A New Statistical Measure of Protein Similarity based on Language Modeling

Alina Bogan-Marta, Marios A. Gavrielides, Ioannis Pitas,*Kleoniki Lyroudia
Artificial Intelligence and Information Analysis Laboratory, Department of Informatics
*Dental School, Department of Endodontology
Aristotle University of Thessaloniki, Greece

***Abstract*— A first attempt using a new strategy for measuring similarity between proteins is introduced. Our approach has its roots in computational linguistics and the related techniques for quantifying and comparing content in strings of characters. The pairwise comparison of proteins relies on the content regularities expected to uniquely characterize each sequence. These regularities are captured by n-gram based modelling techniques and in the sequel are contrasted by cross-entropy related measures. In this attempt to fuse theoretical ideas from language modeling within the field of bioinformatics, we experimented with different implementations having as ultimate goal the development of practical, computationally efficient algorithms.**

I. Introduction

The large volume of genomic and proteomic databases requires the use of tools that search for similarity between sequences. Finding similar sequences is usually the first step in predicting the possible function of a query protein. Several methods have been developed for protein sequence similarity. Most of them are either different adaptations and improvements of pairwise-alignment based methods such as the Needleman-Wunsch [2] and Smith-Waterman [3] algorithms or they are based on principles of Hidden Markov Models [4], [5]. According to Liao and Noble [6], the development of powerful methods for detecting protein similarity can be grouped into four generations. The early methods are represented by the pairwise similarities between proteins, the second by profiles and Hidden Markov Models [7], the third is characterized by the use of information stored in large databases improving results over the profile-based methods [8], [9] and, in the fourth generation, additional accuracy was gained by modeling the difference between positive and negative alignment examples [6]. Despite the maturity of the developed methodologies working towards this direction, the derivation of protein similarity measures is still an active research area proven by works like [10], [11], [12], [13]. The interest is actually renewed due to the continuous growth in size of the widely available databases that calls for alternative cost-effective algorithmic procedures that can reliably quantify protein similarity without resorting to any kind of alignment. Apart from efficiency, a second specification of equal importance for the establishment of similarity measures is the avoidance of parameters that need to be set by the user (a characteristic inherent in the majority of previous methodologies). In this work, we introduce a new method for measuring similarity based on a Markov chains representation known as n-gram in statistical language modeling. A cross entropy estimation procedure was adopted from information theory in order to compute the similarity between the resulting n-grams. The new algorithm was used in the task of identifying protein sequences resulting from gene mutations of a query sequence within a general database of protein sequences.

II. Method

Protein sequences from different organisms can be seen as texts written in a language where the alphabet is composed of the 20 amino acid symbols. The mapping of protein sequences to their structure, dynamics and function then becomes analogous to the mapping of words to their semantic meaning in natural languages. This analogy is exploited here by the application of a statistical language modelling technique, where the conditional probability of observing a word w_n at position n is assumed to be restricted to its immediate m predecessor words $w_{n-m}...w_{n-1}$. The resulting model is that of a Markov chain and is referred as (m+1)-*gram model*. For our data set we generated 2-gram, 3-gram and 4-gram linguistic models for each protein sequence. The models and the frequency information associated to each unique event represent the input for the protein comparison procedure. A protein sequence like WASQVSENRP will take the form: "WA AS SQ QV VS SE EN NR RP" in 2-gram representation, while "WAS ASQ SQV QVS VSE SEN ENR NRP" in 3-gram representation, etc. To compare these linguistic models, we adopted the principles described in [1], where the cross entropy between a random variable X with the probability distribution $p(X)$ and another probability function estimated from a model called q is given by :

$$H(X,q) = \sum_X p(X)\, log\, q(X). \tag{1}$$

Using our notations we consider the maximum likelihood estimates (MLE) from relative frequencies for $p(X)$ as

$$P_{MLE}(x_1...x_n) = \frac{C(x_1...x_n)}{N} \tag{2}$$

and for q(X) the true conditional probability expressed by

$$P_{MLE}(X_n|x_1...x_{n-1}) = \frac{C(x_1...x_n)}{C(x_1...x_{n-1})}, \tag{3}$$

where $C(x_1...x_n)$ is the frequency of the n-gram $(x_1...x_n)$, and N the number of instances considered. In this context, the cross entropy [1] becomes:

$$H(X,P_M) = -\sum_{all\ X_n} P(X_n)\, log\, P_M(X_n|x_1x_2...x_{n-1}), \tag{4}$$

where X_n ranges over all n-grams, $P(X_n)$ is the relative frequency estimate from the query sequence, $P_M(X_n|x_1x_2...x_{n-1})$ is the conditional probability estimated from the model M and $x_1x_2...x_{n-1}$ is the history of the n-gram X_n [1].

III. Experimental analysis

In order to evaluate the proposed algorithm for protein similarity, a data set of 100 protein sequences, consisting of 50 sequences obtained from mutations of the tumor suppressor (p53) gene and 50 unknown random sequences from NCBI [14] genetic bank, were used. The mutated proteins were selected also randomly from the database we created using the descriptions provided by International Agency for Research on Cancer (IARC) [15] Lyon, France. Having the data organized in the form of n-grams, the proposed method estimates the similarity between two proteins as distance from the perfect score and the one obtained in estimation procedure over their model distribution. Applying the cross entropy concept, the expression given in equation (4) was used to get the score over the amino acid n-gram models for one pair of proteins at a time. The similarity measure was evaluated in two different ways: a) considering the whole length of each protein so that P is estimated from the query and P_M from the compared protein sequence; b) estimating measure P from the shorter sequence and P_M from the longer one.This approach is motivated by the sensitivity of estimating the proteins content to the sequences length . The comparison between proteins is extended to a multi sequence comparison in the sense that each protein is compared with the rest from the experimental set. A different experiment was also performed to examine the dependence of the algorithm on the value of n. We evaluated the proposed similarity measure using ROC curves. Comparing different values of n (2,3,4), we found that more accurate representation was achieved using 4-grams. It indicates that the value of n is a significant factor so that a bigger value is preferred. The accuracy of the results depends also on the compared sequences length with the first strategy of unchanged lengths (see a) above) performing better on sequences that have the same or similar length and the second one (b)) performing better when they are unequal. In the specific case of 4-gram representation and following strategy b), protein sequences were successfully identified, by our method, in 99% of all mutations sequences at a false positive rate of 0.7% (see TABLE I).

n-gram	Thresholds	Strategy a)		Strategy b)	
		FPR	TPR	FPR	TPR
2-gram	0.005	0.009	0.298	0.009	0.416
	0.1	0.106	0.698	0.118	0.849
	1.46	0.900	1.000		
	1.74			0.891	1.000
3-gram	0.005	0.007	0.476	0.018	0.512
	0.1	0.256	0.701	0.250	0.834
	0.38	0.885	1.000	0.895	1.000
4-gram	0.005	0.005	0.527	0.005	0.612
	0.1	0.005	0.894	0.007	0.999
	0.25	0.175	1.000		

TABLE I

False positive rates and true positive rates for different n-grams using both evaluation strategies.

IV. DISCUSSION

An advantage of this new method over alignment-based method is that it does not require the selection of any tuning parameters. Moreover, it does not depend on the choice of any score matrix or on the consideration of gaps. Here, the objective of the experiments was to detect only the similar proteins which in this case were the mutated versions of p53 sequences. For this particular application, the proposed method has the advantage that it can work equally well for substitution mutation types (insertions/deletions) whereas alignment-based methods would fail since the whole sequence is shifted. For future work we will examine other metrics of n-grams and compare them with alignment-based methods. The preliminary results are encouraging for further development of our method.

V. ACKNOWLEDGMENTS

This work was supported by the EU project Biopattern: Computational Intelligence for biopattern analysis in Support of eHealthcare, Network of Excellence Project No. 508803.

References

[1] Manning,C.D., and Schutze,H., "Foundations of statistical natural language processing", Massachusetts Institute of Technology Press, 2000, Cambridge, Massachusetts London, England, pp:554 - 556;557 - 588.
[2] Neddelman,S. and Wunsch,C., "A general method applicable to the search for similarities in the amino-acid sequence of two proteins", *J.Mol.Biol.*, vol. 48, pp.443-453, 1970.
[3] Smith,T., and Watermann,M., "Identification of common molecular subsequences", *J. Mol. Biol.*, vol.147, pp.195-197, 1981.
[4] Baldi,P., Chauvin,Y., Hunkapiller,T., and McClure,M.A., "Hidden Markov models of biological primary sequence information", *Proc. Natl.Acad.Sci.USA*, vol.91(3), pp.1059-1036, 1994.
[5] Krogh,A., Brown,M., Mian,I., Sjolander,K., and Haussler,D., "Hidden Markov models in computational biology: Application to protein modeling", *J. Mol. Biol.*, vol.235, pp.1501-1531, 1994.
[6] Liao,L., and Noble,W.S., "Combining pairwise sequence similarity and support vector machines for remote protein homology detection", *J. Comp. Biol.*, vol.10, pp.857-868, 2003.
[7] Gribskov,M., and Robinson,N.L., "The use of receiver operating characteristic (ROC) analysis to evaluate sequence matching", *Comp.and Chemistry*", vol. 20(1), pp.25-33, 1996.
[8] Altschul,S.F., Madden,T.L., Schaffer,A.A., Zhang,J., Zhang,Z., Miller,W., and Lipman,D.J., "Gapped Blast and PSI-Blast: A new generation of protein database search programs", *Nucl.Ac.Res.*, vol.25(17), pp.3389-3402, 1997.
[9] Sequence Alignment and Modeling System http://www.cse.ucsc.edu/research/compbio/sam.html
[10] Katti,M.V., Sami-Subbu,R., Ranjekar,P.K., and Gupta,V.S., "Amino acid repeat patterns in protein sequences: Their diversity and structural-functional implications.", *Protein Sci.*, vol.9, pp.1203-1209, 2000.
[11] Eskin,E., Grundy,W.N., and Singer,Y., " Differential distribution of simple sequence repeats in eukaryotic genome sesquences", *J.Bioinformatics*, vol.17(1), pp.65-73, 2001.
[12] Saigo,H., Vert,J-P., Ueda,N., Akutsu,T., "Protein homology detection using string alignment kernels Bioinformatics", *J. Bioinformatics*, vol.20(11), pp.1682-1689, 2004.
[13] Krasnogor,N., and Pelta,D.A., "Measuring the Similarity of Protein Structures by Means of the Universal Similarity Metric", *J. Bioinformatics*, 20(7), pp.1015-1021, 2004.
[14] National Center for Biotechnology Information (http://www.ncbi.nlm.nih.gov/)
[15] International Agency for Research on Cancer http://www.iarc.fr/p53/Somatic.html