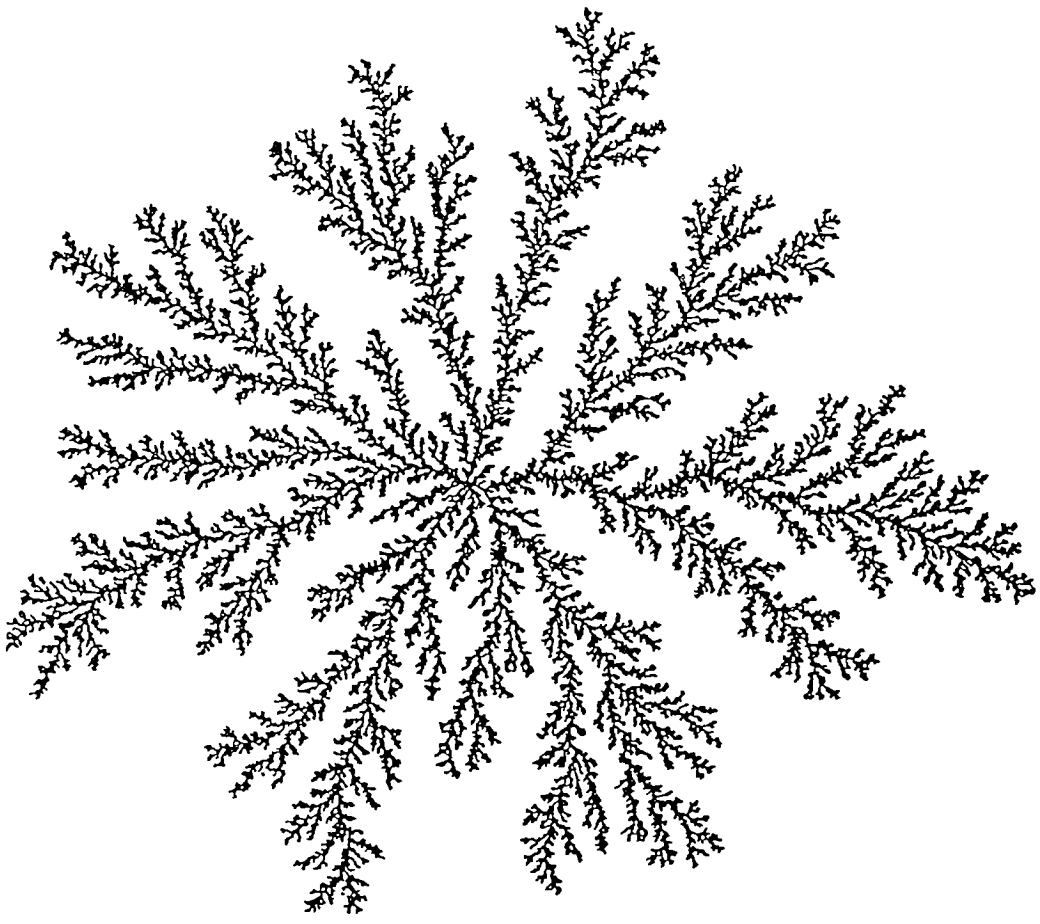
Symmetry: Culture and Science

SPECIAL ISSUE
Fractals

The Quarterly of the
International Society for the
Interdisciplinary Study of Symmetry
(ISIS-Symmetry)

Editors:
György Darvas and Dénes Nagy

Volume 4, Number 3, 1993



DLA fractal cluster
of 10^6 particles

Symmetry: Culture and Science
Vol. 4, No. 3, 1993, 229-241

SYMMETRY: CULTURE AND SCIENCE

LONG RANGE CORRELATION IN HUMAN WRITINGS

Alain Schenkel*, Jun Zhang**, Yi-Cheng Zhang*

Zhang, Yi-Cheng, physicist, (b. Henan, China P. R., 1956).
Addresses: *Institut de Physique Théorique, Université de Fribourg, CH-1700 Fribourg, Switzerland; **Physics Laboratory, H. C. Orsted Institute, Universitesparken 5 DK-2100, Copenhagen, Denmark.
Fields of interest: theoretical physics, fractal geometry, literature.

Abstract: *The long range correlation in the frequency of occurrence of a given string in various human writings is studied by mapping them into a simple 1d random walk model. This approach allows us to obtain better quality scaling data than the traditional power spectrum methods. Optimally written computer programs seem to have highest scaling exponent, i.e., close to that of ideal 1/f noise. Literature, on the other hand, leads to smaller exponents. The* Bible, *however, has the strongest correlation among the Roman letter writings examined.*

1. INTRODUCTION

More than four decades ago, Shannon (1951) already established the fundamental concept of entropy for human writings, which is a meaningful measure of information content. In a general sense, a novel or a piece of poems, a play of music, a computer program, etc., can be regarded as a one dimensional string of symbols. We are interested in the correlation of these strings. However, when a string is very long (a typical novel can easily attain more than 10^6 symbols), it is practically impossible to evaluate the Shannon information entropy. Thus approximate methods are in need. Grassberger (1989) has devised a scheme to estimate Shannon's entropy for a vast selection of literature. However, the entropy itself does not reveal directly the correlation properties of a string of symbols. To probe the correlation in a string of symbols, the mostly used method is to study the Fourier power spectrum. Once a string of symbols is considered a one dimensional signal, it can be readily compared with the traditional study of $1/f$ noise (Press, 1978).

Recently, a new measure of complexity, which is the generalization of the entropy, was introduced by one of us (Zhang, Y-C., 1991). Under two mild assumptions: (1) it should be a linear superposition of the scale dependent entropies; (2) it should be an extensive variable, one arrives at a quantity, called complexity, which is uniquely defined. The complexity is argued to be a relevant measure of correlated systems. In particular, the complexity attains its maximal value when the 1d string

has precisely the $1/f$ power spectrum. Like the entropy itself, the complexity is impractical to measure in reality.

Peng *et al.* (1992) have recently studied the DNA symbol-strings with significant findings. They examined a large class of the DNA data and found generically long range correlation. The exponent of the power spectrum lies between 0 and 1, i.e., between white noise and $1/f$ noise (maximum complexity). What is remarkable, it is that this exponent never exceeds 1, unlike what is often encountered in general $1/f$-like noise in other domains (Press, 1978), in which the exponent deviates from 1 as likely upwards as downwards. These results were later confirmed by Voss (1992) using the more traditional spectrum method.

2. THE MODEL AND THE METHOD

The method used by Peng *et al.* (1992) is to map the string into a random walk model. They then study the correlation of such walks. This method has the advantage over the traditional power spectrum or direct calculation of the string's correlation in that it yields high quality scaling data.

We found that the random walk model is particularly suitable for our calculations. First let us consider various writings in Roman letter. An English text can be seen as a string of 26 letters plus punctuation symbols. We limit the total number of symbols to $2^5 = 32$ by adding to the letters some of the punctuation symbols. Upper cases and lower cases make no difference, and empty space and extra symbols are ignored. We then represent each of these 32 symbols by a binary number of 5 bits. An arbitrary table is thus established *ad hoc*, for instance, '*a*' is represented by 00000, '*b*' 10000, etc. Now the writing is reduced to a string of 0 and 1. Other codes are as valid. We believe for large length scaling the details of coding are not important. In principle we should treat each original symbol independently and deal with a walk in an n-dimensional space (n being the number of independent symbols). Reducing the original problem to binary string introduces correlation among the symbols. Though objections were raised about the appropriateness of the random walk model to the DNA sequences (Voss, 1992), we believe the large length (beyond a word length) scaling behavior is correctly preserved after the dimensional reduction.

Following Peng *et al.* (1992), we interpret the 0's as downwards steps and 1's as upwards steps. Denote these two cases by $u(t) = \pm 1$, respectively. Define the random walk position $f(l)$ after l steps,

$$f(l) = \sum_{i=1}^{l} u(i)\,, \tag{1}$$

and the difference d_l over a distance l by

$$d_l = f(l_0 + l) - f(l_0). \tag{2}$$

The quantity of interest is then the mean square fluctuations of d_l

$$F^2(l) = \overline{d_l^2} - (\overline{d_l})^2\,, \tag{3}$$

where the average is done over l_0. F is expected, in the limit of long range correlation, to be a scaling function of l of the form

$$F(l) \sim l^{\alpha} \tag{4}$$

As it has been emphasized by Peng *et al.* (1992), we have

$$F^2(l) = \sum_{i=1}^{l} \sum_{j=1}^{l} c(i-j)\,, \tag{5}$$

where $c(i-j) = \overline{u(i)u(j)}$.

The above method is more reliable, because that $F^2(l)$ is an average of the correlation functions $c(i-j)$. The exponent α is related to the usual power spectrum exponent β by the following relation

$$\alpha = \frac{\beta+1}{2}\,, \tag{6}$$

where β is defined by the power spectrum of u, $S_u(f) \sim 1/f^{\beta}$. The value of ½ for α implies that the string is not long range correlated and it leads to white noise. On the other hand, $\alpha = 1$ corresponds to the scale invariant $1/f$ noise which is also called the maximal complexity limit (Zhang, Y-C., 1991).

We have selected randomly a few writings, their choice being based on the availability of the electronically stored versions. We want to consider longer texts, whenever available. For all the types of writings studied in this work (literature, human random numbers, computer programs), the averaged absolute value of the walk's displacement scales is

$$\overline{|d_l|} \simeq cl\,, \tag{7}$$

where c is a constant of order one. This implies that the first moment of the walks scales with the exponent 1 exactly. Though the code is symmetric in 0 and 1, the original text usually is not. It may contain, e.g., more *a*'s than *z*'s. This intrinsic asymmetric feature of the texts creates a systematic bias in the walk so that the walks experience a linear drift on average, thus the exponent is 1 for the first moments. This linear drift, however, does not influence the fluctuation properties which are contained in the second moments $F^2(l)$. This was also noted in DNA-sequence study (Peng *et al.*, 1992). In all the following studies, the maximal length for the scaling calculations is always smaller than 1/10 of the total available length. This ensures sufficient average statistics.

3. LITERATURE, FROM THE BIBLE TO SHAKESPEARE

In Figures 1.a and 1.b we show the results of two different texts by the same author (H. G. Wells). We see that the data appear to scale in the large length limit with distinct but well defined exponents (0.52 and 0.62). The value of 0.52 is not very far from the white noise value ½ (should one conclude that the work is almost noise?). Other writings show also scaling behavior. It seems generally true that the same writer does not have a same exponent. Like the two Shakespearean texts (see in

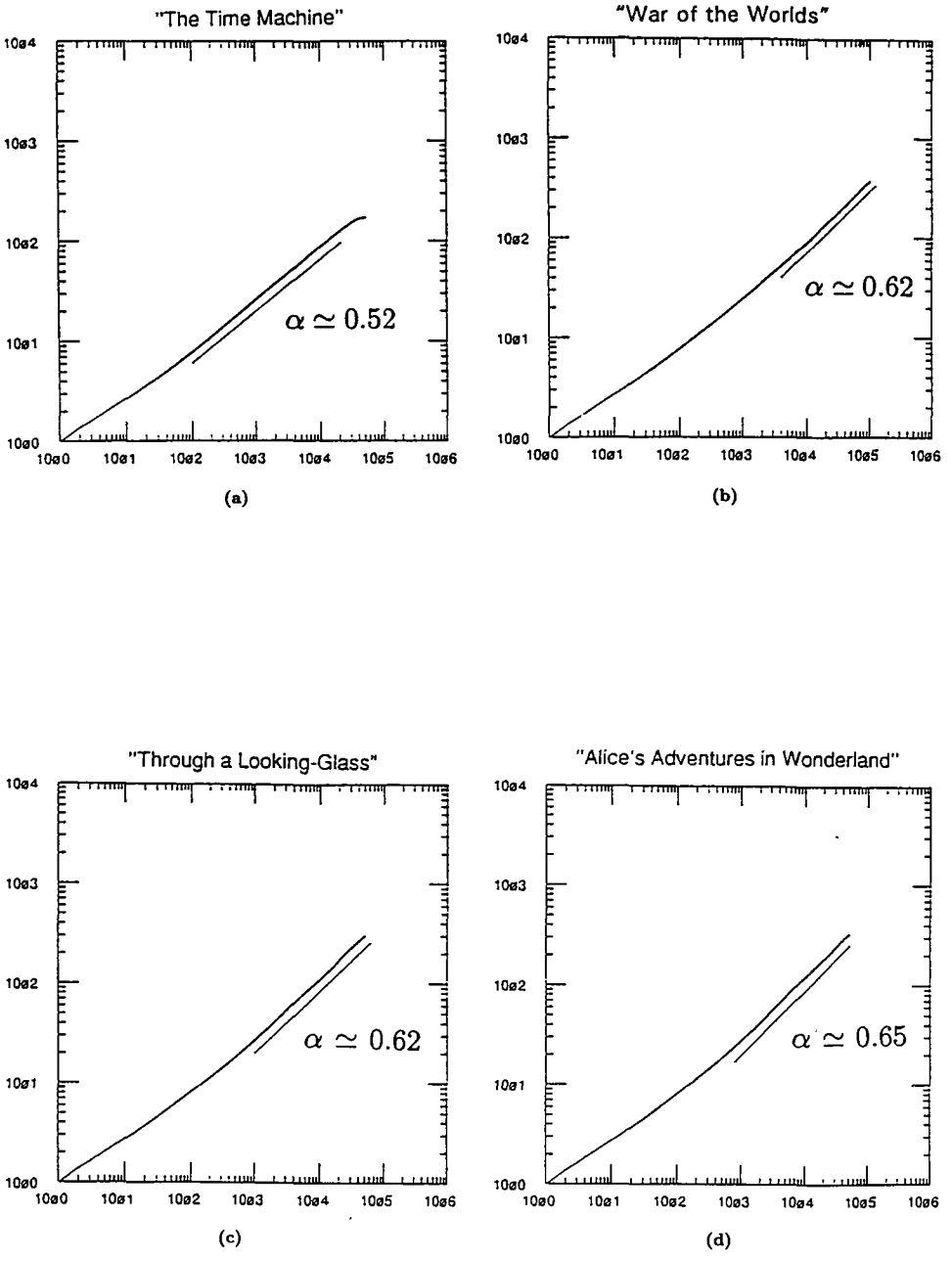


Figure 1: (a) *Time Machine* by Herbert G. Wells; (b) *War of the Worlds*, by Herbert G. Wells; (c) *Through a Looking-Glass*, by Lewis Carroll; (d) *Alice's Adventures in Wonderland*, by Lewis Carroll.

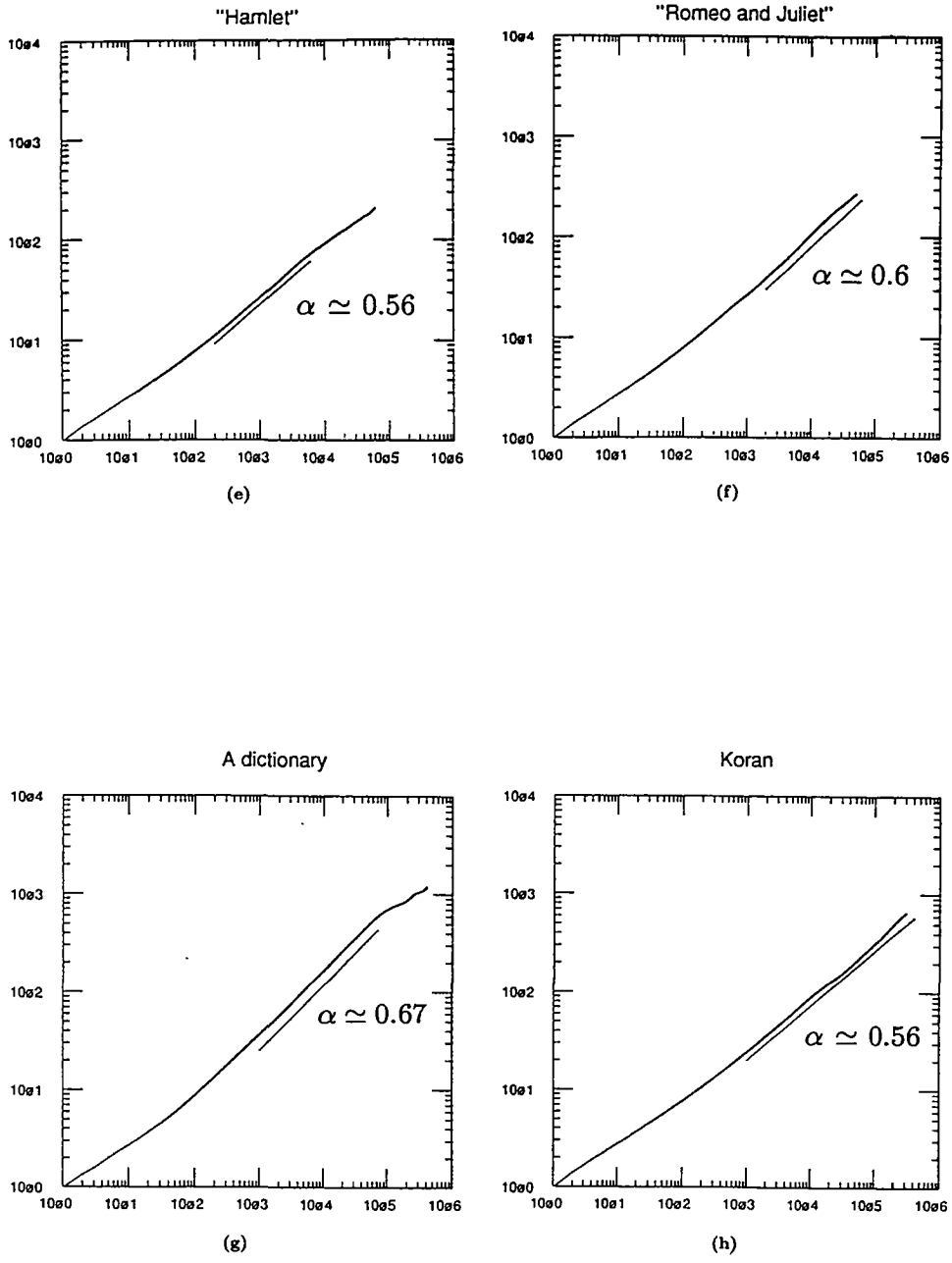
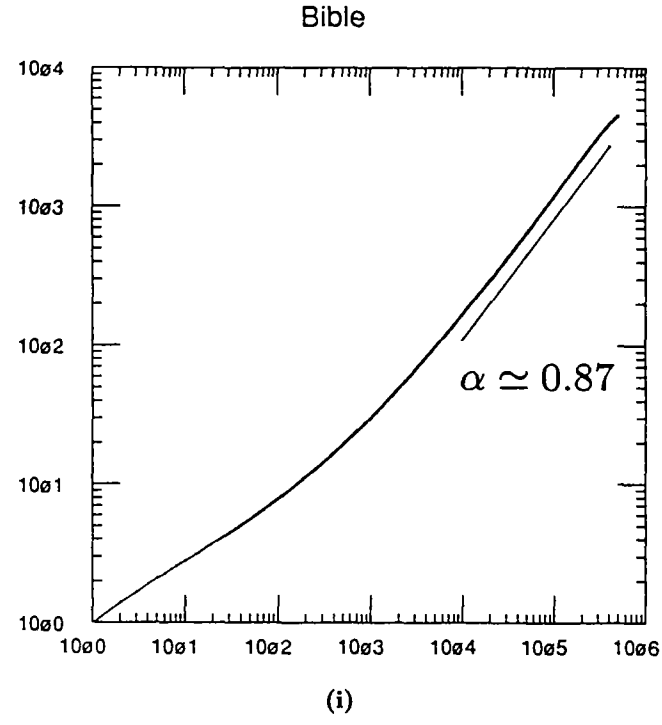


Figure 1 (continued): (e) *Hamlet*, by W. Shakespeare; (f) *Romeo and Juliet*, by W. Shakespeare; (g) Roget's *Thesaurus*, electronic version of the edition published in 1911 by the Crowell Company; (h) *Koran*, translated by M. H. Shakir and published by Tharike Tarsile Qur'an Inc.



(i)

Figure 1 (continued): (i) *Bible*, the first half of the *Old Testament*, electronic version by the university of California at Berkeley.

Figures 1.e and 1.f) have exponents 0.56 and 0.60, respectively. A dictionary of expressions (the Roget's Thesaurus) is also available. In principle, a dictionary does not carry meaningful messages since explanations of items are piled up senselessly. Each item in the dictionary may have perhaps coherent meaning, but it is usually not longer than a few hundreds bits. Therefore it is surprising to find there is consistent long range correlation all the way to 10^5 bits, as it is shown in Figure 1.g. We have carried out tests on long texts (also for computer programs), to check this dictionary effect. We cut a meaningful text into smaller pieces of identical length l (l varies 10-10,000). The pieces are then reshuffled in a completely random way and then chained together again, as if a dictionary is made. However, we have verified that there is a clear crossover of scaling at length l, the data beyond the scale l show the white noise exponent $\alpha = ½$. This indicates that the organization of items in the dictionary may not be random, as we might have assumed. However, the cause of the long range correlation remains to be found.

Religious writings do not have simple origins. Here we study them in a mechanical way without attachment. In Figures 1.h and 1.i the results from the *Koran* and *Bible* are shown. While *Koran* scales with an 'usual' exponent (~ 0.56), the *Bible* seems to have the scaling exponent, about ~ 0.85, largest among the various texts which we have studied. To show the advantage of the random walks model, we plot also the Fourier power spectrum of the *Koran* and *Bible* in Figure 2. It is clear that it is

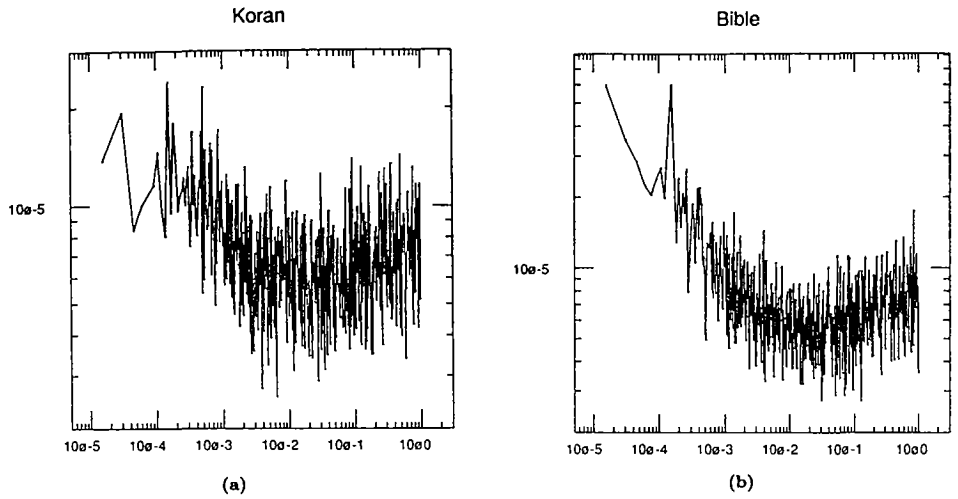


Figure 2: Fourier power spectrum of (**a**) the *Koran* and (**b**) the *Bible*.

hard to draw quantitative conclusions from the spectra. If we look carefully at the Figures 1.a to 1.i, we see a slight upward bend at around 30 ± 10 steps. These correspond approximately to the length of a typical word. The weaker correlation within a word implies that the letters in a word are less correlated than the words in a text.

4. HUMAN RANDOM NUMBER GENERATOR HAS INTRINSIC CORRELATION

Let us turn to a different type of writings. One of us (A. Schenkel) was asked to mimic a random number generator by writing down with equal probability the numbers from 0 to 9. Each digit is then transformed into bits by a simple code. No matter how hard a human being tries to be random and fair, he is intrinsically biased. Aware of this inescapable fate of human-bias-induced correlation, he tried his best to avoid it. What happened is that due to his over-caution of not introducing correlation, he has actually introduced anti-correlation over intermediate ranges. In Figure 3.a we see this clearly from the power spectrum. The downward trend at the high frequency end implies anti-correlation. This reflects the tendency that he tried not to repeat the same number in the next steps. However, over long scales these details do not have influence. The long range correlation is there despite the all the *conscious* efforts against them. As a comparison, we have checked that any reasonable computer can generate a quasi-random series which is

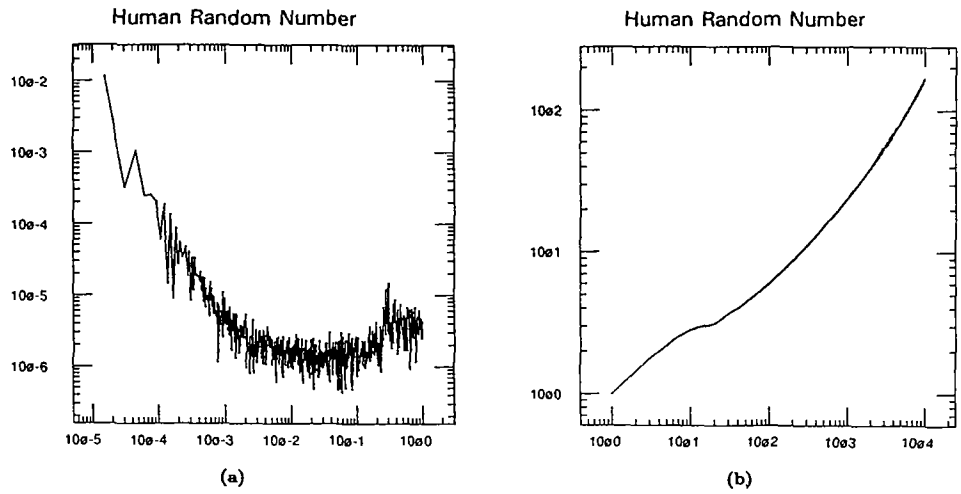


Figure 3: *Fourier power spectrum of (a) mean square fluctuations and (b) of a human generated random sequence.*

basically white. In Figure 3.b we show the result of fluctuations in random walks representation. It is remarkable that we cannot define a scaling exponent since the data continues to bend upwards, though last decade approximately may be fitted with an exponent 1. Human beings when asked to perform a certain task, all possible factors influence their decision, both of physical nature like endurance, pains etc., as well as psychological factors such as feeling bored, personal preferences, habits etc. In writing down a random series, these factors enter most unconsciously. This is to be contrasted with the case of literature or computer programs, which though also written by human beings, are done consciously and with given purposes. However, here we refrain from entering the dangerous domain of philosophical debates.

5. COMPUTER PROGRAMS, OPTIMIZATION YIELDS LARGER EXPONENTS

As another type of writings, though with completely different functions, we consider next the ubiquitous computer programs. A computer program is typically a string of commands, subroutines, etc. Since computer programs are designed to perform certain tasks, usually optimally written (like the subroutines in a software library), much correlation is built-in by construction. This appears in the forms of do-loops, goto statements, usage of same arrays, etc. We consider a program a string of symbols as before. We consider only the programs after compilation, i.e.,

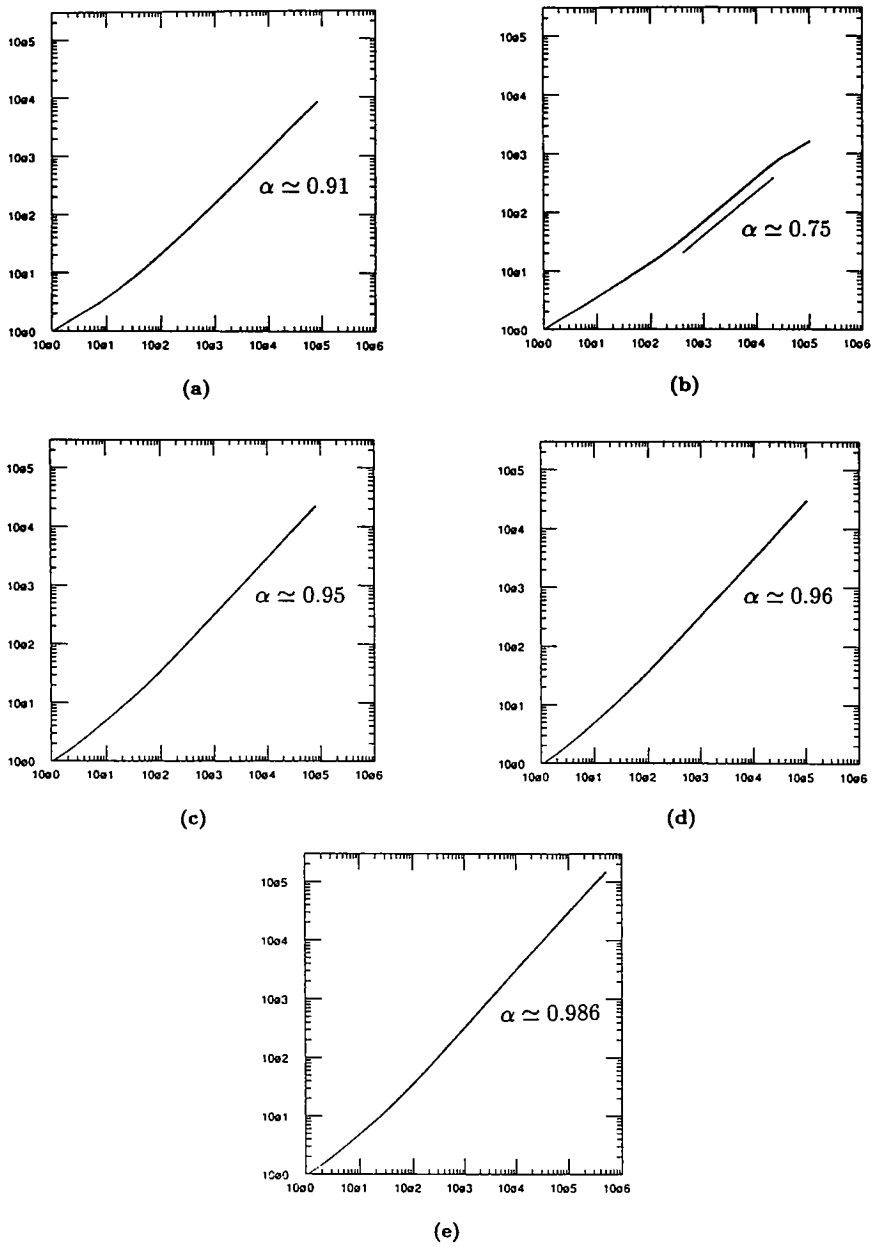


Figure 4: (**a**) A program written with *Turbo Pascal 5.5*; (**b**) The *Knuth's* T_EX version for MS-Dos PC's written with *Turbo Pascal 4.0* by Wayne G. Sullivan and Peter Breitenlohner, Department of Mathematics, University College Belfield, Dublin; (**c**) Concatenation of *Fortran* subroutines (dlsarg, dlftgh, devbhf, dgvccg, dlfdcb, dlsbrr, and devlsb) available at the IMSL MATH/Library; (**d**) Concatenation of *Fortran* subroutines (molch, fps2h and twodq) available at the IMSL MATH/Library; (**e**) Concatenation of 86 *Fortran* subroutines available at the IMSL MATH/Library.

in their executable versions (*.exe* files). A program after compilation consists of a list of commands and memory addresses which can be easily transformed into binary strings.

In Figure 4.a is the result from a program written by one of us (A. Schenkel) which calculates and draws fractal mountains. It contains various statements and also subroutines supplied in the software package Turbo Pascal 5.5 for MS-DOS PC's. We see that it scales with a considerable higher exponent (~ 0.91) than the previous cases. Also the quality of the scaling exponent is remarkably good. One can hardly believe it came from an experiment observation! In Figure 4.b we show the result from the Turbo Pascal version of the TEX compiler written by Wayne G. Sullivan and Peter Breitenlohner. We consider also programs in Fortran for Vax/VMX, mainly subroutines in the Math/Library IMSL. Usually these subroutines are highly optimized and we may expect that the scaling quality is also better. Indeed two sequences (Figs. 4.c and 4.d) assembled from some of the longest subroutines scale very well, with exponents ~ 0.95 and ~ 0.96. We have randomly packed 86 subroutines to form a meaningless superprogram, which represents about 1/10 of the total content of the IMSL Math/Library. The scaling result is shown in Figure 4.e: over more than four decades the data can hardly be told apart from a straight line, with an exponent ~ 0.986. In principle, we can still go for longer length scales, our limitation due only to computer resources rather than to the length of the data.

6. FRACTAL LANDSCAPE OF A WALK (FROM THE BIBLE AND OTHERS)

It is instructive to examine directly the walks. In Figure 5, we present various walks $f(l)$, with the linear drift subtracted. We see that *Romeo and Juliet* and a computer program have similar random shapes. Such configurations are called self-affine fractals, and a random walk with the exponent $\alpha > \frac{1}{2}$ is called a persistent walk (Mandelbrot, 1982). The larger is the α, the more persistent is the walk. Upon a careful inspection of Figures 5.a, and 5.b we may conclude that the program is more persistent than *Romeo and Juliet*. This is not surprising since their α's are different. We note also the particularities in the fine structures in Figures 5.a and 5.b. While the program shows more abrupt turns on all the scales, *Romeo and Juliet* leaves more gentle tracks. These details are not represented in the second moments $F^2(l)$, rather they appeare in still higher moments of the walks. In Figures 5.c-e we also show the configurations of the *Bible* in various resolution scales. The global structure (Fig. 5.c) is about half of the *Old Testament* including *II Chronicles*, seems to be a smooth silhouette. Upon successive magnifications the self-reproducing feature of fractal scaling becomes more and more evident.

One may wonder what is behind the particular shape of a writing such as those above. Is it the intrinsic meaning of a text or the particular language used that dictates the global configurations? We are inclined to believe that the large scale global structures should be independent of the choice of the languages, the carriers of messages. We might use this criterion to judge if a translation is faithful to the original intentions, by checking the overlap of the walks of the translations of a same work. For instance, the same version of the *Bible* in French, say, should have

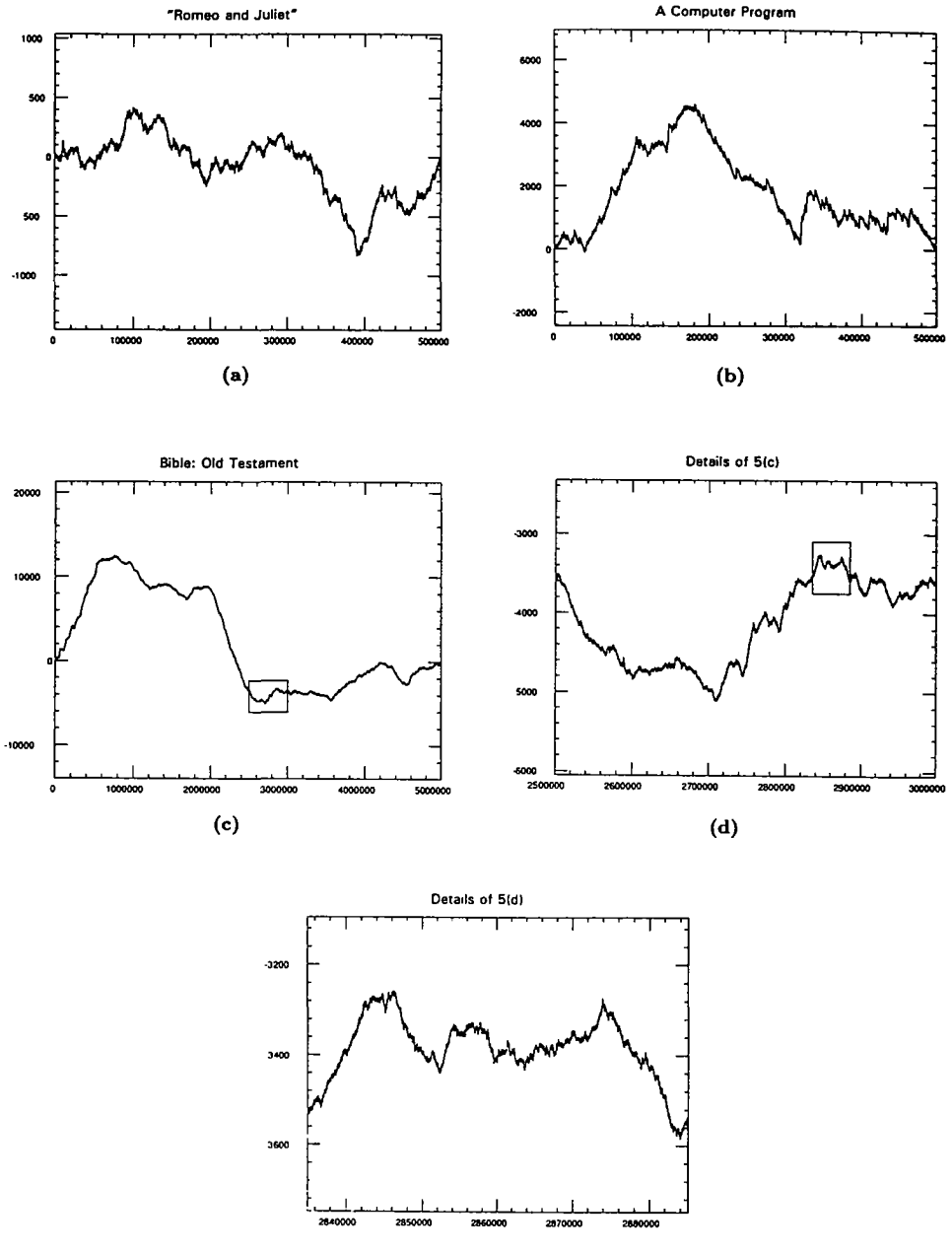


Figure 5: (a) *Romeo and Juliet*, random walk representation; (b) A portion of the superprogram in Fig. 4.e; (c) *Bible, Old Testament* only, in random walk representation; (d) A detail of the *Old Testament*, corresponding to the parts from "Deuteronomy" to "Joshua" in *II Chronicles*; (e) A further detail of Fig. 5.d, corresponding to the part of "Judges".

the same ups and downs in Figure 5.c. We expect that the quasi-overlap persists down to the length scales of phrases, where grammars differ from language to language. This remains to be verified in large scale studies. Computer programs, on the other hand, do not share this overlap property. A program for achieving the same function, when written in two different languages (e.g., Basic and Fortran) may have completely different global structures. This is because the computer languages are *nonlocal*, unlike human languages. More work is needed to better define this concept.

7. (INCONCLUSIVE) CONCLUSION

In conclusion we have experimentally studied various human writings using the random walk model introduced by Peng *et al.* We found intriguing long range correlation in three different types of writings. The Roman letter writings show rather good scaling, but the long range correlation is weak, for they are characterized by smaller exponents. Human random number generator shows more complicated behavior than simple scaling, this may remind us that not all induced correlation can be described by a scaling exponent. Computer programs show remarkably better scaling quality. The long range correlation seems strongest in optimally written Library subroutines. The scaling exponent is very close to 1, the scale-independent $1/f$ noise, or the maximal complexity limit. In all the above studies, the exponent α can approach the value 1 but never reaches it, or alternatively, overshoots it. This is due to the fact that the original signals $u(i)$ are strictly bounded by definition (0, 1). They cannot drift themselves, unlike the physical $1/f$-type noise (Press, 1978). In principle this kind of signals can have also $\alpha > 1$. However if one imposes the translation invariance condition on the signals, one can show that the value 1 is the upper limit of α. What is amusing is that the scaling exponent less than unity implies a self-affine fractal, whereas the value unity implies that the walk is an isotropic fractal (Mandelbrot, 1982). It is remarkable that the maximum complexity principle (Zhang, Y-C., 1991), or the walks from optimized computer programs, take the isotropic fractal scaling as the ideal limit. Maximal complexity configurations arise naturally in nonlinear, nonequilibrium, open and dissipative processes. In fact, one can show that spatial-temporal fluctuations in such processes approach a dynamical attractor in which the complexity measure attains its maximum.

One will do gross injustice without mentioning Zipf (1949), who pioneered in the modern scaling theory by analyzing words frequencies in dozens of languages (at a time when no computer was available), and who also introduced many revealing concepts which still today do not find satisfying answers. However our results on long range correlation in the written texts do not seem to be covered in his seminal work. We did verify that in our texts the so-called Zipf's Law (1949) is approximately satisfied. This should not be a surprise since Mandelbrot already shown that in a random language (which he called Monkey's language) where one can analytically obtain approximately the Zipf's Law for word frequencies (see reference cited in Mandelbrot, 1982).

Like the study of DNA sequences, the explanation and origins of the scaling behavior is still beyond us. The only point we want to make is that unlike DNA or other natural signals, human writings are made by ourselves, *consciously* or *uncon-*

sciously. Therefore, the mystery of their scaling behavior should in principle be easier to decode.

ACKNOWLEDGEMENT

We acknowledge helpful discussions with G. Dietler.

REFERENCES

Grassberger, P. (1989) Estimating the information content of symbol sequences and efficient codes, *IEEE Transactions, Information Theory*, 35, 669.

Mandelbrot, B. B. (1982) *The Fractal Geometry of Nature*, New York: W. H. Freeman and Co., 460 pp.

Peng, C-K., Buldyrev, S. V, Goldberger, A. L., Havlin, S., Sciortino F., Simons, M., and Stanley, H. E. (1992) Long-range correlations in nucleotid sequences, *Nature*, 356, 168.

Press, W. H. (1978) Flicker noise in astronomy and elsewhere, *Comments on Astrophysics*, 7, 103.

Shannon, C. E. (1951) Prediction and entropy of printed English, *Bell System Technical Journal*, 30, 50, and references therein.

Voss, R. F. (1992) Evolution of long-range fractal correlations and $1/f$ noise in DNA base sequences, *Physical Review Letters*, 68, 3805.

Zhang, Y-C. (1991) Complexity and $1/f$ noise: a phase space approach, *Journal de Physique*, (France), I1, 971.

Zipf, G. K. (1949) *Human Behavior and the Principle of Least Effort*, New York: Hafner Publishing Co.