# A NOVEL EFFICIENT PROTEIN SIMILARITY MEASURE BASED ON *N*-GRAM MODELING

**A. Bogan-Marta** **N. Laskaris** **M. A. Gavrielides** **I. Pitas** **K. Lyroudia**
**Auth, Greece** **Auth, Greece** **Auth, Greece** **Auth, Greece** **Auth, Greece**

## ABSTRACT

A new general 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 very first attempt to fuse theoretical ideas from computational linguistic within the field of bioinformatics, we experimented with different implementations having always as ultimate goal the development of practical, computational efficient algorithms. The experimental analysis provides evidence for the usefulness of the new approach and motivates the further development of linguistics-related tools as a means to decipher the biological sequences.

**Keywords**: protein similarity, n-grams, entropy, cross-entropy, maximum likelihood, exploratory data analysis.

## INTRODUCTION

Proteomics refers to the study of the complete collection of cellular proteins (in the same way as genomics refers to the complete set of genes) and finds a wide application in concurrent bioinformatics. Typical questions that apply to almost all genes are the followings. What protein does each gene produce, when is this protein produced, and which is its functional role? Whereas the genomic sequence can inform us about which proteins the cell has the potential to make, and microarrays expression analysis can provide an approximate answer about which proteins are made, it is only proteomic approaches that provide a concrete picture of the fundamental biochemistry of a cell. Among the most important proteomic approaches are the different comparisons among protein sequences. The necessity for such comparisons emerges from the interest in detecting homologies among the proteins, which may, in turn, imply structural and functional similarities. Proteins are large, complex molecules composed of amino acids and their comparison and clustering according to similarity requires specialized algorithms.

The most frequently used methods for measuring protein similarities are based on tedious algorithmic procedures for sequence alignment. According to Liao and Noble (8), the development of powerful methods for detecting protein similarity can be delimited into four generations, with each one representing a step forward in the evolution of these techniques. The early methods are characterized by pairwise similarities between proteins. Smith-Waterman algorithm (Smith and Watermann (1)) remains the standard reference method due to the accuracy of the obtained results. Other heuristic algorithms, within this first generation of methodologies, like the BLAST (2), FASTA (3) or CLUSTAL (4) provide higher computational efficiency at the expense of accuracy. The second generation is characterized by the computation of profiles for whole protein families (Gribskov and Robinson (5)) based on hidden Markov models (Krogh (6), Baldi (7)). These methodologies allow the computational biologist to infer nearly three times as many homologies as a simple pairwise alignment algorithm (Liao (8)). The algorithms included in the third generation, like PSI–BLAST (Altschul et al (9)) and SAM (10), exploit information stored in large databases and improve the results over the profile-based methods by collecting homologous sequences and incorporating the resulting statistics into a central model. The algorithms in the fourth generation, provide additional accuracy by modeling the difference between positive and negative alignment examples (8).

All the above mentioned methods are built over sequence alignment. Despite the maturity of the developed methodologies working towards this direction, the derivation of protein similarity measures is still an active research area. 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). It is often the case with the classical similarity approaches, that the user is faced with a lot of difficulties in the choice of a suitable search algorithm, scoring matrix or function as well as a set of optional parameters for which optimum values correspond to the best possible similarity.

A variety of new alternative methods has already become available for expressing similarity between biological sequences and for use in different applications. In Sjolander et al (11) Dirichlet mixtures are used, where the incorporated densities are designed to be combined with observed amino-acid frequencies to form estimates of expected amino-acid probabilities

for each position in a HMM profile (or any other statistical model). In Katti (12), a set of proteins from the SWISS-PROT database were selected and analyzed for tandem repeats using a sliding window technique. The authors of Eskin et al (13), found a biological motivation for using a mixture model of common ancestors in order to estimate the probability distribution over discrete alphabets from observations. The obtained model was then used to find amino acids probabilities based on observed counts in an alignment and to estimate probability distributions over protein families. Using the approach of support vector machines (SVMs), the authors in Saigo et al (14) apply discriminative methods and prove that their method is the most effective for the problem of superfamily recognition. Latent semantic analysis (LSA) is another method used in Ganapathiraju et al. (15) to capture secondary structure propensities (tendencies) in protein sequences. Finally, in Krasnogor and Pelta (16), the authors propose the use of the universal similarity metric (USM) for structural similarity between pairs of proteins.

Here a new approach for measuring the similarity between two protein sequences is introduced. It is inspired by the successful use of entropy concept for information retrieval in the field of statistical language modeling (Young and Bloothoof (17), Manning and Schütze (18), Jurafsky and Martin (19)). Specifically, *n*-gram modeling is first applied to each protein sequence and cross-entropy measures are then employed to compare pairs of proteins. Since the presented work was actually the first attempt to adopt this dual step for comparing biological sequences, some experimentation was necessary in order to discover the most effective way in which it could be applied in the specific application domain. The final proposal includes detailed algorithmic procedures for implementing the above principles when moderate-sized biological strings (with elements from the restricted vocabulary of 20 aminoacids) are to be compared. Using actual data, from publicly available databases, we validate the suggested similarity measure and show that it provides a very effective way to capture the common characteristics of the compared sequences, while avoiding the annoying task of choosing parameters, additional functions or evaluation methods. This high performance and the ready-to-plug-in character, taken together with the obvious computational efficiency, constitute our approach a promising alternative.

The rest of the paper is structured as follows. An introduction to the employed theoretical concepts is followed by a discussion of the implementation aspects of our proposal. In the sequel we briefly describe the utilized protein data and present the results obtained from the application of two variants of our method. At the end we are concluding with the main aspects of this new approach for measuring protein similarity, while discussing the possible improvements that have to be attempted in a future work.

## METHODS

### Theoretical Background

There are various kinds of language models that can be used to capture different aspects of regularities of natural language (Wang et al (20)). Markov chains are generally considered among the more fundamental concepts for building language models. In this approach the dependency of the conditional probability of observing a word $w_k$ at a position k in a given text is assumed to be depended only upon its immediate *n* predecessor words $w_{k-n} \ldots w_{k-1}$. The resulting stochastic models, usually referred as ***n-grams***, constitute heuristic approaches for building language grammars and their linguistic justification has often been questioned in the past. However, in practice they have turned out to be extremely powerful. Nowadays *n*-gram modeling stands out as superior to any formal linguistic approach (Van Compernolle (21)) and has gained high popularity due to its simplicity.

Closely related with the design of models for textual data are algorithmic procedures for validating them. Apart from the justification of a single model, they can facilitate the selection of the specific one (among competing alternatives) most faithfully representing the available data. ***Entropy*** is a key concept for this kind of procedures. In general, its estimation is considered to provide a quantification of the information in a text and has strong connections to probabilistic language modeling (Durbinn et al (22)). It can also be utilized for expressing how much information is reflected by a particular grammar, how well a given grammar matches a language, how large is the predictive power of a grammar, etc. While relying on the same theoretical principles, the estimation of entropic-measures in the domain of language processing requires some modifications (dictated by the discrete nature of data) with respect to the procedures established in the field of statistics.

As described in (19) and (Brown et al (23)), the entropy of a random variable X that ranges over a domain ℵ, and has a probability density function, P(X) is defined as:

$$H(X) = -\sum_{X \in \aleph} P(X) \log P(X). \qquad (1)$$

The ***cross-entropy*** between the actual probability distribution P(X) (over a random variable X) and the probability distribution Q(X) estimated from a model is defined as:

$$H(X, Q) = -\sum_{X \in \aleph} P(X) \log Q(X) \qquad (2)$$

Two important (for the development of our approach) propositions should be mentioned here. First, the cross entropy of a stochastic process, measured by using a model, is an upper bound on the entropy of the process (i.e. $H(X) \leq H(X,Q)$) (18), (23)). Second, as mentioned in (19), between two given models, the more accurate is

the one with the lower cross-entropy.
Recently, in Van Uytsel and Comparnolle (24), the general idea of entropy has been adopted in the specific case that a written sequence W={$w_1$,$w_2$,…, $w_{k-1}$,$w_k$, $w_{k+1}$,…} is treated as an n-gram based composition and resulted in the following estimating formula

$$H(X) = -\sum_{W^*} p\left(w_i^n\right) \log_2 p\left(w_{i+n} \middle| w_i^{n-1}\right) =$$
$$= -\frac{1}{N} \sum_{W^*} Count\left(w_i^n\right) \log_2 p\left(w_{i+n} \middle| w_i^{n-1}\right) \qquad (3)$$

where the variable X has the form of an n-gram $w_i^n$={ $w_i$,$w_{i+1}$,…,$w_{i+n-1}$} , the summation runs over all the possible n-length combinations of consecutive $w_i$ (i.e W*={{$w_1$,$w_2$,…,$w_n$},{$w_2$,$w_3$,….,$w_{n+1}$},….}) and N is the total number of *n*-grams in the investigated sequence. The second term in the summation 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), it can be estimated by, simply, encountering a counting procedure and expressing the corresponding relative frequencies:

$$P\left(w_{i+n} \middle| w_i^{n-1}\right) = \frac{Count\left(w_{i+n}\right)}{Count\left(w_i^{n-1}\right)} \quad (4)$$

The above entropic estimation (taken together with the general form of eq.1&2 suggesting a direct way to pass from entropy to cross-entropy formulation) was the basis for building our protein similarity measure, which is described in the sequel.

## The *n*-gram Based Protein Similarity Measure

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 analogous to the mapping of words to their semantic meaning in natural languages. Recently (Biological Language Conference, 2003), it was suggested that 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 have become optimist about the identification of Gramar/Syntax rules that could reveal systematics of high importance for biological and medical sciences.
In the presented approach, we adopted a Markov-chain grammar and built for our protein dataset *2*-gram, *3*-gram and *4*-gram models for each protein sequence. To clarify things let a protein sequence WASQVSENR. In the 2-gram modeling the available ‘’words’’ are {WA AS SQ QV VS SE EN NR}, while in the 3-gram representation the words are {WAS ASQ SQV QVS VSE SEN ENR}. Based on the frequencies of these words (estimated by counting) and by forming the appropriate ratios of frequencies, the entropy of a n-gram model can be readily estimated (eq.(3)). This measure is indicative about how well-predicted is a specific protein-sequence by the corresponding 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 couldn’t facilitate the direct comparison of the two proteins (and help us to decide if they are similar or not).
The previous shortcoming led us to devise the corresponding cross-entropy measure, in which the *n*-gram model is, first, built based on the word-counts of one protein sequence (*training-step)* and then the predictability, of the second sequence, by the model is measured (*projection-step*) as a means of contrasting the two proteins. So the common information content is expressed via the formula

$$H\left(X, P_M\right) = -\sum_{all\ w_i^n} P(w_i^n) \log P_M\left(w_{i+n} \mid w_i^{n-1}\right) \quad (5)$$

The first term in the above summation refers to the reference protein sequence (i.e. it results from counting the words of that specific protein). The second term refers to the sequence based on which the model has to be estimated (i.e. it results from counting the words of that protein). Variable X ranges over all the words (that are represented by *n*-grams) of the reference protein sequence.

## Database Searches with the New Similarity Measure

Having introduced the new similarity measure, we proceed here with the description of its use for performing searches within protein databases. The crux of our approach is that both the unknown query-protein (e.g. a newly discovered protein) and each protein 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. By recognizing the most similar proteins within the database, the structure and function of the query-protein can be inferred based on the *principle of assimilation.* In the algorithmic implementation of these ideas, and as a byproduct of the experimentation with actual data, we devised two different ways in which the *n*-gram based cross entropy similarity is engaged in efficient database searches. The most straightforward implementation gave rise to an algorithm called hereafter as *direct method.* A second algorithm, the *alternating method*, was devised in order to cope with fact that the proteins to be compared might be of very different length.

*Direct method.* Let $S_q$ the sequence of a query-protein and {S}={$S_1$, $S_2$, … $S_N$} the given protein database. The

first step is the computation of 'perfect' score (PS) or 'reference' score for the query-protein. This is computed from eq.(5) by using both as reference and model sequence the query-protein (actually the obtained result corresponds to the entropy of the query protein). In the second step, each protein -in turn- from the database serves as the model sequence in the computation of a similarity score using the eq.(5) with the query-protein serving as the reference sequence. In this way, N similarities are computed $H(Xq,P_{Mi})$, i=1,..,N. Finally these similarities are compared against the perfect score PS. By computing the absolute differences $D(S_q,S_i)=|H(Xq,P_{Mi})-PS|$, 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 proteins most resembling the query-protein as those proteins which have been assigned the lowest $D(S_q,S_i)$.

*Alternating method.* The only difference with respect to the direct method is that when comparing the query-protein with each database-protein, the role of reference protein and model protein can be interchanged based on which of the two sequences is the shortest (the shortest sequence playing the role of reference sequence in eq.(5)). The other steps (perfect score estimation, ranking, etc) follow as previously.

## EXPERIMENTS

### Protein Sequence Database

The proposed strategy for measuring protein similarity was demonstrated and validated using a database containing an overall sample of 100 protein sequences. 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 (25). 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 (26). 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.

### Results

First, we followed some classical steps of *Exploratory Data Analysis* in order to validate the two variants of the proposed strategy. In Fig.1 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. It is therefore clearly evident from Fig.1 that *4*-gram modeling followed by 'alternating'-version of our algorithm has an almost excellent performance when searching within the given database.

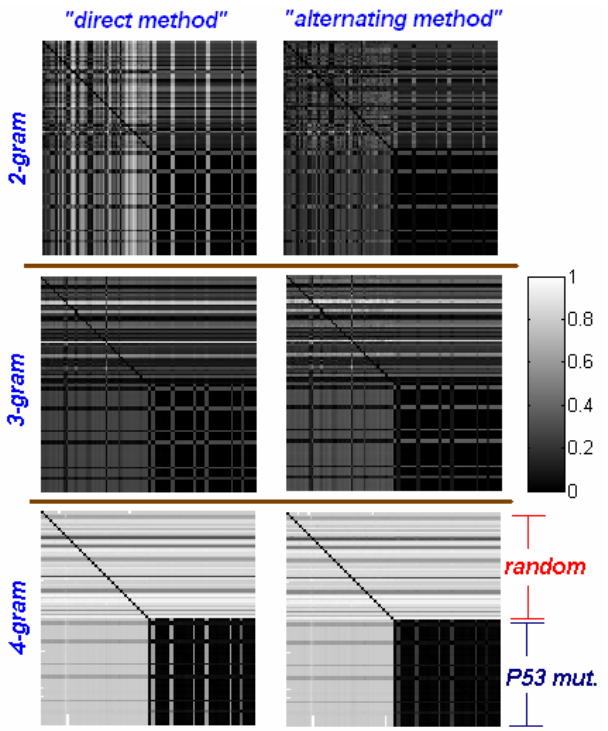


**Fig. 1 Visualization of the matrices containing all the possible pairwise dissimilarities for the 100 proteins in out database.**

Second, in order to provide quantitative measures of performance for the two variants, we adopted an index of search accuracy, that is derived from receiver operating characteristic (ROC) curves and has recently gained popularity when validating protein-databases searches (Liao and Noble (8), Schäffer et al (27)). This index, usually referred as truncated ROC-score, is the ratio of the area under the ROC-curve (in the plot of true- positives versus false positives for different thresholds of dissimilarity). More explicitly, as mentioned in (27), for a number *T* of true positives available to be found and a fixed number of false positives *n*, this index is the proportion of the rectangle [0, *T* ] [ 0, *n* ] that lies under the sensitivity curve. It takes values in the range [0-1], with one corresponding to the highest performance. This ROC-score has been tabulated in Table 1 for different n-grams and both methods.

**Table. 1 The ROC-score**

| ***n*-gram model** | **Normailzed area under ROC curve** | |
|---|---|---|
| | ***Direct method*** | ***Alternating method*** |
| 2-gram | 0.589 | 0.680 |
| 3-gram | 0.723 | 0.817 |
| 4-gram | 0.900 | 0.978 |

## DISCUSSION

The method introduced in this paper represents a first step investigating the engagement of language modelling in characterizing, handling and understanding biological data in the format of sequences. We specifically studied the use of cross-entropy measure applied over n-gram models as a means of searching in protein database in an effective and efficient way. The experimental results indicated the reliability of our algorithmic strategy for expressing similarity between proteins. Given the conceptual simplicity of the introduced approach, it appears as an appealing alternative to previous well-established techniques.

Considering the general dichotomy between ''global'' and ''local'' protein similarity measures, we should mention that our approach belongs to the former category. During the evaluating of our method we observed that from the two introduced variants the better performance is associated with the second one. This means that it is important to come up with improvements that overcome the possible wrong identifications of similar sequences due to the decisively big differences between sequences length. In the exceptional case when all the compared sequences have the same length the *direct method* is equivalent with the *alternating method* and performs excellent.
Regarding the order of the employed n-gram model, after testing with order of 2,3,4,5 we noticed that the performance of the method increases with the order of the model up to 4. After the order of 5 due to lack of data the corresponding maximum likelihood estimates becomes unreasonable uniform and very low. This sets an upper limit for our model order in the specific database (perhaps slightly higher order model could work in different protein databases.

Before continuing the work on the improvement of this method we have to remark that this is a statistical in nature technique. It can be improved by incorporating biological knowledge (e.g. working with functional groups of amino-acids).
Finally, another aspect that deserves further consideration is to test if our method scales well with the size of the protein database.


## ACKNOWLEDGMENT

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. Smith, T., and Watermann, M., 1981, J. Mol. Biol.147, 195-197.
2. Basic Local Alignment Search Tool http://www.ncbi.nlm.nih.gov/BLAST/
3. FAST-All, or fast protein/nucleotide comparison http://www.ebi.ac.uk/fasta33/
4. http://www.ebi.ac.uk/clustalw/
5. Gribskov, M., and Robinson, N.L. 1996, Comp.and Chemistry 20(1), 25-33
6. Krogh, A., Brown, M., Mian, I.,Sjolander, K.,and Haussler,D.,1994, J. Mol. Biol.235, 1501-1531.
7. Baldi, P.,Chauvin, Y.,Hunkapiller, T., and McClure, M.A., 1994, Proc. Natl.Acad.Sci.USA 91(3), 1059-1036
8. Liao, L., and Noble,W.S., 2003, J. Comput. Biology,10, 857-868
9. Altschul, S.F., Madden, T.L., Schaffer, A.A., Zhang J. Zhang, Z., Miller and Lipman D. J., 1997, Nucleic Acids Research, 25(17) , 3389-3402.
10. Sequence Alignment and Modeling System http://www.cse.ucsc.edu/research/compbio/sam.html
11. Sjolander, K. Karplus, M. Brown, R. Hughey, A. Krogh, I.S. Mian, D. Haussler, 1996, J. Bioinformatics,12, 327-345
12. Katti, M.V., Sami-Subbu, R., Ranjekar, P.K., and V.S. Gupta, V.S., 2000, *Protein Sci.* 9, 1203-1209
13. Eskin, E., Grundy, W.N., and Singer, Y., 2001, J.Bioinformatics,17 (1), 65-73
14. Saigo, H., Vert, J-P., Ueda, N., Akutsu**,** T.,2004, *J.* Bioinformatics, 20(11), 1682-1689
15. Ganapathiraju, M., Klein-Seetharaman, J., Rosenfeld, R., Carbonell, J., and Reddy, R., 2002, RECOMB

16. N. Krasnogor, N., and Pelta, D.A., 2004, Bioinf. Adv. Access,20, 1015-1021

17. Young, S., and Bloothooft, G., 2001, Kluwer Academic Publishers,2, 174-179

18. Manning, C.D., and Schütze, H., 2000, "Foundations of statistical natural language processing", Massachusetts Institute of Technology Press, Cambridge, Massachusetts London, England,.554 – 556;557 – 588.

19. Jurafsky, D, and Martin, J, 2000, "Speech and Language Processing", Prentice Hall, Upper Saddle River, New Jersey

20. Wang, S., Schuurmans, D., Peng, F., and Zhao, Y., ICASSP-03, icassp03.ps.gz; http://citeseer.nj.nec.com/575237.html

21. Van Compernolle, D., 2003, "Spoken Language Science and Technology", http://www.esat.kuleuven.ac.be/~compi/pub/spoken_language/TOC.htm

22. Durbinn, R., Eddy, S., Crogh, A., Mitchison, G., 1998, "Biological sequence analysis. Probabilistic models of proteins and nucleic acids", *Cambridge University Press*

23. Brown, P., F., Della Pietra, S., A., Della Pietra, V., J., Mercer, R.L.L., and Lai, J., C., 1992, Assoc.for Comput. Linguistics, Yorktown Heights, NY 10598, P.O. Box 704

24. Van Uytsel D.H., and Van Compernolle, D., 1998, IEEE Benelux Signal Proc. Symp., 227-230.

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

26. http://www.iarc.fr/p53/Somatic.html

27. Schäffer, A., , Aravind, L., Madden, L., Shavirin, S., Spouge, J., Wolf, Y., Koonin, E., Altschul, S., 2001, Nucleic Acids Research, 29, (14), 2994-3005