A Novel Method for Protein 3D-Structure Similarity Measure Based on N-Gram Modeling

J. Razmara, Safaai B. Deris, *Member, IEEE*

***Abstract*—The present paper describes a novel method for measuring structural similarity of proteins in three dimensions. The method gets its roots from computational linguistics and the related techniques for modeling protein structure in string form and pairwise comparison of protein sequences. The method uses n-gram based modeling techniques for capturing regularities in protein structure sequences and joints cross-entropy measures for comparing two protein sequences to do similarity test. In this way, the 3D- structure of protein is represented in string form and, then, a similarity test is performed over these sequences. To find an overlap between two protein structures in 3D-space, a superposition task is also applied. In order to confirm the validity of this method, some experiments were performed using a collection of the protein data sets on publicly available servers which showed that the method is efficient.**

I. Introduction

It is known that the protein structure specifies its characteristics and functions. Therefore, comparison, retrieval and classification of protein structures are indispensable for various applications in bioinformatics, such as the recognition of a new unknown protein function. Many protein structure comparison, retrieval and classification methods have been proposed that are divided into two main categories; sequence comparison and 3D structure comparison [1]. The former can be considered as a sequence alignment problem of amino acids in the primary structure of the proteins. The latter is structure matching process based on three-dimensional structure of the proteins.

Several approaches to protein structure alignment have been explored over the past decade. The techniques used include comparison of distance matrices (DALI) [2], analysis of differences in vector distance plots [3], minimization of the soap-bubble surface area between two protein backbones [4], dynamic programming on pairwise distances between the proteins' residues [5], [6] and secondary-structure elements (SSEs) [7], [8], combinatorial extension of alignment path (CE) [9], vector alignment of SSEs (VAST) [10], Secondary Structure Matching (SSM) [11], and many others.

Manuscript received June 15, 2008. This work was supported in part by the Malaysian Ministry of Science, Technology, and Innovation (MOSTI) and Universiti Technologi Malaysia Research Management Centre (RMC) under Grant Vote 79228.

J. Razmara is with the Faculty of Computer Science and Information Systems, Universiti Technologi Malaysia, 81310 Skudai, Johor, Malaysia, (e-mail: jrazmara@ ucna.ac.ir).

Safaai. B. Deris is with the Faculty of Computer Science and Information Systems, Universiti Technologi Malaysia, 81310 Skudai, Johor, Malaysia, (e-mail: safaai@utm.my).

Despite the maturity of the proposed methods, the study for designing new similarity measures is still an active research area. Due to the continuous growth of protein databases and discover of new unknown proteins, the interest is renewed for designing alternative effective and reliable algorithms. Furthermore, another motivation of equal importance for establishment of similarity measure is proposition of a method without need to parameter setting by the user. The classical similarity approaches such as dynamic programming often needs a set of optional parameters to reach the best possible similarity.

Language modeling and its algorithms is a hybrid research area in protein structure analysis. Furthermore, the amino acid sequence of a protein consists of 20 distinct symbols of alphabet that can be treated as text written in a universal language. The mapping of a protein sequence to its structure, functional and biological role is similar to the mapping of words to their semantic meaning in natural languages. Recently (Biological Language Conference, 2003), it was suggested that this similarity motivates to apply *statistical language modeling* and *text classification techniques* in biological sequences analyzing. Within this hybrid research area, it is believed that the identification of Grammar/Syntax rules could reveal entities/relations of biological and medical sciences [12].

In this paper, a novel method for protein structural similarity measurement based on *n*-gram text modeling techniques is proposed. The method uses entropy concepts for information retrieval in the field of statistical language modeling. Nowadays *n*-gram modeling stands out as superior to any formal linguistics approach and has gained high popularity due to its simplicity [12]. In a very first attempt to fuse theoretical concepts from computational linguistics within the field of bioinformatics, a new general strategy for measuring similarity between primary sequences of proteins was introduced [12]. In this strategy, specifically, *n*-gram modeling is first applied to each protein sequence and cross-entropy measures are then employed to compare pairs of proteins. Based on the fruitful results of this attempt in using *n*-gram modeling, we now extend this approach to protein structural similarity measurement.

The rest of this paper is organized as follows. The next section, describes the protein structure representation in sequence form. In section III, the *n*-gram modeling technique is discussed. Section IV introduces a superposition task to find an overlap between two protein structures. Section V describes the novel method for protein structural similarity measurement based on *n*-gram

modeling. Finally, the experiments results are represented and discussed in section VI.

II. Protein Structure Modeling in String Form

Various kinds of language models can be used to capture different aspects of regularities of natural language. A variety of these alternative methods has already used for expressing similarity between biological sequences. Development of the language models to measure structural similarity of proteins needs protein structure modeling in string form.

There are various databases containing structure details of proteins. The Protein Data Bank (PDB)[1] is the worldwide repository for the processing and distribution of three dimensional biological molecular structure data. From the PDB file of each protein, the position of each residue in 3D space can be extracted using the 3D coordinates of C_α atom of each amino acid. Hence the 3D structure of a protein can be modeled in a sequence form by labeling the position of each residue with respect to the position of its previous residue in 3D coordinate. For labeling each residue *i*, let us suppose that the position of residue *i*-1 is centered at the origin of the spatial coordinate. Thus the position of the residue *i* can be labeled according to its spatial coordinates and can be represented with a specially defined alphabet. Fig.1 shows labels defined for 18 different positions of residue *i* with respect to residue *i*-1. To prevent the ambiguity, the other 8 labels are not shown in the figure. Table 1 represents 26 letters used for 26 position states in spatial coordinate corresponding to its previous residue. In this table, all lengths are expressed in Angstrom.

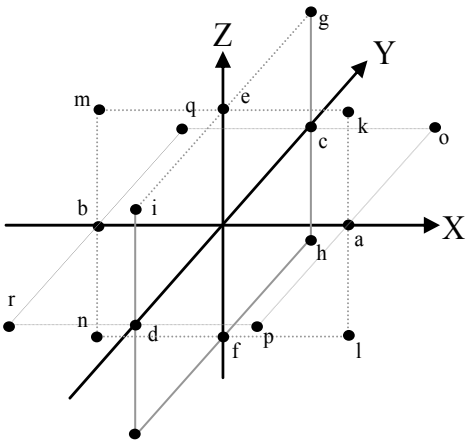


Fig. 1. 3D-space and labels defined for different position of residue i with respect to residue i-1 in the origin of the coordinate.

Accordingly, the protein structure can be represented in two strings sequences: the first string represents amino acids sequence and the second string represents the position label of each amino acid, according to table 1. From now onwards, we call the second sequence as relative residue position sequence. Fig.2 represents the two sequences extracted for 1CRB chain. Having reduced the protein structure to a sequence of characters, we can apply language modeling techniques in protein structure similarity measurement problem.

Table 1.
Letters defined for labeling 3D position of each residue with respect to its previous residue. ((x_2, y_2, z_2) is the position of current residue and (x_1, y_1, z_1) is the position of previous residue)

Conditions for *x*,*y*,*z*			Symbol
$x_2 - x_1 > 0$,	$\lvert y_2 - y_1 \rvert < 1$,	$\lvert z_2 - z_1 \rvert < 1$	'a'
$x_2 - x_1 < 0$,	$\lvert y_2 - y_1 \rvert < 1$,	$\lvert z_2 - z_1 \rvert < 1$	'b'
$\lvert x_2 - x_1 \rvert < 1$,	$y_2 - y_1 > 0$,	$\lvert z_2 - z_1 \rvert < 1$	'c'
$\lvert x_2 - x_1 \rvert < 1$,	$y_2 - y_1 < 0$,	$\lvert z_2 - z_1 \rvert < 1$	'd'
$\lvert x_2 - x_1 \rvert < 1$,	$\lvert y_2 - y_1 \rvert < 1$,	$z_2 - z_1 > 0$	'e'
$\lvert x_2 - x_1 \rvert < 1$,	$\lvert y_2 - y_1 \rvert < 1$,	$z_2 - z_1 < 0$	'f'
$\lvert x_2 - x_1 \rvert < 1$,	$y_2 - y_1 > 0$,	$z_2 - z_1 > 0$	'g'
$\lvert x_2 - x_1 \rvert < 1$,	$y_2 - y_1 > 0$,	$z_2 - z_1 < 0$	'h'
$\lvert x_2 - x_1 \rvert < 1$,	$y_2 - y_1 < 0$,	$z_2 - z_1 > 0$	'i'
$\lvert x_2 - x_1 \rvert < 1$,	$y_2 - y_1 < 0$,	$z_2 - z_1 < 0$	'j'
$x_2 - x_1 > 0$,	$\lvert y_2 - y_1 \rvert < 1$,	$z_2 - z_1 > 0$	'k'
$x_2 - x_1 > 0$,	$\lvert y_2 - y_1 \rvert < 1$,	$z_2 - z_1 < 0$	'l'
$x_2 - x_1 < 0$,	$\lvert y_2 - y_1 \rvert < 1$,	$z_2 - z_1 > 0$	'm'
$x_2 - x_1 < 0$,	$\lvert y_2 - y_1 \rvert < 1$,	$z_2 - z_1 < 0$	'n'
$x_2 - x_1 > 0$,	$y_2 - y_1 > 0$,	$\lvert z_2 - z_1 \rvert < 1$	'o'
$x_2 - x_1 > 0$,	$y_2 - y_1 < 0$,	$\lvert z_2 - z_1 \rvert < 1$	'p'
$x_2 - x_1 < 0$,	$y_2 - y_1 > 0$,	$\lvert z_2 - z_1 \rvert < 1$	'q'
$x_2 - x_1 < 0$,	$y_2 - y_1 < 0$,	$\lvert z_2 - z_1 \rvert < 1$	'r'
$x_2 - x_1 > 0$,	$y_2 - y_1 > 0$,	$z_2 - z_1 > 0$	's'
$x_2 - x_1 > 0$,	$y_2 - y_1 > 0$,	$z_2 - z_1 < 0$	't'
$x_2 - x_1 > 0$,	$y_2 - y_1 < 0$,	$z_2 - z_1 > 0$	'u'
$x_2 - x_1 > 0$,	$y_2 - y_1 < 0$,	$z_2 - z_1 < 0$	'v'
$x_2 - x_1 < 0$,	$y_2 - y_1 > 0$,	$z_2 - z_1 > 0$	'w'
$x_2 - x_1 < 0$,	$y_2 - y_1 > 0$,	$z_2 - z_1 < 0$	'x'
$x_2 - x_1 < 0$,	$y_2 - y_1 < 0$,	$z_2 - z_1 > 0$	'y'
$x_2 - x_1 < 0$,	$y_2 - y_1 < 0$,	$z_2 - z_1 < 0$	'z'

```
1   PVDFNGYWKM LSNENFEEYL RALDVNVALR KIANLLKPDK EIVQDGDHMI
     zwtwxsugu yuauktspjt kvhsqsmqzy wxzywxzlzv ximieuvohh

51  IRTLSTFRNY IMDFQVGKEF EEDLTGIDDR KCMTTVSWDG DKLQCVQKGE
    hkwscuvzvz imuyzustot xtnowiptvj ryzynwxhqz uvsppovssy

101 KEGRGWTQWI EGDELHLEMR AEGVTCKQVF KKVH
    yrxnmzxrqy xckluououo uywnqrxnxn xqh
```

Fig. 2. Two sequences extracted for the 1CRB protein chain.

III. Text-Based Protein Sequence Similarity Measure Based on *N*-gram Modeling

N-gram is one of the various kinds of language models that can be used to capture different aspects of regularities of natural languages. In this approach, the existence of a word w_k at a position *k* in a given text is assumed to depend only upon its immediate *n* predecessor words w_{k-n} … w_{k-1}. *Entropy* is a useful concept in the quantification of information in a textual sequence and making connection with probabilistic language modeling. A common definition of entropy as described in [12], when a written word sequence $w = \{w_1, w_2, ..., w_k, ...\}$ is treated as an *n*-gram:

[1] http://www.wwpdb.org/

$$H(X) = -\sum_{w^*} p(w_i^n)\log_2 p(w_{i+n} \mid w_i^{n-1})$$
$$= -\frac{1}{N}\sum_{w^*} Count(w_i^n)\log_2 p(w_{i+n} \mid w_i^{n-1}) \quad (1)$$

where the variable X is the 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 $W^* = \{\{w_1, w_2,..., w_n\}, \{w_2, w_3,..., w_{n+1}\},...\}$), $Count\ (w_i^n)$ is the number of occurrences of n-gram w_i^n and N is the total number of n-grams in the 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 and can be estimated by a counting procedure.

$$p(w_{i+n} \mid w_i^{n-1}) = Count\ (w_{i+n}) \Big/ Count\ (w_i^{n-1}) \quad (2)$$

As described in [12], the above entropy estimation indicates how a specific protein sequence is well predicted by the corresponding model. In the similarity measuring task, the direct comparison of the two proteins could not be facilitated by applying this measure to two distinct proteins. Cross-entropy measure is the relevant tool for this kind of comparison, where the n-gram model is, first, built based on the word-counts of one protein sequence and then the predictability, of the second sequence, by the model is measured via the formula:

$$H(X, P_M) = -\sum_{allw_i^n} p(w_i^n)\log_2 P_M(w_{i+n} \mid w_i^{n-1}) \quad (3)$$

The term $p(w_i^n)$ refers to the reference protein sequence and results from counting the words of that specific protein. The term $P_M(w_{i+n} \mid w_i^{n-1})$ refers to the sequence which the model has to be estimated (it results from counting the words of this protein). Variable X ranges over all the n-grams of the reference protein sequence [12].

The crux of the applied method in [12] is that both the unknown query-protein and each protein in a given database are represented via n-gram model and the cross-entropy measure is utilized to compare their representations. *Direct* method, a typical implementation of this idea, firstly, computes the perfect score PS from (3) using the query-protein both as reference and model sequence. Then the method uses (3) in the computation of the similarity score between the query-protein as the reference protein and each protein from the database as the model sequence. Therefore, N similarities are computed and applied in the calculation of the absolute differences via the formula:

$$D(S_q, S_i) = |H(X_q, P_{M_i}) - PS| \quad (4)$$

Finally, the most similar protein in the database to the query-protein is easily identified as the one having the lowest $D(S_q, S_i)$. In another implementation of the idea, called *Alternating* method, the only difference with respect to the *Direct* method is that the protein with the shortest sequence plays the role of reference sequence when comparing the query protein with each database-protein. This was devised in order to cope with the more different length of the proteins to be compared.

IV. Secondary Structure Superposition

The application of any one of the structural alignment algorithms requires protein structure representation in some coordinate independent space to make structures comparable. One possible representation is the so-called distance matrix, which is a two-dimensional matrix containing all pairwise distances between all C_α atoms of the protein backbone [14]. This can also be represented as a set of overlapping sub-matrices spanning only fragments of the protein. Another possible representation is the reduction of the protein structure to the level of secondary structure elements (SSEs), which can be represented as vectors and can carry additional information about relationships to other SSEs, as well as about certain biophysical properties [7], [11], [15]. In the case of distance matrix representation, the comparison algorithm breaks down the distance matrices into regions of overlap, which are then again combined if there is overlap between adjacent fragments, thereby extending the alignment. If the SSE representation is chosen, there are several possibilities. One can search for the maximum ensemble of equivalent SSE pairs using algorithms to solve the maximum clique problem from graph theory. Other approaches employ dynamic programming or combinatorial simulated annealing [11].

The proposed method in this paper needs an initial superposition between two proteins before encoding their structure in sequence form. In this way, the method represents the secondary structure elements of proteins as vectors and obtains a match for aligned vector pairs of query and reference proteins by computing angles between them and rotating reference protein in 3D coordinates. The secondary structures that represented in vector form are α-helices and β-strands and all types of helices (α, π, 3-10, and left handed helices) are grouped together in one class. It can easily be altered to use special classes for each type of helix. The SSEs information can be extracted from PDB file of each protein. Following equations are used to compute the beginning and end points of the helix and strand vectors respectively where indices i and j denote the first and last residues in the SSE [7], [11]:

$$r_b = (0.74r_i + r_{i+1} + r_{i+2} + 0.74r_{i+3})/3.48,$$
$$r_e = (0.74r_{j-3} + r_{j-2} + r_{j-1} + 0.74r_j)/3.48 \quad (5)$$

$$r_b = (r_i + r_{i+1})/2,$$
$$r_e = (r_{j-1} + r_j)/2 \quad (6)$$

and then the SSEs are represented by the vectors $r_{SSE} = r_b - r_e$. Helices of length shorter than five residues and strands of length shorter than three residues are neglected [7], [11].

Having reduced the two query and reference proteins to a set of either *Helix* or *Strand* vectors, the method now uses a dynamic programming algorithm to compare these two sets of vectors and find the best matched pairs. The scoring functions used in the algorithm are applied on the SSE type of vector, order of the vector in the protein and angles between matched vectors in 3D coordinates.

Finally, the method computes angles between each pair of matched vectors of query and reference protein and achieves a rotation angle and direction in polar coordinates. Hence, a relevant rotation-translation matrix is produced to achieve an initial overlap between two query and reference protein.

V. Protein Structure Similarity Measurement Based on *N*-gram Modeling

A new approach for 3D-structure of proteins similarity measurement is proposed. This method works based on the above *n*-gram similarity measure over protein structure modeled in sequence form as discussed in section 2. The similarity measurement process uses cross-entropy formula to compute the absolute entropy (4) between each pair of query and reference protein relative residue position sequences and find the most structurally similar protein in the given database to the query-protein.

In this new approach, a modification to the *n*-gram method introduced in [12] is done. In the counting process of the *n*-gram method described in [12], when all of the words have been counted once, the probability by $P_M(w_{i+n} \mid w_i^{n-1})$ become zero, creating problems in the calculation of $H(X, P_M)$. The new method uses a corrected entropy measurement formula:

$$H(X, P_M) = -\sum_{allw_i^n} p(w_i^n)\log_2(2 + P_M(w_{i+n} \mid w_i^{n-1})) \quad (7)$$

Thus, if the estimated term $P_M(w_{i+n} \mid w_i^{n-1})$ is zero, the result of logarithm function will be 1 and the value of $p(w_i^n)$ term will be considered in the summation formula.

The procedure described above for similarity measurement has been implemented in the following steps:

1) Compute the cross-entropy from (7) for the relative residue position sequence of query-protein.
2) For each reference protein in the given database, apply steps 2-1, 2-2 and 2-3.
   2-1) Find the matched pairs of SSE vectors with query protein and compute the rotation-translation matrix as discussed in Section 4. Then, rotate and translate the reference protein to extract the new coordinates of atoms. Then, make the relative residue position sequence of protein as described in section 2.
   2-2) Apply the cross-entropy measure from the (7) to compute the absolute differences via (4), as discussed in section 3.
   2-3) For every atom in the query protein, find the nearest atom (within a threshold distance) on the reference protein and transform the query protein to minimize the RMSD between these pairs of atoms.
3) Therefore an array of *N* extracted similarity is created, where each element of the array contains $D_t(S_q, S_i)$ computed via (4) for the relative residue position sequence. Arrange the array according to D_t.

The input of the algorithm is the unknown query-protein structure modeled in sequence form and a protein database contains the PDB file of each protein. Furthermore, the secondary structure of each protein is represented in collection of some vectors as described in section 4 and used as the input.

VI. Experimental Results

In order to assess the accuracy and efficiency of the proposed method, some experiments were performed. Firstly, to measure the accuracy of the method, 53 proteins are selected from the SCOP database belonging to All Alpha, All Beta, Alpha and Beta and Alpha+Beta categories with less than 40% sequence identity, having more than 7 SSEs. The 3D structure of each selected protein is modeled by the two sequences and vector representation of its secondary structure elements, as described above.

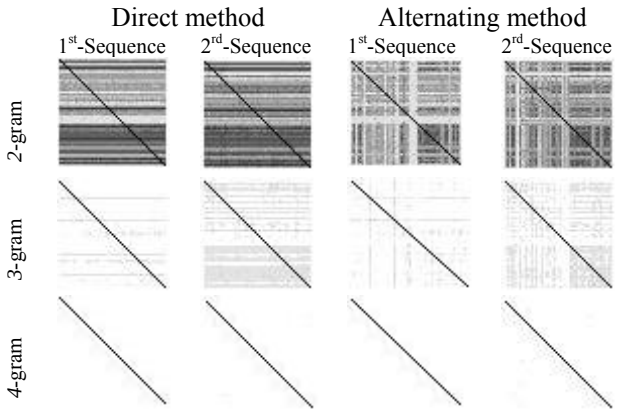


Fig. 3. Gray-scale representation of the output D_p and D_t matrices containing all the possible pairwise dissimilarities for 53 proteins in the database using *Direct* and *Alternating* method.

Fig.3 represents the matrices containing all the measured dissimilarities $D(S_i, S_j)$, i, j=1,2,…,N for each pair of proteins *i*, *j* in the database as grey scale images for the *Direct* and *Alternating* methods of three different *n*-gram models. In the figure, the first and second sequence indicates primary sequence and relative residue position sequence. In each matrix the vertical and horizontal edges represent the query and reference proteins respectively. The white and black colors in the output matrices correspond to the maximum and minimum distances between each pair of proteins. As described in [12], the ideal spatial outlay is a white matrix with only a black diagonal segment. Therefore, it is clearly evident from fig.3 that *4*-gram modeling which uses *Alternating* Method has a better performance in order to distinguish similar and dissimilar proteins. On the other hand, as seen from the figure, *3*-gram modeling outputs represent highly similar, less similar and dissimilar proteins and it is much more informative than 4-gram. Furthermore, fig.3 shows that the results obtained from second sequence are more informative on similarity measurement than the primary sequence.

In order to compare the accuracy and efficiency of the method with other publicly available protein structure

similarity servers, two servers were selected, namely Combinatorial Extension (CE)[2] and Secondary Structure Matching (SSM)[3]. It is believed that none of the scores provides an absolutely reliable measure of structural similarity or statistical significance, and therefore the final decision of accepting a match should be reserved for the user [11]. Hereby, the comparison process is done by calculating three values: RMSD, N_{align} and *Q*-score.An intuitive understanding of structural similarity suggests contradictory requirements of achieving a lower RMSD and a higher number of aligned residues N_{align}. This contradiction may be eliminated, in the first approximation, by a score that represents a ratio of N_{align} and the RMSD. Therefore, the following function is suggested [11]:

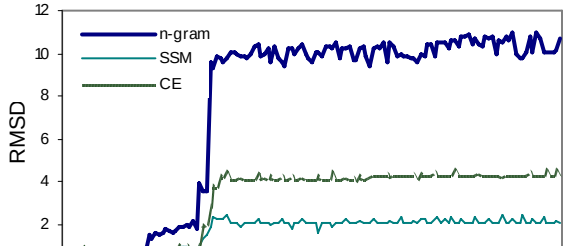
$$Q = N_{align}^2 / ((1 + (RMSD / R_0)^2) N_1 N_2) \quad (6)$$

R_0 is an empirical parameter (chosen at 3 A°) that measures the relative significance of RMSD and N_{align}. N_1 and N_2 are the number of residues in the aligned structures. As seen from the above formula, *Q* reaches 1 only for identical structures ($N_{align} = N_1 = N_2$ and R.M.S.D = 0), and decreases to zero with decreasing similarity (increasing RMSD or/and decreasing N_{align}). Therefore, the higher *Q*, is the better, in general, the alignment [11].

Fig. 4 represents the results of comparing the n-gram based method with SSM and CE methods for the example of protein chain 1sar:A. The experiment is done over whole PDB chains by SSM and CE servers in order to select the top 200 chains from the list and use them to do the same experiments applying the n-gram method. The output results in fig.4 are represented for 150 protein chains ordered by entropy measure of *n*-gram method.

Fig.4 shows that *n*-gram method approximately fully agrees with the other servers in the identification of highly similar, less similar and dissimilar structures. As seen from the figure, all the methods reveal the same RMSD results for the first 30 protein chains, but for the rest of the protein chains there are differences. The differences are because the SSM and CE methods apply some iteration tasks to reduce RMSD value, whereas the *n*-gram method does not perform such a task. RMSD reduction task is a time consuming process. The *n*-gram method, simply, rotates and translates the reference protein in 3D-coordinates to achieve a superposition with the query protein. Therefore, from the viewpoint of running speed, the similarity measurement process has been accelerated in the *n*-gram method.

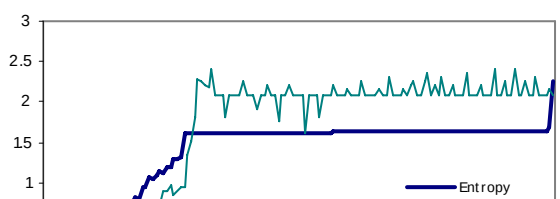
The alignment length of *n*-gram method, represented in fig.4, is approximately the same as SSM. As it is described in [11], longer alignments always come at the expense of higher RMSD and therefore the observed differences between the servers should be mostly due to the different criteria employed to balance these characteristics.



The *Q*-score is an indication of the balance of RMSD and the alignment length [11]. As seen from the *Q*-score plot in fig.4, *Q*-score of the *n*-gram method is lower than those of the two other methods. This is because the *n*-gram method computes high RMSD value compared with the other methods.

We also performed a comparison between entropy measure computed by the *n*-gram method via (4) and RMSD computed by the SSM method. Fig.5 represents that RMSD value increases with the increasing value of entropy. It shows that the similarity measurement results produced by the *n*-gram method are approximately the same as those produced by the SSM method. Therefore, the entropy measure based on *n*-gram modeling is a novel efficient tool for protein structural similarity measurement.

We performed a comparative study, similar to that described above, for a number of structures belonging to different protein folds. The results represent that the outputs showed in figures 4 and 5 are of a common nature.



To evaluate the efficiency of *n*-gram method, an extended dataset of about 2000 proteins was prepared from the various categories in the SCOP database. The algorithm of *n*-gram method is implemented in C++ programming language and done on Pentium IV 2.8GHz machine with 512MB RAM running Windows- XP. Average time of similarity measurement for each query is about 30 seconds. Because the source code of SSM method was not accessible, run-time comparison of two methods could not be conceived. However, including related experiments [7], [11], [15], [16], the efficiency of the method is established compared with the other similar methods.

VII. CONCLUSION

The proposed method in this paper uses the introduced method in [12] to apply entropy concept for information retrieval in the field of statistical language modeling for measuring the structural similarity of proteins. Specifically, the studied method, simply, applies a superposition task to achieve an initial overlap between the secondary structure elements of two proteins and then, creates relative residue position sequence for them and uses cross-entropy measure over *n*-gram model to compare their structures. In order to confirm the validity of the proposed method, some experiments on similar protein retrieval methods were performed which demonstrates the applicability and efficiency of this method. Also, the results of experiments represent the method is comparable with the publicly available web servers namely SSM and CE. Moreover regarding the conceptual simplicity of the approach, the preference and applicability of the method to other applied techniques is indicated.

ACKNOWLEDGMENTS

We would like to thank our research grant sponsor, Malaysian Ministry of Science, Technology, and Innovation (MOSTI) and Universiti Technologi Malaysia Research Management Centre (RMC) for their support (research grant number: Vote 79228)

REFERENCES

[1] T. Ohkawa, S. Hirayama, H. Nakamura, "A Method of Comparing Protein Structures Based on Matrix Representation of Secondary Structure Pairwise Topology", in *Proc. of the IEEE Int. Conf. on Information Intelligence and Systems*, 1999.

[2] L. Holm, C. Sander., "Protein structure comparison by alignment of distance matrices", *J. of Molecular Biology*, 233, pp.123–138, 1993.

[3] C.A. Orengo, W.R. Taylor, "SSAP: sequential structure alignment program for protein structure comparison", *Methods Enzymol*, 266, pp. 617–635, 1996.

[4] A. Falicov, FE. Cohen, "A surface of minimum area metric for the structural comparison of proteins", *J. of Molecular Biology*, May 24, 258, pp.871-92, 1996.

[5] S. Subbiah, D. V. Laurents, M. Levitt, "Structural Similarity of DNA-binding Domains of Bacteriophage Repressors and the Globin Core", *Current Biol.*, 3, pp.141-148, 1993.

[6] M. Gerstein, M. Levitt, "Comprehensive Assessment of Automatic Structural Alignment against a Manual Standard, the SCOP Classification of Proteins", *Protein Science*, 7, pp.445-456, 1998.

[7] A. P. Singh, D. L. Brutlag, "Hierarchical Protein Structure Superposition using both Secondary Structure and Atomic Representations", in *Proc. of the 5th Int. Conf. on Intelligent Systems for Molecular Biology*, 1997.

[8] C. H. Chionh, Z. Huang, K. L. Tan, Z. Yao, "Augmenting SSEs with Structural Properties for Rapid Protein Structure Comparison", in *Proc. of the 3rd IEEE Symp. on BIBE*, 2003.

[9] I. N. Shindyalov, P. E. Bourne, "Protein Structure Alignment by Incremental Combinatorial Extension (CE) of the Optimal Path", *Protein Engineering*, Vol.11, pp.739-747, 1998.

[10] J. F. Gibrat, T. Madej, J. L. Spouge, S. H. Bryant, "The VAST protein structure comparison method", *Biophysical Journal*, 72:Pt2, pMP298, 1997.

[11] E. Krissinel, K. Henrick, "Secondary-structure matching (SSM), a new tool for fast protein structure alignment in three dimensions", *Acta Crystallographica Section D: Biological Crystallography*, Vol.60, No.1, 2004.

[12] 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. of the 2nd Int. Conf. on CIMED*, 2005.

[13] S. C. Chen, T. Chen, "Retrieval of 3D Protein Structures", in *Proc. of the IEEE Int. Conf. on ICIP*, 2002.

[14] P. Chi, G. Scott and C. R. Shyu, "A Fast Protein Structure Retrieval System Using Image-Based Distance Matrices and Multidimensional Index", in *Proc. of the 4th IEEE Symp. on BIBE*, 2004.

[15] A. C. R. Martin, "The Ups and Downs of Protein Topology; Rapid Comparison of Protein Structure", *Protein Engineering*, Vol.13, No.12, 2000.

[16] Z. Aung, K. L. Tan, "Automatic Protein Structure Classification through Structural Fingerprinting", in *Proc. of the 4th IEEE Symp. on BIBE*, 2004.