Hierarchical Structures in Complex Systems: from DNA to Financial Markets

Giovanni Bonanno, Fabrizio Lillo, Salvatore Micciché and Rosario N. Mantegna [1]

Istituto Nazionale per la Fisica della Materia, Unità di Palermo
and
Dipartimento di Fisica e Tecnologie Relative, Università di Palermo,
Viale delle Scienze, I-90128, Palermo, Italy

Abstract. In this paper we discuss the concepts of short-range and long-range correlated stochastic processes and we investigate the presence of such variables in two model complex systems. The selected model systems are DNA sequences of complete genomes and financial time series of equities traded in a stock market. Specifically, by starting from our research results, we discuss the statistical properties of (i) coding and non-coding regions of DNA and (ii) equity returns and volatility in financial markets. The stylized facts about these variables are presented and discussed with a focus on the statistical tools already used and/or still needed to better characterize these model complex systems.

INTRODUCTION

The interest of some physicists in the analysis and modeling of complex systems by using paradigms and tools of statistical physics has increased during the past years [1–5]. Complex systems are open systems whose dynamic depends on so many control parameters that a deterministic description is pragmatically hopeless. By investigating complex systems, physicists attempt to model hierarchical systems where different elements are interacting in a nonlinear way often in the absence of constants of motion.

The achievements of the past years in the theoretical modeling of physical systems at the critical state [6], nonlinear dynamic systems [7], physical systems in the presence of temporal and/or quenched randomness [8] and self-organized critical systems [9] provide the theoretical background used to model complex systems. These systems are disparate and different the one from the other. However, there

[1] e-mail: mantegna@unipa.it, web page: http://lagash.dft.unipa.it

are aspects which are in common to several of them. For some of these systems a statistical description turns out to be effective and fruitful. It provides the "stylized" facts that are observed in similar complex systems under different conditions supplying statistical regularities that can be used as basis for a theoretical modeling.

In these lectures we will focus on some aspects and properties observed in two well monitored complex systems: the DNA and financial markets. In both systems, we focus on variables that can be modeled either in terms of short-range correlated stochastic processes (i.e. stochastic processes having a typical scale) or in terms of long-range correlated stochastic processes. The simultaneous presence of short-range and long-range correlated stochastic variables which are linked the ones with the others in the same system is rather common in complex systems. The aim of these lectures is to show that methods of statistical physics can be fruitfully used to investigate statistical regularities in complex systems as disparate as DNA and financial markets.

It is worth pointing out that the studies of these systems connote applied aspects of primary importance for our society. The investigation of the selected complex systems is motivated today by a series of facts:

1. The sequencing of complete genomes of various species is occurring in the present period. In fact, since the sequencing of the bacterium *Haemophilus influenzae* completed in 1996, the genomes of several organisms have been completely sequenced and new complete genomes are released approximately each week. We are then in a crucial period for the investigation of genome organization. More than 35 genomes of prokaryotes are today available and, among eukaryotes, the genome of *Yeast*, *C. elegance*, *Drosophila* and *Homo sapiens* (from Celera Corp.) have been completely sequenced. The Human Genome project is speedily proceeding and more than 200 new sequencing projects of disparate organisms are under realization.

2. During the 1990s the empirical investigations of financial processes begun to be performed by using high-frequency data (namely data recorded with time intervals as short as few seconds). Physicists have been concurring to perform empirical studies of financial markets and have published results in the physics literature since 1991 [10,11]. Today the amount of financial data easily available is ever increasing. These high-frequency data are allowing for studies that can focus on details of the price formation that have been impossible before [12,13]. These studies are possible either in mature markets (as, for example the New York Stock Exchange) or in emerging markets (as, for example, the Budapest Stock Exchange).

In summary the genomic DNA of living organisms and financial markets are two *complex systems* which are overally well defined and which are controlled by a (broad) set of constraints which are relatively stable in time. Moreover a huge amount of electronic data is available allowing for the search of statistical regularities that can be used in the theoretical modeling.

STATISTICAL INDEPENDENCE AND CORRELATION IN MODEL COMPLEX SYSTEMS

The analysis and modeling of diffusive stochastic processes has a long tradition dating back to 1827 when Robert Brown first observed what is nowadays called an example of Brownian motion. In the physical sciences, the theoretical modeling of Brownian motion was first formalized by Einstein and Smoluchowski. On the pure mathematical side, essential contributions were provided by mathematicians such as Wiener and Doob. The stochastic process describing the Brownian motion proposed by Einstein and Smoluchowski, today often addressed as Wiener process, provides a satisfactory physical description only for relatively large values of t. Such limitation is overcome by the Ornstein-Uhlenbeck (OU) process [14]. In the OU process the stochastic evolution of a diffusive particle can be described at any time. For this reason, the OU model has assumed the role of paradigmatic model for Gaussian distributed diffusive physical systems.

Diffusive stochastic processes, i.e. stochastic processes $x(t)$ characterized by a linear growth in time of the variance

$$< x^2(t) > \propto t \tag{1}$$

are quite common in physical systems. A linear growth in time of the variance implies that successive steps of the stochastic process are short-range correlated.

Deviations from a diffusive process are sometime observed in several systems. In fact, superdiffusive,

$$< x^2(t) > \propto t^\beta \quad ; \quad \beta > 1 \tag{2}$$

and subdiffusive,

$$< x^2(t) > \propto t^\beta \quad ; \quad \beta < 1 \tag{3}$$

random processes have been detected and investigated in physical and complex systems. A classical example of superdiffusive random process is Richardson's observation that two particles moving in a turbulent fluid which at time $t = 0$ are originally placed very close the one with the other have a relative separation ℓ at time t that follows the relation $< \ell^2(t) > \propto t^3$ [15]. Most recent examples include anomalous kinetics in chaotic dynamics due to flights and trapping [16,17], anomalous diffusion in aggregate of amphiphilic molecules [18] and anomalous diffusion in a two-dimensional rotating flow [19]. Subdiffusive stochastic processes have also been detected and investigated. Examples include charge transport in amorphous semiconductors [20,21] and the dynamics of a bead in polymers [22]. Another class of stochastic processes which are not diffusive in a simple way is the one characterized by a variance with a stretched exponential time dependence. When such a process is Gaussian distributed the probability of return to the origin $P_0(t)$ is

described by the Kohlrausch law $P_0(t) \simeq \exp[-t^\beta]$. Similar behaviors are observed in glassy systems and in random walks in ultrametric spaces [23].

Stochastic processes with anomalous diffusion are long-range correlated (when the stochastic process is superdiffusive) or long-range anti-correlated (when the stochastic process is subdiffusive).

The modeling of some of the above discussed anomalous diffusing stochastic processes has been done by using a variety of approaches. To cite some examples, we recall that superdiffusive and subdiffusive processes have been modeled by writing down a generalized diffusion equation [15,24–26], by introducing Lévy walks models [27], by using a fractional Fokker-Planck equation approach [28] or by using "ad hoc" stochastic models such as, for example, the fractional Brownian motion [29].

STYLIZED FACTS

Statistical properties of biological and financial complex systems have been investigated by several research groups during the past years. This research activity has produced an accumulation of a number of results that may be referred to as "stylized" facts. Here we shortly summarize the ones we consider most relevant to our research.

I. DNA

DNA is the primary source of hereditary information for almost all the living species. The main role of DNA is to store information about the sequences of amino acids characterizing the proteins necessary for the life of any living organism. The DNA performing this function is addressed as coding DNA. This information is stored by using the universal genetic code. Information about amino acid sequence of a protein is not stored in a standardized way in the DNA of various organisms. Specifically regions of non-coding DNA are observed within the region of DNA codifying a specific protein. This means that in the DNA we find regions of coding and non-coding DNA. The non-coding regions may be classified as introns (non-coding DNA found within a gene) and intergenic DNA (non-coding DNA found between two genes in a given region of DNA). Non-coding DNA is a great mystery at the present moment. Scholars have different views about its purpose and function. Some researchers believe it is just a by-product of evolution and sometime address it as "junk" DNA. Others think it has some regulatory purposes [30]. It may be useful to recall that in human DNA 97% of DNA is non-coding.

To illustrate with a very simple example the degree of complexity of genome organization even in very simple organisms, in Fig. 1 we show the location of coding and non-coding regions observed in the first 1,000,000 base pairs of the complete genome of *E. coli*. This model eubacterium shows an intricate location of the coding regions which are present in both strands of DNA and are separated

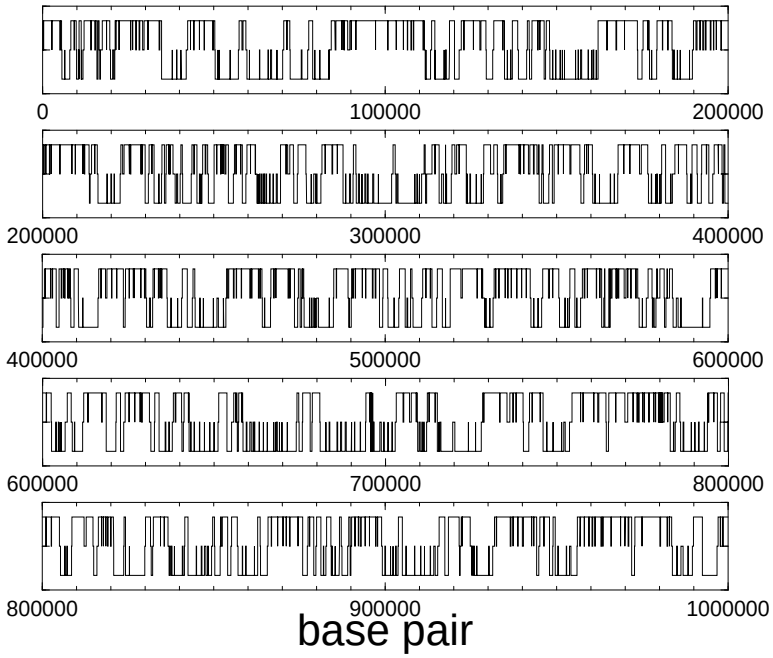

FIGURE 1. Scheme of the location of coding regions and non-coding regions in the first 1,000,000 base pairs of the complete genome of *E. coli*. A horizontal line at the "up" level indicates a coding regions located in the 5' to 3' DNA coding strand, a horizontal line at the "middle" level indicates non-coding regions whereas a horizontal line at the "bottom" level indicates a coding regions codified in the complementary coding strand.

by intergenic non-coding regions. In complete genomes of prokaryotes non-coding DNA is mostly intergenic. Very few introns are observed.

If a similar degree of complexity is already observed in relatively simple organisms the complexity observed in higher organisms is probably much higher. This observation motivates an investigation of the statistical properties of complete genomes.

Several studies have shown that the statistical properties of coding and non-coding DNA are different. Specifically, long-range correlations and a higher degree of compositional complexity (namely the existence of several length scales in the base compositional fluctuations) have been observed in non-coding DNA [31–33]. On the other hand, coding DNA regions are characterized by short-range correlated fluctuations when the bias due to isochores and/or base pair concentration fluctuations of DNA is taken into account.

This kind of studies have been performed by investigating the correlation properties of DNA sequences by introducing the so-called DNA-walk. A DNA-walk is a walk obtained by mapping the symbolic sequence of the nucleotides composing DNA into a one-dimensional random function [31,32]. This formal mapping has provided robust statistical indication that a DNA-walk obtained from coding regions is a short-range correlated random walk characterized by normal diffusion [33]. On the contrary, long-range correlations are observed in the DNA-walks ob-

tained from non-coding DNA regions [31,33].

The reasons for the presence of long-range correlation in non-coding regions are still under debate and several models and interpretations have been proposed to explain this empirical behavior.

Additional differences are observed in the degree of redundancy observed in the symbolic sequences of base pairs in coding and non-coding regions and in the statistical properties of n-gram words detected in the sequences. Specifically, non-coding regions are more redundant than coding regions and the statistical properties of the most frequent n-gram words show a Zipf-like behavior in non-coding regions while are well approximated by a logarithmic function of the rank in coding regions [34,35].

In conventional Zipf analysis [36], the frequency of occurrence of words present in a given text is measured by counting the number of occurrences of each word throughout the text and dividing this value by the number of words. The frequency of occurrence f of each word is then ordered from the most frequent to the least frequent value. The position of each word in this ordered list is called its rank R.

By studying log-log plots of word frequency versus word rank, Zipf discovered a heuristic power-law relation between them

$$f = \frac{a}{R^{\zeta}} \tag{4}$$

The exponent ζ was found to be close to 1 in several texts written in different natural languages [36,37]. Equation (1) is called the Zipf law. From the publication of Zipf's seminal work there have been several attempts to prove, as well as disprove, the Zipf law [37–39].

Zipf behavior has been universally observed in analyses of natural and technical languages. It is important to note that since Zipf analysis is a statistical technique, it can be performed on texts of unknown languages (with the only limitation of being able to recognize the basic semantic unit: the word). No knowledge of the investigated language is required.

On the other hand, conventional Zipf analysis has been criticized [37] since Zipf scaling can emerge in a purely random symbolic sequence if one character is defined as a "word" delimiter [37,39]. Hence, while the observation of power-law behavior in a conventional Zipf analysis is *necessary* in natural and formal languages, it is *not sufficient* to prove the existence of non-Markovian correlations in the analyzed symbolic sequence.

Here we report about n-tuple Zipf analysis used to investigate the statistical properties of coding and non-coding regions of DNA [35]. The n-tuple Zipf analysis of a symbolic text differs from the conventional Zipf analysis performed in natural languages. In a symbolic text, the elementary semantic unit (the word) is not immediately recognizable (if present). In the study of the complexity of symbolic sequences the usual approach is to investigate the statistical properties of the substrings of length n obtained from the symbolic text by progressively shifting over the text a window of n characters. Zipf scaling does not emerge in n-tuple

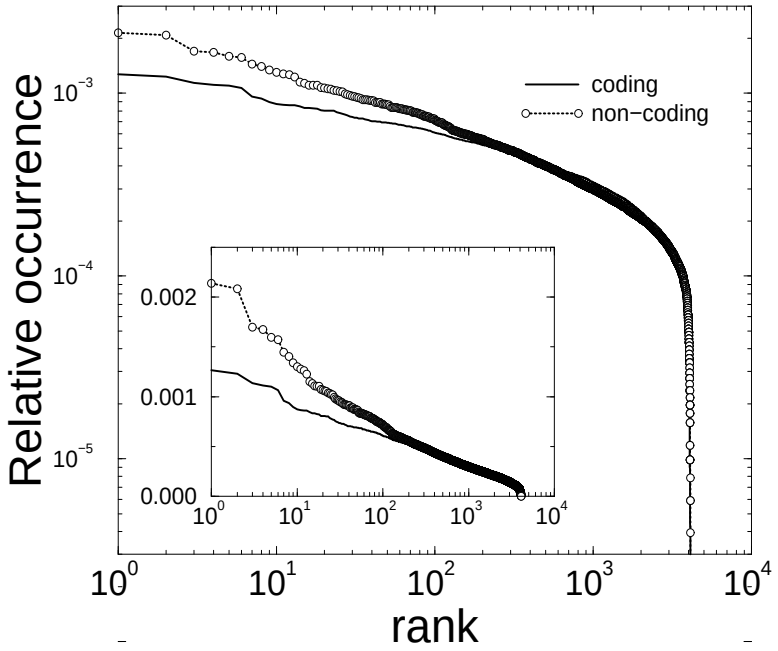

FIGURE 2. Zipf plot of the 6-grams observed in the coding (solid line) and non-coding (circles and dotted line) regions of the complete genome of *E. coli*. The coding regions comprise 4,093,640 6-grams whereas we count 507,863 6-grams in non-coding regions. The nucleotide concentration in this genome is $c_A = 24.6$, $c_C = 25.4$, $c_G = 25.4$ and $c_T = 24.6$. In the inset the same curves are shown in a semi-logarithm plot making evident the logarithmic dependence between the relative occurrence of 6-grams and their rank in coding regions.

Zipf analysis of a pure random symbolic sequence if the occurrence frequency is the same for all symbols, while if the occurrence frequency is unequal, a log-normal-like Zipf plot can emerge.

The practical usefulness of the n-tuple analysis of natural languages (in spite of the theoretical possible shallowness which is still also present in the n-tuple analysis) is exemplified in a paper in which information derived from n-tuple frequency combined with a simple vector-space technique allows a language-independent categorization of similarity in unrestricted text [40].

The sequencing of a great number of complete genomes recently performed provides a unique opportunity to test these findings in complete organisms. In Fig. 2 we provide the Zipf plot obtained by investigating the 6-gram observed in coding and non-coding regions of the *E. coli* complete genome. The difference between the Zipf plot in the two subsets of the genome is clear and it manifests that the statistical properties of non-coding regions are markedly different from the statistical properties of coding regions. In fact, a logarithm relation is observed in coding regions (a straight line is observed in the semilogarithmic plot of the inset of Fig. 2) whereas a deviation from this dependence is observed in non-coding regions. We observe a similar behavior in several completely sequenced genomes (*Aquifex aeolicus*,

Bacillus halodurans, *Bacillus subtilis*, *Borrelia burgdorferi*, *Campylobacter jejuni*, *Chlamydia pneumoniae*, *Chlamydia trachomatis*, *Escherichia coli*, *Haemophilus influenzae*, *Helicobacter pylori 26695*, *Helicobacter pylori J99*, *Rickettsia prowazekii*, and *Vibrio cholerae*). A deviation from this general pattern is observed in complete genomes of prokaryotes with a high GC content.

In summary, complete genomes are complex systems presenting two subsets which are characterized by different statistical properties. The first set is the set of coding regions which is a highly non-redundant symbolic set. It is characterized by a relative occurrence of n-grams which is a logarithmic function of the rank of n-grams [34]. Non-coding regions have different statistical properties. They are more redundant, long-range correlated and characterized by a Zipf-like profile which is closer to the one observed in natural and formal languages. Further studies are needed to find other "stylized" statistical facts that might concur to the important task of modeling the general structure of genomes of simple and complex living organisms.

II. Financial markets

The other complex system considered in these lectures is the financial market. Financial markets are essential to the life of our society and their influence is growing up in all the aspects of the society in the present period.

The time evolution of the price (to be precise, a nonlinear function of the price such as its natural logarithm it is often investigated) looks erratic and very similar to a diffusive process. In Fig. 3 we present an example of such time evolution. It is the time evolution of the logarithm of the price of General Electric (GE) which is one of the most capitalized assets traded in the US equity markets. The time series is recorded at high-frequency. The time interval between two successive records being 1/20 of a trading day (equals to 1170 s). The time series is recorded during the time interval ranging from January 1995 to December 1998. In the top panel of Fig. 3 the time evolution of the logarithm of the price is shown whereas in the bottom panel its successive variations are drawn. The random character of the time evolution is quite evident (even if a clear overall positive bias is present for the considered asset in the investigated time period).

In any financial market—either well established and highly active as the New York Stock Exchange, *"emerging"* as the Budapest Stock Exchange, or *"regional"* as the Milan Stock Exchange—the autocorrelation function of returns (i.e. the relative change of the price over a given time horizon) is a monotonic decreasing function with a very short correlation time. High frequency data analyses have shown that correlation times can be as short as a few minutes in highly traded stocks or indices (see, for example, [41,42]).

In Fig. 4 we show the autocorrelation function of the changes of the logarithm of price of GE. The autocorrelation function is a very fast decaying autocorrelation function with a characteristic time shorter than 1170 s. Indeed for $\tau = 1170$ s a

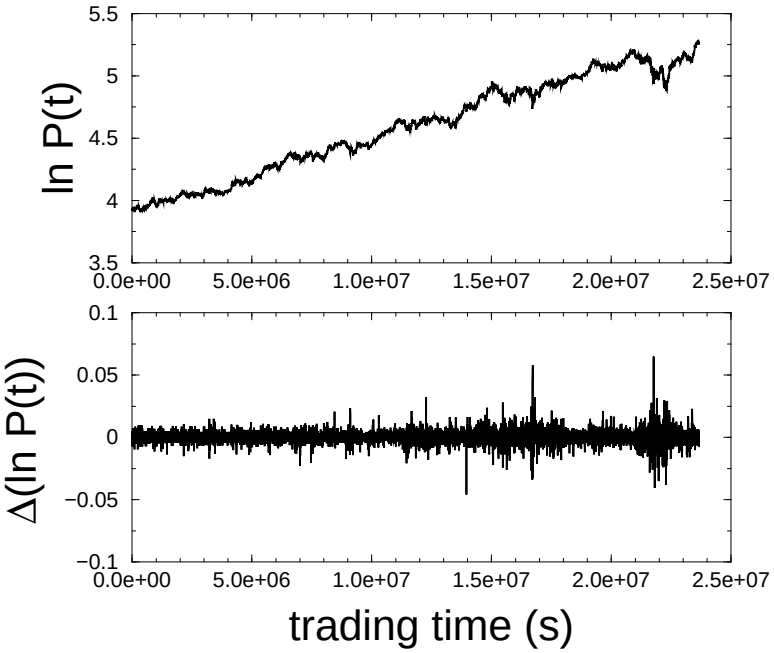

FIGURE 3. Time evolution of the logarithm of the price (top) and of its successive variations (bottom) of General Electric stock recorded with a time interval of 1170 s during the time period form January 1995 to December 1998.

slight degree of negative correlation is detected. The autocorrelation is down to noise level for longer time intervals.

The short-range memory between returns is directly related to the necessity of absence of continuous arbitrage opportunities in efficient financial markets. In other words, if correlation were present between returns (and then between price changes) this would allow for devising trading strategies that would provide a net gain continuously and without risk. The continuous search for and the exploitation of arbitrage opportunities from traders focused on this kind of activity drastically diminishes the redundancy in the time series of price changes. Another mechanism reducing the redundancy of stock price time series is related to the presence of so-called "noise traders". With their action, noise traders add into the time series of stock price information, which is unrelated to the economic information decreasing the degree of redundancy of the price changes time series.

It is worth pointing out that not all the economic information present in stock price time series disappears due to these mechanisms. Indeed the redundancy that needs to be eliminated concerns only price change and not any nonlinear functions of it [44].

In fact, the absence of time correlation between returns does not mean that returns are identically distributed over time. Several authors have observed that nonlinear functions of price return such as the absolute value or the square are correlated over a time scale much longer than a trading day. Moreover the func-

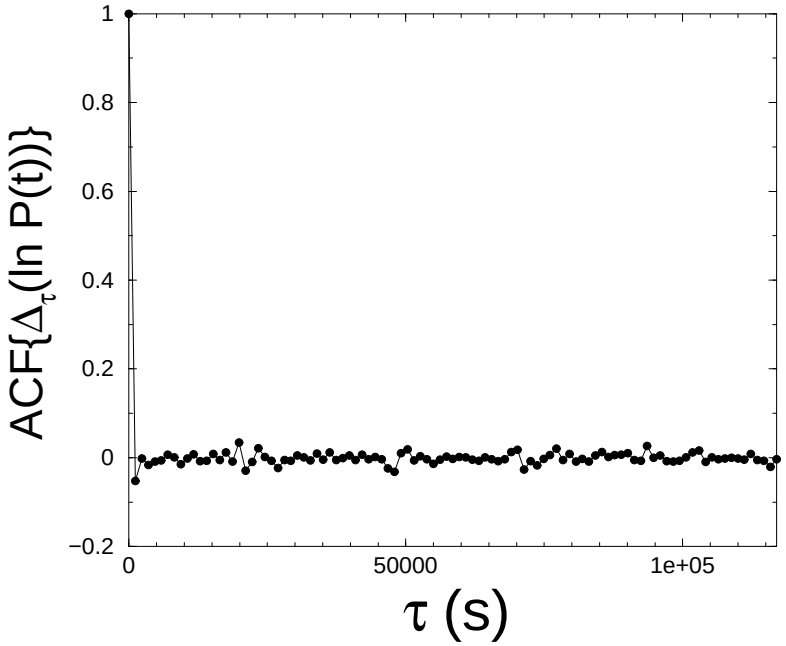

FIGURE 4. Autocorrelation function of the 1170 s successive variations of the logarithm of price of General Electric. The investigated time period is form January 1995 to December 1998. The picture presents an interval of τ of 5 trading days.

tional form of this correlation seems to be power-law up to at least 20 trading days approximately [45–49,42,50]. In the financial literature the standard deviation of price return is called the *volatility* of the financial asset. Volatility of liquid stocks is approximately distributed as a lognormal and it shows a persistent memory. There are indications that the memory is a long-range memory but a conclusive demonstration of this hypothesis is still lacking.

To illustrate this concept in Fig. 5 we show the time evolution of the volatility of GE. The daily historical volatility of GE is calculated by using high-frequency intraday data and by considering the standard deviation of the changes of the logarithm of price observed at a 1950 s time horizon. The time evolution of the volatility is itself a random process. The autocorrelation function of the volatility is not a fast decaying function. In fact it needs almost 20 trading days to reach a value of the order of 0.05. Moreover, a behavior consistent with a power-law decaying function is observed when the autocorrelation function is plotted in a log-log plot. In Fig. 6 we show the autocorrelation function of the volatility of GE both in a linear and in a log-log plot.

Another key observation concerns the degree of stationary behavior of the stock return dynamics. Empirical analysis shows that returns are not strictly-sense stationary stochastic processes. Indeed we have seen that the volatility (standard deviation of log price changes) is itself a stochastic process. Although a general proof is still lacking, empirical analyses performed on financial data of different financial

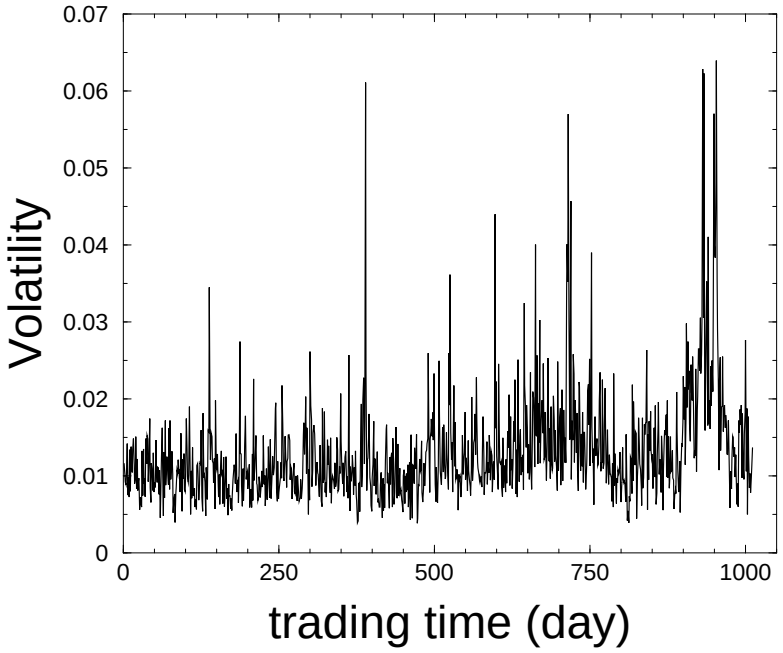

FIGURE 5. Time evolution of the daily volatility of General Electric during the time period from January 1995 to December 1998.

markets suggest that the stochastic process is locally non-stationary but asymptotically stationary. By asymptotically stationary one means that the probability density function (pdf) of returns measured over a wide time interval exists and it is uniquely defined. A paradigmatic example of simple stochastic processes which are locally non-stationary but asymptotically stationary is provided by ARCH [51] and GARCH [52] processes.

The pdf of returns shows some "universal" aspects. By "universal" aspects we mean features that are observed in different financial markets at different periods of time provided that a sufficiently long time period is used in the empirical analysis. The first of these "universal" or stylized facts is the leptokurtic nature of the pdf. A symmetric leptokurtic pdf is a pdf which is more peaked than a Gaussian pdf around the mean value and with distribution tails which are 'fatter' than in the case of a Gaussian pdf. Leptokurtic pdfs have been observed in stocks and indices time series by analyzing both high-frequency and daily data. Thanks to the recent availability of transaction-by-transaction data, empirical analyses on a single-transaction time-scale have also been performed. One of these studies performed by analyzing stock price in the Budapest Stock Exchange show that return pdfs of highly traded Hungarian stocks are leptokurtic down to a "single-transaction" time-scale [43].

The origin of the observed leptokurtosis is still debated. There are several models trying to explain it. Just to cite (rather arbitrarily) a few of them: (i) a model of Lévy stable stochastic process [53]; (ii) a model assuming that the non-Gaussian behavior occurs as a result of the uneven activity during market hours

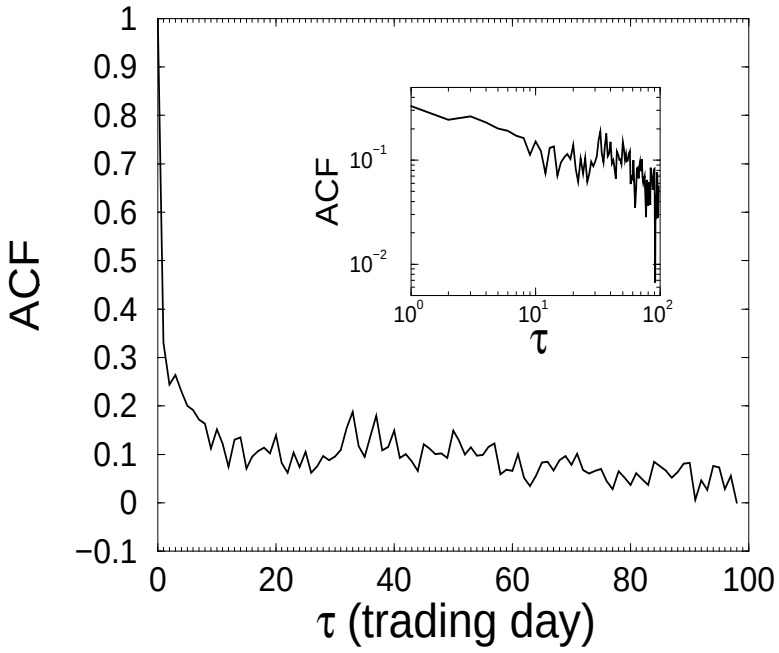

FIGURE 6. Autocorrelation function of the daily volatility of General Electric. The autocorrelation function is slowly decaying. The main panel presents an interval of τ of 100 trading days. The same function and interval is shown in the inset as a log-log plot. The observed autocorrelation function is compatible with a power-law decaying function.

[54]; (iii) a model where a geometric diffusive behavior is superimposed to Poissonian jumps [55]; (iv) a quasi-stable stochastic process with finite variance [56]; and (v) a stochastic process with rare events described by a power-law exponent not falling into the Lévy regime [57–59]. The above processes are characterized by finite or infinite moments. In the attempt to find the stochastic process that better describes the stock price dynamics, it is then important to try to preliminarily conclude about the finiteness or infiniteness of the second moment.

The answer to the above question is not simply obtained [60] and careful empirical analyses must be performed to reach a reliable conclusion. It is our opinion that an impressive amount of empirical evidence has been recently found supporting the conclusion that the second moment of the return pdf is finite [61,57–59,62,63]. This conclusion has a deep consequence on the stability of the return pdf. The finiteness of the second moment and the independence of successive returns imply that the central limit theorem asymptotically applies. Hence the form expected for the return pdf must be Gaussian for very long time horizons. We then have two regions: at short time horizons we observe leptokurtic distributions whereas at long time horizons we expect a Gaussian distribution. A complete characterization of the stochastic process needs an investigation performed at different time horizons. When such a kind of investigation has been performed, non-Gaussian scaling and its breakdown has been detected [61,41].

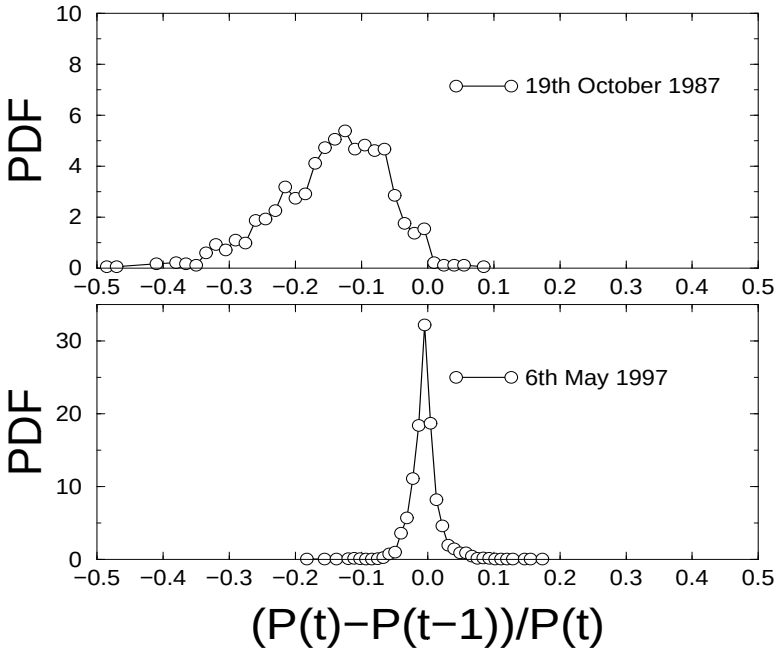

FIGURE 7. *Ensemble* return distribution of all the equities traded in the New York Stock Exchange at the dates 19th October 1987 (the worst crashes in the last 50 years, top panel) and 6th May 1997 (a typical trading day, bottom panel). The shape of the ensemble return distribution occurring at 6th May 1997 is non-Gaussian and approximately symmetrical. This is the typical shape observed during "normal" trading days. During crises (and rallies) variance of the pdf is increased and the symmetry of the distribution is lost (top panel).

In summary, "universal" facts suggest that the stock return dynamics in financial markets is well described by an unpredictable time series. However, this does not imply that the stochastic dynamics of stock return time series is a random walk with independent identically distributed increments. Indeed the stochastic process is much more complex than a customary random walk.

Another important aspect in the analysis and modeling of a financial market concerns the independence of the return time series of different stocks traded simultaneously in the same market. The presence of cross-correlations between pairs of stocks has been known since long time and it is one of the basic assumptions of the theory of the selection of the most efficient portfolio of stocks [64]. Recently, physicists have also started to investigate empirically and theoretically the presence of such cross-correlations.

In a recent study [65] it has been observed that the collective behavior of the market during "normal" days of activity can be quite different from the behavior observed during crashes and rallies. Specifically, Lillo and Mantegna find that the daily return distribution of an *ensemble* of stocks traded simultaneously is non-Gaussian and approximately symmetric during "normal" period of times whereas during crashes and rallies the symmetry property changes dramatically. To provide

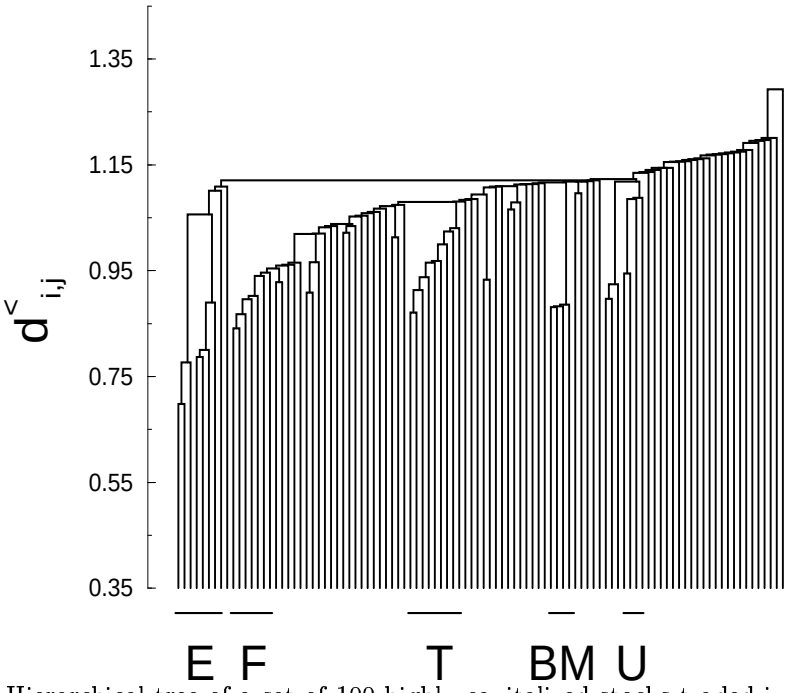

FIGURE 8. Hierarchical tree of a set of 100 highly capitalized stocks traded in the US equity markets obtained with the correlation based clustering method of Ref. [66] starting from the return time series computed with a $\Delta t = 6$ h and 30 min time horizon (1 trading day) during the time period Jan 1995-Dec 1998. Each stock is indicated by a vertical line. The presence of several clusters of stocks belonging to the same economic sector is seen. Two stocks (lines) links when a horizontal line is drawn between two vertical lines. The letters E, F, T, BM and U indicates the five prominent clusters of energy, financial, technology, basic material and utility sectors. Adapted from Ref. [68] where the complete characterization of all clusters is provided with a color code.

an example of this change, we show the daily *ensemble* return distribution for two different days in Fig. 7. Specifically, we consider the "normal" day of 6th May 1997 and the crucial day of 1987 Black Monday (19th October 1987). The change in the symmetry property of the *ensemble* return distribution is clear. The *ensemble* considered is the *ensemble* of all equities traded in the New York Stock Exchange at the considered date.

It has also been found that a meaningful economic taxonomy may be obtained by starting from the information stored in the time series of stock returns only. This has been achieved by some of us with a correlation based clustering procedure performed between the synchronous time evolution of a set of stocks traded in a financial market. The specific clustering procedure is performed under the essential *ansatz* that the most relevant economic information stored in the time evolution dynamic of returns is present in the subdominant ultrametric space obtained by using a metric distance determined starting from the correlation coefficient matrix [66–68].

In Fig. 8 we show the hierarchical tree obtained with the correlation clustering method of Ref. [66] when we investigate a set of 100 highly capitalized stocks traded in the US equity markets. The clusters of some economic sectors are clearly detected (prominent example are the clusters of energy, financial, technology, basic material and utility stocks indicated in the figure with the first letter of each sector). For a complete inspection of the hierarchical tree of Fig. 8 the interested reader can refer to [68].

Another kind of study is devoted to detect the statistical properties of eigenvalues and eigenvectors of the covariance matrix of n stocks simultaneously traded. Also with this approach the hypothesis that the dynamics of stock price in a portfolio of n stocks is described by independent random walks is falsified [69–71]. Moreover information about the number of relevant eigenvectors can be detected.

The observation of the presence of a certain degree of statistical synchrony in the stock price dynamics suggests the following statement. A study of the time evolution of only a *single* stock price dynamics might be insufficient to reach a complete modeling of all essential aspects of a financial market.

DISCUSSION

In summary, complex systems can be fruitfully investigated with tools and methods of statistical physics. DNA and financial markets simultaneously present short-range correlated and long-range correlated variables. This implies that a complete characterization of the statistical properties of the system requires the characterization of some of the higher-order conditional probability density functions of the basic stochastic process. In spite of this complexity some "stylized facts" can be discovered and formalized as statistical rules in both cases.

A complete picture of the problem also require to cite that the studies performed make clear that new tools of statistical physics need to be developed if we aim at describing complex systems in a complete way. In fact, one of the theoretical problems encountered in the modeling of long-range correlated systems is the absence of strict-sense stationary behavior in the investigated processes. This implies that a successful modeling of these processes needs to solve the theoretical problem of the modeling of stochastic processes, which are non-stationary on a local scale but asymptotically stationary. Similar models are not widespread in physical sciences while successful examples, such as ARCH or GARCH classes of stochastic processes, have been proposed in the financial literature.

Acknowledgements

The research work described in these lectures has been supported by funding from: (i) the Istituto Nazionale per la Fisica della Materia (INFM)-G4 Palermo unit research line; (ii) INFM-Applied research fund special project "Volatility in financial markets"; (iii) INFM-FSE joint tutoring projects "Software development

for a liquidity indicator" and "Modeling of financial markets with advanced statistical physics tools"; (iv) INFM-PAIS project "Statistical modeling of non-coding DNA" and (v) MURST (ex 60%) Palermo University funding.

REFERENCES

1. P. W. Anderson, J. K. Arrow and D. Pines, eds., *The Economy as an Evolving Complex System* (Addison-Wesley, Redwood City, 1988).
2. G. Weisbuch, *Complex systems dynamics* (Perseus Press, 1991).
3. J. H. Holland, *Adaptation in natural and artificial systems* (Bradford Books, 1992).
4. F. Mallamace and H.E. Stanley, eds., *The Physics of Complex Systems* (IOS, Amsterdam, 1997).
5. H. Haken, *Information and self-organization*, 2nd edition, (Springer Verlag, Berlin, 2000).
6. H. E. Stanley, *Introduction to Phase Transitions and Critical Phenomena* (Oxford University Press, Oxford, 1971).
7. M.C. Gutzwiller, *Chaos in Classical and Quantum Mechanics* (Springer-Verlag, Berlin, 1990).
8. M. Mézard, G. Parisi, and M. A. Virasoro, *Spin Glass Theory and Beyond* (World Scientific, Singapore, 1987).
9. P. Bak, C. Tang and K. Wiesenfeld, Phys. Rev. Lett. **57**, 381 (1987).
10. R. N. Mantegna, *Physica* A **179**, 232 (1991).
11. W. Li, *International Journal of Bifurcation and Chaos* **1**, 583 (1991).
12. R.N. Mantegna and H.E. Stanley, *An Introduction to Econophysics: Correlations and Complexity in Finance* (Cambridge University Press, Cambridge, 2000).
13. J.-P. Bouchaud and M. Potters, *Theory of Financial Risks* (Cambridge University Press, Cambridge, 2000).
14. G.E. Uhlenbeck and L.S. Ornstein, Phys. Rev. **36**, 823 (1930).
15. L.F. Richardson, Proc. R. Soc. London Ser. A **110**, 709 (1926).
16. T. Geisel, J. Nierwetberg, and A. Zacherl, Phys. Rev. Lett. **54**, 616 (1985).
17. M.F.Shlesinger, G.M. Zaslavsky, and J. Klafter, Nature **363**, 31 (1993).
18. A. Ott, J.-P. Bouchaud, D. Langevin, and W. Urbach, Phys. Rev. Lett. **65**, 2201 (1990).
19. T.H.Solomon, E.R. Weeks, and H.L. Swinney, Phys. Rev. Lett. **71**, 3975 (1993).
20. H. Scher and E. Montroll, Phys. Rev. B **12**, 2455 (1975).
21. Q. Gu, et al., Phys. Rev. Lett. **76**, 3196 (1996).
22. F. Amblard, et al., Phys. Rev. Lett. **77**, 4470 (1996).
23. A. T. Ogielski and D. L. Stein, Phys. Rev. Lett. **55**, 1634 (1985).
24. G.K. Batchelor, Proc. Cambridge Philos. Soc. **48**, 345 (1952).
25. H.G.E. Hentschel and I. Procaccia, Phys. Rev. A **29**, 1461 (1984).
26. F. Lillo and R.N. Mantegna, Phys. Rev. E **61**, R4675 (2000).
27. M.F. Shlesinger, J. Klafter, and Y.M. Wong, J. Stat. Phys. **27**, 499 (1982).
28. R. Metzler, E. Barkai, and J. Klafter, Phys. Rev. Lett. **82**, 3563 (1999).
29. B. B. Mandelbrot and J. W. van Ness, SIAM Review **10**, 422 (1968).

30. S.B. Primrose, *Principles of Genome Analysis*, second edition, (Blackwell Science, Oxford, 1998).
31. C.-K. Peng, S. V. Buldyrev, A. L. Goldberger, S. Havlin, F. Sciortino, M. Simons and H. E. Stanley, Nature **356**, 168 (1992).
32. W. Li and K. Kaneko, Europhys. Lett. **17**, 655 (1992).
33. S.V. Buldyrev, A.L. Goldberger, S. Havlin, R.N. Mantegna, M.E. Matsa, C.-K. Peng, M. Simons, and H.E. Stanley, Phys. Rev. E **51**, 5084 (1995).
34. M. Yu. Borodovsky and S. M. Gusein-Zade, J. Biomol. Struct. Dyn. **6**, 1001 (1989).
35. R. N. Mantegna, S. V. Buldyrev, A. L. Goldberger, S. Havlin, C.-K. Peng, M. Simons, and H. E. Stanley, Phys. Rev. Lett. **73**, 3169 (1994); Phys. Rev. E **52**, 2939 (1995).
36. G. K. Zipf, *Human Behavior and the Principle of Least Effort* (Addison-Wesley Press, Redwood City CA, 1949).
37. B. B. Mandelbrot, *The Fractal Geometry of Nature* (Freeman and Co., New York, 1983).
38. H. A. Simon, Biometrika **42**, 435 (1955).
39. W. Li, IEEE Trans. on Inf. Theory **38**, 1842 (1992).
40. M. Damashek, Science **267**, 843 (1995).
41. R. N. Mantegna and H. E. Stanley, Nature **383**, 587 (1996).
42. Y. Liu, P. Gopikrishnan, P. Cizeau, M. Meyer, C.-K. Peng, and H. E. Stanley, Phys. Rev. E **60**, 1390 (1999).
43. Z. Palágyi and R. N. Mantegna, Physica A **269**, 132 (1999).
44. R. Baviera, M. Pasquini, M. Serva, D. Vergni, and A. Vulpiani, 'Efficiency in Foreign Exchange Markets', cond-mat/9901225.
45. M. M. Dacorogna, U. A. Müller, R. J. Nagler, R. B. Olsen, and O. V. Pictet, J. Int'l Money and Finance **12**, 413 (1993).
46. R. Cont, M. Potters, and J.-P. Bouchaud, in *Scale Invariance and Beyond*, edited by B. Dubrulle, F. Graner, and D. Sornette (Springer, Berlin, 1997).
47. P. Cizeau, Y. Liu, M. Meyer, C.-K. Peng, and H. E. Stanley, Physica A **245**, 441 (1997).
48. Y. Liu, P. Cizeau, M. Meyer, C.-K. Peng, and H. E. Stanley, Physica A **245**, 437 (1997).
49. M. Pasquini and M. Serva, Physica A **269**, 140 (1999).
50. M. Raberto, E. Scalas, G. Cuniberti, and M. Riani, Physica A **269**, 148 (1999).
51. R. F. Engle, *Econometrica* **50**, 987 (1982).
52. T. Bollerslev, *J. Econometrics* **31**, 307 (1986).
53. B. B. Mandelbrot, J. Business **36**, 394 (1963).
54. P. K. Clark, Econometrica **41**, 135 (1973).
55. R. C. Merton, J. Financial Econ. **3**, 125 (1976).
56. R. N. Mantegna and H. E. Stanley, Phys. Rev. Lett. **73**, 2946 (1994).
57. T. Lux, Applied Financial Economics **6**, 463 (1996).
58. T. Lux, J. Econ. Behav. Organ **33**, 143 (1998).
59. P. Gopikrishnan, M. Meyer, L. A. N. Amaral, and H. E. Stanley, Eur. Phys. J. B **3**, 139 (1998).
60. A. L. Tucker, J. Business & Econ. Stat. **10**, 73 (1992).
61. R. N. Mantegna and H. E. Stanley, *Nature* **376**, 46 (1995).

62. P. Gopikrishnan, V. Plerou, L. A. N. Amaral, M. Meyer, and H. E. Stanley, Phys. Rev. E **60**, 5305 (1999).
63. V. Plerou, P. Gopikrishnan, L. A. N. Amaral, M. Meyer, and H. E. Stanley, Phys. Rev. E **60**, 6519 (1999).
64. H. Markowitz, *Portfolio Selection: Efficient Diversification of Investment* (J. Wiley, New York, 1959).
65. F. Lillo and R. N. Mantegna, *Eur. Phys. J. B* **15**, 603 (2000).
66. R. N. Mantegna, *Eur. Phys. J. B* **11**, 193 (1999).
67. G. Bonanno, N. Vandewalle and R. N. Mantegna, Phys. Rev. E **62**, R7615 (2000).
68. G. Bonanno, F. Lillo and R. N. Mantegna, High-frequency Cross-correlation in a Set of Stocks, cond-mat/0009350, to appear on Quantitative Finance
69. S. Galluccio, J.-P. Bouchaud, and M. Potters, Physica A **259**, 449 (1998).
70. L. Laloux, P. Cizeau, J.-P. Bouchaud, and M. Potters, Phys. Rev. Lett. **83**, 1468 (1999).
71. V. Plerou, P. Gopikrishnan, B. Rosenow, L. A. N. Amaral, and H. E. Stanley, Phys. Rev. Lett. **83**, 1471 (1999).