Clustering Sequences with a Statistical Content Evaluation Method

Alina Bogan-Marta

University of Oradea, Faculty of Electrotechnics and Information Technology, Universitatii 1, 410087 Oradea, Romania, e-mail: alinab@uoradea.ro

Nicolae Robu

"Politehnica" University of Timisoara, Faculty of Automation and Computer Science and Engineering, Vasile Pârvan No. 2, 1900 Timisoara, Romania
e-mail: nicolae.robu@rectorat.utt.ro

Abstract: The clustering of biological sequences into biologically meaningful classes denotes two computationally complex challenges: the choice of a biologically pertinent and computable criterion to evaluate the clusters homogenity, and the optimal exploration of the solution space. Here we are analysing the clustering potential of a new method of sequence similarity based on statistical sequence content evaluation. Its computational efficiency and high accuracy of the results is encouraging for further development that could make it an appealing alternative to the existent methods.

Keywords: biological sequence, n-grams, entropy, dissimilarity matrix, exploratory data analysis

1 Introduction

In bioinformatics, **sequence clustering** algorithms attempt to group sequences that are somehow related. Generally, the clustering algorithms are single linkage clustering, constructing a transitive closure of sequences with a similarity over a particular threshold. The similarity score is often based on sequence alignment. Usually, sequence clustering is used to make a non-redundant set of representative sequences [1] and sequence clusters are often synonymous with (but not identical to) protein families. Determining a representative structure for each *sequence cluster'* is the aim of many structural genomics initiatives [1]. The general purpose of grouping proteins into families leads to more sensitive detection of new

members and improved discrimination against spurious hits based on the essential conserved features in a family [2].

The most obvious measure of the similarity (or dissimailarity) between two samples is the distance between them. One way to begin a clustering investigation is to define a suitable metric and compute the matrix of distances between all pairs of samples. If distance is a good measure of dissimilarity, then one would expect the distance between samples in the same cluster to be significantly less then the distance between the samples in different clusters [3].

There are numerous algorithms and associated programs to perform cluster analysis, for example, hierarchical methods [4], self-organizing maps [5], k-means [6], and model-based approaches [7], [8], [9]. Existing clustering approaches that have been applied to biological sequences, mostly proteins, are reviewed in [2]. Many of them are based on manual or semi-manual procedures; others are fully automatic but less reliable. To our knowledge, there is no generally accepted method that is able to produce automatically an accurate clustering of a large biological sequence database. Conventional clustering algorithms employ distance (or similarity) measure to form the clusters [9] when graph partitioning algorithms exploit the structure of a graph to find highly connected objects. Hence, the biologist wishing to perform cluster analysis is faced with a dyzzing array of algorithmic choices and little basis on which to make a choice.

Having proposed a new similarity measure for protein sequences in a previous work [11] we come here to analyse it in clustering process. The new method is based on Markov chains representation known as *n*-gram in statistical language modeling. A similarity measure estimation derived from cross entropy was adopted from information theory field in order to compute the similarity between the resulting *n*-grams. The new strategy was applied for the task of clustering protein sequences using a geometrical representation, based on the dissimilarity matrices dertived from sequence comparisons within two different databases of proteins.

2 Method

2.1 The New Statistical Similarity Metod

Protein sequences from all different organisms can be treated as texts written in a universal language in which the alphabet consists of 20 distinct symbols, the amino-acids. The mapping of a protein sequence to its structure, functional dynamics and biological role then becomes analog to the mapping of words to their semantic meaning in natural languages. This analogy can be exploited by

applying *statistical language modeling* and *text classification techniques* for the advancement of biological sequences understanding. Scientists within this hybrid research area believe that the identification of Grammar/Syntax rules could reveal entities/relations of high importance for biological and medical sciences.

In the presented method, we adopted a Markov-chain grammar to build for our protein dataset *2*-gram, *3*-gram and *4*-gram models. To clarify things we chose a hypothetical protein sequence WASQVSENR. In the *2*-gram modeling the available tokens/words were {WA AS SQ QV VS SE EN NR}, while in the *3*-gram representation they were {WAS ASQ SQV QVS VSE SEN ENR}. Based on the frequencies of these tokens/words (estimated by counting) and by forming the appropriate ratios of frequencies, the entropy of an *n*-gram model can be readily estimated using (1) as comes from Van Uytsel and Compernolle's work [12].

$$\hat{H}_L(X) = -\frac{1}{N}\sum_{W^*} Count\left(w_1^n\right)\log_2 p_L\left(w_n \middle| w_1^{n-1}\right) \quad , \tag{1}$$

where the variable *X* has the form of an *n*-gram $X = w_1^n \Leftrightarrow \{w_1, w_2, \ldots, w_n\}$ and $Count(w_1^n)$ is the number of occurrences of w_1^n. The summation runs over all the possible *n*-length combinations of consecutive *w* (i.e $W^*=\{\{w_1,w_2,\ldots,w_n\},\{w_2,w_3,\ldots,w_{n+1}\},\ldots.\}$) and *N* is the total number of *n*-grams in the investigated sequence. The second term, $p\left(w_n \middle| w_1^{n-1}\right)$, in (1) is the conditional probability that relates the *n*-th element of an *n*-gram with the preceding *n*-1 elements. Following the principles of maximum likelihood estimation (MLE) [13], it can be estimated by using the corresponding relative frequencies:

$$\hat{p}\left(w_n \middle| w_1^{n-1}\right) = \frac{Count\left(w_n\right)}{Count\left(w_1^{n-1}\right)}. \tag{2}$$

This measure is indicative about how well a specific protein sequence is modeled by the corresponding *n*-gram model. While this measure could be applied to two distinct proteins (and help us to decide about which protein is better represented by the given model), the outcomes cannot be used for a direct comparison of them. Thus, the common information content between two proteins *X* and *Y* is expressed via the formula:

$$E(X,Y) = -\sum_{all\ w_1^n} P_X(w_1^n)\log P_Y\left(w_n \mid w_1^{n-1}\right) \tag{3}$$

The first term $P_X\left(w_1^n\right)$ in (3) corresponds to the reference protein sequence *X* (i.e. it results from counting the words of that specific protein). The second term corresponds to the sequence *Y* based on which the model has to be estimated (i.e. it results from counting the tokens of that protein). Variable w_1^n ranges over all the words (that are represented by *n*-grams) of the reference protein sequence.

2.2 Sequence Comparison Strategies with the New Similarity Method

Having introduced the new similarity measure, we proceed here with the description of its use in order to perform comparisons within protein databases. The essential point of our approach is that the compared proteins in a given database (containing annotated proteins with known functionality, structure etc.) are represented via *n*-gram encoding and the above introduced similarity is utilized to compare their representations.

We considered two different ways in which the *n*-gram based similarity is engaged in efficient database searches. The most direct implementation is called hereafter as *direct method.* A second algorithm, the *alternating method*, was devised in order to cope with the fact that the proteins to be compared could be of very different length. It is easy to observe the need of having two methods if sequences of very different length are compared. The procedure of experimenting with both methods and contrasting their performances gave the opportunity to check the sensitivity of the proposed measure regarding the length of the sequences.

Direct method. Let S_q be the sequence of a query-protein and $\{S\}=\{S_1, S_2, \ldots S_N\}$ the given protein database. The first step is the computation of 'perfect' score (PS) or 'reference' score for the query-protein. This is done by computing $E(S_q,S_q)$ using the query-protein both as reference and model sequence (we call here "model" the sequence compared with the query) in equation (3). In the second step, each protein S_i, i=1…N, from the database serves as the model sequence in the computation of a similarity score $E(S_q,S_i)$, with the query-protein serving as reference sequence. In this way, N similarities are computed $E(S_q,S_i)$, i=1,..,N. Finally, these similarities are compared against the perfect score PS by computing the absolute differences $D(S_q,S_i)=|E(S_q,S_i)-PS|$. The 'discrepancies' in term of information content between the query-protein and the database-proteins are expressed. By ranking these N measurements, we can easily identify the most similar proteins to the query-protein as those which have been assigned the lowest distance $D(S_q,S_i)$.

Alternating method: The only difference with respect to the direct method is that when comparing the query-protein with those from the database, the role of reference and model protein can be interchanged based on the shortest (the shortest sequence plays the role of reference sequence in (3)). The other steps, perfect-score estimation, ranking and selection, follow as previously.

3 Experiments

3.1 Sequence Database

The proposed strategy based on measuring protein similarity was demonstrated and validated using two experimental databases. A small one, containing an overall sample of 100 protein sequences where two distinct groups of protein data had been selected as follows. The first 50 entries of the database corresponded to proteins selected at random from the NCBI public database [14]. The last 50 entries corresponded to proteins resulted from different mutations of the p53 gene. The mutations were selected randomly from the database we created using the descriptions, provided by the International Agency for Research on Cancer (IARC) Lyon, France [15]. This set of 50 proteins, denoted hereafter as p53-group, is expected to form a tight-cluster of textual-patterns in the space of biological semantics. On the contrary, the rest 50 proteins should appear as textual-patterns in the same space that differ not only with other, but also (and mainly) from the p53-group. It could be formulated as the problem of two class recognition.

The second database is a set of 1460 proteins extracted from Astral SCOP 1.67 sequence resources [16]. From the available/original corpus of data, which is a structured one, only those families containing at least 10 protein sequences were included in our new database. In this way, 31 different families unequally populated were finally included. We mention that the annotation of our database follows the original annotation relaying on the biological meaning of similarity concept (and therefore can be considered as providing the 'ground-truth' for the protein classification). As in the small database set, we expected that all the proteins belonging to the same family would appear as a tight cluster of textual patterns and having a proper similarity measure so as we could differentiate the existent families.

This database (of 1460 proteins) was organized in 3 different sets, in order to observe at a smaller scale the behavior of the applied similarity technique.

3.2 Results

The geometrical consideration, according to which the patterns are represented by points (i.e. the end tails of corresponding vectors), in a multidimensional space, is very useful in order to conceptualize morphological relationships between patterns, to search for natural groupings inside the sample patterns, etc. The key idea is that similar patterns are mapped onto nearby points [17].

In order to validate the two variants of the strategy we proposed, are followed

some classical steps of *Exploratory Data Analysis*. Generating the procedure of similarity search between the sequences in each data set we have, we built the corresponding dissimilarity matrix used by the representation technique to illustrate the geometrical distribution of our data. In Figure 1, 2, 3 and 4 the matrix containing all the possible dissimilarity measures $D(S_i,S_j)$, i,j=1,2,…N is depicted as a grey scale image, for both algorithmic variants of our method and three different *n*-gram models. In the adopted visualization scheme all the shown matrices (after proper normalization) share a common scale in which the 1 (white) corresponds to the maximum distance in each matrix. It is worth mentioning here that the 'ideal' spatial outlay is a white matrix with only a black segment at the lower right corner. Therefore, it is evident from all these figures that *4*-gram modeling has a very good representation for searching sequence similarity within the given database.

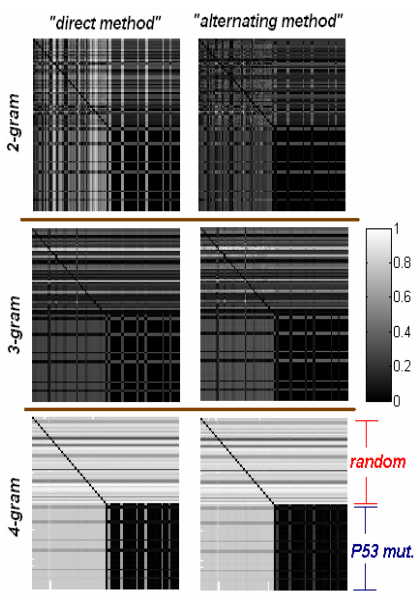

Figure 1
Visualization of the matrices containing all the possible pairwise dissimilarities of the 100 proteins for 2,3,4-gram models

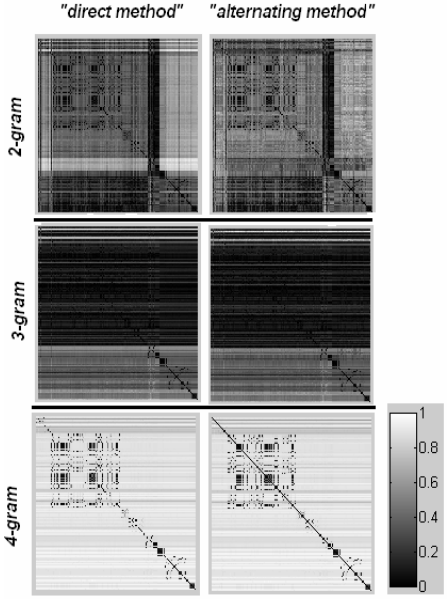

Figure 2
Visualization of the matrices containing all the possible pairwise dissimilarities for the 497 proteins of Set1, for 2,3,4-gram models

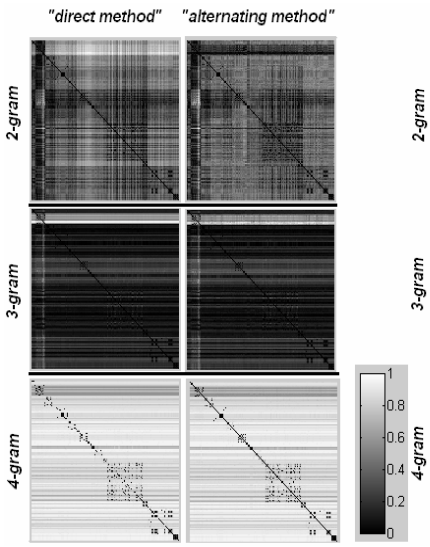

Figure 3
Visualization of the matrices containing all the possible pairwise dissimilarities for the 497 proteins of Set2, for 2,3,4-gram models

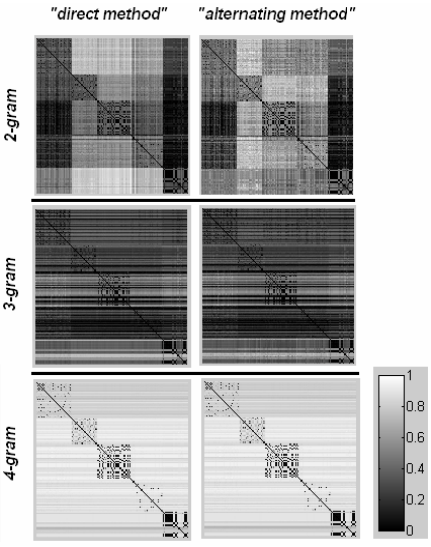

Figure 4
Visualization of the matrices containing all the possible pairwise dissimilarities for the 466 proteins of Set3, for 2,3,4-gram models

Due to the obvious separation of sequences in the small database we tried to identify the affiliation of sequences grouped as tight cluster in the dark corner of the Fig. 1. The results are shown in Figure 5 which is a low-dimensional representation of protein sequences using dissimilarity measure for the small set of experimental data. Here, it is obvioius the fact that we obtained a very god solution to the two class identification problem.

Regarding the cluster identification in the second experiments, we used a strategy based on Huberts's statistics [18] in determining the correlation factors between the clusters we obtained and the 'ground-truth' offered by the original protein sequence families/superfamilies structure. The walues are tabulated in Table 1.

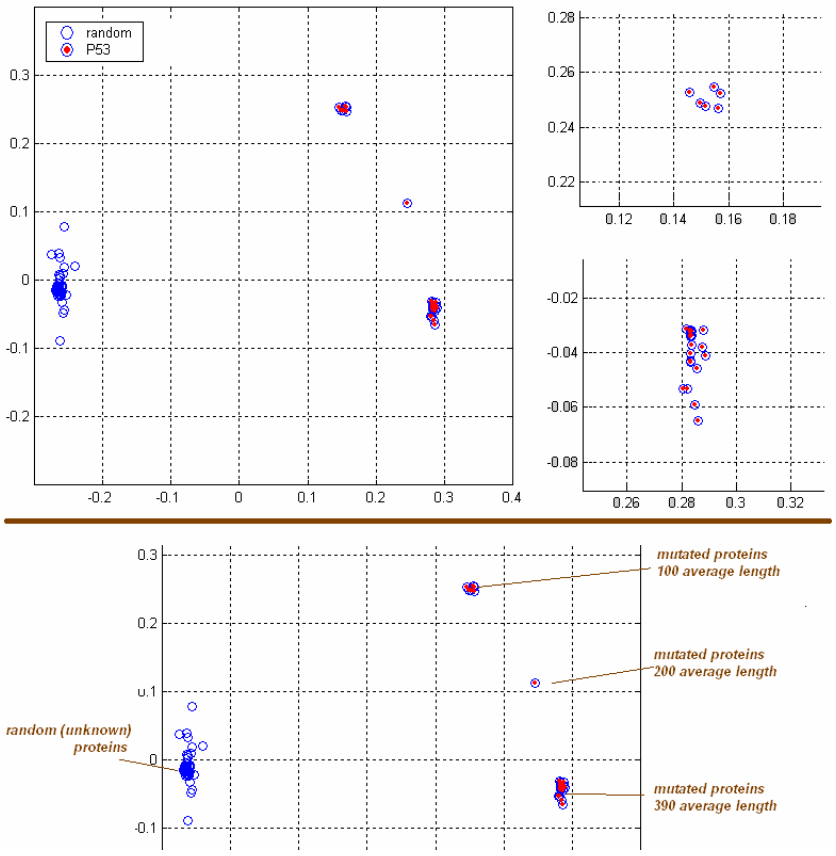

Figure 5

Low-dimensional representation of protein sequences using dissimilarity measure for the small set of experimental data

Table 1

Correlation values for set1, 2 and 3 of the second protein database using Huberts's statistics

Set	Direct method	Variant method
1	0.1230	0.1110
2	0.0347	0.0354
3	0.3339	0.3622

In the interpretation of the correlation values there are considered as good scores the high values and it is not the case achieved with the assumption we made (the

possible identification of biological structural classification). Under this circumstances the biologist reasoning helps in elucidating the clusters representation meaning. The explataion comes from the fact that many times sequences of different length and with partial identity in content may belong to the same biological family. So, the clusters we get are representing the similar sequences in textual representation. This conclusion is already justified by the very good performance of mutated proteins identification in the small database.

Conclusions

The method experimented in this paper constitutes a step forward in investigating the engagement of language modelling for characterizing, handling and understanding biological data in the format of sequences. Specifically, we studied the efficiency of this new method in revealing the context relatedness between sequences. The experimental results indicate the reliability of our algorithmic strategy for expressing similarity between proteins according to the sequence text content. Given the conceptual simplicity of the introduced approach, it appears as an encouraging alternative to previous well-established techniques.

Here won't be discussed the two methods comparative performance as the results of the similarity search are geometrically represented but regarding the order of the employed *n*-gram model, after testing with order of 2, 3, 4, 5 we noticed, as can be seen in Figures 1-4 that the performance of the method increases with the order of the model up to 4. After the order of 5 due to the lack of data, the corresponding maximum likelihood estimates become unreasonable uniform and very low.

Analysing the meaning of clusters identified in visual representation of the dissimilarity matrices we may consider that this content evaluation similarity measure performs well for sequences having related textual representation. This aspect can lead to a general clustering strategy. Despite the corelation values that didn't confirm our biological expectation we are motivated to adjust the similarity method by working with functional groups of amino acids. It has to be made the observation that till now we worked only with concepts from information theory field applied on protein sequence.

Considering the algorithmic simplicity and computational efficiency of our approach, in this form, we are justified to suggest it as a first choice when searches in large databases are required. In terms of time complexity in absence of a detailed analysis we are motivated to consider this method efficient especially when search procedure is running over large databases containing long sequences. This motivate us to pursue further on how to achieve even higher performance.

Acknowledgement

Experiments of this work were supported by the EU project Biopattern: Computational Intelligence for biopattern analysis in Support of eHealthcare, Network of Excellence Project No. 508803. One of the authors is very grateful to

dr. Nikos Laskaris for invaluable help in understanding and appying exploratory data algorithms used to represent and conclude over the results.

References

[1] Wikipedia, the free encyclopedia: http://en.wikipedia.org/wiki/Sequence_clustering

[2] Heger A, Holm L: Towards a covering set of protein family profiles, Progress in Biophysics and Molecular Biology, Vol. 73, 2000, pp. 321-337

[3] Duda O Richard, Hart E Peter, Stork G David: Unsupervised Learning and Clustering, in Pattern Classification, 2nd edition, John Wiley & Sons, Inc, USA, 2001, pp. 538

[4] Hartigan JA: Clustering Algorithms, John Wiley & Sons, New York, 1975

[5] Kohonen T: Self-organizing Maps. Springer-Verlag, Berlin/Heidelberg, 1997

[6] MacQueen J: Some methods for classification and analysis of multivariate observations, in Proceedings of the 5th Berkeley Symp. Math. Stat. Probability. Ed, Cam LML and Neyman J. University of California Press, 1965, 281-297

[7] Fraley C, Raftery AE: Model-based clustering, discriminant analysis, and density estimation. J Am Stat Assoc 2002, Vol. 97, pp. 611-631

[8] McLachlan GJ, Basford KE: Mixture Models: Inference and Applications to Clustering, Ed. Marcel Dekker, 1988

[9] McLachlan GJ, Peel D: Finite Mixture Models, Ed. Willey, New York, 2000

[10] Kirsten M., Wrabel, S., & Horvath, T: Distance-based approaches to relational learning and clustering, in Relational data mining, Springer-Verlag New York, Inc. 2000, pp. 213-230

[11] Bogan-Marta A, Gavrielides A Marios, Pitas Ioannis, Lyroudia Kleoniki: A New Statistical Measure of Protein Similarity based on Language Modeling, GENSIPS 2005, May 22-24, in Newport, RI, USA

[12] Van Uytsel DH, and Van Compernolle D: Entropy-based context selection in variable-length n-gram language models, *IEEE Benelux Signal Proc. Symp.*, 1998, pp. 227-230

[13] Manning CD, and Schütze H: Foundations of statistical natural language processing, Massachusetts Institute of Technology Press, Cambridge, Massachusetts London, England, 2000, pp. 554-556; 557-588

[14] National Center for Biotechnology Information: (http://www.ncbi.nlm.nih.gov/)

[15] International Agency for Researc on Cancer: http://www.iarc.fr/p53/Somatic.html

[16] Structural Classification Of Proteins official site: http://scop.mrc-lmb.cam.ac.uk/scop/

[17] Laskaris AN: Algorithms for Vectorial Pattern-Analysis, in Basics of Geometrical Data-Analysis, under publishing

[18] Laskaris NA, Ioannides AA: Clinical Neurophysiology, Nr. 113, 2002, 1209-1226