International Journal of Biological, Life and Agricultural Sciences
ISSN: 2415-6612
Vol:2, No:8, 2008

SAF: A Substitution and Alignment Free Similarity Measure for Protein Sequences

Abdellali Kelil, Shengrui Wang, and Ryszard Brzezinski

***Abstract*—**The literature reports a large number of approaches for measuring the similarity between protein sequences. Most of these approaches estimate this similarity using alignment-based techniques that do not necessarily yield biologically plausible results, for two reasons.

First, for the case of non-alignable (i.e., not yet definitively aligned and biologically approved) sequences such as multi-domain, circular permutation and tandem repeat protein sequences, alignment-based approaches do not succeed in producing biologically plausible results. This is due to the nature of the alignment, which is based on the matching of subsequences in equivalent positions, while non-alignable proteins often have similar and conserved domains in non-equivalent positions.

Second, the alignment-based approaches lead to similarity measures that depend heavily on the parameters set by the user for the alignment (e.g., gap penalties and substitution matrices). For easily alignable protein sequences, it's possible to supply a suitable combination of input parameters that allows such an approach to yield biologically plausible results. However, for difficult-to-align protein sequences, supplying different combinations of input parameters yields different results. Such variable results create ambiguities and complicate the similarity measurement task.

To overcome these drawbacks, this paper describes a novel and effective approach for measuring the similarity between protein sequences, called SAF for Substitution and Alignment Free. Without resorting either to the alignment of protein sequences or to substitution relations between amino acids, SAF is able to efficiently detect the significant subsequences that best represent the intrinsic properties of protein sequences, those underlying the chronological dependencies of structural features and biochemical activities of protein sequences. Moreover, by using a new efficient subsequence matching scheme, SAF more efficiently handles protein sequences that contain similar structural features with significant meaning in chronologically non-equivalent positions. To show the effectiveness of SAF, extensive experiments were performed on protein datasets from different databases, and the results were compared with those obtained by several mainstream algorithms.

***Keywords*—**Protein, Similarity, Substitution, Alignment.

Abdellali Kelil is with ProspectUS Laboratory, Faculty of Sciences, Department of Computer Sciences at the University of Sherbrooke, J1K 2R1 Canada (corresponding author - phone: 819-571-2945; fax: 819-821-8200; e-mail: Abdellali.Kelil@USherbrooke.ca).

Shengrui Wang is with ProspectUS Laboratory, Faculty of Sciences, Department of Computer Sciences at the University of Sherbrooke, J1K 2R1 Canada (e-mail: Shengrui.Wang@USherbrooke.ca).

Ryszard Brzezinski is with Laboratory of Molecular Biotechnology, Faculty of Sciences, Department of Biology at the University of Sherbrooke, J1K 2R1 Canada (e-mail: Ryszard. Brzezinski@USherbrooke.ca).

I. INTRODUCTION

THE literature reports a large number of approaches developed for measuring the similarity between protein sequences. Prominent among these are alignment-based approaches, which, for a pair of protein sequences, find the best matching by inserting "*gaps*" in the appropriate positions, so that the positions from both sequences with identical or similar amino acids are aligned. The alignment-based approaches have a major drawback due to the fact that they are based on the matching of subsequences in chronological order. These approaches breakdown when applied to protein sequences comprising similar structural features (i.e., subsequences characterizing the intrinsic sequential nature of related protein sequences) that do not occur in the same chronological order, such as multi-domain, circular permutation, and tandem repeat proteins. In fact, protein sequences often have similar and conserved domains in non-equivalent positions when viewed in terms of primary structure, which makes them difficult to align and match in chronological order. However, these domains might well be in equivalent positions when viewed in terms of three-dimensional structure. Moreover, these approaches yield similarity measures that depend heavily on the "*substitution matrix*" used as well as the costs assigned by the user to the "*opening gap*" and the "*extension gap*". This creates ambiguities and complicates the similarity measurement task, especially for sequences of significantly different length, and even more so when it comes to hard-to-align protein sequences.

The literature also reports another type of approach that does not rely on alignment (for a review see [1]). Most of these approaches map protein sequences to vectors, for which Linear Algebra and Statistical Theory had useful analytical tools already available. These produced vectors are defined by the frequencies of the N-grams within the corresponding protein sequences. The N-grams are the set of all possible subsequences of a fixed length N. However, the N-grams approach has a major drawback, because the value of the fixed length N for collecting the subsequences from the protein sequences is set independently of the intrinsic structure of the sequences, such as their length and the distribution of the amino acids within them. Depending on the value of N, this results in either the collection of subsequences constituting noise or the exclusion of significant subsequences. Moreover, all subsequences of length N are collected without distinguishing between significant and non-significant subsequences, which increases the probability of collecting a number of noise motifs.

In the aim of overcoming the drawbacks cited above, a novel and original approach for measuring the similarity between protein sequences named SAF is proposed.

Without resorting either to the alignment of protein sequences or to substitution relations between amino acids, SAF allows us to extract hidden relations between protein sequences, by capturing structural and chronological dependencies using global information extracted from a large number of sequences rather than merely comparing pairs of sequences. SAF detects and makes use of the significant subsequences underlying the chronological dependencies of the structural features that can reveal biochemical properties shared between protein sequences, by filtering out noise through the collection of the significant patterns (i.e., subsequences) that best represent the intrinsic structural properties of protein sequences and by discarding those patterns that occur by chance and merely noise.

In addition, SAF allows us to measure similarity in a way that more adequately reflects the structural and biochemical relationships between the protein sequences, and yields a linear worst-case computational cost with respect to sequence length. Moreover, by taking advantage of an efficient subsequence matching scheme, SAF simultaneously addresses the "within" chronological order and the "between" non-chronological order of the structural features. This makes it possible to handle protein sequences containing similar structural features with significant meaning in non-equivalent positions, such as multi-domain, circular permutation, and tandem repeat protein sequences.

SAF constitutes an effective method for measuring the similarity between protein sequences. To show this, extensive experiments on different types of proteins from different databases were performed. Furthermore, the obtained results were compared with those obtained by different mainstream approaches. Experiments on these types of sequences have shown that the patterns used in SAF are more significant in terms of representing the biochemical properties of protein sequences.

II. SAF Overview

By applying a new pairwise sequence matching scheme, FAS extracts from a set of protein sequences a set of patterns with significant meaning, and filters out noise patterns. This is done by looking at each pair of sequences for shared identical patterns, as well as those that are slightly different, known as "*Paronyms*" and "*Cognates*". In natural language text, paronyms such as "*affect*" and "*effect*" are words that are related and derive from the same root, while cognates such as "*shirt*" and "*skirt*" are words that have a common origin. Taking into account identical patterns, paronyms and cognates makes it possible to improve the extraction of significant patterns.

Following the extraction of significant patterns, the *N*-grams algorithm is applied on the set of collected patterns obtained from the pairwise sequence matching, instead of on the original input protein sequences. Then, by performing spectral decomposition, the sequences are mapped onto a new vector space of reduced dimension, in which each sequence is represented by a vector. Finally, the similarity between different sequences is computed by applying the cosine distance between the corresponding vectors. The development of this idea is shown in the next sections.

III. The Main Idea of FAS

Very often, in natural language text processing [2], methods such as Latent Semantic Analysis are used to extract hidden relations between documents by capturing semantic relations, using global information extracted from a large number of documents, rather than merely comparing pairs of documents. These methods usually make use of a word-document matrix $T(W \times L)$, in which rows correspond to words and columns correspond to documents, where W is the number of possible words and L is the number of documents. The term $T_{i,j}$ represents the occurrence of word i in document j. Although protein sequences do not contain distinctive patterns like words in natural language text, protein sequence analysis is in many respects similar to natural language text analysis. However, the challenge is to be able to identify those patterns that map to a specific meaning in terms of sequence structure and distinguish significant patterns from patterns resulting from random phenomena.

Similar to the use of a word-document matrix to extract the hidden relations between documents in natural language text, a pattern-sequence matrix is used on protein sequences to extract the hidden relations between these sequences. This will be done by capturing structural relations, using global information extracted from a large number of sequences, rather than merely comparing pairs of sequences. Henceforth, the pattern-sequence matrix $T(W \times L)$ is used, in which the term $T_{i,j}$ represents the frequency of pattern i in sequence j, while W is the number of possible patterns, and L is the number of sequences. The significant patterns used to construct T are detected and collected using the matching approach described in the next section.

IV. Significant Patterns

In the work described here, a significant pattern is obtained from the matching of a pair of sequences. Let F be a set of protein sequences, from which X and Y are a pair of sequences. Let x and y be a pair of subsequences belonging respectively to X and Y. Here, the symbols x and y are simply used as variables: they represent any subsequence belonging to the sequences X and Y, respectively.

Now, the set of significant patterns are detected and collected by building a matching set $E_{X,Y}$. This is performed by collecting all the possible pairs of subsequences x and y that satisfy the following conditions:

$$E_{X,Y} = \left\{ x, y \;\middle|\; \begin{array}{c} |x| = |y| \\ |x \cap y| \geq N_{X,Y} \\ |x \setminus y| \leq N_{X,Y} \\ \forall x', y' \in E_{X,Y} \Rightarrow (x \not\subset x') \vee (y \not\subset y') \end{array} \right\} \quad (1)$$

The symbols x' and y' in (1) are simply used as variables, in the same way as x and y. The expression $(.\not\subset.)$ means that the element to the left of the symbol $\not\subset$ is not included in the one to the right, either in terms of the composition of the patterns or in terms of their respective positions in their sequence. The parameter $N_{X,Y}$ is used to represent the minimum number of matched positions with similar amino acids between x and y, at the same time, $N_{X,Y}$ is also used to represent the maximum number of matched positions with different amino acids allowed. A detailed discussion on the choice of $N_{X,Y}$ is provided in the next section. Here are a few explanations about the previous formula:

1. $|x| = |y|$: means that x and y have the same length.
2. $|x \cap y| \geq N_{X,Y}$: means that x and y include at least $N_{X,Y}$ matched positions with similar amino acids.
3. $|x \setminus y| \leq N_{X,Y}$: means that x and y include at most $N_{X,Y}$ matched positions with different amino acids.
4. $\forall x', y' \in E_{X,Y} \Rightarrow (x \not\subset x') \vee (y \not\subset y')$: means that, for any pair of matched subsequences x' and y' belonging to $E_{X,Y}$, at least one of x or y is not included in x' or y', respectively, either in terms of their compositions or in terms of their respective positions in their corresponding sequences according to the partial order induced by set inclusion.

By looking for similar patterns in X and Y, the aim of the matching set $E_{X,Y}$ is to capture shared information between X and Y related to their intrinsic structural features that manifest certain chronological dependencies. At the same time, by taking into account multiple occurrences of patterns in non-equivalent positions, the matching set $E_{X,Y}$ seeks to capture the structural features in non-chronological order. In fact, with this formula, $E_{X,Y}$ captures pairs of patterns x and y that show a "*within*" chronological similarity, even if they are in non-chronological order according to their respective positions within the sequences X and Y. The choice of the length $N_{X,Y}$ is described in the next section.

V. Length of Significant Patterns

Our aim is to detect and make use of the significant patterns best representing the natural structure of protein sequences and to minimize the influence of those patterns that occur by chance and represent only noise. This motivates one of the major statistical features of our similarity measure, the inclusion of all long similar patterns (i.e., multiple occurrences) in the matching, since it is well known that the longer the patterns, the smaller the chance of their being identical by chance, and vice versa. For each pair of compared sequences X and Y, the statistical theory developed by Karlin *et al.* [3] is used. This very useful theory makes possible calculating the expected length of the longest common pattern (i.e., subsequence) present by chance at least a number of times out of a set of sequences made up of a given number of categories (i.e., m-letters alphabet).

This theory is used in this paper to calculate the minimum length of matched significant patterns, which is the value to be assigned to $N_{X,Y}$.

According to the theorem 1 developed by Karlin *et al.* [3], the expected length $K_{X,Y}$ of the longest common pattern present by chance at least 2 times out of 2 m-letters X and Y (i.e., here m=20), is calculated as follows:

$$K_{X,Y} = \frac{log(|X|^2 + |Y|^2) + log\,\lambda_{X,Y}(1 - \lambda_{X,Y}) + 0.57}{-\,log\,\lambda_{X,Y}} \quad (2)$$

$$\lambda_{X,Y} = max\left(\sum_{i=1}^{m} (p_i^X)^2, \sum_{i=1}^{m} (p_i^Y)^2\right) \quad (3)$$

$$\sigma_{X,Y} \approx \frac{1.28}{|log\,\lambda_{X,Y}|} \quad (4)$$

Here, p_i^X and p_i^Y are generally the i^{th} amino acid frequency of the observed X and Y sequences respectively, while $\sigma_{X,Y}$ is the asymptotic standard deviation of $K_{X,Y}$.

According to the conservative criterion proposed by Karlin *et al.* [3], for a pair of sequences X and Y, a pattern observed 2 times is designated statistically significant if it has a length that exceeds $K_{X,Y}$ by at least two standard deviations. Thus, in building the matching set $E_{X,Y}$, all the common patterns that satisfy this criterion are extracted. This means that, for the pair of sequences X and Y, a specific and appropriate value of $N_{X,Y}$ is calculated such that $N_{X,Y} = K_{X,Y} + 2\sigma_{X,Y}$. This criterion guarantees that a matched pattern designated as statistically significant has probability less than a 1/100 probability of occurring by chance.

VI. The Pattern Sequence Matrix

Let F be a set of protein sequences, among which X and Y are two different sequences for which $N_{X,Y}$ is the minimum length of the significant patterns, and $E_{X,Y}$ is the set of collected pairs of significant patterns. Let E be the set of all possible matching sets, such that:

$$E = \bigcup_{X,Y \subset F} E_{X,Y} \quad (5)$$

And

$$N_{min} = \min_{X,Y \subset F} N_{X,Y} \quad (6)$$

Now, to compute the pattern-sequence matrix T, all the N_{min}-grams from each significant pattern included in E are collected. Thus, for a set of sequences made up of m possible amino acids, a maximum of $m^{N_{min}}$ possible N_{min}-grams (i.e., m=20 amino acids) could be obtained.

Let E_X be the subset of all possible matching sets involving the protein sequence X, such that:

$$E_X = \bigcup_{Y \subset F} E_{X,Y} \quad (7)$$

The value of the term $T_{i,X}$ (initially set to zero) representing the intersection of row i^{th} with the column corresponding to the sequence X, is simply augmented by the occurrence of the i^{th} collected N_{min}-grams belonging to the subset E_X.

After building the matrix T, all the rows corresponding to N_{min}-grams that exist at most in only one sequence are removed. In our experiments, the number of remaining rows W is found to be much smaller than $m^{N_{min}}$ (i.e., $\ll m^{N_{min}}$). This property is very important for the next section.

The most important advantage of this new sophisticated approach is that each member of the set of protein sequences contributes to the capture of structural features and chronological dependencies of all other sequences in the set. And the more often a pattern is present in the sequences, the more heavily it is represented in the pattern-sequence matrix T. Moreover, the matrix T is filled by using only the N_{min}-grams corresponding to the collected significant patterns, not all the N-grams as in the classical approach.

VII. Spectral Decomposition

In the pattern-sequence matrix T, each sequence is expressed as a column vector and each pattern as a row vector. This representation is known as a vector space model. The sequences represented in this way are seen as points in the multidimensional space spanned by patterns. However, this representation does not recognize related patterns or sequences and the dimensions are too large [4]. To take advantage of this representation, the theorem of spectral decomposition in linear algebra will be utilized, which states that any $W \times L$ matrix T with total rank R, whose number of rows W is greater than or equal to its number of columns L can be written as the product of an $R \times L$ column orthogonal matrix U, an $L \times L$ diagonal matrix Σ with non-negative elements, which are the singular values, and the transpose of an $L \times L$ row orthogonal matrix V. This decomposition is named the singular value decomposition (SVD). The matrix T can be written as follows:

$$T = U \times \Sigma \times V^T \tag{8}$$

VIII. Similarity Measure

According to the singular value decomposition theory [2], the sequences expressed as column vectors in the matrix T are projected via the spectral decomposition onto a new multidimensional space of reduced dimension L spanned by the column vectors of the matrix V^T. The representation of the sequences in the new L-dimension space corresponds to the column vectors of the $L \times L$ matrix $\Sigma \times V^T$. Now, the similarity measure $S_{X,Y}$ between the pair of sequences X and Y, is simply computed by using the cosine product of their corresponding column vectors in the matrix $\Sigma \times V^T$.

IX. Time Complexity

At the stage of collecting the significant patterns, the fast string matching approach developed by Amir *et al.* [5] is used, which allows us to find all the locations of any pattern from a protein sequence X in a protein sequence Y in time $O\left(|Y|\sqrt{N_{X,Y}\log N_{X,Y}}\right)$.

For the singular value decomposition, the fast, incremental, low-memory and large-matrix SVD algorithm recently developed by Brand [6] is used, which allows performance of the SVD for a R rank matrix $W \times L$, the SVD can be performed in $O(WLR)$ time with $R \leq \sqrt{min(W,L)}$.

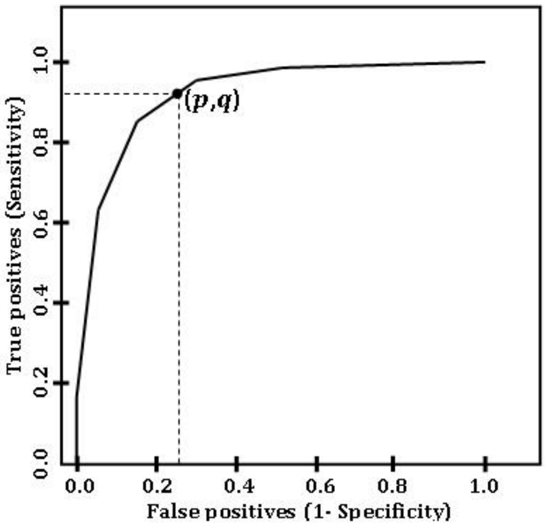

Fig. 1 ROC Curve

X. Experiments

To evaluate the performance of our new similarity measure approach on both easy-to-align and hard-to-align protein sequences, SAF was tested on a variety of protein datasets from different databases, in the aim to assess its discrimination power between proteins with different biochemical activities. Furthermore, the obtained results were compared with those obtained by various mainstream algorithms from two types of approaches.

The first type are the alignment-based approaches, for instance the widely used algorithms BLAST [7], which detects isolated regions of similarity by making use of high-scoring segment pairs; FASTA [8], which employs exact matches as seeds, known as k-tuples, which are used to build local alignments to capture the similarities; and CLUSTAL [9], which calculates the global best match between protein sequences and lines them up so that the similarities can be captured. Also the recent algorithm Scoredist introduced by Sonnhammer *et al.* [10] was used in our experiments, which makes use a logarithmic correction of observed divergence based on the alignment score according to substitution scores.

The second type are the alignment-free approaches; for instance SMS, introduced by Kelil *et al.* [11], based on a strict matching scheme that captures the most significant patterns in chronological and non-chronological order; tSMS introduced by Kelil *et al.* [12]; which is an improved version of SMS that allows mismatches; and the approach introduced by Wu *et al.* [13], based on short patterns used analogously to the index terms in information retrieval; and the one introduced by Bogan-Marta *et al.* [14], based on the cross-entropy measure applied over the collected N-grams patterns with a fixed length value N.

To evaluate the quality of the results obtained, the well-known Receiver Operating Characteristic method was used, known also as the ROC Curve. This method allows us to evaluate the discriminative power of each of the similarity measure approaches studied in our experiments. The ROC

TABLE I
SIMILARITY MEASURE QUALITY ON COG DATABASE

Set			Alignment-based				Alignment-free			
		SAF	*BLAST*	*FASTA*	*CLUSTAL*	*Scoredist*	*tSMS*	*SMS*	*Wu*	*Bogan*
C_1	*Q*	0.96	0.70	0.72	0.91	0.93	**0.97**	0.93	0.78	0.84
	S	0.01	0.06	0.11	0.06	0.03	**0.01**	0.05	0.07	0.04
C_2	*Q*	0.95	0.61	0.67	0.89	0.94	**0.96**	0.95	0.84	0.88
	S	**0.01**	0.06	0.22	0.03	0.04	0.02	0.04	0.13	0.09
C_3	*Q*	0.91	0.77	0.78	0.87	0.88	**0.98**	0.95	0.88	0.82
	S	0.02	0.02	0.05	0.03	0.10	**0.01**	0.03	0.06	0.04
C_4	*Q*	0.93	0.74	0.73	0.85	0.87	**0.98**	0.89	0.77	0.82
	S	0.04	0.05	0.13	0.12	0.03	**0.01**	0.06	0.04	0.04
C_5	*Q*	0.92	0.60	0.68	0.90	0.95	**0.95**	0.93	0.81	0.84
	S	0.05	0.16	0.22	0.04	**0.01**	0.03	0.05	0.02	0.07
C_6	*Q*	0.94	0.68	0.75	0.92	0.94	**0.97**	0.95	0.77	0.86
	S	0.04	0.28	0.09	0.02	0.05	**0.01**	0.03	0.04	0.03
Av.	*Q*	0.94	0.68	0.72	0.89	0.92	**0.97**	0.93	0.81	0.84
	S	0.03	0.11	0.14	0.05	0.04	**0.02**	0.04	0.06	0.05

TABLE II
SIMILARITY MEASURE QUALITY ON KOG DATABASE

Set			Alignment-based				Alignment-free			
		SAF	*BLAST*	*FASTA*	*CLUSTAL*	*Scoredist*	*tSMS*	*SMS*	*Wu*	*Bogan*
K_1	*Q*	0.91	0.65	0.69	0.85	0.88	**0.92**	0.91	0.68	0.66
	S	0.04	0.26	0.17	0.06	0.09	**0.02**	0.07	0.16	0.17
K_2	*Q*	0.91	0.55	0.61	0.88	0.90	**0.94**	0.91	0.67	0.71
	S	0.04	0.19	0.23	0.02	0.04	**0.02**	0.08	0.11	0.12
K_3	*Q*	0.92	0.58	0.67	0.91	0.92	**0.96**	0.93	0.74	0.69
	S	**0.01**	0.29	0.12	0.05	0.08	0.04	0.04	0.12	0.16
K_4	*Q*	0.86	0.54	0.63	0.79	0.80	**0.92**	0.86	0.62	0.61
	S	0.06	0.41	0.16	**0.04**	0.08	0.05	0.04	0.21	0.13
K_5	*Q*	0.88	0.70	0.71	0.72	0.76	**0.94**	0.84	0.68	0.71
	S	0.10	0.10	0.11	**0.01**	0.08	0.03	0.11	0.13	0.07
K_6	*Q*	0.88	0.75	0.76	0.74	0.79	**0.91**	0.84	0.58	0.69
	S	0.10	0.03	0.20	0.10	**0.03**	0.04	0.10	0.13	0.08
Av.	*Q*	0.89	0.63	0.68	0.82	0.84	**0.93**	0.88	0.66	0.68
	S	0.06	0.21	0.17	0.05	0.07	**0.03**	0.07	0.14	0.12

TABLE III
SIMILARITY MEASURE QUALITY ON PC DATABASE

Set			Alignment-based				Alignment-free			
		SAF	*BLAST*	*FASTA*	*CLUSTAL*	*Scoredist*	*tSMS*	*SMS*	*Wu*	*Bogan*
P_1	*Q*	0.94	0.78	0.78	0.89	0.92	**0.96**	0.93	0.81	0.76
	S	0.04	0.14	0.04	0.10	0.06	**0.02**	0.02	0.09	0.07
P_2	*Q*	0.95	0.76	0.81	0.84	0.89	**0.98**	0.92	0.90	0.79
	S	0.02	0.11	0.12	0.03	0.05	**0.01**	0.03	0.02	0.12
P_3	*Q*	0.93	0.62	0.65	0.88	0.91	**0.95**	0.94	0.68	0.83
	S	0.03	0.16	0.10	0.05	**0.02**	0.04	0.05	0.03	0.11
P_4	*Q*	0.94	0.79	0.80	0.81	0.87	**0.95**	0.91	0.80	0.80
	S	0.05	0.14	0.06	0.14	0.07	0.05	0.04	**0.03**	0.06
P_5	*Q*	0.93	0.73	0.77	0.83	0.85	**0.95**	0.92	0.79	0.78
	S	0.04	0.13	0.12	0.09	0.12	**0.01**	0.03	0.10	0.17
P_6	*Q*	0.91	0.80	0.81	0.90	0.94	**0.98**	0.94	0.87	0.93
	S	0.02	0.18	0.15	0.05	0.02	**0.01**	**0.01**	0.13	0.03
Av.	*Q*	0.93	0.75	0.77	0.86	0.90	**0.96**	0.93	0.81	0.82
	S	0.03	0.14	0.10	0.08	0.06	**0.02**	0.03	0.07	0.09

Curve makes it possible to quantify the Quality Index of the similarity measures obtained between a sequence X and all the sequences in a set F, by making use of the known classification of X in F. Below, a brief description of how the Quality Index is computed.

After sorting the sequences belonging to F according to the decreasing order of their similarities with the sequence X, and by considering the subset of sequences belonging to F that have the same biochemical activity of X as "*true positives*", and the remaining sequence as "*false positives*", the ROC Curve can be represented by plotting the fraction of true positives rate *vs.* the fraction of false positives rate. A plotted point in this curve with the coordinates (p,q) (i.e., see Fig. 1) means that the subset of sequences from the sorted set F that includes the first p percent of true positives, includes also q percent of false positives. The best possible similarity measures of X with the sequences in F, would yield a point near the upper left corner of the ROC space, representing 100% sensitivity, corresponding to p=1.0 (i.e., all sequences from the same class of X have the highest similarity measures) and 100% specificity, corresponding to q=0.0 (i.e., all sequences from different classes of X have the lowest similarity measures). In our experiments the value of the area under the ROC curve is defined as the Quality Index of the similarity measures obtained with a given protein sequence X, since the larger this area is, the greater the discriminative power of the similarity measure.

A. *Easy-to-Align Protein Sequences*

To illustrate the effectiveness of our new approach in measuring the similarity between easy-to-align protein sequences according to their functional annotations and biological classifications, extensive tests were performed on the widely known databases including the well-aligned protein sequences COG [15], KOG [15] and PC [16]. The six randomly generated subsets [12] from each database were used: C_1 to C_6 from the COG database, containing 509, 448, 546, 355, 508, and 509 protein sequences, respectively; K_1 to K_6 from the KOG database, containing 317, 419, 383, 458, 480, and 388 sequences; and P_1 to P_6 from the PC database, containing 538, 392, 442, 595, 561, and 427 sequences. Each generated subset includes non-orphan protein sequences (i.e., each sequence has at least one sequence from the same biochemical activity) with at least 20 biochemical activities.

To evaluate our new similarity measure approach efficiently, all-against-all similarity measures of the protein sequences within each generated subset were computed. Then, the Quality Index for each protein sequence was evaluated. Finally, the "*Mean*" and the "*Standard Deviation*" of all the Quality Indexes obtained for each generated subset were computed. Below, the results obtained for the different generated subsets are reported with support from the literature and functional annotations.

In TABLE I, TABLE II, and TABLE III the results obtained by each algorithm on each protein subset are summarize. Each table shows the Quality Index average (i.e., "*Q*" row) and the Standard Deviation (i.e., "*S*" row) obtained by each approach (i.e., column) for each subset of protein sequences (i.e., row). The last row in each table contains the Quality Index average and the Standard Deviation obtained by each approach with all the generated subsets.

The results, illustrated in TABLE I, TABLE II, and TABLE III, show that tSMS obtained the best Quality Indexes on all generated subsets. The results with tSMS are closely followed by those obtained by SAF and SMS, and a bit farther behind those obtained by Scoredist and CLUSTAL followed by those of the approaches developed by Wu *et al.* [13] and Bogan-Marta *et al.* [14], while BLAST and FASTA obtained the weakest results. These results warrant further comments.

First, among alignment-based approaches, Scoredist and CLUSTAL obtained better Quality Indexes (slightly better with Scoredist, since in our experiments the alignment generated by CLUSTAL for each test is used as the input alignment for the Scoredist algorithm, which allowed Scoredist to improve on the results already obtained by CLUSTAL). FASTA and BLAST obtain less significant Quality Indexes, since both of them are approximate and simplified alignment approaches, which allow them to run much faster than a conventional alignment-based algorithm at the cost of some sensitivity.

Second, among the alignment-free approaches, tSMS, FAS, and SMS obtained better results over all generated subsets, with a small relative advantage for tSMS. We believe strongly this is due to the fact that, apart from the approach proposed in this paper, tSMS and SMS are the only algorithms among those used here that significantly address the non-chronological order of structural features of protein sequences. However, tSMS and SMS need a substitution matrix as input parameter, to decide which amino acids should be matched and compute the weights of the significant patterns. In our experiments, the results obtained by tSMS and SMS were made possible by the use of the substitution matrix that maximizes the quality measure for each test. This means that one needs prior knowledge about the classes of the protein sequences in order to choose the appropriate matrix for tSMS and SMS. This is the very reason why SAF is proposed in this paper: SAF does not depend on the use of a substitution matrix or any other input parameter.

B. Hard-to-Align Protein Sequences

To show the performance of our new similarity measure with multi-domain protein sequences which are known to be hard-to-align and have not yet been definitively aligned, experimental tests were performed on the 33 $(\alpha/\beta)_8$-barrel proteins studied recently by Côté *et al.* [17] and Fukamizo *et al.* [18], which form a group in "*Glycoside Hydrolases*" family 2 (GH2) from the "*Carbohydrate Active Enzymes*" database (CAZy) [19]. The periodic character of the catalytic module known as "*$(\alpha/\beta)_8$-barrel*" makes these sequences hard to align using classical alignment approaches. The difficulties in aligning these modules are comparable to the problems encountered with the alignment of tandem repeats. Another reason for the difficulty of aligning these proteins is that these sequences are multi-modular, with various types of modules. The problems encountered with aligning tandem repeat and multi-modular protein sequences have been exhaustively discussed by Higgins [20]. This group of 33 protein sequences includes '*β-galactosidase*', '*β-mannosidase*', '*β-glucuronidase*' and '*exo-β-D-glucosaminidase*' enzymatic activities, all extensively studied at the biochemical level. The database names and entries of the 33 $(\alpha/\beta)_8$-barrel group are indicated in [21].

To be able to evaluate efficiently our new similarity measure approach, all-against-all similarity measures of the 33 $(\alpha/\beta)_8$-barrel proteins were computed, then the Quality Index for each protein sequence were evaluated. Below, the results obtained for the different protein sequences are reported with support from the literature and functional annotations.

The TABLE IV shows the Quality Index obtained by each algorithm (i.e., column) for each protein sequence (i.e., row). The last row contains the Quality Index average obtained by each approach on the 33 proteins group. SAF and tSMS obtained the best Quality Indexes over all protein sequences. For all protein sequences, they obtained 100% sensitivity, meaning that they assigned all proteins from the same biochemical classification the highest similarity measures, and 100% specificity, which means they assigned all proteins from different biochemical classifications the lowest similarity measures. The other approaches tested obtained less significant Quality Indexes. The results warrant further comments.

As in the previous experiment, BLAST and FASTA obtained the less significant Quality Indexes. However, unlike the previous experiment, Scoredist and CLUSTAL obtained relatively close results comparable with those obtained by the alignment-free approaches developed by Wu *et al.* [13] and Bogan-Marta *et al.* [14], with a relatively small advantage to the latter. This can be explained by the fact that, in this experiment, the $(\alpha/\beta)_8$-barrel protein group used as benchmark contains

TABLE IV
SIMILARITY MEASURE QUALITY ON 33$(\alpha/\beta)_8$ BARREL GROUP

		Alignment-based				*Alignment-free*			
Protein	*SAF*	*BLAST*	*FASTA*	*CLUSTAL*	*Scoredist*	*tSMS*	*SMS*	*Wu*	*Bogan*
UnA	***1.00***	*0.67*	*0.81*	*0.95*	*0.97*	***1.00***	*0.93*	*0.92*	*0.99*
UnBv	***1.00***	*0.65*	*0.79*	*0.83*	*0.90*	***1.00***	*0.94*	*0.89*	*0.97*
UnBc	***1.00***	*0.57*	*0.76*	*0.98*	*0.98*	***1.00***	*0.94*	*0.98*	*1.00*
UnBm	***1.00***	*0.54*	*0.71*	*0.89*	*0.91*	***1.00***	*0.96*	*0.90*	*0.97*
UnBp	***1.00***	*0.56*	*0.74*	*0.84*	*0.90*	***1.00***	*0.97*	*0.98*	*0.96*
UnR	***1.00***	*0.63*	*0.63*	*0.98*	*0.98*	***1.00***	*0.98*	*0.97*	*0.91*
MaA	***1.00***	*0.55*	*0.60*	*0.92*	*0.96*	***1.00***	*0.91*	*0.99*	*1.00*
MaB	***1.00***	*0.77*	*0.75*	*0.89*	*0.95*	***1.00***	*0.83*	*0.94*	*1.00*
MaH	***1.00***	*0.63*	*0.64*	*0.90*	*0.94*	***1.00***	*0.84*	*0.92*	*0.99*
MaM	***1.00***	*0.65*	*0.66*	*0.93*	*0.95*	***1.00***	***1.00***	*0.96*	*0.93*
MaC	***1.00***	*0.77*	*0.81*	*0.87*	*0.90*	***1.00***	***1.00***	*0.97*	*0.95*
MaT	***1.00***	*0.66*	*0.61*	*0.98*	*0.98*	***1.00***	***1.00***	*0.95*	*0.98*
GlC	***1.00***	*0.76*	*0.71*	*0.98*	*0.98*	***1.00***	***1.00***	*0.97*	*0.95*
GlE	***1.00***	*0.63*	*0.77*	*0.99*	*0.99*	***1.00***	***1.00***	*0.92*	*0.93*
GlH	***1.00***	*0.64*	*0.81*	*0.92*	*0.93*	***1.00***	***1.00***	*0.95*	*1.00*
GlL	***1.00***	*0.78*	*0.71*	*0.97*	*0.99*	***1.00***	***1.00***	*0.90*	*1.00*
GlM	***1.00***	*0.72*	*0.70*	*0.96*	*0.99*	***1.00***	***1.00***	*0.92*	*0.99*
GlF	***1.00***	*0.62*	*0.74*	*0.90*	*0.97*	***1.00***	***1.00***	*0.92*	*0.96*
GlS	***1.00***	*0.61*	*0.74*	*0.96*	*0.99*	***1.00***	***1.00***	*0.93*	*0.91*
GaEco	***1.00***	*0.66*	*0.70*	*0.96*	*0.96*	***1.00***	***1.00***	*0.97*	*0.99*
GaA	***1.00***	*0.77*	*0.78*	*0.99*	***1.00***	***1.00***	***1.00***	*0.94*	*0.90*
GaK	***1.00***	*0.56*	*0.63*	*0.94*	*0.97*	***1.00***	***1.00***	*0.96*	*0.95*
GaC	***1.00***	*0.55*	*0.76*	*0.91*	*0.96*	***1.00***	***1.00***	*0.89*	*0.92*
GaEcl	***1.00***	*0.78*	*0.59*	*0.96*	*0.99*	***1.00***	***1.00***	*0.97*	*0.92*
GaL	***1.00***	*0.63*	*0.80*	*0.92*	*0.99*	***1.00***	***1.00***	*0.97*	*0.94*
CsAo	***1.00***	*0.55*	*0.62*	*0.99*	*0.99*	***1.00***	***1.00***	*0.91*	*0.93*
CsS	***1.00***	*0.58*	*0.81*	*0.90*	*0.92*	***1.00***	***1.00***	*0.94*	*0.97*
CsG	***1.00***	*0.62*	*0.80*	*0.93*	*0.96*	***1.00***	***1.00***	*0.93*	*0.99*
CsM	***1.00***	*0.70*	*0.80*	*0.94*	*0.97*	***1.00***	***1.00***	*0.92*	*0.96*
CsN	***1.00***	*0.68*	*0.72*	*0.86*	*0.87*	***1.00***	***1.00***	*0.92*	*0.99*
CsAn	***1.00***	*0.76*	*0.64*	*0.93*	*0.98*	***1.00***	***1.00***	*0.96*	*0.95*
CsH	***1.00***	*0.67*	*0.72*	*0.95*	*0.96*	***1.00***	***1.00***	*0.96*	*0.99*
CsE	***1.00***	*0.76*	*0.77*	*0.94*	*0.99*	***1.00***	***1.00***	*0.90*	*0.91*
Av.	***1.00***	*0.66*	*0.72*	*0.93*	*0.95*	***1.00***	***0.97***	*0.94*	*0.96*

TABLE V
SIMILARITY MEASURE QUALITY OF CIRCULAR PERMUTED PROTEINS

		Alignment based				Alignment free			
Protein	*SAF*	*BLAST*	*FASTA*	*CLUSTAL*	*Scordist*	*tSMS*	*SMS*	*Wu*	*Bogan*
H_PLADU	***0.96***	*0.48*	*0.48*	*0.69*	*0.62*	***0.96***	*0.90*	*0.83*	*0.76*
H_BOVIN	***0.98***	*0.56*	*0.51*	*0.54*	*0.59*	*0.97*	*0.93*	*0.85*	*0.72*
L_BOWMI	*0.95*	*0.52*	*0.55*	*0.58*	*0.51*	***0.97***	*0.95*	*0.79*	*0.80*
L_DIOGR	*0.97*	*0.59*	*0.50*	*0.51*	*0.57*	***0.99***	*0.94*	*0.81*	*0.79*
G_THEFU	***0.99***	*0.43*	*0.57*	*0.63*	*0.55*	***0.95***	*0.93*	*0.87*	*0.81*
G_CELFI	*0.96*	*0.40*	*0.54*	*0.50*	*0.54*	***0.98***	*0.95*	*0.83*	*0.83*
Av.	***0.97***	*0.50*	*0.53*	*0.58*	*0.56*	***0.97***	*0.93*	*0.83*	*0.79*

sequences for which the alignment causes difficulties for classical alignment-based approaches such as CLUSTAL. This shows the clear advantage of the alignment-free approaches compared to alignment-based approaches.

The comment in the first experiment about the use of a substitution matrix by tSMS and SMS also applies for this experiment.

C. Circular Permutation Protein Sequences

In a protein's structure, the positions of certain amino acids in the primary structure could be rearranged such that the N terminal and C terminal regions are swapped. The three-dimensional structure remains almost unaffected by the permutation, and the native structure and biological function are usually retained. Protein sequences that have been subjected to this type of transformation may escape detection from biochemical activity prediction algorithms based on sequence alignment alone. Moreover, alignment-based approaches for measuring the similarity of protein sequences breakdown when handling sequences of this type. For more details see [22].

In this experiment, our new approach were evaluated on selected three pairs of well-characterized protein sequences previously identified as circular permuted: H1B_PLADU and H11_BOVIN, known as *Histones*; LEC_BOWMI and LECA_DIOGR, known as *Lectins*; and GUN2_THEFU and GUNA_CELFI, known as *β-Glucanases*. For more details see [23].

To evaluate the different approaches efficiently, the similarity measure between each of the selected circular permuted protein sequences were computed with all the 6 863 213 proteins included in "*nr database*" version of august 9th 2008, the non-redundant protein database maintained by NCBI [24] as a target for BLAST search services. Before performing this test, of course the selected protein sequences were verified if they really exist in this database. The quality of the results is evaluated using the method described below.

Let X and Y be a pair of circular permuted protein sequences with similar biochemical activities. The quality of the similarity measures Q_X^Y obtained between the sequence X and all the sequences in the nr database, in terms of Y, is defined as follows:

$$Q_X^Y = \frac{N_{Z=X,Y}^Y}{N_{Z=X,Y}^Y + N_{Z\neq X,Y}^Y} \tag{9}$$

With,

$$N_{Z=X,Y}^Y = \left|\forall Z \in nr \middle| S_{X,Z} > S_{X,Y};\ Z = X,Y\right| \tag{10}$$

$$N_{Z\neq X,Y}^Y = \left|\forall Z \in nr \middle| S_{X,Z} > S_{X,Y};\ Z \neq X,Y\right| \tag{11}$$

Here are a few explanations about the previous formulas:

1. $Z = X,Y$: Means that the sequence Z has similar biochemical activity of the sequences X and Y.
2. $Z \neq X,Y$: Means that the sequence Z has different biochemical activity of the sequences X and Y.
3. $S_{i,j}$: Defines merely the similarity measure between the sequences i and j.
4. $N_{Z=X,Y}^Y$: Defines the number of sequences from the nr database that have the same biochemical activities of X and Y, and more similar to X than is Y.
5. $N_{Z\neq X,Y}^Y$: Defines the number of sequences from the nr database that have different biochemical activities of X and Y, and more similar to X than is Y.

In the computing of the Q_X^Y only well-characterized protein sequences were considered in the previous formulas.

This quality measure aims to assess the discrimination power of the similarity measure by looking for the sequences that are more similar to X than is Y, and have a different biochemical activities of X and Y. In other words, Less the number of these sequences larger the quality of the similarity is, and vice-versa.

In TABLE V, the quality of the similarity measures obtained by each algorithm tested (i.e., column) on each selected protein sequence (i.e., row) is shown. The last row contains the quality average obtained by each approach. The table shows that SAF and tSMS obtained the best quality results on all protein sequences, closely followed by SMS, and a bit farther behind the approaches developed by Wu *et al.* [13] and Bogan-Marta *et al.* [14]. The alignment-based approaches obtained less significant results. These results also warrant further comments.

First, unlike the previous experiment, CLUSTAL obtained less significant results than the alignment-free approaches tested. Here, the results obtained by CLUSTAL are relatively comparable to those obtained by FASTA and BLAST, while Scoredist does not succeed in improving on the results obtained by CLUSTAL. We believe this is due to the circular permuted nature of the sequences used in this experiment. These sequences have similar and conserved domains in non equivalent positions when viewed in terms of primary structure, which makes them difficult to align and match in chronological order. However, these domains might well be in equivalent positions when viewed in terms of three-dimensional structure.

Second, in the previous experiments, the approaches developed by Wu *et al.* [13] and Bogan-Marta *et al.* [14] obtain results poorer than or, at best, equivalent to those obtained by CLUSTAL. However, in this experiment, these approaches obtained significantly better results than CLUSTAL, FASTA, and BLAST. This is due to the efficiency of alignment-free approaches and their advantage over existing mainstream alignment-based approaches for measuring the similarity between hard-to-align protein sequences.

Third, the comment about the use of a substitution matrix by tSMS and SMS again applies for this experiment.

D. Biochemical Activity Prediction of Protein Sequences

In this experimentation, our new similarity approach is used to predict biochemical activities of two sets of selected protein sequences, obtained from the NCBI website [24]. The first set includes well characterized proteins, all extensively studied at the biochemical level. The second set includes none yet characterized proteins that we aim to predict the biochemical activities. The database entries and the corresponding organisms of the selected protein sequences are indicated in the TABLE V.

To be able to predict the biochemical activities of the selected protein sequences, our new approach is used to measure the similarity between each of these sequence with all the 6 863 213 protein sequences included in the nr database. Then, the most significantly similar sequences, whose the choice is discussed below, are selected and used as input dataset for the alignment free clustering algorithm CLUSS [25], developed by Kelil *et al.* [11], given that it proves its accuracy to highlight the biochemical activities of proteins than do the alignment based algorithms, especially for sequences that are hard to align. Therefore, a biochemical activity can be attributed with high confidence to the uncharacterized protein sequence, if a well-characterized protein within the same cluster is already known. Below are provided more details about the systematic technique used to select the most significant similar sequences for each selected sequence.

First, the sequences from the nr database are sorted in decreasing order of their similarities with the sequence to be predicted. Second, the maximum interclass inertia is computed, based on the Koenig-Huygens theorem, which gives the relationship between the total inertia and the inertia of each group relative to the centre of gravity. In this case, merely two groups are concerned, the high similarity group and the low similarity group. The procedure is described as follows:

Let R be the uncharacterized protein sequence to be predicted. And, let F be the set of obtained similarity measures between the sequence R and all the sequences from the nr database, with F_L the subset of low similarity measures, and F_H the subset of high similarity measures, such that:

$$F_L \cup F_H = F \quad and \quad F_L \cap F_H = \emptyset \tag{12}$$

$$\forall X, Y \in F \middle| X \in F_L, Y \in F_H \Longrightarrow S_{R,X} < S_{R,Y} \tag{13}$$

Where, $S_{R,X}$ and $S_{R,Y}$ are the similarity measures obtained between the sequences R with X and R with Y, respectively. The symbols F_L and F_H are simply used as variables representing all possible separations of F according to previous Equations (12) and (13). By making use of the Koenig-Huygens theorem, the total inertia I is calculated as follows:

$$I = \sum_{i \in F_L} \left(S_{R,i} - \bar{S}_{F_L}\right)^2 + \sum_{j \in F_H} \left(S_{R,j} - \bar{S}_{F_H}\right)^2 + \left(\bar{S}_{F_L} - \bar{S}_{F_H}\right)^2 \tag{14}$$

Where, $S_{R,i}$ and $S_{R,j}$ the obtained similarity measures of sequences R with i and R with j, such that i and j are belonging to the subsets F_L and F_H, all respectively; and $\bar{S}_{F_L}$ and $\bar{S}_{F_H}$ are the means (i.e., centers of gravity) of subsets F_L and F_H, respectively. The best separation of F is the subsets F_L and F_H that maximize the value of the total inertia I in the previous Equation (14). Then, the most significant similar sequences to be used as input data for the clustering process, is the subset of protein sequences corresponding to the subset F_H maximizing I the total inertia.

In TABLE VI, are shown the predicted biochemical activities of the selected protein sequences. For the set of well characterized sequences, the clustering has predicted exactly the adequate biochemical cluster of each protein. For the set of protein sequences with unknown biochemical activities, the clustering has classified each uncharacterized sequence within the same cluster of an already well characterized protein, which the activity is assigned to the uncharacterized protein sequence. Please see TABLE VI.

XI. Discussion

The excellent results obtained in this paper on different protein datasets and databases clearly demonstrate the efficiency of our new approach and its advantage over existing mainstream approaches are shown, both alignment-based and alignment-free, for measuring the similarity between protein sequences. Our experiments show that the new measure efficiently extracts the significant hidden information behind the biochemical activities of protein sequences, without resorting either to the alignment of protein sequences or to substitution relations between amino acids. This also constitutes an important advantage compared to alignment-free approaches that need a substitution matrix as input parameter, such as tSMS and SMS.

Our new approach makes it possible to detect more efficiently the significant patterns that best represent the intrinsic properties of protein sequences, those underlying the chronological dependencies of structural features that can reveal biochemical activities of protein sequences. Moreover, by using a new efficient subsequence matching scheme, our approach more effectively handles protein sequences that contain similar structural features with significant meaning in chronologically non-equivalent positions.

So far, the performance and the effectiveness of our new approach were shown on different types of protein sequences such as those from the COG, KOG, and PC databases, the group of multi domain 33 $(\alpha/\beta)_8$ proteins, and also on different circular permutation protein sequences. Furthermore, the prediction of biochemical activities of several and different well-characterized as well as non-characterized protein sequences was performed. This prediction was done using a new and systematic technique that revealed hidden relations between proteins for which the alignment based approaches have not been able to detect.

In future work, the study and the analysis of our new similarity measure will be deepen by further experimenting on hard-to-align protein sequences, such as tandem repeat and circular permutation protein sequences. More evidence on the ability of our new similarity measure to capture and make use of important structural features as well as the information hidden in the chronological and non-chronological order of the protein sequences will also attempt to be discovered. The significant patterns detected by SAF will be compared with those biochemically identified as conserved domains, involved in biochemical activities of proteins, with support from literature and functional annotations. The study and the analysis by further testing of the systematic technique presented in this

TABLE VI
BIOCHEMICAL ACTIVITIES PREDICTION OF THE SELECTED PROTEIN SEQUENCES

	Protein	*Organism*	*Known Activity*	*Predicted activity*
Protein Sequences with Known Activities	*AAA24053*	Bacteria	*β-Galactosidase*	*β-Galactosidase*
	AAA69907	Bacteria		
	AAA35265	Eukaryota		
	AAA23216	Bacteria		
	BAA07673	Bacteria		
	AAK06078	Bacteria		
	AAC48809	Eukaryota	*β-Glucuronidase*	*β-Glucuronidase*
	AAC74689	Bacteria		
	AAA52561	Eukaryota		
	AAK07836	Bacteria		
	AAA37696	Eukaryota		
	AAD01498	Eukaryota		
Protein Sequences with Unknown Activities	*AAV32104*	Eukaryota	Unknown	*Ribonucleotide-Diphosphate Reductase*
	XP_960828	Eukaryota	Unknown	
	NP_249831	Bacteria	Unknown	
	ACB94306	Bacteria	Unknown	*FMRFamide*
	XP_001675807	Eukaryota	Unknown	
	YP_869103	Bacteria	Unknown	
	NP_718648	Bacteria	Unknown	
	YP_001473371	Bacteria	Unknown	
	ZP_02158382	Bacteria	Unknown	
	YP_001831795	Bacteria	Unknown	
	ABK18067	Bacteria	Unknown	*ATP-Binding Cassette*
	YP_846502	Bacteria	Unknown	
	XP_001636168	Eukaryota	Unknown	
	YP_908731	Bacteria	Unknown	
	YP_672786	Bacteria	Unknown	
	XP_001632468	Eukaryota	Unknown	
	ABS67555	Bacteria	Unknown	*Neuropeptides Precursor*
	YP_429591	Bacteria	Unknown	
	YP_001417212	Bacteria	Unknown	
	XP_001621143	Eukaryota	Unknown	
	YP_429591	Bacteria	Unknown	
	YP_605034	Bacteria	Unknown	*Methyltransferase Type 12*
	YP_342594	Bacteria	Unknown	
	YP_049838	Bacteria	Unknown	
	ZP_01904033	Bacteria	Unknown	
	ZP_01627072	Bacteria	Unknown	
	ZP_02134734	Bacteria	Unknown	
	YP_206188	Bacteria	Unknown	

paper to predict the biochemical activities of protein sequences will be also deepen. Its web application server will be also provided.

REFERENCES

[1] S. Vinga, and J. Almeida, "Alignment-free sequence comparison – a review," *BIOINFORMATICS*, 4, vol. 19, 2003, pp. 513-523.

[2] M.W. Berry, and R.D. Fierro, "Low-rank orthogonal decompositions for information retrieval applications," *Numerical Linear Algebra with Applications*, 3, vol. 4, 1996, pp. 301-327.

[3] S. Karlin, and G. Ghandour, "Comparative statistics for DNA and protein sequences: Single sequence analysis," *Proc. Natl. Acad. Sci. USA*, 17, vol. 82, 1985, pp. 5800-5804.

[4] M. Ganapathiraju, J. Klein-Seetharaman, N. Balakrishnan, and R. Reddy, "Characterization of protein secondary structure," *Signal Processing Magazine*, IEEE, 3, vol. 21, May 2004, pp. 78-87.

[5] A. Amir, M. Lewenstein, and E. Porat, "Faster algorithms for string matching with k mismatches," *J. Algorithms*, 2, vol. 50, 2004, pp. 257-275.

[6] M. Brand, "Fast low-rank modifications of the thin singular value decomposition," *Linear Algebra and Its Applications*, 1, vol. 415, 2006, pp. 20-30.

[7] http://www.ncbi.nlm.nih.gov/BLAST.

[8] http://www.ebi.ac.uk/fasta33.

[9] http://www.ebi.ac.uk/clustalw.

[10] E. L. L. Sonnhammer, and V. Hollich, "Scoredist: A simple and robust protein sequence distance estimator," *BMC Bioinformatics* 2005, vol. 6 pp. 108.

[11] A. Kelil, S. Wang, and R. Brzezinski, "A new alignment-independent algorithm for clustering protein sequences," in *Proc. 7th IEEE International Conference on BioInformatics and BioEngineering*. Cambridge, Harvard University, Massachusetts, USA, 2007, pp. 27-34.

[12] A. Kelil, S. Wang, and R. Brzezinski, "CLUSS2: An alignment-independent algorithm for clustering protein families with multiple biological functions," *International Journal of Computational Biology and Drug Design*, 2008. (In press).

[13] K.P. Wu, H.N. Lin, T.Y. Sung, and W.L. Hsu, "A New Similarity Measure among Protein Sequences," in *Proc. Computational Systems Bioinformatics*, Stanford, CA, USA, 2003, pp. 347-352.

[14] A. Bogan-Marta, N. Laskaris, M. A. Gavrielides, I. Pitas, K. Lyroudia, "A novel efficient protein similarity measure based on n-gram modeling," in *Proc. 2nd International Conference on Computational Intelligence in Medicine and Healthcare*. Costa da Caparica, Lisbon, Portugal, 2005.

[15] R.L. Tatusov, N.D. Fedorova, J.D. Jackson, A.R. Jacobs, B. Kiryutin, E.V. Koonin, D.M. Krylov, R. Mazumder, S.L. Mekhedov, A.N. Nikolskaya, B.S. Rao, S. Smirnov, A.V. Sverdlov, S. Vasudevan, Y.I. Wolf, J.J. Yin, and D.A. Natale, "The COG database: An updated version includes eukaryotes," *BMC Bioinformatics*, 4, vol. 41, 2003.

[16] K. ONeill, W. Klimke, and T. Tatusova, "Protein clusters: A collection of proteins grouped by sequence similarity and function," *NCBI*, October 04, 2007.

[17] N. Côté, A. Fleury, E. Dumont-Blanchette, T. Fukamizo, M. Mitsutomi, and R. Brzezinski, "Two exo-β-D-glucosaminidases/exochitosanases from actinomycetes define a new subfamily within family 2 of glycoside hydrolases," *Biochem. J.*, vol. 394, 2006, pp. 675–686.

[18] T. Fukamizo, A. Fleury, N. Côté, M. Mitsutomi, and R. Brzezinski, "Exo-β-D-glucosaminidase from Amycolatopsis orientalis: Catalytic residues, sugar recognition specificity, kinetics, and synergism," *Glycobiology*, vol. 16, 2006, pp. 1064-1072.

[19] www.cazy.org.

[20] D. Higgins, "Multiple Alignment," in *The Phylogenetic Handbook*, M. Salemi, and A. M. Vandamme, Ed. Cambridge University Press, 2004, pp. 45-71.

[21] A. Kelil, S. Wang, R. Brzezinski, and F. Alain, "CLUSS: Clustering of protein sequences based on a new similarity measure," *BMC Bioinformatics*, 2007, vol. 8, pp. 286.

[22] W. C. Lo, and P. C. Lyu, "CPSARST: An efficient circular permutation search tool applied to the detection of novel protein structural relationships," *Genome Biology*, vol. 9, 2008.

[23] S. Uliel, A. Fliess, and R. Unger, "Naturally occurring circular permutations in proteins," *Protein Eng.*, vol. 14, 2001, pp. 533-542.

[24] http://www.ncbi.nlm.nih.gov.

[25] http://prospectus.usherbrooke.ca/CLUSS.