entropy

MDPI

Article

# Non-Linear Dynamics Analysis of Protein Sequences. *Application to CYP450*

**Xavier F. Cadet [1,2], Reda Dehak [2], Sang Peter Chin [3] and Miloud Bessafi [4,*]**

1 PEACCEL, Protein Engineering Accelerator, 6 square Albin Cachot, box 42, 75013 Paris, France
2 LSE laboratory, EPITA, Paris 94276, France
3 Learning Intelligence Signal Processing Group, Department of Computer Science, Boston University, Boston, MA 02215, USA
4 LE2P-Energy Lab, Laboratory of Energy, Electronics and Processes EA 4079, Faculty of Sciences and Technology, University of La Reunion, 97444 St Denis CEDEX, France
* Correspondence: miloud.bessafi@univ-reunion.fr

Received: 30 June 2019; Accepted: 29 August 2019; Published: 31 August 2019


**Abstract:** The nature of changes involved in crossed-sequence scale and inner-sequence scale is very challenging in protein biology. This study is a new attempt to assess with a phenomenological approach the non-stationary and nonlinear fluctuation of changes encountered in protein sequence. We have computed fluctuations from an encoded amino acid index dataset using cumulative sum technique and extracted the departure from the linear trend found in each protein sequence. For inner-sequence analysis, we found that the fluctuations of changes statistically follow a −5/3 Kolmogorov power and behave like an incremental Brownian process. The pattern of the changes in the inner sequence seems to be monofractal in essence and to be bounded between Hurst exponent [1/3,1/2] range, which respectively corresponds to the Kolmogorov and Brownian monofractal process. In addition, the changes in the inner sequence exhibit moderate complexity and chaos, which seems to be coherent with the monofractal and stochastic process highlighted previously in the study. The crossed-sequence changes analysis was achieved using an external parameter, which is the activity available for each protein sequence, and some results obtained for the inner sequence, specifically the drift and Kolmogorov complexity spectrum. We found a significant linear relationship between activity changes and drift changes, and also between activity and Kolmogorov complexity. An analysis of the mean square displacement of trajectories in the bivariate space (drift, activity) and (Kolmogorov complexity spectrum, activity) seems to present a superdiffusive law with a 1.6 power law value.

**Keywords:** power law; Brownian process; Kolmogorov complexity; entropy; chaos; monofractal; non-linear; cumulative sum; sequence analysis; protein engineering

## 1. Introduction

From the information viewpoint, a protein sequence can be considered as a distribution of successive symbols extracted with a rule from a dictionary. Conceptually, it means that the protein sequence is simply encoded to a set of symbol combinations. Moreover, the number of the symbols used is usually very small in comparison to the length of the protein sequence. Consequently, there is a huge variety of combinations of symbols to encode a protein sequence in the real world. It is well-known that the molecular mechanism (stability, structure function, disorder) is often triggered by complex interactions [1–3]. Like the emerged part of an iceberg, the intricated symbol set of an encoded protein sequence can be seen as a footprint of a wide range of covert biochemical interactions within the protein. Then, there are numerous encoder models that try to reflect the reality accurately

using a conversion rule related to physicochemical and biochemical properties [4–6]. Beyond the symbol combination and arrangement of the protein sequence, understanding the nature and the organization of the symbols is very challenging in protein biology. Therefore, analyzing the encoded protein sequence by means of nonlinear analysis can provide some insights about the dynamics of the changes within the dataset. Searching for similarities between encoded protein sequences in a dataset is one of the important advantages of morphological analysis of protein sequences. There are many approaches to extract groups, which are conceptually based on a clustering method of global or local information about the protein sequence [7–13]. The prediction of disorder of the protein sequence is often related to the ability to track the degree of randomness, the stochasticity, and the complexity embedded in the whole encoded dataset. There are studies which focus on randomness, chaos, long-range interaction between sequences for classification, and predictability. For example, Yu et al. [14] have made a comparative study of structure and intrinsic disorder between 10,000 natural and random protein sequences and found that natural sequences have more long disordered regions than random sequences. In addition, Gök et al. [5] have used the Lyapunov exponent and test four classifier algorithms (Bayesian network, Naïve Bayes, k-means, and SVM) to identify the disordered protein regions. Long short-term memory (LSTM) recurrent neural networks is a deep learning algorithm that has gained some interest for tracking the long-range interactions between sequences [1,15]. These studies reveal that there is potential information about degree of randomness, disorder, and stochasticity in protein sequences and beyond some degree of predictability. It means that the protein sequence exhibits some order within disorder and changes are not a likelihood for this set of symbols. To find out what kind of information and properties of disorder or complexity we are able to extract from protein sequences, we propose to scan the changes inside the protein sequences and between sequences using a multidisciplinary approach. It means that we intend, at the same time, to use tools from information theory field (entropy of information, Kolmogorov complexity), physical theory (chaos, fractional Brownian processes, drift-diffusion processes), and signal processing (multifractality, Fourier analysis). To our knowledge, the use of multidisciplinary tools to analyze the dynamics of the changes within a protein sequence and between sequences is new. As mentioned previously, the encoded protein sequence contains successive numerical values and can also be considered as a time series. The aim of this paper is to encompass the variability of the inner changes hidden behind the encoded protein sequence using nonlinear tools, and to assess the predictability of the underlying non-stationary protein sequence activity.

The study is organized as follows. Section 2 presents the experimental dataset and the encoded protein sequence. Section 3 describes the algorithm used to analyze the time series (i) entropy and chaos, (ii) Kolmogorov complexity and Turing machine, (iii) law-scaling and stochastic process, and (iv) surrogated and shuffled data. Finally, Section 4 includes both presentation of the results obtained and discussion. The concluding remarks are given in Section 5.

## 2. Experimental Dataset

To facilitate the understanding of readers outside the realm of life sciences, we will provide a brief definition of a polypeptide/protein sequence. A protein sequence is a chain made of residues of amino acids. Twenty amino acids are the basic building blocks for proteins. We will provide an application example as well.

### *2.1. Alphabetical Dictionary*

Each amino acid is represented by a letter corresponding to the one-letter code for an amino acid. The global sequence has a biological meaning. A single variation in the sequence could have a huge impact on the activity of the protein. An example of a protein sequence (Cytochrome P450) is given below:

MTIKEMPQPKTFGELKNLPLLNTDKPVQALMKIADELGEIFKFEAPGRVTRYLSSQRLIKEACDES
RFDKNLSQALKFVRDFAGDGLATSWTHEKNWKKAHNILLPSFSQQAMKGYHAMMVDIATQLI
QKWSRLNPNEEIDVADDMTRLTLDTIGLCGFNYRFNSFYRDSQHPFITSMLRALKEAMNQSKRL
LRLWPTAPAFSLYAKEDTVLGGEYPLEKGDELMVLIPQLHRDKTIWGDDVEEFRPERFENPSAIPQ
HAFKPFGNGQRACIGQQFALHEATLVLGMILKYFTLIDHENYELDIKQTLTLKPGDFHISVQSRH
QEAIHADVQAAE

*2.2. An Application Example: Cytochrome P450*

Cytochrome P450 is a protein, i.e., a polypeptidic sequence of 464 or 466 amino acids. It is used to generate products of significant medical and industrial importance. Three parental cytochromes P450, i.e., CYP102A1(P1), CYP102A2(P2), and CYP102A3(P3) were used to generate 242 chimeric sequences of cytochrome P450 [16]. Further, 242 thermostable protein sequences were created by recombination of stabilizing fragments. For each variant, the thermostability (defined herewith as: Activity) was analyzed by the measurement of the $T_{50,}$ $T_{50}$ being the temperature at which 50% of the protein was irreversibly denatured after incubation for 10 min. The result is a decrease in activity. Activity ranges from 39.2 °C to 64.48 °C. Chimeras are written according to fragment composition: 23121321 represents a protein that inherits the first fragment from parent P2, the second from P3, the third from P1, and so on.

## 3. Methodology

In this study, the questions are: "*Can statistical, nonlinear, and complexity analysis give us some information about the pattern in a protein sequence and its changes along the sequence and also the next, or other sequences? Can we group sequences according to their activity but also their morphological pattern?*". To assess the ability of the statistical chaos and complexity tools, we have transformed each protein sequence into numerical or binary time series according to the need of the use of the tool.

First of all, there exist different conversion tables to transform protein residues (letters) to numerical sequences. We have used the freely available one, namely AA index database [17,18]. This database contains a huge number of ascribed numerical values for each protein residue. There are 566 numerical values, which are for each index in the sequence univocally in correspondence with physicochemical and biochemical properties of the residues. In this case, we have selected the index 532 in the dataset, which allows us to rank and encode 20 standard amino acids.

*3.1. Entropy and Chaos*

Entropy is a concept that was first discovered in physics. Nevertheless, this concept is also encountered in other fields and especially in the theory of information. In 1948, Shannon [19] formalized the concept of entropy of the information $H$ of a string of length $N$, which contains Q repeated symbols $S = \left(s_1, s_2, \ldots, s_Q\right)$. $H$ is shown by the well-known formula:

$$H = -\sum_{i=1}^{Q} \hat{p}_i log \hat{p}_i \quad (1)$$

where $\hat{p}_i = \frac{N_{s_i}}{N}$.

$N_{s_i}$ is the number of appearances of the symbol $s_i$ in the string of length $N$. Thus, $p_i$ is the probability of occurrence within the range value ]0 1]. As we suppose that all Q symbols exist in the string, the probability 0 is excluded. The minus sign is to ensure a positive value of the entropy $H$ as the logarithm is always negative. $H$ is a global measure of the total amount of information in an entire probability distribution contained in a sequence.

Another measure of entropy is the sample entropy [20]. Let us consider a set of $N$ symbols $s_{i,k}$ in a sequence $S_i$ chosen among $M$ sequences in the dataset. From the sequence $S_i$ we extract two subsets of $m$ symbols $S_{i,p}^m = \left\{s_{i,p}, s_{i,p+1}, \ldots, s_{i,p+m}\right\}$ and $S_{iq}^m = \left\{s_{i,q}, s_{i,q+1}, \ldots, s_{i,q+m}\right\}$ where $p \neq q$. The parameters $p$ and $q$ correspond to the index position of the first symbol of respectively the subset

$S_{i,p}^m$ and $S_{i,q}^m$ within the sequence $S_i$. The sample entropy (*SampEn*) of the sequence $S_i$ is defined as $SampEn(m,r,N)_i = -log\left(\frac{A_i}{B_i}\right)$, where $A_i$ is the number of pair-wise subset symbols $\left(s_{ip}^{m+1}, s_{iq}^{m+1}\right)$ of length $m+1$ with a distance $d\left(s_{ip}^{m+1}, s_{iq}^{m+1}\right) < r$ while $B_i$ is the number of pair-wise subset symbols $\left(s_{ip}^{m}, s_{iq}^{m}\right)$ of length $m$ with a distance $d\left(s_{ip}^{m}, s_{iq}^{m}\right) < r$. The $r$ is a threshold value of similarity between the pair-wise subset symbols $\left(s_{ip}^{m}, s_{iq}^{m}\right)$. In our study, the sequence is a set of numbers. Then, the distance $d\left(s_{ip}^{m}, s_{iq}^{m}\right)$ is a Euclidian distance and the tolerance value threshold value $r$ is chosen between 0.1 and 0.2 of the standard deviation of the sequence $S_i$ [20]. Moreover, the embedding dimension $m$ is usually taken to be 2. Finally, the sample entropy is a positive value, which can be 0 for a regular sequence and roughly 2.2 or 2.3 for a strongly irregular sequence. The sample entropy is a measure of the regularity within a sequence.

In addition, sometimes an irregularity pattern in a time series could be related to the chaos process within a sequence. The largest Lyapunov exponent is the most common parameter used to characterize chaos in a dynamical system. The sign and the value of this parameter give an indication of the response of a system to amplify, damp, or oscillate a small perturbation. In our case, it means that if the largest Lyapunov exponent is (i) positive, then the process is chaotic, (ii) close to zero, then the process is periodic or quasi-periodic, and finally (iii) negative, the process is damping and has an attractor. In our study, to achieve the search for chaos pattern in a sequence $S_i$, we have used Wolf's algorithm [21] to compute the Lyapunov exponent spectrum and the largest Lyapunov exponent $(LLE)$.

### *3.2. Kolmogorov Complexity and Turing Machine*

Let us assume we have a set of $M$ sequences $S = \{S_1, S_2, \ldots, S_M\}$. Then, we suppose that we have for each sequence $i$ of string $S_i$, a set of $N$ values defined as $S_i = \left\{{p_i}^1, {p_i}^2, \ldots, {p_i}^N\right\}$. To assess disorder within a sequence, we use the Kolmogorov complexity method [22]. This method is based on the concept of Turing machine and the mathematical expression of the algorithmic complexity can be written $K_T(s) = min\left\{\left|p\right|, T(p) = s\right\}$. This states that the algorithmic complexity of a string $s$ is the shortest program $p$ computed with a Turing's machine $T$ to gather output $s$ [23,24]. To compute the Kolmogorov complexity (*KC*), there are three processes: (i) Convert the sequence $S_i$ to binary sequence $B_i$ using a threshold method, (ii) compress the sequence $B_i$ with Lempel-Ziv compressor to a compressed sequence $C_i$, and (iii) compute and normalize the Kolmogorov complexity number associated with the original sequence $S_i$.

Binarizing the sequence $S_i$ is based on the particular value used as threshold value ${p_i}^T$ to assign each number ${p_i}^k$ in the sequence $S_i$ with the value of 0 if ${p_i}^k$ is less than the threshold value ${p_i}^T$, or conversely assigned with the value of 1 if ${p_i}^k$ exceeds the threshold value ${p_i}^T$. The mathematical expression of the binary value of the number ${p_i}^k$ in the sequence $S_i$ is:

$$B_i^k \Big|_{\substack{i=\{1,2,\ldots,M\} \\ k=\{1,2,\ldots,N\}}} = \begin{cases} 0 & if\ {p_i}^k < {p_i}^T \\ or & \\ 1 & if\ {p_i}^k \geq {p_i}^T \end{cases} \quad (2)$$

where ${p_i}^T$ is a threshold value of sequence $S_i$.

Usually, the mean of the set $\left\{{p_i}^1, {p_i}^2, \ldots, {p_i}^N\right\}$ is used as a threshold value of the sequence $S_i$. Nevertheless, we will take into account the amplitude of the numbers ${p_i}^k$ to compute the optimum threshold value ${p_i}^T{}_{opt}$ associated with the sequence $S_i$. Thus, we introduce the Kolmogorov complexity spectrum (*KCS*), which is an iterative procedure to compute the Kolmogorov complexity for various

threshold values within the range values ${p_i}^k$ of the sequence $S_i$ [25]. The encoding number to binary value is presented as:

$$B_i^k\big|_m = \begin{cases} 0 & if\ {p_i}^k < {p_i}^{T_m} \\ or & \\ 1 & if\ {p_i}^k \geq {p_i}^{T_m} \end{cases} \quad m = 1, 2, \ldots, K \tag{3}$$

where ${p_i}^{T_L} = min_k\left(\left\{{p_i}^k\right\}\right) + m\left\{\frac{max_k\left(\left\{{p_i}^k\right\}\right) - min_k\left(\left\{{p_i}^k\right\}\right)}{K-1}\right\}$.

Thus, for each sequence $S_i$, the Kolmogorov complexity spectrum is a set of $K$ Kolmogorov complexity values ${KC_i}^K = \left\{{KC_i}^1,\ {KC_i}^2,\ \ldots,\ {KC_i}^K\right\}$. The optimum threshold ${p_i}^{T_{opt}}$ is chosen among the set of threshold values $\left\{{p_i}^{T_1},\ {p_i}^{T_2},\ \ldots,\ {p_i}^{T_K}\right\}$ using the condition ${p_i}^{T_{opt}} = \left\{{p_i}^{T_j} \,\middle|\, {KC_i}^j = max_k\left({KC_i}^k\right)\right\}$.

The compression method used in this study is the Lempel-Ziv compressor [26]. This is an iterative search in the binary series $B_i$ of the overall possible subset sequences, which are different from each other. The result is a compressed sequence $C_i$. If $|C_i|$ represents the length of the compressed binary sequence $C_i$, then Kolmogorov complexity $KC_i$ associated with the sequence $S_i$ is:

$$KC_i = |C_i| log_2 N/N. \tag{4}$$

The term $log_2 N/N$ in the expression of $KC_i$ insures the normalization of the Kolmogorov complexity.

### *3.3. Law-Scaling and Stochastic Process*

As previously mentioned, a sequence is defined as a set of alphabetic letters, which could be converted to other symbols (numerical, binary, etc.). Nevertheless, the changes of symbols along the chain are usually related to the real world of biochemical activities along the protein sequence. The question is "*Do those changes present a regular or irregular pattern within a sequence which can provide some information about an underlying dynamic in a sequence?*" First, we have to define the changes in a sequence $i$ of pairwise symbols separated by a distance, namely an increment of position. Let us assume $d$ is the increment pairwise symbols and the quantity $\Delta p_{di} = \left|{p_i}^j - {p_i}^k\right|_{d=|k-j|}$ is the magnitude of changes of the pairwise symbols separated by an increment of $d$. We define the structure function $S_{q_i}(d)$ for a sequence $i$ defined by the expression $S_{q_i}(d) = \frac{1}{N_{di}} \sum_{m=1}^{N_{di}} \left|{p_i}^j - {p_i}^k\right|^q_{d=|k-j|}$ where $N_{di}$ is the number of pairwise symbols separated with a distance $d$. By extension, this function can also be used to track the existence of scaling law in the data $S_{q_i}(d) \propto d^{\xi(q)}$. $\xi(q)$ is the generalized Hurst exponent, which is indicative of the nature of pairwise symbol changes and the stochasticity of processes like long-term memories, Brownian motion, self-similarity pattern [27]. The probability function (PDFs) of the distribution of the normalized changes of pair-wise symbols $\Delta p_{di}/\sigma\left(\Delta p_{di}\right)$ within a sequence $i$ can be computed to analyze the normality of the changes in a sequence. Additionally, kurtosis or flatness is another measure of the normality of the changes of the pairwise symbols. For sequence $i$, the kurtosis $F_i = S_{4i}(d)/(S_{2i}(d))^2$. The terms $S_{4i}(d)$ and $S_{2i}(d)$ are, respectively, the fourth- and second-order moment of the pairwise distribution.

### *3.4. Surrogated and Shuffled Data*

The methods to surrogate and shuffle the data are very popular tools to assess the existence of nonlinearities and the scaling properties of a process. Both algorithms are based on the generation of randomized synthetic data using specific constraint rule to generate the synthetic data. Surrogated data used in this study are the iterative amplitude-adjusted Fourier transform (IAAFT). This method preserves the statistical properties of the original data but randomizes the phase spectrum of the Fourier transform of the original data. The synthetic data generated with this method lead to removing nonlinearities in the original data. Shuffled data are obtained by a random permutation between values of the original data. This method is a bootstrapping algorithm without repetition of the indices'

permutation. Variants of the protein (synthetic sequences) are obtained by variation of any position in the sequence and not by variation of the fragments constitutive of the protein (described in the Section 2.2 "An application example: Cytochrome P450"). The data obtained are a set of values that do not exhibit any linear correlation in the synthetic data and preserve the amplitude distribution. For more information about these two algorithms, the reader can refer to the review of Schreiber and Schmitz [28].

## 4. Normalized Detrended Cumulative Sum (NDCS) Method

Fluctuations or changes along the protein sequence are of interest in this study but we need to show how we extract this information from the original data. Cumulative sum is a sequential method that is widely used to detect changes in a time series and to track the self-similarity in a dataset [29]. In this study, we have applied this algorithm for each sequence and generated a new sequence of fluctuations defined as a departure from the linear trend. Within the 242 protein sequences of a length of 466 for each one, each index in a sequence is originally labelled with an alphabetical letter. There are 20 letters used (A, C, D, E, F, G, H, I, K, L, M, N, P, Q, R, S, T, V, W, and Y) corresponding to the one-letter code for amino acid. In this study, the D PRIFT index is chosen from the AA index catalog to convert the alphabetical symbols to numerical values [30]. It allows us to distinguish each of the 20 amino acid residues by a unique value related to its hydrophobicity property. The encoding process, which converts the original alphabetical letters to numerical values within the [−5.68 6.81] range, is shown in Table 1.

**Table 1.** Conversion rule of protein sequence of AA index 532—D PRIFT index [30].

| **AA Index 532 D PRIFT Index (Cornette et al. 1987)** | | | | | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Letter | A | C | D | E | F | G | H | I | K | L | M | N | P | Q | R | S | T | V | W | Y |
| Value Index | −5.68 | −5.62 | −5.30 | −4.47 | −3.99 | −3.86 | −1.94 | −1.92 | −1.28 | 0.96 | 0.62 | 0.21 | 0.75 | 3.34 | 4.54 | 4.76 | 5.06 | 5.39 | 5.54 | 6.81 |

We are aware that this description by their hydrophobicity values is oversimplified and does not account (i) for many other properties of amino acids that are well known to strongly affect pattern changes in protein sequences along families, such as volume, aromaticity, and different charge states for the same amino acid in distinct positions or, (ii) for the fact that the exposure of continuous amino acids sequences to solvent or their occlusion in protein cores is a fundamental requirement for proteins to fold in functional arrangements, giving importance to hydrophobic and polar amino acids and their distribution. However, whatever the choice among all the possible amino acid indexes that are able to distinguish between the 20 amino acid residues, the index will be insufficient.

As shown in Figure 1a, the distribution values show a non-normal distribution, which is indicative of the non-gaussian process along the protein sequence. Roughly, the distribution looks like a U-shape where the highest probability of occurrence is obtained for the extreme values and the lowest for the mean value of the available D PRIFT index. Then, the pattern of the encoded protein sequence appears like complex bounced stairs with randomness as a sharp jump (Figure 1b).

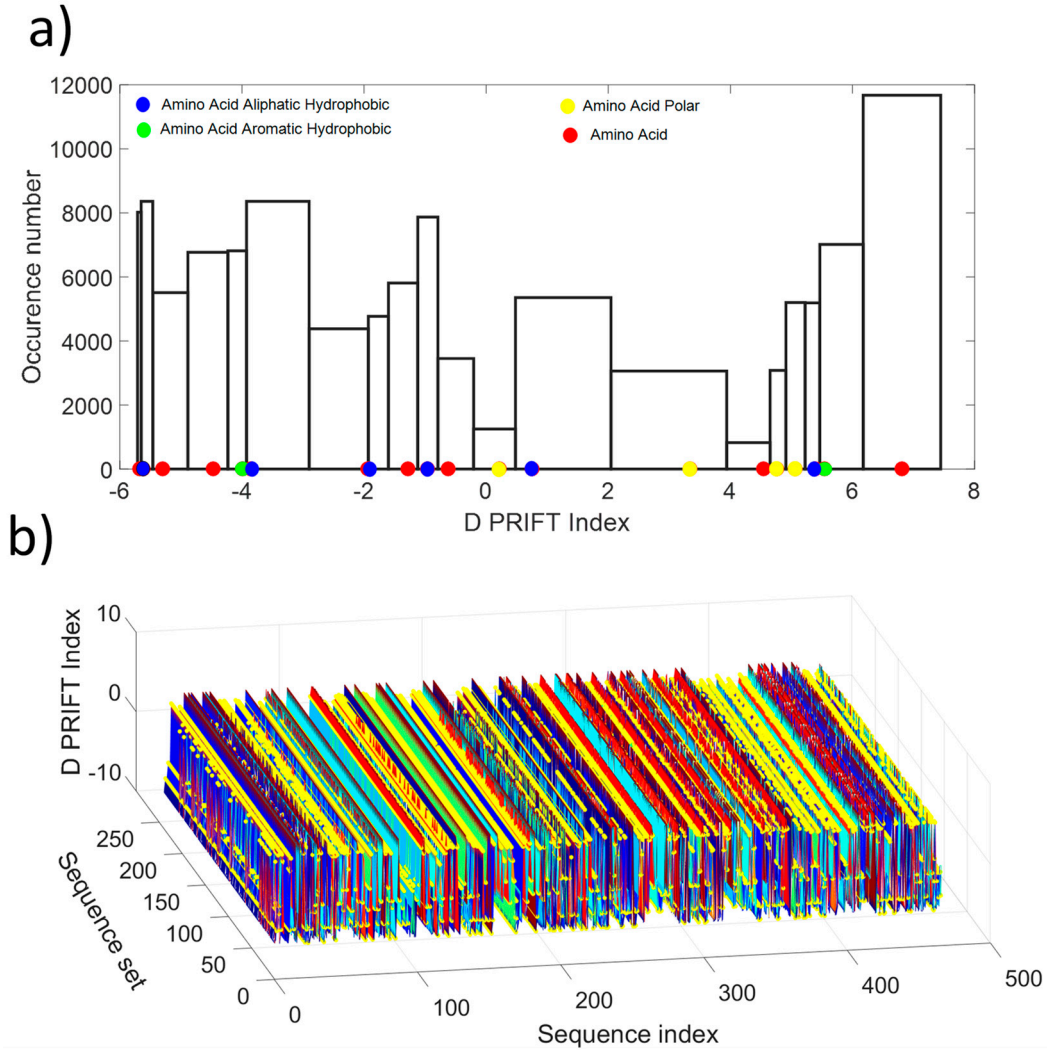


**Figure 1.** (**a**) Histogram of the D PRIFT index for 242 protein sequences. Red, blue, green, and yellow dots along the x-axis corresponds to the 20 values of the D PRIFT index. (**b**) Global view of the converted dataset (i.e., 242 protein sequences) using D PRIFT index rule. Yellow circle is indicative of the position within each sequence of the aliphatic hydrophobic, aromatic hydrophobic, and polar amino acids.

To target the jump stair pattern analysis within the protein sequence, we have used the normalized detrended cumulative sum (NDCS) method. The cumulative sum is a well-known and widely used algorithm to detect changes and shifts in time series [31]. In this study, we have extracted the linear long-term and normalized the cumulative sum of each sequence to (i) focus on the local change and (ii) have the same scale to compare transformed data. Figure 2 presents an example of transforming the original data (Sequence 1) into a detrended cumulative sum data. For clarity, we only present here the cumulative sum and linear detrending of the data. The normalized process is shown in the next figure. The trend of the cumulative sum is considered to be a linear trend for all the 242 protein sequences. The negative drift of the cumulative sum is related to the mean of a sequence. In our dataset, the average of the D PRIFT index is negative for each sequence and explains the downward drift of the cumulative sum.

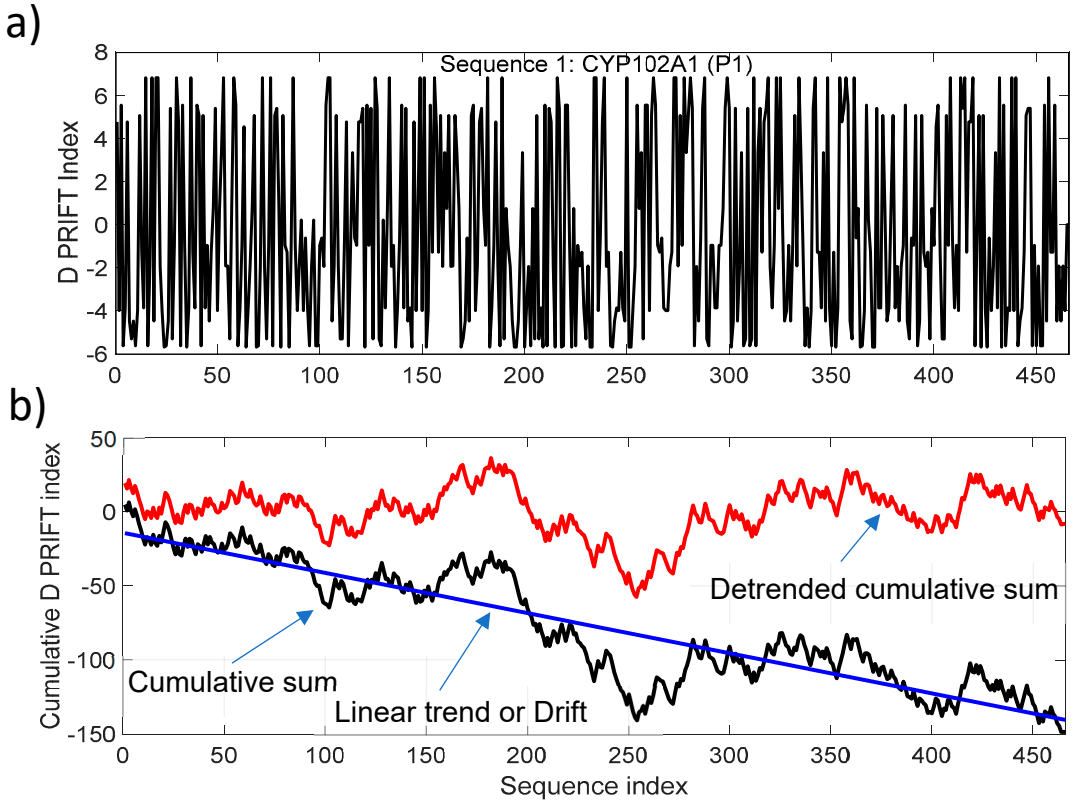


**Figure 2.** (**a**) D PRIFT index of sequence 1, which is parent CYP102A1 (P1); (**b**) Cumulative index (black line) and detrended cumulative sum (red line) of D PRIFT index of sequence 1. The blue line corresponds to the linear trend or drift of the cumulative sum of D PRIFT index.

Figure 3a depicts the NDSC plot in comparison with the original data (Sequence 1). Fluctuations reflect the local changes along the sequence and also a significant change pattern around the middle of the sequence. The fluctuation pattern relying on the cumulative sum transformation involves continuous distribution, conversely to the discrete distribution of the original D PRIFT index (Figure 3b).

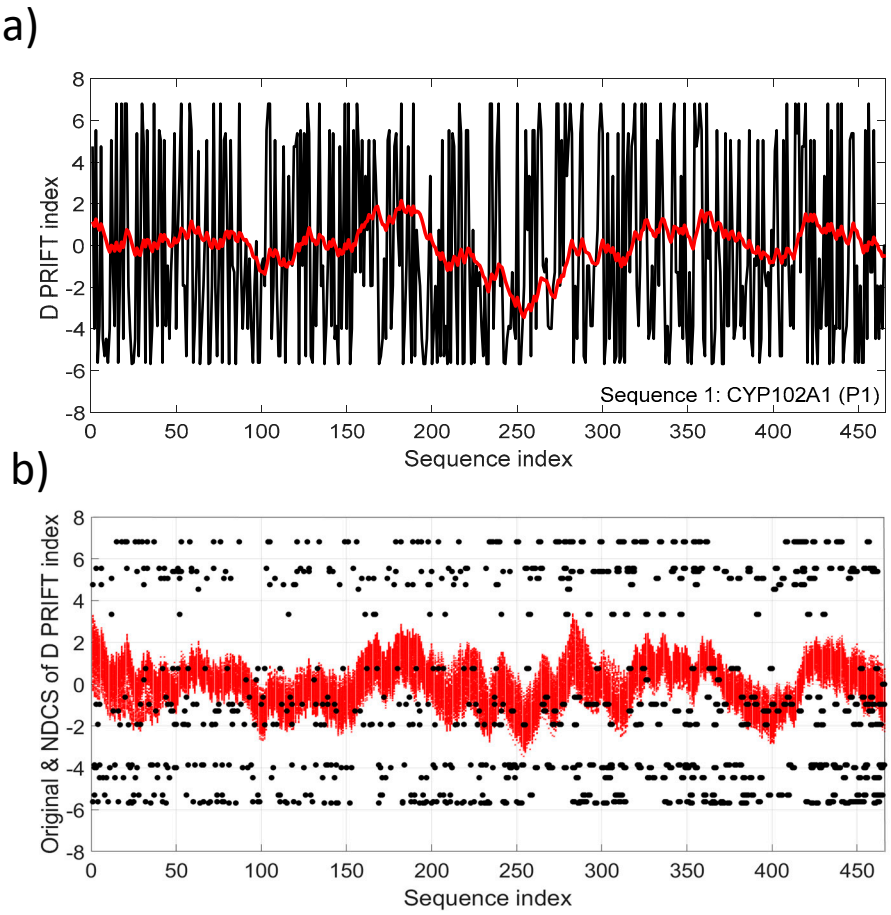


**Figure 3.** (**a**) D PRIFT index of sequence 1, which is parent CYP102A1 (P1). A superimposed red line corresponds to the normalized detrended cumulative sum (NDCS) of D PRIFT index; (**b**) Original (black dot) and normalized detrended cumulative sum (small red cross) of D PRIFT index for 242 protein sequences.

Figure 4a shows that the fluctuations of the NDCS of the D PRIFT index changes are normally distributed, with skewness close to 0 and kurtosis close to 3, which are the expected values for a normal distribution. In addition, the QQ-plot displayed in Figure 4b reveals that the observed distribution is close to a normal distribution and the two samples' (dataset values and generated normal data values) Kolmogorov–Smirnov test applied to this distribution does not reject the null hypothesis at the 5% significance level.

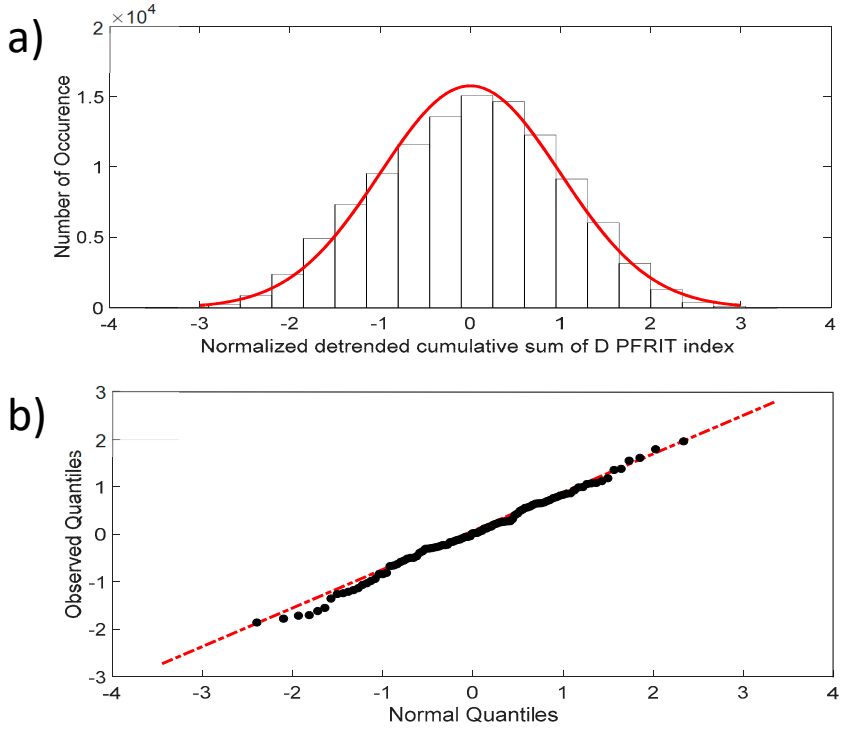


**Figure 4.** (**a**) Distribution of the NDCS of D PRIFT index changes for all sequences (black dots). Red line corresponds to Gaussian distribution; (**b**) QQ-plot of the NDCS of D PRIFT index changes quantiles and Gaussian quantiles. Red dotted line is a linear fitting of observed quantile distribution versus normal quantile distribution.

## 5. Results and Discussion

### *5.1. Normality and Intermittency*

The changes along the protein sequence for four different pairwise distances show a platykurtic nature (Figure 5a). The average distribution exhibits large amplitude for fluctuations greater than 2.5 times the standard deviation of NDCS of D PRIFT index changes. The average is computed using 242 protein sequences. Below this threshold value, the distribution is close to the Gaussian distribution. This kind of departure from the Gaussian distribution in fluctuations is indicative of intermittency. Moreover, Figure 5b highlights that the platykurtic nature of the fluctuations covers a wide range of pairwise distances, but it is more pronounced with the [30–60] pairwise distance and for distances less than 10 pairwise. To summarize, this flat distribution indicates more diversity of changes for the large amplitude of pairwise distance within the protein sequence.

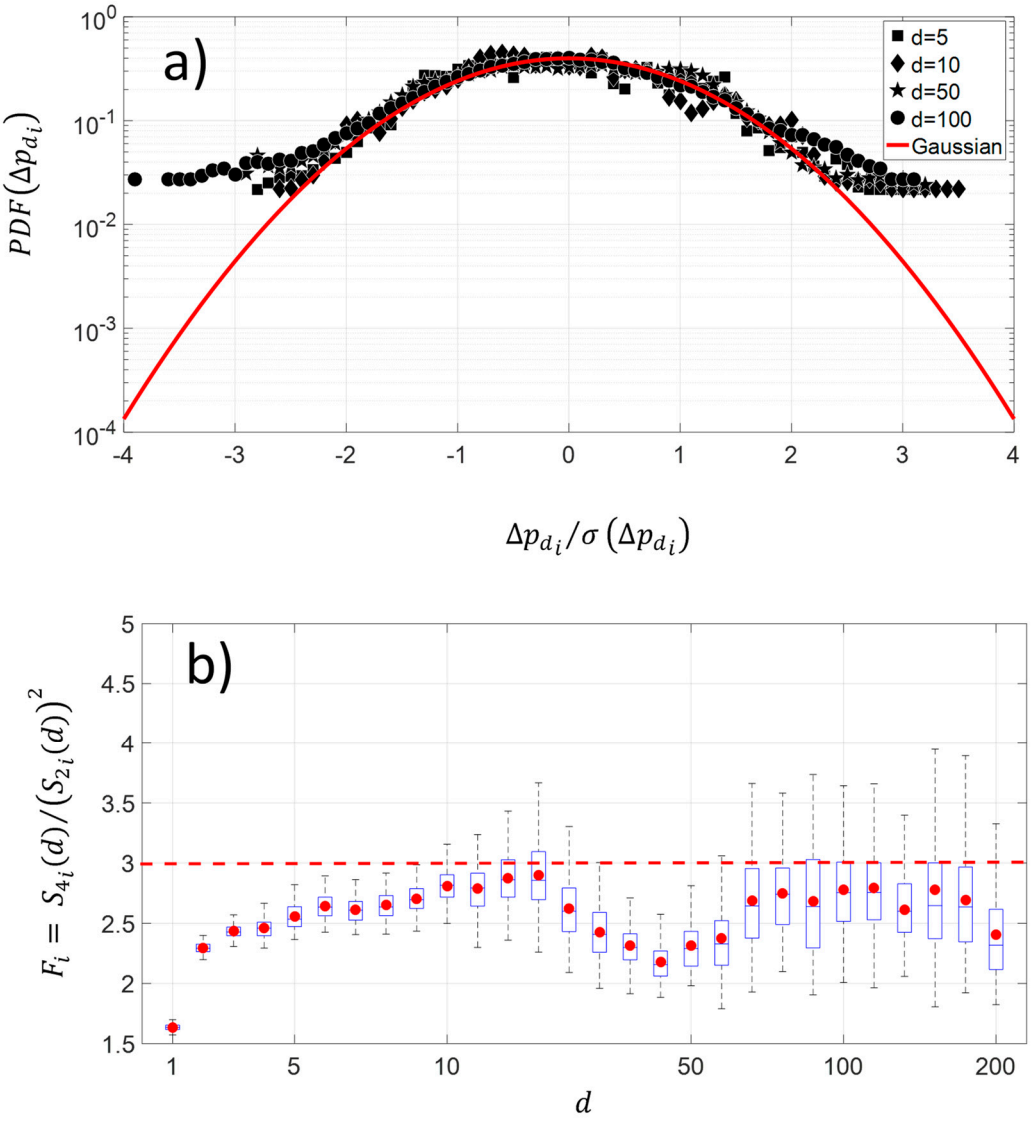


**Figure 5.** (**a**) Shape of average and normalized experimental probability functions (PDFs) of the increment of NDCS of PRIFT index changes at different distances in pairwise sequence $d = 5$, $d = 10$, $d = 50$, and $d = 100$ of 242 protein sequences. (**b**) Deviation of NDCS of PRIFT index changes distribution with respect to the Gaussian distribution at different pairwise sequence $d$.

### *5.2. Kolmogorov's Law and Brownian Process*

We have conducted a Fourier analysis to focus on the fluctuation of the NDCS of D PRIFT index changes. Surprisingly, scale invariance can be detected in the log-log presentation of the Fourier spectra (Figure 6a). An average of −1.68 based on power law is obtained, which is very close to the Kolmogorov power law result of −5/3. This highlights that the fluctuations of the NDCS of D PRIFT index changes along a sequence are similar to a non-stationary process and obey the famous Kolmogorov's law of the energy cascade for turbulence in the inertial scale range [22]. In addition, as shown in Figure 6b, the range scale value for each sequence is rather close to −5/3, with an observed minimum slope value of −1.56 and a maximum slope value of −1.84. This means that the changes within the protein sequence can be formulated according to Fourier transform as $E(f) = f^{\beta}$ where $\beta$ is the slope of the law and is close to the Kolmogorov spectrum. In addition, we can use criteria to check if the changes of protein are stationary or not [32]. This is summarized by the following test:

- $\beta < 1$, the changes are stationary,
- $\beta > 1$, the changes are non-stationary,

- $1 < \beta < 3$, the changes are non-stationary with stationary increments.

Thus, the changes in the sequence protein follow a non-stationary process. Moreover, the coefficient of variation of the fluctuations of the NDCS of D PRIFT index changes computed for all 242 sequences is less than 3%, confirming that this similarity with the Kolmogorov spectrum seems to be reproducible for each protein sequence as confirmed by the distribution of the spectrum slope obtained randomly with surrogated and shuffled data.

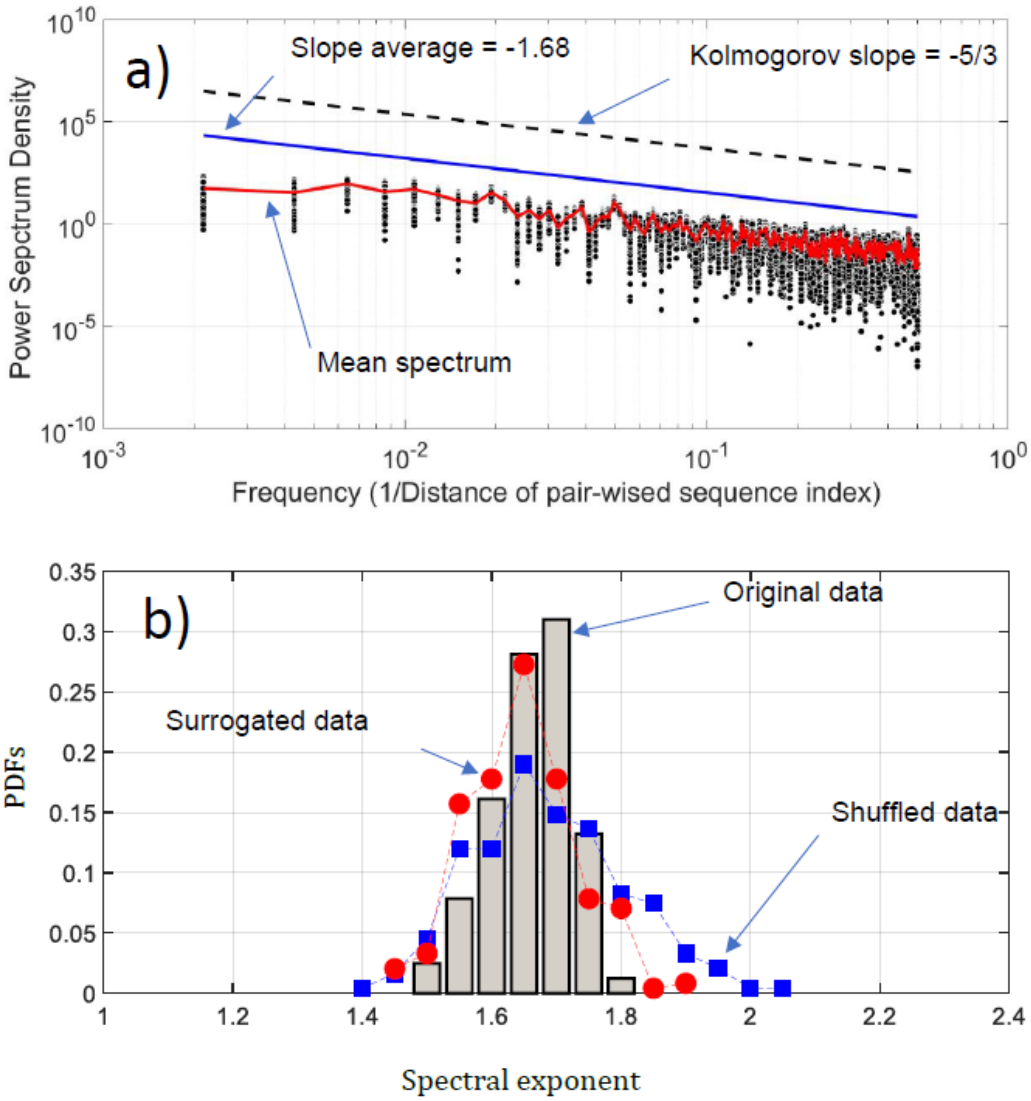


**Figure 6.** (**a**) Power spectrum density (PSD) of the NDCS of D PRIFT index changes of all 242 protein sequences $S_i$ (black dots); (**b**) PDFs of the spectral exponent estimated from Fourier analysis. We have superimposed the PDFs obtained with surrogated (red spots) and shuffled data (blue squares).

As shown previously in Figure 3b, the fluctuations of the NDCS of D PRIFT index changes appear to show seemingly organized fluctuations. The question is "*Is there some dynamic pattern of these change fluctuations along a sequence $S_i$ and is there some randomness of changes within the protein sequence?*". A first approach is to analyze the behavior of the fluctuation of the pairwise protein index. Figure 7a shows that on average, the second-order moment $S_{2i}(d)$ of the pairwise protein sequence index separated by a distance $d$ is linearly scaled in a sequence between pairwise protein sequence indexes separated by a distance $d$ roughly below 50. We found a power law of 0.87, which is close to the Brownian power law process. Then, the behavior of the change fluctuations along each protein sequence $S_i$ seems to be close

to a Brownian process. Furthermore, we found for each protein sequence a power law between a range of [0.69 0.99] and a coefficient of variation less than 7%, which reveals that the fluctuations of NDCS of the D PRFIT index changes along a sequence $S_i$ statistically have a behavior close to a Brownian process in regard to the results obtained with the surrogated and shuffled data (Figure 7b).

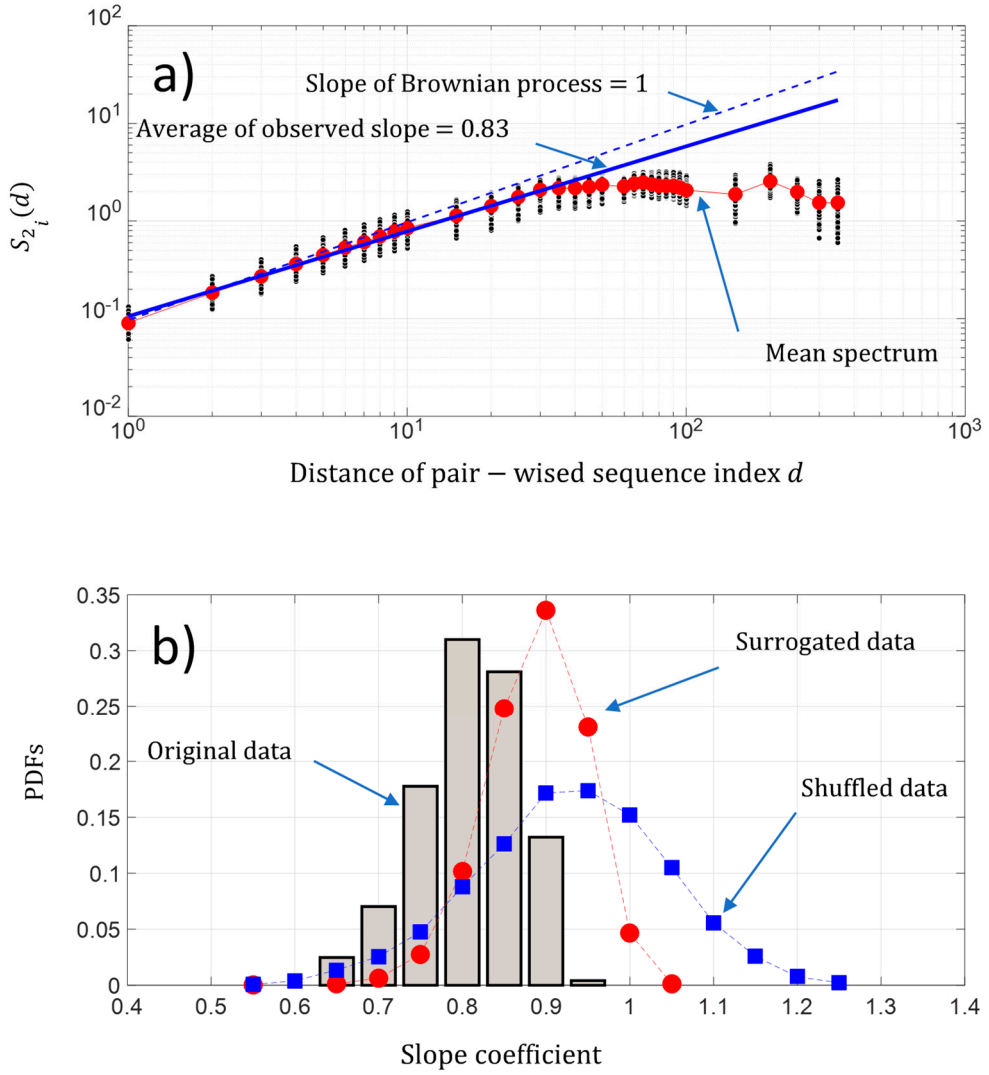


**Figure 7.** (**a**) Log-log presentation of the second-order moment $S_{2i}(d)$ of the NDCS of D PRIFT index changes of all 242 protein sequences $S_{2i}$ versus the distance $d$ of the pairwise protein sequence index (black dots); (**b**) PDFs of the slope of the scaling law distribution of the second-order moment $S_{2i}(d)$ of the NDCS of D PRIFT index changes estimated for each protein sequence $S_i$. We have superimposed the PDFs obtained with surrogated (red spots) and shuffled data (blue squares).

In addition, we have also computed the *q-order* moment for each protein sequence $S_i$. The result is shown in Figure 8a. As observed with second-order moment $S_{2i}(d)$ analysis, we again have a scaling law distribution between pairwise protein sequence index $S_i$ below $d = 50$ for a higher-order moment. This result reveals the existence of a monofractal feature along the protein sequence $S_i$. Figure 8b shows that the fluctuations of NDCS of D PRIFT index changes of each protein sequence $S_i$ contain a monofractal feature with $\xi(q) = 0.43\,q$, which is a linear law of $q$ and reveals monofractal behavior. The slope of the linear law is called the Hurst exponent $H$. As a reminder, if the value of $H = \frac{1}{2}$, it means the changes in a sequence contain no memory as for the Brownian motion. If the changes of the sequence are anti-persistent $\left(0 < H < \frac{1}{2}\right)$, then the main pattern of the changes shows that a decrease is followed by an increase and vice-versa. Finally, if the Hurst exponent is as $\frac{1}{2} < H < 1$,

then there is a persistent behavior in the changes and an increase or decrease will be maintained in a sequence. In our case, the changes are anti-persistent and they are statistically embedded between Kolmogorov process $\xi(q) = \frac{q}{3}$ [22] and the Brownian process $\xi(q) = \frac{q}{2}$. Thus, there is a potential stochastic model like the fractional Brownian model to predict the changes along the protein sequence.

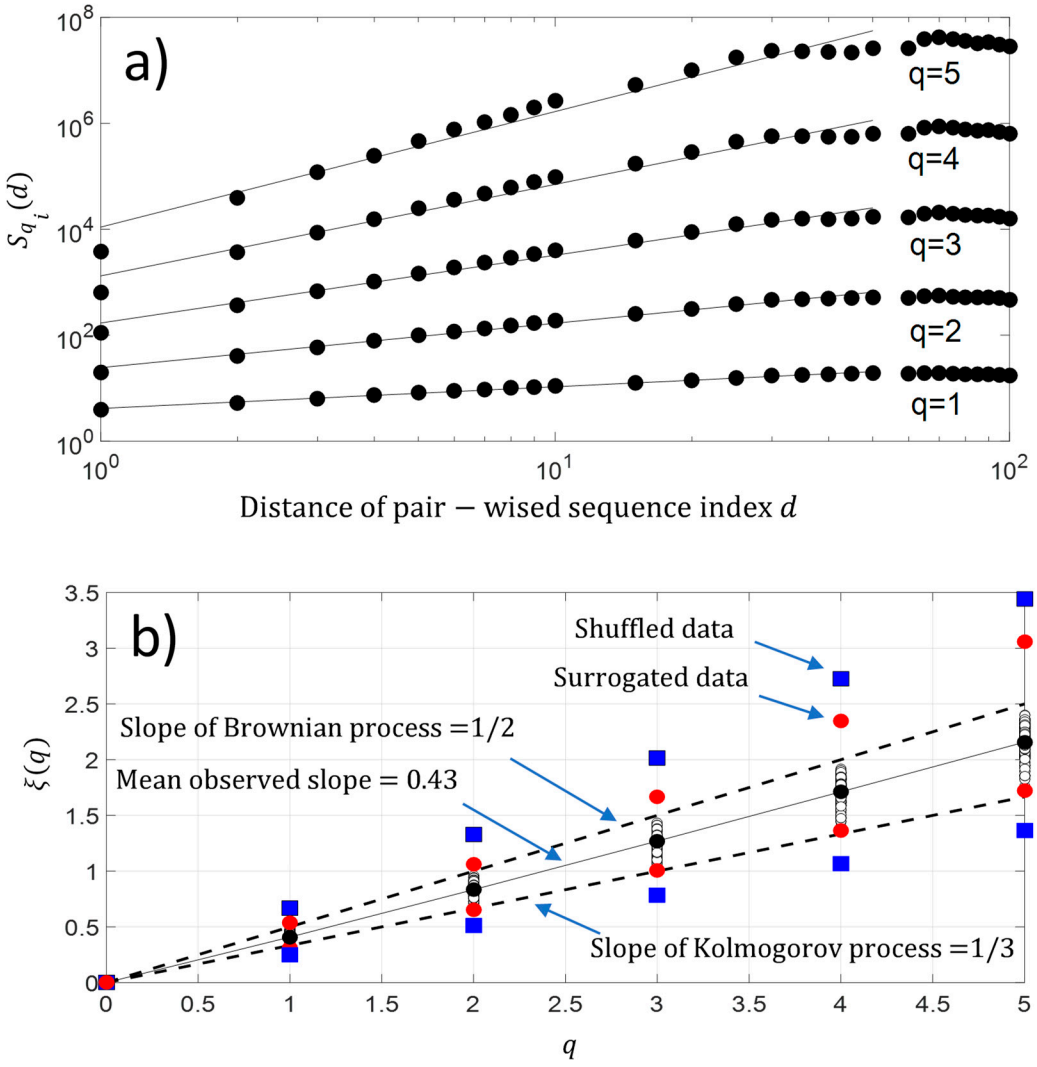


**Figure 8.** (**a**) Experimental high-order structure functions $S_{q_i}(d)$ with varying moments for $q = 1$, 2, 3, 4, and 5; (**b**) Generalized Hurst exponent $\xi(q)$. We have added the maximum and minimum value of $\xi(q)$ obtained with surrogated and shuffled data.

### *5.3. Entropy, Chaos, and Complexity*

As previously mentioned, a sequence is defined as a set of alphabetic letters, which could be converted to other symbols (numerical, binary, etc.). Nevertheless, the changes of symbols or numerical values along the sequence are usually related to the real world of biochemical activities inside the whole protein sequence. The question is "*Do those changes present regular, irregular, chaotic and complex pattern within a sequence*?" Furthermore, nonlinear analysis is one approach to estimate the changes in features along a sequence. In this study, we have used five algorithms to assess the degree of the randomness or the disorder and complexity in protein sequences: (i) The Shannon entropy (*ShEn*); (ii) the sample entropy (*SampEn*); (iii) the largest Lyapunov exponent (*LLE*); (iv) Kolmogorov complexity (*KC*); and (v) the Kolmogorov complexity spectrum (*KCS*) algorithm. Table 2 presents the descriptive statistics of the NDCS of D PRIFT index changes for 242 protein sequences. On average, there is a significant amount of information in an entire probability distribution contained in a sequence. We observe that

*SampEn* and *LLE* values are close to one. Moreover, the *KC* method underestimates the complexity in comparison to the *KCS* method, which takes into account the amplitude of the changes. Following the comparison with the surrogated and shuffled data generated from the original data, we found that the NDCS of D PRIFT index changes for 242 protein sequences used in this study include stochastic and moderate chaotic processes and show apparent embedding between the Kolmogorov ($H = 1/3$) and Brownian ($H = 1/2$) monofractal processes.

**Table 2.** Descriptive statistics of entropy, chaos, and complexity of the NDCS of D PRIFT index changes for 242 protein sequences.

| D PRIFT Index | | Entropy | | Chaos | Complexity | | Fractal |
|---|---|---|---|---|---|---|---|
| | | Information | Regularity | | | | |
| NDCS Data | | Shannon Entropy | Sample Entropy | Largest Lyapunov Exponent | Kolmogorov Complexity | Kolmogorov Complexity Spectrum | Hurst Exponent |
| Minimum | Original | 3.671 | 1.251 | 0.930 | 0.247 | 1.008 | 0.347 |
| | Surrogate | 3.514 | 1.051 | 0.730 | 0.152 | 1.046 | 0.332 |
| | Shuffled | 3.498 | 0.600 | 0.332 | 0.095 | 1.046 | 0.273 |
| Mean | Original | 3.880 | 1.433 | 1.277 | 0.475 | 1.071 | 0.432 |
| | Surrogate | 3.875 | 1.289 | 1.070 | 0.399 | 1.105 | 0.481 |
| | Shuffled | 3.911 | 1.147 | 0.911 | 0.328 | 1.103 | 0.498 |
| Median | Original | 3.888 | 1.436 | 1.286 | 0.475 | 1.065 | 0.436 |
| | Surrogate | 3.895 | 1.296 | 1.072 | 0.399 | 1.103 | 0.482 |
| | Shuffled | 3.933 | 1.154 | 0.906 | 0.323 | 1.103 | 0.498 |
| Maximum | Original | 4.066 | 1.618 | 1.601 | 0.647 | 1.141 | 0.481 |
| | Surrogate | 4.131 | 1.547 | 1.501 | 0.646 | 1.179 | 0.615 |
| | Shuffled | 4.188 | 1.604 | 1.469 | 0.627 | 1.160 | 0.690 |
| Standard deviation | Original | 0.084 | 0.063 | 0.117 | 0.084 | 0.031 | 0.027 |
| | Surrogate | 0.117 | 0.094 | 0.143 | 0.081 | 0.023 | 0.033 |
| | Shuffled | 0.130 | 0.188 | 0.220 | 0.109 | 0.022 | 0.058 |
| 1st quartile | Original | 3.833 | 1.389 | 1.207 | 0.418 | 1.046 | 0.420 |
| | Surrogate | 3.805 | 1.226 | 0.969 | 0.342 | 1.084 | 0.459 |
| | Shuffled | 3.842 | 1.017 | 0.750 | 0.228 | 1.084 | 0.459 |
| 3rd quartile | Original | 3.940 | 1.470 | 1.351 | 0.533 | 1.103 | 0.450 |
| | Surrogate | 3.963 | 1.355 | 1.160 | 0.456 | 1.122 | 0.503 |
| | Shuffled | 4.005 | 1.297 | 1.045 | 0.399 | 1.122 | 0.538 |

### *5.4. Drift (DRF), Kolmogorov Complexity Spectrum (KCS), and Activity (ACT): Linear Correlation and Superdiffusive Process between Sequences*

The activity as defined in Section 2.2 (Thermostability) is also freely available for each protein sequence. Figure 9a shows the cumulative sum of activity, entropy, chaos, complexity, fractal, and drift parameters for 242 protein sequences. In order to track the biochemical activity changes through an invariant sequence arrangement, we have sorted, in ascending order, each sequence with increasing activity. Then, we have also sorted the remaining parameters in respect to the increasing activity and applied the cumulative sum. For clarity, we have presented the 10th of the entropy, chaos, complexity, fractal, and drift parameters, and the 1000th for activity. Most of the curves show a slightly linear shape, which is the average mode through increasing sequence activity. Nevertheless, the dynamic of changes through this increasing activity highlights that *NDCS*'s activity changes are well correlated with the *NDCS* of Kolmogorov complexity spectrum and drift (Figure 9b). There are pronounced parabola with an open upwards shape for activity (*ACT*) changes and a conversely open downwards shape for the Kolmogorov complexity spectrum (*KCS*) and drift (*DRF*) changes. The correlation coefficient is very high between *ACT*, *KCS*, and *DRF* as shown in Figure 9c.

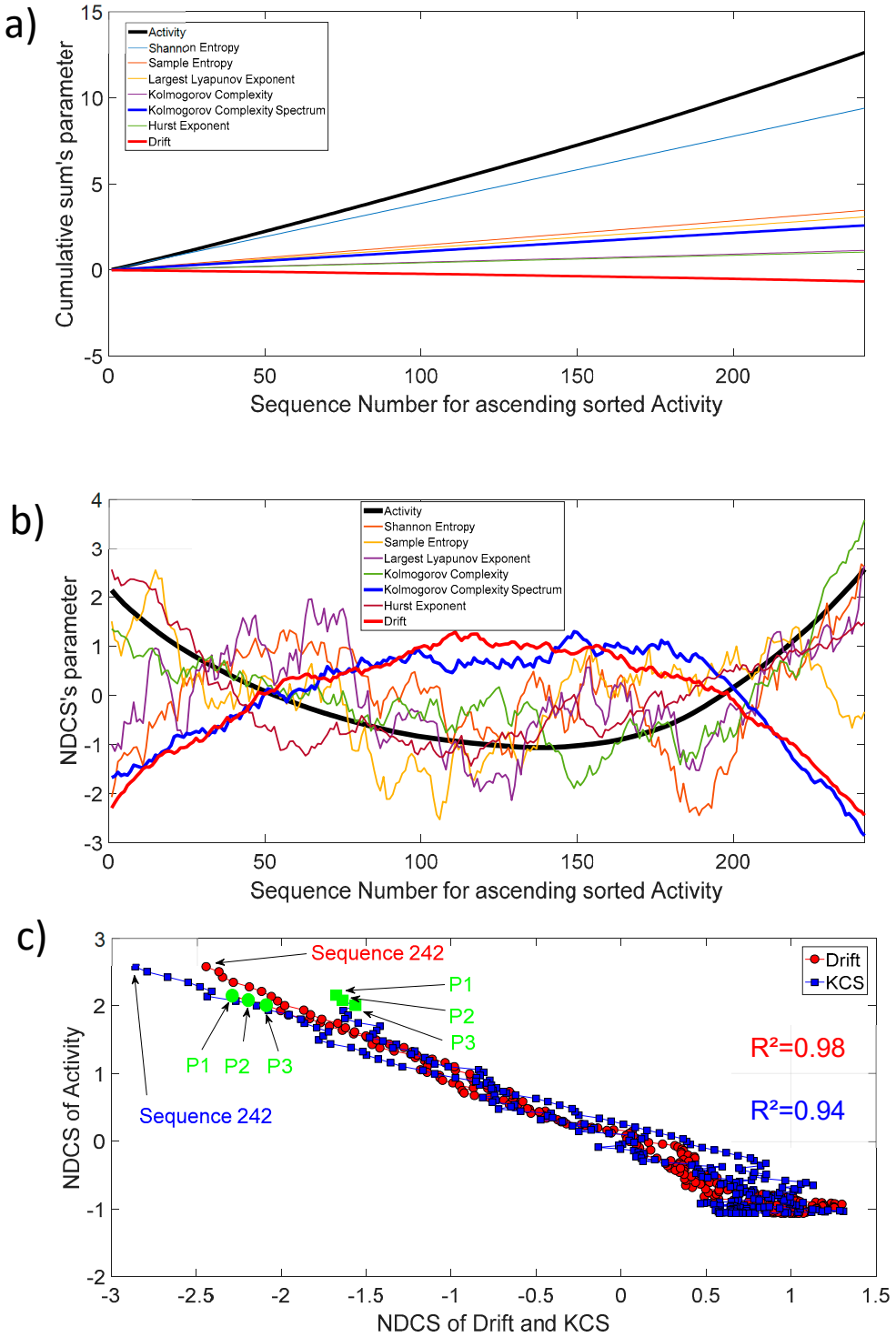


**Figure 9.** (**a**) Cumulative sum of activity, entropy, chaos, complexity, fractal, and drift parameters for ascending sorted activity; (**b**) Normalized detrended cumulative sum of activity, entropy, chaos, complexity, fractal, and drift parameters for ascending sorted activity; (**c**) NDCS of activity (*ACT*) versus NDCS of drift (*DRF*) and Kolmogorov complexity spectrum (*KCS*). The square of the correlation coefficient $R^2$ for both curves is added on the figure. The first and last sequence positions of the 242 ordered sequences are also shown. The green circle and square symbol indicate the position of the parents CYP102A1 (P1), CYP102A2 (P2), and CYP102A3 (P3) in this diagram.

We found a relationship between the inner-sequence changes drift, the complexity, and the activity throughout crossed 242 rearranged increasing activity protein sequences. As shown in Figure 9c, the trajectories of the bivariate parameter (drift, activity) or (complexity, activity) exhibits trajectories with jump between sequences, which leads to the question: "*Are these successive jumps related to variable changes ruled by a power law?*". Then, we have analyzed these trajectories by calculating the mean square displacement of changes $\langle (\Delta d_S)^2 \rangle$ in the bivariate parameter (drift, activity) or (complexity, activity) space where $d_S$ is the distance between two sequences. Moreover, we defined the mean square displacement as $\langle \Delta(d_S)^2 \rangle = \frac{1}{N_{dS}} \sum_{m=1}^{N_{dS}} \left[ \left( X^j - X^k \right)^2 + \left( ACT^j - ACT^k \right)^2 \right]_{d_S=|k-j|}$ where $N_{dS}$ is the number of pairwise sequences separated by a distance $d_S$ and $X$ is the drift (*DRF*) or Kolmogorov complexity spectrum (*KCS*). Figure 10 shows $\langle \Delta(d_S)^2 \rangle \sim {d_S}^{\alpha}$ with $\alpha \sim 1.7$ for the drift and $\alpha \sim 1.6$ for the complexity. We found that there is a scaling law of the bivariate (*DFT*, *ACT*) or (*KCS*, *ACT*) parameter that is similar to a super diffusive process with an exponent coefficient $\alpha > 1$ [33]. Here, we have plotted $\langle \Delta(d_S)^2 \rangle / \langle \Delta(d_{S_c})^2 \rangle$ where $d_{S_c}$ is the characteristic distance between two sequences computed with the correlation function $\langle \delta(d_S) \rangle = \frac{1}{N_{dS}} \sum_{m=1}^{N_{dS}} \left[ X^j X^k + ACT^j ACT^k \right]_{d_S=|k-j|}$.

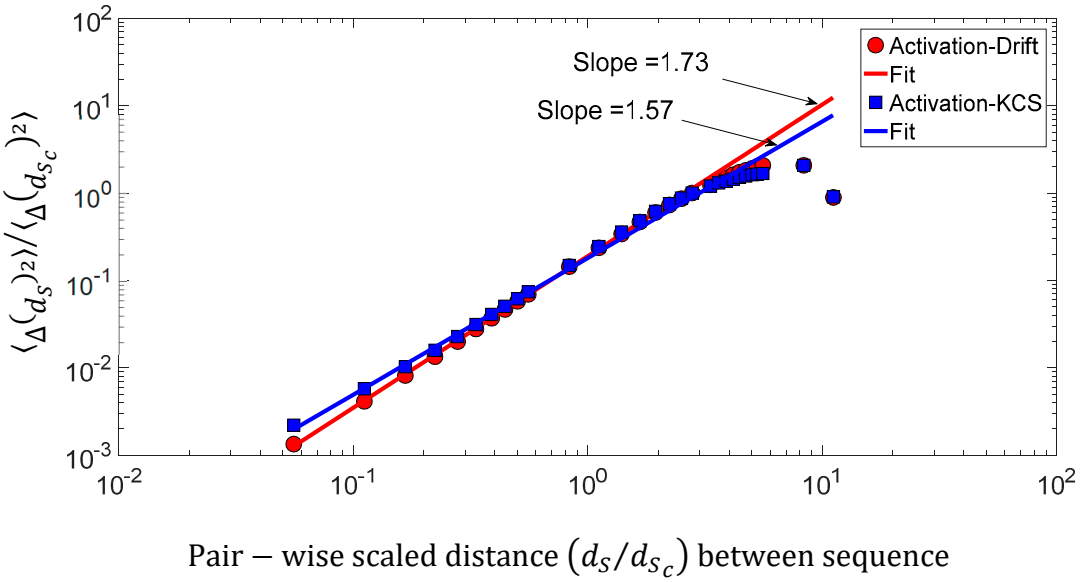


**Figure 10.** Log–log presentation of the mean square displacement $\langle \Delta(d_S)^2 \rangle / \langle \Delta(d_{S_c})^2 \rangle$ of the bivariate (*KCS*, *ACT*) parameter versus the pairwise scaled d distance $(d_S/d_{S_c})$ between sequences.

## 6. Conclusions

In this work, we analyze the nonlinear behavior of the D-PRIFT index changes around the overall linear trend scale of the protein sequence. To assess the nonlinear analysis, we have used protein residue values that are freely available, namely the AA index database. The protein dataset used contains 242 sequences and each sequence has 466 numerical values, one per amino acid residue. A protein sequence corresponds to a combination of encoding symbols from a dictionary of 20 standard amino acids symbols.

We have applied to each sequence a normalized detrended cumulative sum algorithm to extract the fluctuations of the numerical signal in the protein sequence. We analyzed these fluctuations with different tools, which are related to (i) entropy (information and regularity); (ii) chaos (largest Lyapunov exponent); (iii) complexity (Kolmogorov complexity and Kolmogorov complexity spectrum); and (iv) fractal (Hurst exponent). First, we showed that the change fluctuations of all the studied 242 protein sequences in the dataset seem to be non-stationary and follow on average a $-5/3$ Kolmogorov power-law. This result seems to be statistically significant in regard to a coefficient of variation less than 2% and a test done with randomly generated synthetically obtained data with surrogate and shuffle

technique. To understand the nature of the inner changes within the protein sequence, we achieved the analysis of the variance of the changes through the scope of the spatial correlation: Here, the index position within the protein sequence. We found an invariance of pairwise scale index $d$, which is ruled by a $S_{2i}(d) \propto d^{\alpha}$ with $\alpha = 0.87$, a coefficient close to one of the well-known stochastic Brownian processes. The dispersion of the slope obtained for all 242 protein sequences is statistically coherent in comparison with the results obtained with synthetic data. Following the local analysis of the changes along the protein sequence, we have performed a systematic q-order moment of the fluctuations in order to track if there is a self-similar repeating pattern in the inner sequence. We showed that change fluctuations within the protein sequence have a monofractal behavior, which is an average among the 242 sequences embedded between the Kolmogorov and Brownian monofractal processes with a Hurst exponent ranging between $1/3$ and $1/2$. To encompass the local analysis and to have an overview of the nonlinearity analysis, we have computed statistical parameters related to entropy, chaos, complexity, and fractality. We demonstrated that the NDCS of D PRIFT index changes for the 242 protein sequences used in this study exhibit statistically moderate complexity, and low chaotic fluctuations.

Moreover, to integrate these results in the analysis of the protein activity changes for each sequence, we have conducted a study of the relationship between the linear-trend (drift) computed with the cumulative sum algorithm, the Kolmogorov complexity spectrum, which is indicative of computational complexity, and the activity of each protein sequence. As this analysis focused on the dynamics of the changes, we also applied the normalized detrended cumulative sum for these three parameters as done for the inner-sequence analysis. The results show a strong linear relationship between the bivariate (drift, activity) and (complexity, activity) parameters, which provides insight into the potential use of drift and complexity as a predictor in a linear model. Moreover, the analysis of the trajectories in the bivariate space highlights superdiffusive behavior of the change fluctuations with a power-law around $-1.6$ of the mean square displacement for each chosen bivariate parameter. This study demonstrates that the changes in the inner sequence and throughout the crossed inter-sequence are nonstationary, stochastic, irregular, complex, weakly chaotic, and monofractal. To conclude, there is some predictability of protein sequence changes, which can be modelled using a stochastic model. Linear law and scale invariance features found in this study should be explored in future work to study for classification, regression predictive model, and could be useful in the field of protein engineering.

**Author Contributions:** Data curation, X.F.C. and M.B.; formal analysis, M.B., X.F.C., S.P.C., and R.D.; investigation, X.F.C. and M.B.; methodology, M.B., and X.F.C.; project administration, M.B. and X.F.C.; resources, X.F.C. and M.B.; software, M.B. and X.F.C.; supervision, M.B.; validation, M.B., X.F.C., S.P.C., and R.D.; visualization, M.B. and X.F.C.; writing—original draft, X.F.C. and M.B.; writing—review and editing, M.B., X.F.C., S.P.C., and R.D.

**Funding:** Peaccel gratefully acknowledge support from a research program co-funded by the European Union (UE) and Region Reunion (FEDER). The funding agencies had no influence on the conduct of this research.

**Conflicts of Interest:** X.F.C. is linked to Peaccel. M.B., X.F.C., S.P.C. and R.D. declare no competing interests.

## References

1. Hanson, J.; Yang, Y.; Paliwal, K.; Zhou, Y. Improving protein disorder prediction by deep bidirectional long short-term memory recurrent neural networks. *Bioinformatics* **2016**. [CrossRef] [PubMed]
2. Kovacs, A.; Telegdy, G. Modulation of active avoidance behavior of rats by ICV administration of CGRP antiserum. *Peptides* **1994**, *15*, 893–895. [CrossRef]
3. Niessen, K.A.; Xu, M.; George, D.K.; Chen, M.C.; Ferré-D'Amaré, A.R.; Snell, E.H.; Cody, V.; Pace, J.; Schmidt, M.; Markelz, A.G. Protein and RNA dynamical fingerprinting. *Nat. Commun.* **2019**, *10*, 1026. [CrossRef] [PubMed]
4. Qi, Z.-H.; Jin, M.-Z.; Li, S.-L.; Feng, J. A protein mapping method based on physicochemical properties and dimension reduction. *Comput. Biol. Med.* **2015**, *57*, 1–7. [CrossRef] [PubMed]
5. Gök, M.; Koçal, O.H.; Genç, S. Prediction of Disordered Regions in Proteins Using Physicochemical Properties of Amino Acids. *Int. J. Pept. Res. Ther.* **2016**, *22*, 31–36. [CrossRef]

6. Wang, Y.; You, Z.H.; Yang, S.; Li, X.; Jiang, T.H.; Xi, Z.X. A High Efficient Biological Language Model for Predicting Protein–Protein Interactions. *Cells* **2019**, *8*, 122. [CrossRef] [PubMed]
7. Plötz, T.; Fink, G.A. Pattern recognition methods for advanced stochastic protein sequence analysis using HMMs. *Pattern Recognit.* **2006**, *39*, 2267–2280. [CrossRef]
8. Chattopadhyay, A.K.; Nasiev, D.; Flower, D.R. A statistical physics perspective on alignment-independent protein sequence comparison. *Bioinformatics* **2015**, *31*, 2469–2474. [CrossRef] [PubMed]
9. Vinga, S. Information theory applications for biological sequence analysis. *Brief. Bioinform.* **2014**, *15*, 376–389. [CrossRef]
10. Zhao, J.; Wang, J.; Hua, W.; Ouyang, P. Algorithm, applications and evaluation for protein comparison by Ramanujan Fourier transform. *Mol. Cell. Probes* **2015**, *29*, 396–407. [CrossRef]
11. Czerniecka, A.; Bielińska-Wąż, D.; Wąż, P.; Clark, T. 20D-dynamic representation of protein sequences. *Genomics* **2016**, *107*, 16–23. [CrossRef] [PubMed]
12. Zhu, X.-J.; Feng, C.-Q.; Lai, H.-Y.; Chen, W.; Hao, L. Predicting protein structural classes for low-similarity sequences by evaluating different features. *Knowl. Based Syst.* **2019**, *163*, 787–793. [CrossRef]
13. Yang, L.; Wei, P.; Zhong, C.; Meng, Z.; Wang, P.; Tang, Y.Y. A Fractal Dimension and Empirical Mode Decomposition-Based Method for Protein Sequence Analysis. *Int. J. Pattern Recognit. Artif. Intell.* **2019**. [CrossRef]
14. Yu, J.F.; Cao, Z.; Yang, Y.; Wang, C.L.; Su, Z.D.; Zhao, Y.W.; Wang, J.H.; Zhou, Y. Natural protein sequences are more intrinsically disordered than random sequences. *Cell. Mol. Life Sci.* **2016**, *73*, 2949–2957. [CrossRef] [PubMed]
15. Cao, C.; Liu, F.; Tan, H.; Song, D.; Shu, W.; Li, W.; Zhou, Y.; Bo, X.; Xie, Z. Deep Learning and Its Applications in Biomedicine. *Genom. Proteom. Bioinform.* **2018**, *16*, 17–32. [CrossRef]
16. Li, Y.; Drummond, D.A.; Sawayama, A.M.; Snow, C.D.; Bloom, J.D.; Arnold, F.H. A diverse family of thermostable cytochrome P450s created by recombination of stabilizing fragments. *Nat. Biotechnol.* **2007**, *25*, 1051–1056. [CrossRef] [PubMed]
17. Kawashima, S.; Ogata, H.; Kanehisa, M. Aaindex: Amino Acid Index Database. *Nucleic Acids Res.* **1999**, *27*, 368–369. [CrossRef]
18. Kawashima, S.; Pokarowski, P.; Pokarowska, M.; Kolinski, A.; Katayama, T.; Kanehisa, M. Aaindex: Amino acid index database, progress report 2008. *Nucleic Acids Res.* **2008**, *36*, D202–D205. [CrossRef]
19. Shannon, C.E. A Mathematical theory of Communication. *Bell Syst. Tech. J.* **1948**, *27*, 379–423. [CrossRef]
20. Richman, J.S.; Moorman, J.R. Physiological time-series analysis using approximate entropy and sample entropy. *Am. J. Physiol. Heart Circ. Physiol.* **2000**, *278*, H2039–H2049. [CrossRef]
21. Wolf, A.; Swift, J.B.; Swinney, H.L.; Vastano, J.A. Determining Lyapunov exponents from a time series. *Phys. Nonlinear Phenom.* **1985**, *16*, 285–317. [CrossRef]
22. Kolmogorov, A.N. The local structure of turbulence in incompressible fluid for very large Reynolds numbers. *Dokl. Akad. Nauk. SSSR* **1941**, *30*, 299–303. [CrossRef]
23. Chaitin, G.J. On the Length of Programs for Computing Finite Binary Sequences: Statistical considerations. *J. ACM* **1969**, *16*, 145–159. [CrossRef]
24. Lempel, A.; Ziv, J. On the Complexity of Finite Sequences. *IEEE Trans. Inf. Theory* **1976**, *22*, 75–81. [CrossRef]
25. Mihailović, D.T.; Mimić, G.; Nikolić-Djorić, E.; Arsenić, I. Novel measures based on the Kolmogorov complexity for use in complex system behavior studies and time series analysis. *Open Phys.* **2015**, *13*, 1–14. [CrossRef]
26. Ziv, J.; Lempel, A. Compression of individual sequences via variable-rate coding. *IEEE Trans. Inf. Theory* **1978**, *24*, 530–536. [CrossRef]
27. Monin, A.S.; Yaglom, A.M. *Statistical Fluid Mechanics: Mechanics of Turbulence*; MIT Press: Cambridge, MA, USA, 1987; Volume 1, p. 784.
28. Schreiber, T.; Schmitz, A. Surrogate time series. *Phys. Nonlinear Phenom.* **2000**, *142*, 346–382. [CrossRef]
29. Peng, C.; Buldyrev, S.V.; Havlin, S.; Simons, M.; Stanley, H.E.; Goldberger, A.L. Mosaic organization of DNA nucleotides. *Phys. Rev. E* **1994**, *49*, 1685–1689. [CrossRef] [PubMed]
30. Cornette, J.L.; Cease, K.B.; Margalit, H.; Spouge, J.L.; Berzofsky, J.A.; DeLisi, C. Hydrophobicity scales and computational techniques for detecting amphipathic structures in proteins. *J. Mol. Biol.* **1987**, *195*, 659–685. [CrossRef]

31. Regier, P.; Briceño, H.; Boyer, J.N. Analyzing and comparing complex environmental time series using a cumulative sums approach. *MethodsX* **2019**, *6*, 779–787. [CrossRef]
32. Marshak, A.; Davis, A.; Cahalan, R.; Wiscombe, W. Bounded cascade models as nonstationary multifractals. *Phys. Rev. E* **1994**, *49*, 55–69. [CrossRef] [PubMed]
33. Richardson, L.F. Atmospheric Diffusion Shown on a Distance-Neighbour Graph. *Proc. R. Soc. Math. Phys. Eng. Sci.* **1926**, *110*, 709–737. [CrossRef]