

## Research Article

# Can Power Laws Help Us Understand Gene and Proteome Information?

J. A. Tenreiro Machado,<sup>1</sup> António C. Costa,<sup>1</sup> and Maria Dulce Quelhas<sup>2</sup>

<sup>1</sup> *Institute of Engineering, Polytechnic of Porto, Department of Electrical Engineering, Rua Dr. António Bernardino de Almeida 431, 4200-072 Porto, Portugal*

<sup>2</sup> *National Health Institute, Biochemical Genetics Unit, Medical Genetics Center “Jacinto de Magalhães”, Praça Pedro Nunes 88, 4099-028 Porto, Portugal*

Correspondence should be addressed to J. A. Tenreiro Machado; [jtm@isep.ipp.pt](mailto:jtm@isep.ipp.pt)

Received 11 February 2013; Accepted 27 February 2013

Academic Editor: Dumitru Baleanu

Copyright © 2013 J. A. Tenreiro Machado et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Proteins are biochemical entities consisting of one or more blocks typically folded in a 3D pattern. Each block (a polypeptide) is a single linear sequence of amino acids that are biochemically bonded together. The amino acid sequence in a protein is defined by the sequence of a gene or several genes encoded in the DNA-based genetic code. This genetic code typically uses twenty amino acids, but in certain organisms the genetic code can also include two other amino acids. After linking the amino acids during protein synthesis, each amino acid becomes a residue in a protein, which is then chemically modified, ultimately changing and defining the protein function. In this study, the authors analyze the amino acid sequence using alignment-free methods, aiming to identify structural patterns in sets of proteins and in the proteome, without any other previous assumptions. The paper starts by analyzing amino acid sequence data by means of histograms using fixed length amino acid words (tuples). After creating the initial relative frequency histograms, they are transformed and processed in order to generate quantitative results for information extraction and graphical visualization. Selected samples from two reference datasets are used, and results reveal that the proposed method is able to generate relevant outputs in accordance with current scientific knowledge in domains like protein sequence/proteome analysis.

## 1. Introduction

Tyers and Mann [1] identified the future importance of proteomics (the study of the proteome) and the requisites needed to fulfill its potential. The proteome concept has been studied by researchers like Nicodeme et al. [2], Bock and Gough [3], and Nabieva et al. [4], just to mention a few. Nowadays most of proteome research uses alignment methods and focuses on portions of the protein sequence code.

While chromosome sizes range from tens of thousands to thousands of million base nucleotides, protein sizes range from half a dozen up to tens of thousands of amino acids. Another difference between genome and proteome codification is in the alphabets used of: in the genome the DNA base nucleotides belong to a 4 symbols of alphabet {A, C, G, T}; in the amino acids sequences of the proteome, the alphabet

contains at least 20 symbols [5]. In this study the following set of 21 amino acids was adopted: alanine: A; Cysteine: C; aspartic acid: D; glutamic acid: E; phenylalanine: F; glycine: G; histidine: H; isoleucine: I; lysine: K; leucine: L; methionine: M; asparagine: N; proline: P; glutamine: Q; arginine: R; serine: S; threonine: T; selenocysteine: U; valine: V; tryptophan: W; and tyrosine: Y [6].

Inspired by the work of Vinga and Almeida [7] on alignment-free comparison methods, in [8] the authors describe how the nuclear and chromosomal genomes are analyzed as DNA sequences of symbols from the {A, C, G, T} nucleotide alphabet and how information processing methods are applied to generate several types of data visualizations depicting distinct levels of structural organization. To be able to cope with different DNA sequence lengths, the authors adopted a histogram-based approach, converting the sequence information into tuples and then counting relative

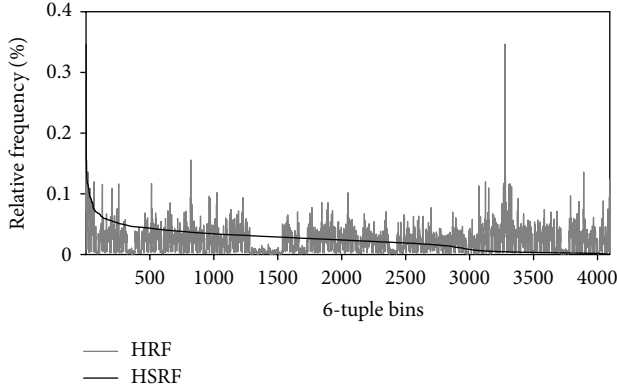


FIGURE 1: HRF and HSRF of the human chromosome 1.

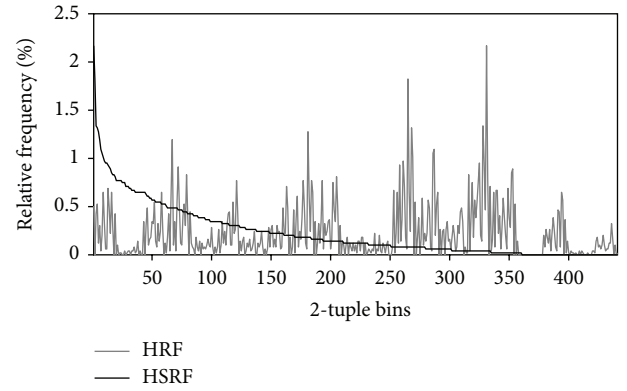


FIGURE 2: HRF and HSRF of the PCLO\_HUMAN\_Q9Y6V0-6 protein.

tuple frequencies along the whole sequence. After generating the histograms for the input sequences, those histograms are processed by mathematical tools and further information about chromosomes, genomes, and organisms is produced.

Weiss et al. [9] stated that protein sequences can be regarded as slightly edited random strings. Dai and Wang [10] introduced the “protein sequence space” concept to explore similar sequences using statistical measures. Another amino acid sequence approach is described by Hemmerich and Kim [11] based on correlation measures and is able to classify proteins without alignment information.

In this paper a histogram-based approach is proposed for dealing with and analyzing the amino acid sequences of proteins. Before counting relative frequencies of amino acid tuples, the tuple length to use must be defined, as well as the process of moving from one tuple to the next one. Because the adopted amino acid alphabet contains 21 symbols and amino acid sequences typically do not exceed 40000 symbols, only tuples of length  $n = \{1, 2, 3\}$  were considered, knowing that the total number of different tuples is  $21^n$  for a certain  $n$  ( $n = 4$  allows 194481 different tuples, much larger than 40000). As such, when using  $n > 2$ , most of many protein’s relative frequencies tend to be zero. For moving from one  $n$ -tuple to the next, the one amino acid sliding window was adopted (i.e., overlap of  $n - 1$  amino acids).

A histogram of relative frequencies (HRF) of a sequence containing symbols from a certain alphabet may be considered a digest or hash representation of that sequence. The size of an HRF does not depend on the sequence length, but only on the tuple size adopted for counting the relative frequencies, which facilitates the comparison of sequences with different lengths and does not require previous assumptions about the sequences’ contents.

Figure 1 shows the HRF of human chromosome 1 ([12], 6-tuple bins) and the same frequencies sorted from left to right in decreasing mode. Figure 2 (4935 amino acids, 2-tuple bins) corresponds to the PCLO\_HUMAN\_Q9Y6V0-6 protein (Swiss-Prot:Q9Y6V0-6). In both cases the HRFs merely reveal large variations between relative frequencies, while the histograms of sorted relative frequencies (HSRF) show a pattern similar to the “Pareto principle,” commonly associated with Power-Law (PL) relationships [13, 14].

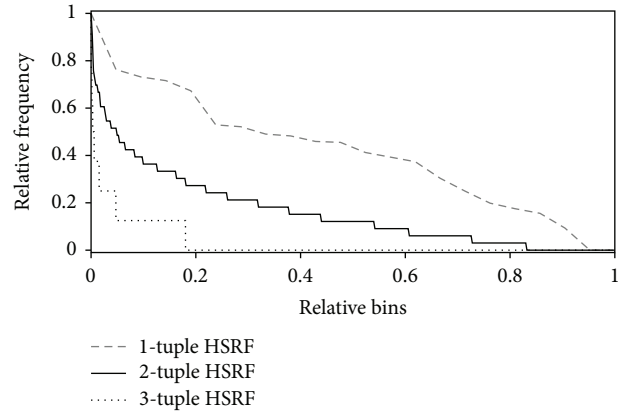
FIGURE 3: HSRFs of the FINC\_RAT\_P04937-4 protein ( $n = \{1, 2, 3\}$ ) with normalized horizontal and vertical axes.

Figure 3 shows the effect of using distinct values of  $n$  for generating several HSRFs of the sample FINC\_RAT\_P04937-4 protein (Swiss-Prot:P04937-4). For 1-tuple bins, the PL relationship does not show up, but for 2 tuples it is clearly visible. Although the PL pattern is still noticeable for 3 tuples, most of the bins in the HSRF are zero due to the small size of the protein sequence ( $\sim 2400$  amino acids  $\ll 21^3$ ).

## 2. Methods

**2.1. Datasets.** The protein sequences used in this study were downloaded in the first week of January 2012 from the Universal Protein Resource database [15], namely, the archives “Complete UniProtKB/Swiss-Prot data set” (Dataset 1) and “Additional sequences of the UniProtKB/Swiss-Prot data set that represent all annotated splice variants” (Dataset 2), both using the FASTA format. Dataset 1 was selected because it contains a very large sample of proteins (associated to genes and organisms) and Dataset 2 because it mostly contains isoforms of a large number of proteins in Dataset 1.

**2.2. Implementation.** From the analysis of Figures 1–5 we decided to adopt statistical methods for studying protein sequence information and for highlighting underlying relationships between protein sequences. As such, we identified the histogram of relative frequencies (HRF) and the histogram of sorted relative frequencies (HSRF) as the main tools to use. Both in HRF and HSRF, each bin is associated with an  $n$ -tuple of amino acids, chosen from an alphabet of 21 distinct amino acids. Considering that the size of longest protein sequence is less than 40000 and that, for a certain  $n$ , the number of different  $n$ -tuples is  $21^n$ , the choice of adequate values for  $n$  is  $\{1, 2, 3\}$  (larger values of  $n$  originate HRF and HSRF mostly containing null bins).

The sorting technique adopted to transform a HRF into an HSRF is usually associated with the statistical analysis of phenomena characterized by Power-Law (PL) relationships. In fact, there is a broad class of natural and manmade phenomena whose statistical description includes histograms with long tails and may be approximated by expressions like

$$f(x) = ax^b, \quad a > 0, \quad b < 0, \quad (1)$$

where  $(a, b)$  are parameters related to the analyzed phenomena.

In this study about protein sequences, there was no significant *a priori* knowledge about the type of resulting histograms. After initial experiments with the HRFs and HSRFs of many protein sequences and the detection of PL patterns in those HSRFs, we were convinced that this approach could lead to an assertive characterization of proteins, genes, and organisms. Therefore, by using the UniProtKB/Swiss-Prot protein datasets, many HSRFs were computed along with their respective PL regression using the “*hrfpl*” application. Because proteins are associated with genes and the chosen datasets contain many homologous proteins of the same gene, the “*hrfplg*” was built to compute the PL regression of proteins per gene.

The process of generating an HSRF from an HRF is done by sorting of the  $n$ -tuple bins so that they become listed by decreasing values of relative frequency. During the sorting, the initial bin sequence, numbered from 1 to  $21^n$ , is transformed into a distinct bin sequence, with a bin numbering different from the initial one (the final bin numbering is a permutation of the initial one). Both the initial and final bin numberings may be treated as “ranked lists” and processed by any method able to compute a “distance” between ranked lists, which behaves like another parameter ( $c$ ) related to protein sequence analysis. To compute the  $(a, b, c)$  parameters at once from a set of HRFs, the following methods were implemented:

- (i) PL regression + Kendall-Tau distance “*hrfplkt*” [16];
- (ii) PL regression + Spearman FR distance “*hrfpls*” [17];
- (iii) PL regression + Canberra distance “*hrfplcd*” [18].

After computing the  $(a, b, c)$  parameters for a set of protein sequences, eventually gene or organism related, the results can be visualized by means of 2D graphics involving two of the aforementioned parameters:  $a$  versus  $b$ ,  $a$  versus

$c$ , and so forth. Another possibility is a 3D visualization with the three parameters by means of 3D rendered graphics using shadows, reflections, and other visual artifacts or 3D videos.

Observing some regularities in 2D graphics-relating PL parameters  $(a, b)$ , we decided to calculate the trend-line parameters  $(p, q)$  of a set of related protein sequences by using the following regression:

$$b = p \log(a) + q \quad (2)$$

which is aiming to improve the perception of underlying regularities.

In order to compare the results of the methods described in this study with the methods presented in [8], we include a brief description of the HRF-based methods in that paper. After computing several HRFs, the first step is to build a square correlation matrix relating each HRF to all others. Typically a correlation matrix entry varies between 0 (no correlation) and 1 (100% correlated). To compute distance measures between HRFs, many techniques may be applied—the following one was adopted: Jensen-Shannon divergence “*hrfcorrjsd*” [19, 20]. The multidimensional scaling technique (MDS, [21]) can be used for the visualization of correlation matrix data in 2D or 3D graphics by means of the GGobi software package [22].

**2.3. Testing.** Although the application of PL to research-involving proteins was firstly described by Huynen and Nimwegen [23], Qian et al. [24], and Karev et al. [25], the recent availability of new data sets of protein sequences related to genes and organisms opens up new research possibilities. For instance, Dataset 1 contains 131771 protein sequences belonging to 21231 genes or 6751 organisms as follows (Figure 4):

- (i) 43 genes, each having 500 or more proteins;
- (ii) 185 genes, each having 100 or more proteins;
- (iii) 1379 genes, each having 10 or more proteins,

or

- (i) 19 organisms, each having 500 or more proteins;
- (ii) 256 organisms, each having 100 or more proteins;
- (iii) 1400 organisms, each having 10 or more proteins.

Figure 4 shows the frequencies of protein sizes up to 6000 amino acids, with most of the proteins inside the [20–1500] interval. In Dataset 2, there are 30800 protein/isoform sequences, corresponding to 11960 genes (of which 300 have 10 or more protein/isoforms) or 614 organisms (of which 31 have 10 or more proteins/isoforms). Figure 5 shows the frequencies of protein sizes up to 6000 amino acids, with most of the proteins inside the [30–3000] interval.

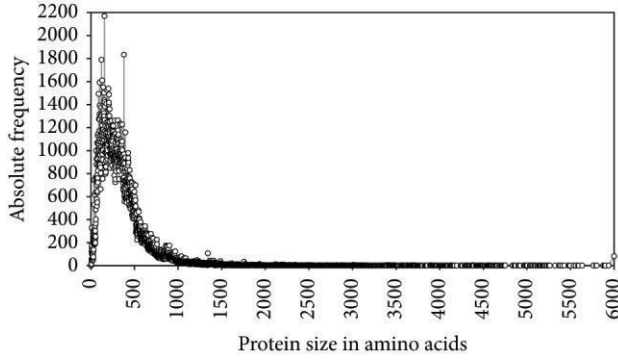


FIGURE 4: Frequencies of protein sizes for 131771 proteins (Dataset 1).

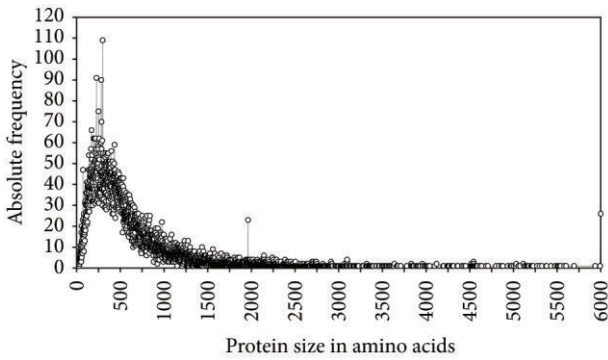


FIGURE 5: Frequencies of protein sizes for 30800 proteins (Dataset 2).

### 3. Results and Discussion

**3.1. Regression-Based-Only Approaches.** In Figure 5 the absolute frequencies of the 30800 protein sequences of Dataset 2 were shown. Figure 6 shows the PL regressions  $(a, b)$  of all those protein sequences. The spatial distribution of all  $(a, b)$  values clearly denotes a linear pattern of organization around the diagonal line from the top-left corner to the bottom-right corner of that figure, with some spread around that line.

Because Figure 6 suggests the possibility that the *locus*  $(a, b)$  of HSRFs may have underlying regularities, in Figure 7(a) we plot the *locus*  $(a, b)$  of 27204 proteins belonging to 40 genes (each gene having at least 500 proteins from several organisms), in which each symbol represents a protein from a certain gene. It is clearly visible that almost all the genes have a linear protein distribution along the  $\log(a)$  versus  $b$  top-left bottom-right diagonal. Figure 7(b) details the PL regression *locus*  $(a, b)$  of genes APT, ARGB, and AROA present in Figure 7(a). These three genes contain 1803 protein sequences.

The 638 protein HSRF PL regressions of the APT gene (diamond-shaped symbols) generate a trend line with an 81.4%  $R^2$  goodness of fit. For the PL regressions of the 585 protein HSRFs of the ARGB gene (circle-shaped symbols), the trend-line has an 84.7%  $R^2$  goodness of fit. Lastly, for the PL regressions of the 580 protein HSRFs of the AROA

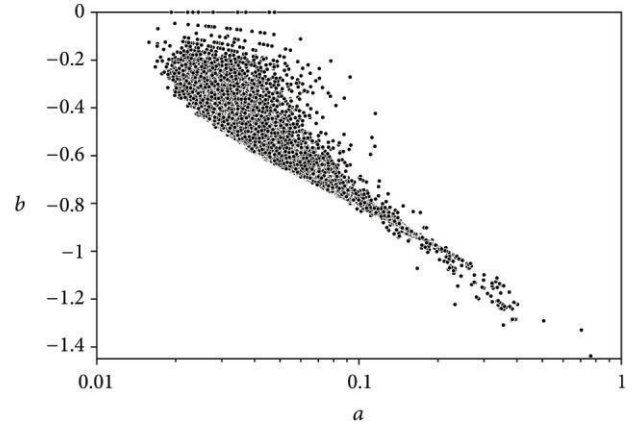


FIGURE 6: *Locus*  $(a, b)$  of PL regressions of all 2-tuple HSRFs in Dataset 2.

gene (square-shaped symbols), the resulting trend line has a 98.0%  $R^2$  goodness of fit. In all these genes the goodness of fit is very high and their proteins HSRF PL regressions are remarkably aligned in “lines.”

In Figure 7(b) the HSRF PL regressions were obtained for each gene, but regressions can also be derived for each organism, as represented in Figure 7(c). This figure depicts four organisms (ARATH, HUMAN, MOUSE, and TREPA), each including at least 15 proteins from Dataset 1, with 32 genes involved (82% of the genes shared between organisms). In Figure 7(c), no organism-protein regularities are detected (this was confirmed by other test cases).

Figure 8 shows the *locus*  $(a, b)$  of 2463 proteins/isoforms belonging to 100 genes (each gene having at least 14 proteins/isoforms from several organisms), with each symbol representing a protein from a certain gene. A similar “line alignment” regularity is depicted by most of the 100 genes. When each gene is subjected to a trend-line regression, only 9 of the 100 genes have  $R^2$  goodness of fit values below 80%, with 27 genes having  $R^2$  goodness of fit values above 99%.

In Figures 7(a), 7(b), and 8, it was observed that proteins of the same gene tend to be aligned in straight lines. This motivated the application of a second-level abstraction to the PL regression *locus*  $(a, b)$  in order to facilitate the perception of new regularities: the  $(p, q)$  trend-line regression from *locus*  $(a, b)$  data. Figure 9(a) shows the *locus*  $(p, q)$  of the trend-line regression for the 40 genes and the associated protein data used in Figure 7(a). Each circle in Figure 9(a) represents a gene with an area inversely proportional to its PL regression  $R^2$  goodness of fit. The *locus*  $(p, q)$  is remarkably close to a linear fitting, which is confirmed by its significant  $R^2$  goodness of fit: 99.0%. The discrepancy in the PL regression  $R^2$  goodness of fit value for the 40 genes in Figures 7(a) and 9(a) is depicted by the distinct circle areas and detailed in Figure 9(b), which shows the distribution of  $R^2$  goodness of fit values (16 values in 40 below 90%).

Figure 10(a) shows the *locus*  $(p, q)$  of the trend-line regression for the 1544 genes (at least 5 proteins/gene within 12412 protein sequences/isoforms from Dataset 2) and



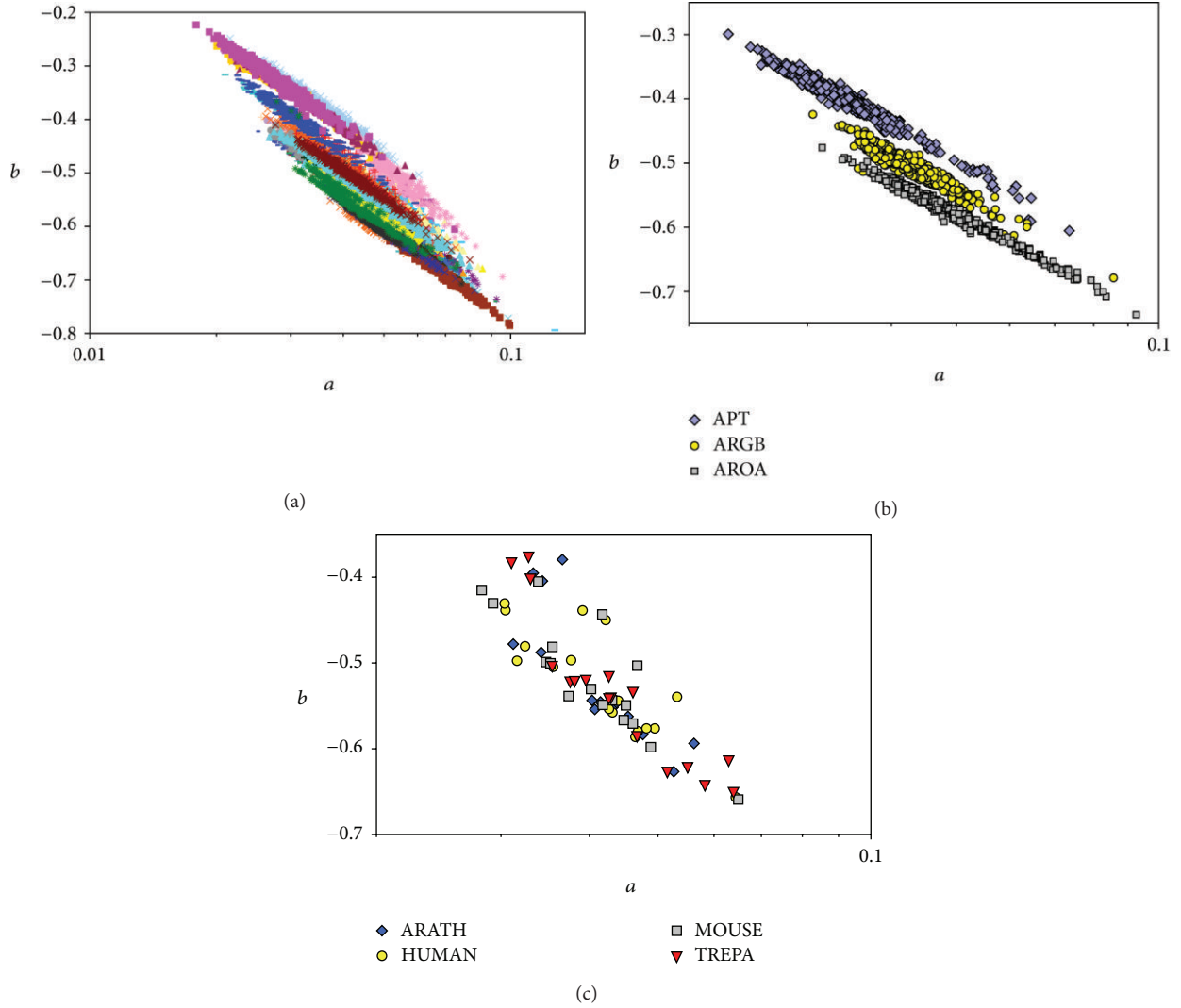


FIGURE 7: (a) PL regression locus  $(a, b)$  of 27204 2-tuple HSRFs in Dataset 1. (b) PL regression locus  $(a, b)$  of genes APT, ARGB and AROA (1803 2-tuple HSRFs) in Dataset 1. (c) PL regression locus  $(a, b)$  of organisms ARATH, HUMAN, MOUSE, and TREPA (66 2-tuple HSRFs) in Dataset 1.

Figure 10(b) depicts the corresponding distribution of  $R^2$  goodness of fit values (434 values in 1544 below 90%).

**3.2. Regression-Based and Distance Approaches.** As described in Section 2, whenever an HRF is sorted by decreasing relative frequencies and generates an HSRF, the bin numbering sequence is modified. Assuming that all HRFs use the initial bin numbering sequence  $\{1, 2, \dots, 21^n\}$ , each HSRF will contain a permuted sequence of the initial one. The particular permutation depends on the sorting process and the HRF relative frequencies. Being the sorting process universal, an HSRF bin numbering sequence only depends on the contents of its associated HRF.

Any bin numbering sequence can be considered a ranked list of integers in the range  $[1, 21^n]$ , so any method that

computes the distance between ranked lists can be used for finding the distance between bin numbering sequences. This means that an HSRF can be used to extract three parameters as follows:

- (i)  $(a, b)$ —PL regression on the sorted relative frequencies;
- (ii) (c)—distance between the HRF and HSRF ranked lists.

An immediate implication of having  $(a, b, c)$  instead of merely  $(a, b)$  is that PL regression plots become 3-dimensional, allowing for the detection of new and previously not described regularities.

Previously three techniques were described to simultaneously compute the locus  $(a, b, c)$  from a set of protein's

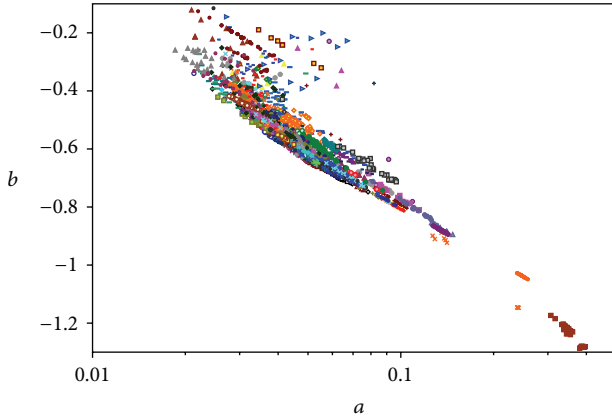


FIGURE 8: PL regression locus ( $a, b$ ) of 2463 2-tuple HSRFs in Dataset 2, including canonical protein sequences and isoforms.

HRFs: PL regression + Kendall-Tau distance, PL regression + Spearman FR distance, and PL regression + Canberra distance. Using the 1083 proteins of genes APT, ARGB, and AROA (see Figure 7(b)), the three methods were tested. It was found that the method with the best visual performance is “PL regression + Canberra distance,” whose results are presented in Figure 11(a), which depicts a 3D rendering, with shadows, of the locus ( $a, b, c$ ) of all 1083 proteins (colored by gene). In this figure the existence of three gene clusters, clearly separated from each other is very noticeable. Beyond being elongated, each cluster is also mostly planar. The “shadow in the floor” is an approximation of Figure 7(b).

One important task is the identification of proteins belonging to the aforementioned 3D clusters. The GGobi interactive software package is a versatile tool for the analysis and exploration of complex data and was used to create Figure 11(b). There we can see a 2D projection of the 3D clusters and some labeled proteins. Using GGobi, we verified that each cluster is composed by proteins codified by the same gene (APT, ARGB, or AROA) and not by organism type (also verified in Figures 7(b) and 7(c)).

Figures 12(a) and 12(b) show the resulting 3D gene clustering when the “PL regression + Canberra distance” method is applied to 40 genes (27204 2-tuple HRFs in Dataset 1), previously represented in Figure 7(a). The 3D gene clusters of Figures 12(a) and 12(b) basically exhibit the same patterns previously described for Figure 11(a). Nevertheless, some new regularities can now be detected when viewing the clusters at close: interspersed clusters, crossing clusters, and nonplanar clusters.

**3.3. Correlation and MDS Approaches.** In [8] is described an HRF based approach for the analysis and visualization of nuclear/mitochondrial genomic data. That approach is not genome specific and can be applied to other HRFs.

Figure 13(a) depicts a rendering of the protein/gene 3D clustering that results from applying the Jensen-Shannon divergence correlation technique, followed by the MDS tool of GGobi package, to the HRFs of 1803 proteins from genes

APT, ARGB, and AROA. The existence of three spatially separated gene clusters is clearly visible: the “spherical” APT gene cluster involved by the ARGB cluster and both surrounded by the more spread AROA cluster. When compared to Figure 11(a), Figure 13(a) shows distinct regularities regarding the three genes and their related proteins.

Figure 13(b) displays a rendering of the protein/gene 3D clustering after applying the Jensen-Shannon divergence correlation technique, followed by the MDS tool of GGobi package, to the HRFs of 782 proteins from 36 genes (each represented by at least 15 proteins/isoforms).

In the upper part of Figure 13(b), many gene clusters are clearly visible, while in the middle of the figure the clusters are more spread and mixed. There are also “globular” and “linear” clusters, as well as regions where no clusters are identifiable.

## 4. Conclusions

According to Murray et al. [26], in biochemistry the structure of proteins is divided into four categories as follows:

- (1) primary—the amino acid sequence;
- (2) secondary—regularly repeating local structures;
- (3) tertiary—the protein’s overall shape;
- (4) quaternary—a complex formed by several proteins.

In terms of information content, a protein is derived from a gene or a group of genes. But even for a single gene, several distinct protein representations can be generated by means of “alternative splicing,” a process in which symbols that codify amino acids are manipulated and changed before protein synthesis. Alternative splicing is one of the mechanisms that increase the biodiversity of genome encoded proteins [27], but its implications are still incompletely understood.

Contrary to the DNA, the alphabet of proteins contains a large number of symbols ( $\sim 21$ ) and, although protein sequence lengths are small when compared to chromosomes, the number of possible proteins is almost infinite. For the eukaryotic DNA primary sequence Arneodo et al. [28] have shown that it contains a multiscale information encoding and a hierarchical structure (from tens of DNA bps up to hundreds of millions of DNA bps).

The primary protein sequence also seems to exhibit those two characteristics: multiscale encoding and hierarchical structure. Figure 2 shows that, for a sample protein, the decreasing frequency of 2 tuples is compatible with a PL pattern (this pattern also occurs for 3 tuples with protein sequences larger than 8 k amino acids). As PL distributions have successfully contributed to the modeling of real phenomena, our main motivation was to apply PL-based methods to the study of protein sequences in search of clues for protein multiscale encoding and hierarchical structure.

Using two datasets from the Universal Protein Resource KnowledgeBase repository, several experiments were designed and performed, based on the concepts of HRF and HSRF. Figure 6 shows a structured locus ( $a, b$ ) for the PL regressions of 30800 HSRF protein sequences from Dataset 2. Figure 7(a) shows a more structured locus of PL

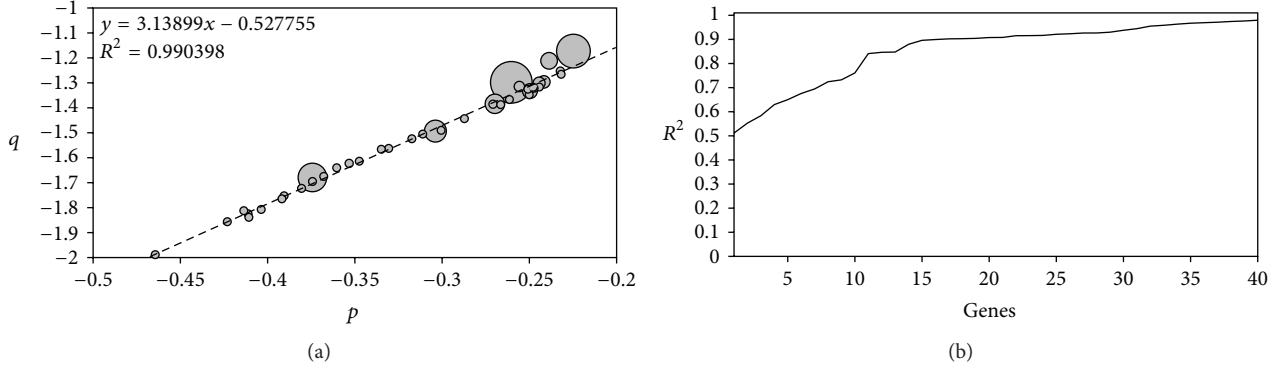


FIGURE 9: (a) Trend-line regression of the *locus* ( $a, b$ ) of 40 genes from 27204 2-tuple HSRFs in Dataset 1.  $R^2 = 99.0\%$ . (b) Distribution of  $R^2$  goodness of fit values for the PL regression of 40 genes from 27204 2-tuple HSRFs in Dataset 1.

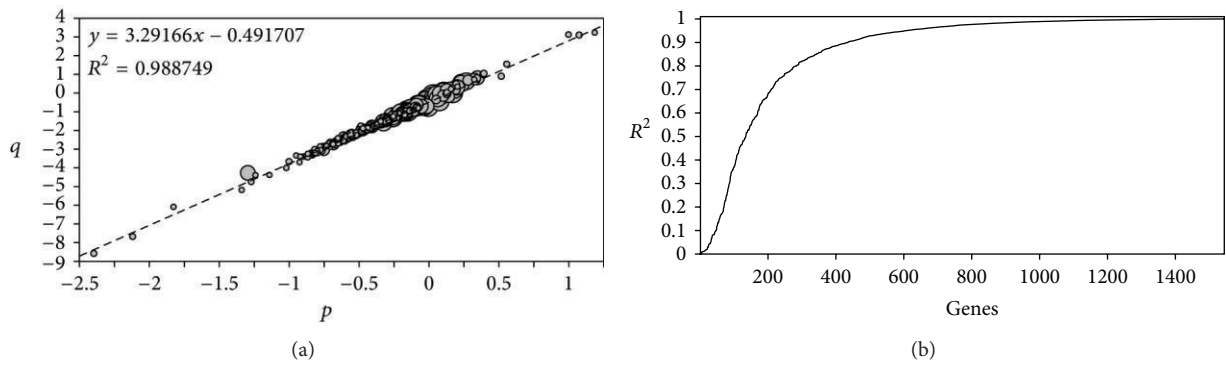


FIGURE 10: (a) Trend-line regression of the *locus* ( $a, b$ ) of 1544 genes from 12412 2-tuple HSRFs in Dataset 2.  $R^2 = 98.9\%$ . (b) Distribution of  $R^2$  goodness of fit values for the PL regression of 1544 genes from 12412 2-tuple HSRFs in Dataset 2.

regressions for 27204 proteins belonging to 40 genes, with a closeup of three genes depicted in Figure 7(b) pointing to a clear gene-protein association. Contrarily, in Figure 7(c) no organism-protein relationship is noticeable. Another well-structured *locus* ( $a, b$ ) of PL regressions for 100 genes (2463 proteins/isoforms from Dataset 2) is shown in Figure 8.

The “line-oriented” spatial distribution of gene-related PL regressions depicted in Figure 7(a) (also present in Figures 7(b) and 8) was captured in Figure 9(a) by means of a trend-line regression of the *locus* ( $a, b$ ). In that figure the 40 genes (20204 protein sequences) are clearly line-aligned, most of them with very high goodness of fit values (Figure 9(b)). Although Figure 9 is just another way of presenting the information contained in Figure 7(a), it facilitates the perception of the underlying structuring between genes and proteins. The same phenomena is displayed and verified in Figures 10(a) and 10(b) using 1544 genes (12412 proteins/isoforms) from Dataset 2.

Figures 11(a), 11(b), 12(a), and 12(b) are the consequence of another observation: the process of sorting an HRF into an HSRF transforms the numbering sequence of the relative frequency bins. In the end of that process we get two “ranked lists,” one for the HRF another for the HSRF, and a “distance measure” can be computed between them. Figure 7(b) has its three-dimensional version in Figure 11(a),

which uses a new *locus* ( $a, b, c$ ), with  $c$  being the aforementioned “distance measure” parameter. In the vertical axis of Figure 11(a) (labeled “ $c$ ”), it is clearly visible that genes/proteins have another structuring pattern, which can be confirmed in Figure 11(b). Figures 12(a) and 12(b) are the three-dimensional extended version of Figure 7(a) and the vertical axis separation between the color-coded genes is also noticeable.

In order to test the adopted datasets with another methodology, it was decided to use the HRF-based approach described in [8]: HRF calculation from protein sequences, correlation between HRFs and finally clustering using the correlation matrix. One equivalent of Figure 11(a) using the correlation-clustering approach previously described is shown in Figure 13(a), in which the proteins of the three genes are grouped and the genes themselves are spatially separated (look at the “shadows on the floor”). Figure 13(b) is similar to Figure 13(a), with 36 genes and 782 proteins taken from Dataset 2. With at least 15 proteins/isoforms per gene, in that figure are visible many gene clusters with distinct spatial organization.

One of the major benefits of using histograms of  $n$ -tuples relative frequencies (HRF) for the analysis of variable length categorical data sequences is that it simplifies the process of comparing those sequences, making it less dependent

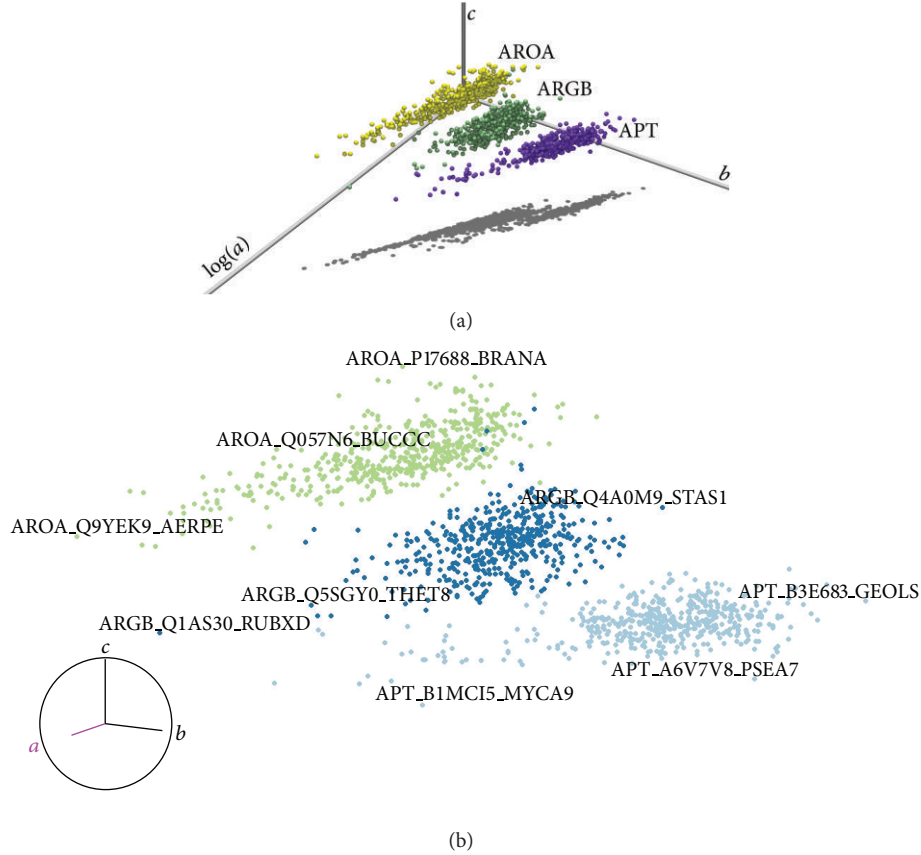


FIGURE 11: (a) *Locus*  $(a, b, c)$  of genes APT, ARGB, and AROA (1803 2-tuple HSRFs) in Dataset 1—HRF PL regression + Canberra distance. (b) GGobi made 2-dimensional projection of *locus*  $(a, b, c)$  of genes APT, ARGB, and AROA (1803 2-tuple HSRFs) in Dataset 1, with some proteins labeled.

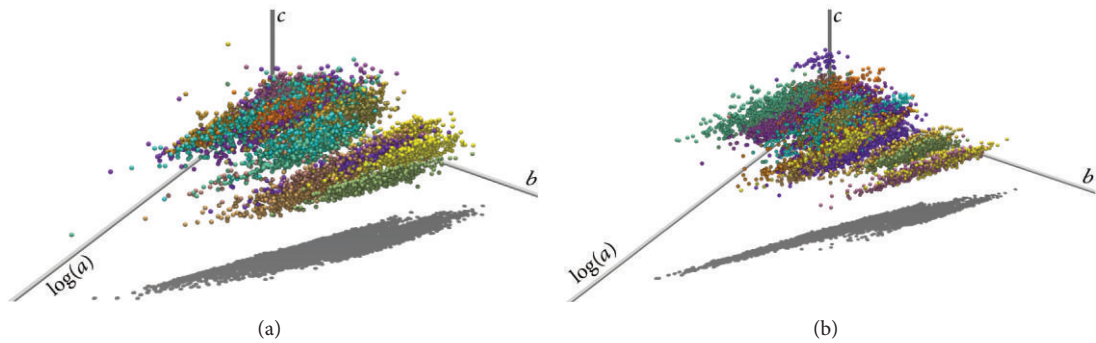


FIGURE 12: (a) *Locus*  $(a, b, c)$  of genes 1–20 for 40 genes (1803 2-tuple HSRFs) in Dataset 1—HRF PL regression + Canberra distance. (b) *Locus*  $(a, b, c)$  genes 21–40 for 40 genes (1803 2-tuple HSRFs) in Dataset 1—HRF PL regression + Canberra distance.

on their length. To be a free-alignment method is another important benefit, because it requires almost no *a priori* knowledge about the sequences.

This is the case with chromosomal/mitochondrial sequences (genome) and amino acid sequences (proteome). Nevertheless, for amino acid sequences, the large variation in sequence size (displayed in Figures 4 and 5) may create

difficulties in dealing with HRFs, especially with sequences smaller than one hundred symbols. With this type of sequences, an HRF may contain mostly null bins and this can adversely affect subsequent data processing. This is the reason why amino acids smaller than 100 symbols were avoided in our experiments. The open source code of developed tools is freely available for download.



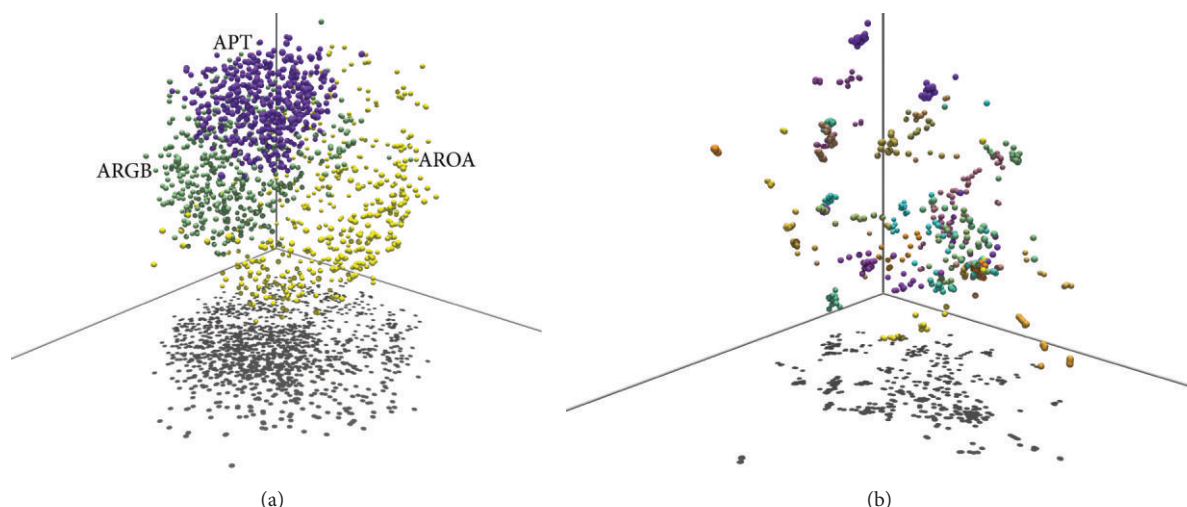


FIGURE 13: (a) Locus  $(a, b, c)$  of genes APT, ARGB, and AROA (1803 2-tuple HRFs) in Dataset 1—HRF Jensen-Shannon correlation and MDS 3D clustering. (b) Locus  $(a, b, c)$  of 36 genes (782 2-tuple HRFs) in Dataset 2—HRF Jensen-Shannon correlation and MDS 3D clustering.

**4.1. Open Issues and Future Work.** One very relevant open issue is the low availability of good-quality amino acid sequences in protein repositories like UniProt [29] or similar. This issue severely limits experiments involving large numbers of proteins per gene and proteins per organism, which are necessary when one is trying to detect and identify multiscale regularities.

Another challenging issue is the application of the described HRF/HSRF methodology to other protein processing frameworks as a preprocessor for data validation or as a tool for result verification, just to mention two examples. This is a promising issue for future research work.

The HRF concept is not limited to the processing of  $n$ -tuple “successive” symbols, or even to the existence of only one HRF per sequence. The concept may be generalized, extended, and applied in novel ways, which we are already actively researching.

Another interesting open issue is the existence of a “physical” interpretation of the HRF/HSRF-derived  $(a, b, c)$  or  $(p, q)$  parameters and how they relate to measurable quantities belonging to the problem domain. A good answer to this question may be the key to improve and increase the application of HRF methodology in problems in biology and other related areas.

## Abbreviations

HRF: Histogram of relative frequencies  
 HSRF: Histogram of sorted relative frequencies  
 PL: Power law.

## Acknowledgments

The authors thank the following organizations for access to input data: (1) Genome Reference Consortium (Human genome), <http://www.ncbi.nlm.nih.gov/projects/genome/assembly/grc/>. (2) Universal Protein Resource, <http://www.uni->

[prot.org/](http://prot.org/). This work was supported by FEDER Funds through the “Programa Operacional Factores de Competitividade—COMPETE” program and by National Funds through FCT “Fundação para a Ciência e a Tecnologia” under Project FCOMP-01-0124-FEDER-PEst-OE/EEI/UI0760/2011.

## References

- [1] M. Tyers and M. Mann, “From genomics to proteomics,” *Nature*, vol. 422, pp. 193–197, 2003.
- [2] P. Nicodeme, T. Doerks, and M. Vingron, “Proteome analysis based on motif statistics,” *Bioinformatics*, vol. 18, no. 2, pp. S161–S171, 2002.
- [3] J. Bock and D. Gough, “Whole-proteome interaction mining,” *Bioinformatics*, vol. 19, no. 1, pp. 125–134, 2003.
- [4] E. Nabieva, K. Jim, A. Agarwal, B. Chazelle, and M. Singh, “Whole-proteome prediction of protein function via graph-theoretic analysis of interaction maps,” *Bioinformatics*, vol. 21, no. supplement 1, pp. i302–i310, 2005.
- [5] D. Nelson and M. Cox, *Lehninger Principles of Biochemistry*, Worth Publishers, 3rd edition, 2000.
- [6] International Union of Pure and Applied Chemistry, <http://www.iupac.org/>.
- [7] S. Vinga and J. Almeida, “Alignment-free sequence comparison—a review,” *Bioinformatics*, vol. 19, no. 4, pp. 513–523, 2003.
- [8] A. Costa, J. Machado, and M. Quelhas, “Histogram-based DNA analysis for the visualization of chromosome, genome and species information,” *Bioinformatics*, vol. 27, no. 9, pp. 1207–1214, 2011.
- [9] O. Weiss, M. Jimenez-Montano, and H. Herzog, “Information content of protein sequences,” *Journal of Theoretical Biology*, vol. 206, no. 3, pp. 379–386, 2000.
- [10] Q. Dai and T. Wang, “Comparison study on k-word statistical measures for protein: from sequence to ‘sequence space,’” *BMC Bioinformatics*, vol. 9, no. 394, pp. 1471–2105, 2008.
- [11] C. Hemmerich and S. Kim, “A study of residue correlation within protein sequences and Its application to sequence

- classification,” *EURASIP Journal on Bioinformatics and Systems Biology*, vol. 2007, Article ID 87356, 2007.
- [12] NCBI Genome Download/FTP, [ftp://ftp.ncbi.nlm.nih.gov/genomes/H\\_sapiens/CHR\\_01/](ftp://ftp.ncbi.nlm.nih.gov/genomes/H_sapiens/CHR_01/).
  - [13] A. Clauset, C. R. Shalizi, and M. E. J. Newman, “Power-law distributions in empirical data,” *SIAM Review*, vol. 51, no. 4, pp. 661–703, 2009.
  - [14] C. M. A. Pinto, A. Mendes Lopes, and J. A. T. Machado, “A review of power laws in real life phenomena,” *Communications in Nonlinear Science and Numerical Simulation*, vol. 17, no. 9, pp. 3558–3578, 2012.
  - [15] Universal Protein Resource, [ftp://ftp.uniprot.org/pub/databases/uniprot/current\\_release/knowledgebase/](ftp://ftp.uniprot.org/pub/databases/uniprot/current_release/knowledgebase/).
  - [16] M. Kendall, “A new measure of rank correlation,” *Biometrika*, vol. 30, no. 1-2, pp. 81–89, 1938.
  - [17] D. Sculley, “Rank Aggregation for Similar Items,” in *Proceedings of the 7th SIAM International*, SIAM, Philadelphia, Pa, USA, 2007.
  - [18] G. Jurman, S. Riccadonna, R. Visintainer, and C. Furlanello, “Canberra distance on ranked Lists,” in *Proceedings of Advances in Ranking and Neural Information Processing Systems Workshop (NIPS '09)*, S. Agrawal, C. Burges, and K. Crammer, Eds., pp. 22–27.
  - [19] J. Lin, “Divergence measures based on the Shannon entropy,” *Transactions on Information Theory*, vol. 37, no. 1, pp. 145–151, 1991.
  - [20] S. H. Cha, “Taxonomy of nominal type histogram distance measures,” in *Proceedings of the American Conference on Applied Mathematics 2008*, pp. 325–330, WSEAS.
  - [21] I. Borg and P. Groenen, *Modern multidimensional scaling*, Springer Series in Statistics, Springer, New York, NY, USA, 1997, Theory and applications.
  - [22] GGobi software package, <http://www.ggobi.org/>.
  - [23] M. Huynen and E. Nimwegen, “The frequency distribution of gene family sizes in complete genomes,” *Molecular Biology and Evolution*, vol. 15, no. 5, pp. 583–589, 1998.
  - [24] J. Qian, N. Luscombe, and M. Gerstein, “Protein family and fold occurrence in genomes: power-law behaviour and evolutionary model,” *Journal of Molecular Biology*, vol. 313, no. 4, pp. 673–681, 2001.
  - [25] G. Karev, Y. Wolf, A. Rzhetsky, F. Berezovskaya, and E. Koonin, “Birth and death of protein domains: a simple model of evolution explains power law behavior,” *BMC Evolutionary Biology*, vol. 2, no. 18, 2002.
  - [26] R. Murray, D. Bender, V. Rodwell, K. Botham, P. Kennelly, and P. A. Weil, *Harper's Illustrated Biochemistry*, McGraw-Hill, 28th edition, 2009.
  - [27] Q. Pan, O. Shai, L. J. Lee, B. J. Frey, and B. J. Blencowe, “Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing,” *Nature Genetics*, vol. 40, no. 12, pp. 1413–1415, 2008.
  - [28] A. Arneodo, C. Vaillant, B. Audit, F. Argoul, Y. Aubenton-Carafa, and C. Thermes, “Multi-scale coding of genomic information: from DNA sequence to genome structure and function,” *Physics Reports*, vol. 498, pp. 45–188, 2010.
  - [29] UniProtn, <http://www.uniprot.org/>.