A Novel Text Modeling Approach for Structural Comparison and Alignment of Biomolecules

JAFAR RAZMARA, SAFAAI B. DERIS
Faculty of Computer Science and Information Systems
Universiti Teknologi Malaysia
81310, UTM, Skudai, Johor Bahru, Malaysia
razmaraj@gmail.com, safaai@utm.my

Abstract: - Within this paper, a novel strategy for structural alignment of proteins based on text modeling techniques is introduced. The method summarizes the protein secondary and tertiary structure in two textual sequences. The first sequence is used to initial superposition of secondary structure elements and the second sequence is employed to align the 3D-structure of two compared structures. The comparison technique used by the method has been inspired from computational linguistics for analysing and quantifying textual sequences. In this strategy, the cross-entropy measure over n-gram models is used to capture regularities between sequences of protein structures. The performance of the method is evaluated and compared with CE and SSM methods. The results of the experiments reported here provide evidence for the preference and applicability of the new approach in terms of efficiency and effectiveness.

Key-Words: - protein structure alignment; n-gram modeling; cross-entropy

1 Introduction

Structural comparison and alignment of proteins are the fundamental problem in structural biology. There are several motivations that made the problem as of interest. Structure comparison gives a powerful tool for searching homologous proteins to classify proteins. Furthermore, it can be utilized to predict function of new unknown proteins based on structural similarity with the known proteins. Finally, it facilitates to study evolutionary relationships between protein families. Accordingly, it has wide applications in protein structure analysis which have been interested extensively and studied widely within the recent years.

Protein structure comparison, retrieval and classification have been explored in two main categories: sequence comparison and 3D-structure comparison. The methods in former category investigate the sequence alignment of amino acids in the primary structure of the proteins [1]. The latter category includes the methods to align the three dimensional structure of the proteins [2]. Due to determination of 3D-structure of a protein by its amino acid sequence, which in turn determines the protein function, it might be think that the sequence similarity is also very good predictor of the functional similarity. However, structural biologists highly believed that different amino acid sequences can yield very different structures and also similar sequences can sometimes yield dissimilar structures. Therefore, sequence similarity is much less reliable evident for functional similarity prediction than structural similarity. As a result, developing efficient and effective structure comparison and similarity measures have been highly paid attention and approached in the past years.

The structural comparison tool, generally, is used to distinguish the differences among various proteins functionalities. The algorithm proceeds to find the best superposition between the atoms in two structures with a minimal distance between the matched pairs [3]. There is no information which pairs of atoms are corresponding. Therefore, it needs an exhaustive heuristic search to find the best correspondence between the atoms of two structures. Output of the algorithm is a list of matched pair of residues from two compared proteins.

The protein structure comparison has two main problems: *Complexity* and *Curse of dimensionality* [4]. In the view of complexity, the structure comparison is an NP-hard problem and there is not exists any exact solution for structural comparison of proteins. There are several algorithms proposed for optimizing the results, however, none of them can guarantee optimality within any given precision. Rapidly growing of the discovered protein databases, also, provides the dimensionality problem. The Protein Data Bank (PDB)[1], currently[2],

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

[2] May 2009

contains 57,706 known protein structure. The increasing number of entries in the PDB requires more efficient methods to search and find structural similar proteins.

Different numbers of methods for protein structure comparison and alignment have been explored within the past decade and are available currently for the structural biologists. These approaches can be grouped in three different categories: pairwise structure alignment methods, multiple structure alignment methods and indexing of protein databases [5]. Pairwise structure alignment is an algorithm to find a one-to-one map between the elements of two proteins. Multiple structure alignment algorithms are an extended version of the pairwise algorithms that simultaneously detect similarities between *n* different molecules. Obviously, it is necessary to highlight partial similarities between the subsets of these *n* molecules. Finally, database indexing algorithms are designed for online searching in a reference database. Regarding the time complexity of the structural alignment methods, algorithms in the third group meet the requirements for online processing of the queries [5].

The structure comparison and alignment methods, generally, perform a comparison between the geometry of the residues, but the basics of their algorithms are different. The employed techniques are comparison of distance matrices (DALI) [6], vector alignment of SSEs (VAST) [7], combinatorial extension of alignment path (CE) [8], structure alignment using environmental profiles (SHEBA) [9], flexible structure alignment by chaining aligned fragment pairs with twists (FATCAT) [10], Markov transition of protein structure evolution (MATRAS) [11], Secondary Structure Matching (SSM) [12], combination of TM-score rotation matrix and dynamic programming to identify the best structural alignment (TM-Align) [13], genetic algorithm for non-sequential gapped protein structure alignment (GANGSTA) [14], alignment of protein structure in the presence of conformational changes (RAPIDO) [15] and many others. Several comprehensive reviews are reported [16, 17, 18] for analysis and comparison of protein structure alignment methods that are useful for detailed evaluation of them.

None of the above introduced methods provides a complete solution for the problem and the study for designing new efficient methods is still an active research area. Continuous growth of the protein databases renews the interests for designing alternative powerful and reliable tools. Moreover, another important objective in these studies is proposition of a strategy without need to parameter setting by the user. The classical structure comparison approaches such as dynamic programming based methods often needs a set of optional parameters to reach the best possible results.

Biological data, naturally, are large organic compound made of amino acids arranged in a linear chain. This chain can be seen as a text in a natural language. The alphabets of this language are amino acids that include 20 distinct symbols. The protein sequence mapping to its structure, functional and biological properties is highly similar to the mapping of words to their semantic meanings in natural languages [19]. This analogy has opened a new perspective in the evolution of structural biology. Consequently, it is not a surprise to apply *statistical text modeling* and *classification concepts and techniques* in analyzing biological sequences.

In this paper, a novel text modeling approach for structural comparison and alignment of proteins is introduced. The method is influenced by the fruitful applications of entropy concept in the field of statistical language modeling for information retrieval problems [20, 21]. *N*-gram modeling is also a preferable concept to any formal linguistics approach [19]. In a very preliminary study to fuse computational linguistics theoretical concepts in the field of computational biology, a novel general approach for similarity measurement between primary sequences of proteins was introduced by Bogan-Marta et al. [19]. Based on the desirable results of this attempt, now an extended implementation of this approach to protein structure alignment is represented.

2 Methods

2.1 Protein Structure Modeling in Textual Sequence

Most of the protein structure comparison methods use a simplified representation of these macromolecules to reduce the complexity of the problem. The introduced method in this paper uses textual representation of protein structure. Specifically, protein secondary and tertiary structures are summarized in two different sequences of alphabets.

The first sequence represents secondary structure elements (SSEs) as a string of symbols. Each type of these regular substructures is encoded by an alphabetic symbol which is listed in table 1. A sample sequence of SSEs for 1CRB PDB chain extracted from PDB database is showed in figure 1.

Table 1. Symbols defined for secondary structure elements

3-turn helix (3_10 helix). Min length 3 residues	'G'
4-turn helix (alpha helix). Min length 4 residues	'H'
5-turn helix (pi helix). Min length 5 residues	'I'
hydrogen bonded turn (3, 4 or 5 turn)	'T'
beta sheet in parallel and/or anti-parallel sheet conformation (extended strand)	'E'
residue in isolated beta-bridge (single pair beta-sheet hydrogen bond formation)	'B'
bend (the only non-hydrogen-bond based assignment)	'S'
regions that do not form a regular known secondary structure element	' '

ESHT H ETE S ET E T ETEST ETETE

Fig.1 Secondary structure elements sequence extracted for 1CRB pdb chain.

The second sequence models 3D-structure of protein in a string of letters. Having protein structure details extracted from their PDB file, 3D-structure of a protein can be represented in a sequence form by labeling the position of C_α atom of each residue with respect to the position of its previous C_α atom in 3D coordinates. For labeling each residue i, suppose that the position of $C_{\alpha\ i-1}$ is centered at the origin of the spatial coordinates. Thus, the position of $C_{\alpha\ i}$ can be labeled according to its spatial coordinates and represented with a specially defined alphabet. Figure 2 represents sample labels defined for different positions of $C_{\alpha\ i}$ with respect to $C_{\alpha\ i-1}$ in 3D-coordinates. To prevent the ambiguity, labels are shown only for 8 different positions. Table 2 represents 26 letters used for 26 position states in 3D coordinates corresponding to its previous residue. In this table, the parameter *t*, after some experiments and comparison between the results, was evaluated by 0.1 Angstrom. Figure 3 represents this sequence extracted for 1CRB PDB chain.

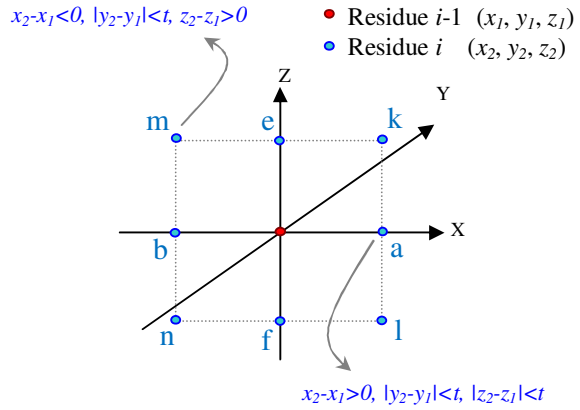

Fig.2 Secondary structure elements sequence extracted for 1CRB pdb chain.

Table 2 - Labels defined for 3D position of each residue with respect to its previous residue. $((x_2,y_2,z_2)$ and (x_1,y_1,z_1) are the position of current and previous residues respectively

Conditions for x,y,z			Symbol
$x_2-x_1>0$,	$\|y_2-y_1\|<t$,	$\|z_2-z_1\|<t$	'a'
$x_2-x_1<0$,	$\|y_2-y_1\|<t$,	$\|z_2-z_1\|<t$	'b'
$\|x_2-x_1\|<t$,	$y_2-y_1>0$,	$\|z_2-z_1\|<t$	'c'
$\|x_2-x_1\|<t$,	$y_2-y_1<0$,	$\|z_2-z_1\|<t$	'd'
$\|x_2-x_1\|<t$,	$\|y_2-y_1\|<t$,	$z_2-z_1>0$	'e'
$\|x_2-x_1\|<t$,	$\|y_2-y_1\|<t$,	$z_2-z_1<0$	'f'
$\|x_2-x_1\|<t$,	$y_2-y_1>0$,	$z_2-z_1>0$	'g'
$\|x_2-x_1\|<t$,	$y_2-y_1>0$,	$z_2-z_1<0$	'h'
$\|x_2-x_1\|<t$,	$y_2-y_1<0$,	$z_2-z_1>0$	'i'
$\|x_2-x_1\|<t$,	$y_2-y_1<0$,	$z_2-z_1<0$	'j'
$x_2-x_1>0$,	$\|y_2-y_1\|<t$,	$z_2-z_1>0$	'k'
$x_2-x_1>0$,	$\|y_2-y_1\|<t$,	$z_2-z_1<0$	'l'
$x_2-x_1<0$,	$\|y_2-y_1\|<t$,	$z_2-z_1>0$	'm'
$x_2-x_1<0$,	$\|y_2-y_1\|<t$,	$z_2-z_1<0$	'n'
$x_2-x_1>0$,	$y_2-y_1>0$,	$\|z_2-z_1\|<t$	'o'
$x_2-x_1>0$,	$y_2-y_1<0$,	$\|z_2-z_1\|<t$	'p'
$x_2-x_1<0$,	$y_2-y_1>0$,	$\|z_2-z_1\|<t$	'q'
$x_2-x_1<0$,	$y_2-y_1<0$,	$\|z_2-z_1\|<t$	'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.3 Amino acids sequence and relative residue position sequence extracted for 1CRB PDB chain.

Accordingly, the protein structure can be represented in two different sequences: the first sequence denotes the SSEs of protein secondary structure and the second string represents the position label of each amino acid in 3D-space, according to table 2. From now onwards, the second sequence is called as relative residue position sequence. Having reduced protein structure to textual sequences, now we can apply language modeling techniques in protein structure comparison and alignment problem.

2.2 Text Modeling by N-gram Method and Entropy Concept

Variety numbers of language modelling techniques have been explored to capture and analyze regularities of universal languages [22, 23]. Several numbers of these techniques have already been used

for similarity measurement of biological sequences. Hidden Markov Models (HMM) are the more fundamental concept to combine information theory with statistics that highly used in language modeling. HMM assumes that the existence of a word W_i at a position i in a text is related to its urgent n previous words $W_{i-n},\ldots, W_{i-1}$ [19]. The model, usually called as n-gram, has been more popular and widely used in formal linguistics approaches due to its simplicity. *Entropy* is also a useful concept in information quantification in a textual sequence and making connection with probabilistic language modeling. The entropy estimation, as described in [19], indicates how a specific sequence is well predicted by the given model. In the similarity measuring task, direct comparison of the two sequences could not be facilitated by applying this measure. Cross-entropy is the relevant tool for this kind of comparison, where the n-gram model is firstly made by counting the words of one sequence in the training phase. Then the second sequence predictability is measured in the recall phase via the following formula.

$$H(X, P_M) = -\Sigma_{all\ w^*} p(w_i^n) \log_2 P_M(w_{i+n}|w_i^{n-1}) = -(1/N)\Sigma_{all\ w^*} Count(w_i^n) \log_2 P_M(w_{i+n}|w_i^{n-1}) \quad (1)$$

The variable X is in the form of n-gram and ranges over all the words of the reference sequence. N is the number of n-gram words. The term $p(w_i^n)$ refers to the words count of reference sequence. The conditional probability in the second term relates the n-th word with the preceding n-1 words and can be computed by counting the words of the query sequence which the model has to be estimated:

$$P(w_{i+n}|w_i^{n-1}) = Count(w_{i+n}) / Count(w_i^{n-1}) \quad (2)$$

The crux of the applied method in [19] to measure similarities between two protein sequences is that both the unknown query-protein and each protein in a reference database are modeled in the n-gram form and then cross-entropy measure is utilized to compare them. *Direct* method, a typical implementation of this idea, firstly, computes the perfect score *PS* from (1) over the query protein as reference and model sequence (training). Then the method uses (1) to compute the similarity score between the query protein as the reference protein and each protein in the database as the model sequence (recall). Therefore, N similarities are computed and applied in the calculation of the absolute differences via the formula:

$$D(S_q, S_i) = |H(X_q, P_{Mi}) - PS| \quad (3)$$

S_q and S_i in this equation denote the query and i-th reference proteins respectively. Finally, the most similar proteins from the database is easily identified as the one having the lowest $D(S_q,S_i)$. In another implementation of the idea, called *alternating* method, the protein with the shortest sequence is considered as reference sequence when comparing the query protein with each database-protein [19]. This was experimented to overcome the more different length of the proteins to be compared.

A modified version of the cross-entropy formula (1) is used in the implementation of the new proposed method. In the counting process of the n-gram method, when all of the words have been counted once, the probability by $P_M(w_{i+n}|w_i^{n-1})$ become zero that causes lose of data in the calculation of $H(X, P_M)$. The proposed method uses a corrected entropy measurement formula as following:

$$H(X,P_M) = -\Sigma_{all\ w^*} p(w_i^n) \log_2 (2+P_M(w_{i+n}|w_i^{n-1})) \quad (4)$$

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

2.3 Secondary Structure Superposition

Due to the availability of 3D-coordinates of any protein structure in an arbitrary relative orientation, the matched parts of a protein pair may not be correspond. Therefore, it is necessary to do a superposition task between two structures in order to make them comparable. Distance matrix is one of the possible commonly used solutions, which is a two-dimensional matrix including pairwise distances between all C_α backbone atoms [6]. To compare proteins in this strategy, the distance matrices are broken down into matched regions, which are combined again with matched adjacent fragments to extend alignment. Another prevalent scheme is protein structure reduction to the secondary structure elements (SSEs), which can be managed as a set of vectors to extract geometrical relationships with the other SSEs [12]. The algorithm in this case, searches for the maximum set of matched SSE pairs and applies graph theory techniques to solve the maximum clique problem.

In order to achieve an initial superposition between two proteins before encoding their 3D-structure in relative residue position sequence, the proposed method in this paper uses vector representation of secondary structure elements

(SSEs) technique introduced by Singh and Brutlag [24] to find matched SSEs of two proteins. The beginning and end points of the helix and strand vectors are computed via the following equations where indices i and j denote the first and last residues in each element [12, 24] and then the SSEs are represented by $r_{SSE}=r_b-r_e$:

$$r_b = (0.74r_i + r_{i+1} + r_{i+2} + 0.74r_{i+3}) / 3.48,$$
$$r_e = (0.74r_{j-3} + r_{j-2} + r_{j-1} + 0.74r_j) / 3.48, \quad (5)$$
$$r_b = (r_i + r_{i+1}) / 2,$$
$$r_e = (r_{j-1} + r_j) / 2 \quad (6)$$

Having protein secondary structure information reduced in a SSEs sequence and a set of either *Helix* or *Strand* vectors, now a procedure is applied to match SSEs of two proteins. Firstly, SSEs sequences are represented via n-gram model and words of the first sequence are matched with the second sequence. An iterative loop is employed for n-gram size from n (defined empirically) decreasing to 2. Apparently, it is possible that a word from the first sequence is matched with two or more words in the second sequence. In the sequel, a dynamic programming algorithm is employed to refine the initially matched SSEs using their vectors and find the best matched pairs. The scoring functions used in the algorithm are the SSE type of vector, order of the vector in the protein structure and angles between the matched vectors in 3D coordinates. Finally, the method computes the average distance and angle between matched vectors and produces a relevant rotation-transformation matrix in order to achieve an initial overlap between two proteins structure.

2.4 Structural Alignment by N-gram modeling

After the initial superposition of two protein structures, the second sequence of protein structure called relative residue position sequence is created as discussed in section A. Then, the cross-entropy measure is employed to measure similarity between two structures. Also an alignment procedure is performed simultaneously to establish equivalencies between the pairs of residues from the compared proteins. This alignment is initially obtained while computing similarity between two relative residue position sequences using the n-gram modeling. In this procedure, the identical words from two proteins are marked as matched. Therefore, each word in the reference protein points to the corresponding words in the query protein. We also note the matched SSEs acquired in previous section. In the sequel the method uses a dynamic programming algorithm to refine and complete the alignment by the following steps:

1. Inside each pair of the matched SSEs, locate the pairs of the matched words and mark their corresponding residues as aligned. Expand the alignment to the ends of the SSEs for each pair of residues, leaving no unmatched pair of residues between the matched ones.
2. For each pair of exclusively matched words from two structures, if the connectivity of the aligned residues is not violated and the distance of the residues is less than the maximum distance of already aligned residues, mark the corresponding residues as aligned.
3. For the reference protein words that matched with more than a word in the query protein, consider their connectivity with the aligned words in previous step and the general order of them along the protein chain be the same in both structures. Then mark the residues of the selected matched words as aligned. Note that any number of missing residues between the identical words is ignored.
4. Finally, try to align all remained unaligned residues, if there is a corresponding residue in the other structure that their distance is less than the maximum distance between the aligned residues.

In steps 2, 3 and 4, pairs of residues are not marked as aligned if they belong to different types of secondary structure.

2.5 Database Search Algorithm

A new approach for structural alignment of proteins is proposed. The method works based on the above introduced n-gram similarity measure over protein structure modeled in sequence form. The similarity measurement process uses cross-entropy formula to compute the absolute entropy (3) between each pair of query and reference proteins sequences and find the most structural similar protein in the given database to the query-protein. The procedure has been implemented in the following algorithm. In this algorithm, the *RotateTransformProtein* and *CreateRelativeResiduePositionSequence* functions work based on the procedure introduced respectively in sections 2.3 and 2.1. The algorithm produces an array of N extracted similarity, where each element of the array contains a value computed via (4) for relative residue position sequence. The minimum value in this array denotes the most similar structure to the query protein.

Input: *Structure information of the query protein and each reference protein in the database including protein primary, secondary and tertiary structure.*

Output: *An array of computed similarity scores by the n-gram method.*

Algorithm:
Let S_q and T_q be the secondary and relative residue position sequences of query protein and S_i and T_i have the same role for each protein in the reference database.

$PS_t = H(T_q, T_q)$ *// Perfect Score*

for each protein i in the reference database do
RotateTransformProtein()
$T_i \leftarrow$ *CreateRelativeResiduePositionSequence()*
$D_t[i] = |H(T_q, T_i) - PS_t|$

3 Results

3.1 Determine the best form of the algorithm

The performance of the proposed method is studied by several experiments. The first experiment is established in order to empirically specify the relevant form of the algorithm to balance accuracy and sensitivity against computational efficiency. In this experiment, 53 proteins are selected belonging to All Alpha, All Beta, Alpha and Beta and Alpha+Beta categories in SCOP database with less than 40% sequence identity, having more than 7 SSEs. The proteins are compared all-against-all by the above introduced method.

Figure 4 represents the matrices containing all the measured dissimilarities D(Si, Sj), for each pair of proteins i, j in the dataset as grey scale images for the Direct and Alternating methods of three different n-gram models. The vertical and horizontal edges represent the query and reference proteins respectively. In the output matrices, the white and black colors correspond to the maximum and minimum similarity between each pair of proteins. The ideal outline is an image with white background and only a black diagonal line which represents that the method can distinguish similar and dissimilar structures. Therefore, it is clearly shown from the figures that 4-gram modeling which uses Alternating Method has a better performance in order to recognize similar and dissimilar proteins. On the other hand, as seen from the figures, 3-gram modeling output represent highly similar, less similar and dissimilar proteins and it is much more informative than 4-gram.

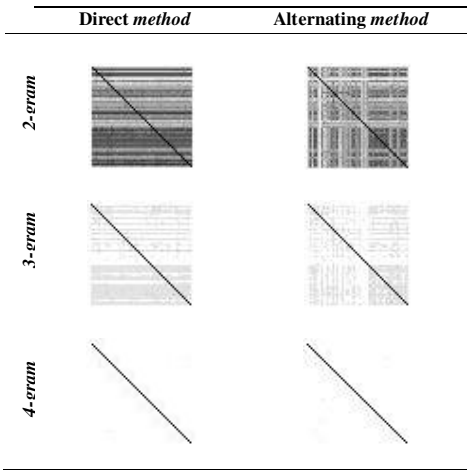

Fig.4 Gray-scale representation of the outputs including all the pairwise similarities for 53 proteins using Direct and alternating methods.

3.2 Represent an alignment sample

In this section, a typical alignment result using the above introduced method is represented and compared with the alignment result produced by CE [8] as the base for protein structure comparison. The experiment is performed between two protein chains 1AKT:_ (147 residues) and 1CRP:_ (166 residues) with less than 9% sequence identity. Figure 5 shows the structural alignment results in primary sequence level for CE and primary and relative residue position sequence for n-gram method. In figure 5(b), the letters with gray background identifies the similar words in two relative residue position sequences of proteins that are aligned. As seen from the figure, the alignment result of CE has 134 aligned reside with 4.8Å for RMSD whereas n-gram method aligns 138 residues with an RMSD of 5.7Å.

The proposed n-gram based method for structural alignment does not perform root mean square deviation (RMSD) minimization task between two structures. RMSD is one of the most commonly used measures of structural alignment quality. In order to obtain a high quality of alignment, the structure alignment algorithms apply different strategies to reduce RMSD which is a time consuming procedure. The *n*-gram method, simply, rotates and transforms the reference protein in 3D-coordinates to achieve a superposition with the query protein. However, a decision can be made by the user to achieve the optimal RMSD between the two structures.

a) CE alignment result:

```
1AKT:_    PKALIVYGS--TTGNTEYTAET------------------IARELADAGYEVDSRDAA-------SVEAGGLFEGFDLVLLGCSTWNDDSIELQ
1CRP:_    TEYKLVVVGAGGVGKSALTIQLIQNHFVDEYDPTIEDSYRKQVVIDGETCLLDILDTAGQEEYSAMRDQYMRTG---EGFLCVFAINNTKSFED

1AKT:_    DDFIPLFDSLEETGAQGRKVACFGCGDS----SYEYFCGAVDAIEEKLKNLGAEIVQDGLRIDGDPRAARDDIVGWAHDVRGAI
1CRP:_    --IHQYREQIKRVKDSDDVPMVLVGNKCDLAARTVESRQAQDLARSYG--------IPYIETSAKTRQGVEDAFYTLVREIRQH
```

b) The n-gram method alignment result:

```
1AKT:_    PKALIVYGSTTGNTEYTA-ETI---------------ARELADAGYEVDSRDAAS------VEAGGLFEGFDLVLLGCSTWNDDSIELQDDFIP
           ywyqyqmnmxlokjhsp-jhk--------------vkopkpspiimywymkrv------susptxtplqtqmqmqmqgywikplxvzivoip
           mqwqgmhqgxloklijokjhkpjhgxwnqrinizpppkopkpskgkjrn-mkrvgnqmqnzmiililsipogilokmqqywghg-xvzipkmj
1CRP:_    TEYKLVVVGAGGVGKSALTIQLIQNHFVDEYDPTIEDSYRKQVVIDGETC-LLDILDTAGQEEYSAMRDQYMRTGEGFLCVFAIN-NTKSFEDI

1AKT:_    LFDSLEETGAQGRKVACFGCG---DSSYEYFCGAVDAIEEKLKNLGAEIVQDGLRIDGDPRAARDDIVGWAHDVRGAI
          tgivtkjhogstnxmqmnmnw---qgmusphvkvtsivogilointnhqhqwrnmnqzjlngkvokilouvtsivoip
          kmivtkjlirjknxmqmlqhshojlmnxqphvkvhqgbimukoinj-------kmzpkmjkhohhkilohjhjhkognjk
1CRP:_    HQYREQIKRVKDSDDVPMVLVGNKCDLAARTVESRQAQDLARSYGI-------PYIETSAKTRQGVEDAFYTLVREIRQH
```

Fig.5 Structural alignment of 1AKT:_ and 1CRP:_ PDB chains using (a) CE (Combinatorial Extension) and (b) The n-gram method. The second and third lines (shown with lowercase letters) in (b) represent the relative residue position sequence for two aligned proteins. Also, the letters with gray background identifies the similar words in two sequences.

3.3 Investigate the retrieval effectiveness of the method

Another comparative study was performed to determine how well the methods detect members of the same protein families. This experiment is similar to that described in [25] and assesses the retrieval effectiveness of the four schemes: BLAST [26], SCALE [25], SSM [12] and the proposed *n*-gram based method. BLAST is a basic local alignment search tool for comparing primary sequences of proteins. The NCBI version of the BLAST[3], algorithm was used in the experiments. SCALE is also a comparison algorithm that applies secondary structure elements for protein structure alignment. The third method in the experiments is SSM, a powerful publicly available protein structure alignment server[4] that works based on secondary structure matching. Except for BLAST, the other three methods apply secondary structure elements as a basic strategy for structure superposition.

The experiments were done over the same dataset of 90 proteins that used in [25] with the same properties mentioned in our first experiment using SSM and *n*-gram methods and combined with the results of the BLAST and SCALE methods reported in [25]. The 90 proteins were compared all-against-all and two *precision* and *recall* values were computed for each protein as follows:

$$Precision = m / n \quad (7)$$

$$Recall = m / N \quad (8)$$

where m is the number of top n proteins from the result list that belong to the SCOP category of the selected protein and N is the number of the proteins in the relevant SCOP category. The average value of *precision* and *recall* parameters for each category of SCOP database is computed.

Figure 6 represents the experiments results for the All Alpha, All Beta, Alpha / Beta and Alpha + Beta categories of SCOP database. Due to the different number of proteins selected from the categories, the *x*-axis begins at different value of n for the categories and increases by 5 each time.

Several interesting observations can be extracted from the figures. Firstly, with increasing n, precision decreases and recall increases for all the schemes. This denotes that the number of additional picked similar proteins with increasing n is less than the number of non-similar proteins. In the other words, all of the schemes can pick the relevant structures in the low value of n. Secondly, the figures represent that the results of BLAST are not well compared with the three other schemes. Due to low sequence similarity of the dataset used in these experiments and disability of BLAST to detect structural similar proteins, these results are normally expected.

Comparing the results of the experiments for different categories shows that both SSM and *n*-gram methods perform equally well and better than SCALE method. Moreover, a further study of the figures demonstrate that *n*-gram method results are lower than SSM in the initial values of n for All Alpha, All Beta and Alpha / Beta categories. However, with increasing n, *n*-gram method produces better results than SSM relatively. Both methods SSM and *n*-gram recall also incline towards 1 for all categories except for Alpha + Beta category. This shows that both methods can pick the

[3] ftp://ncbi.nlm.nih.gov/blast/db/README

[4] http://www.ebi.ac.uk/msd-srv/ssm/ssmstart.html

relevant proteins in displayed ranges in the figures for these categories.

The results produced by all the methods are not appropriate for Alpha + Beta category in comparison with the other categories. But SSM and *n*-gram methods give better results compared with two others. The recall figure of this category shows that the results of SSM and *n*-gram methods also do not incline towards 1 in experimented range of *n*. SSM performs better than other methods for this category because it uses flexible connectivity options for secondary structure elements. Due to possibility of a match between a separated beta region of Alpha + Beta category protein with a protein from All Beta category, there may be structural similarity between these categories. Therefore, SSM be able to distinguish well the proteins belonging to this category by considering connectivity and non-connectivity information of secondary structure elements.

Another tool used for effectiveness retrieval of the schemes is the *F-measure* that is commonly applied in information retrieval systems evaluation. The *F-measure* combines *precision* and *recall* parameters into a unique value as the harmonic mean of them and defined as following:

$$F = 2 \times (Recall \times Precision)/(Recall + Precision) \quad (9)$$

The range of the *F-measure* falls between 0 and 1 that 1 is the best score. Figure 6 also represents the *F-measure* values calculated for the previous *precision* and *recall* values of the schemes for the categories. The results demonstrate that SSM and *n*-gram methods work better than two others. It is also observed that the optimal value for the *F-measure* is obtained in most of the cases in low values of *n* and it is decreased with increasing *n*.

4 Discussion

A new approach for protein structure comparison based on language modelling techniques was introduced. The major difference between the method presented here and other methods is that the method uses a symbolic representation of protein structure to do comparison procedure. Therefore, regarding the other comparison methods that generally compares the geometry of the C_α backbone atoms in 3D-coordinates, our method reduces a three-dimensional problem to a one-dimensional. This has a distinct speed advantage that needs only a textual sequence comparison algorithm and improves the time complexity of the problem into a linear function.

Although the experiments provide evidence for the high efficiency of the method compared with the similar methods, however, in the present study, we did not perform a direct comparison on running speed and CPU time. Due to several obstacles, the direct and objective comparison is hardly provided. Most of the available services use a pre-calculated alignments database or some representative structures, therefore, the number of alignments actually is different. Moreover, some of the services [12] run on a CPU cluster and employ different numbers of processors based on the complexity of the query, while incomplete information is available about the hardware and running environment.

The competitive performance of the new method is its capability in retrieval effectiveness and accuracy with respect to the other studied methods [12, 25, 26]. Experimental shown in the previous section demonstrate high accuracy and applicability of the method to search and retrieve the structural similar proteins.

Another advantage of the introduced method is its independency from any parameter setting by the user. The application of three-dimensional alignment algorithms [10, 11, 12] typically involves a number of empirical parameters and heuristic elements which could be set by the user. Choosing different values for these parameters causes naturally differences in results, so that it needs more attention to choose the best setting to reach the best possible alignment.

The implemented algorithm uses PDB file of each protein to extract its structure information. In common with other comparison and alignment methods, it is most likely to fail when an input structure is poorly defined.

5 Conclusion

The introduced method models protein structure in textual sequences in order to apply entropy concept and statistical language modeling for structural comparison and alignment of proteins. Specifically, protein structure is represented in two different sequences. The first sequence shows secondary structure elements and used for superposition of two structures. Then, the second sequence is made that represents relative residue positions in 3D-space. In the sequel, cross-entropy measure over *n*-gram model is used to capture regularities in the second sequences and compare them. Moreover, in the alignment procedure, the identical words in this sequence are marked as aligned and used to expand the alignment to other residues.

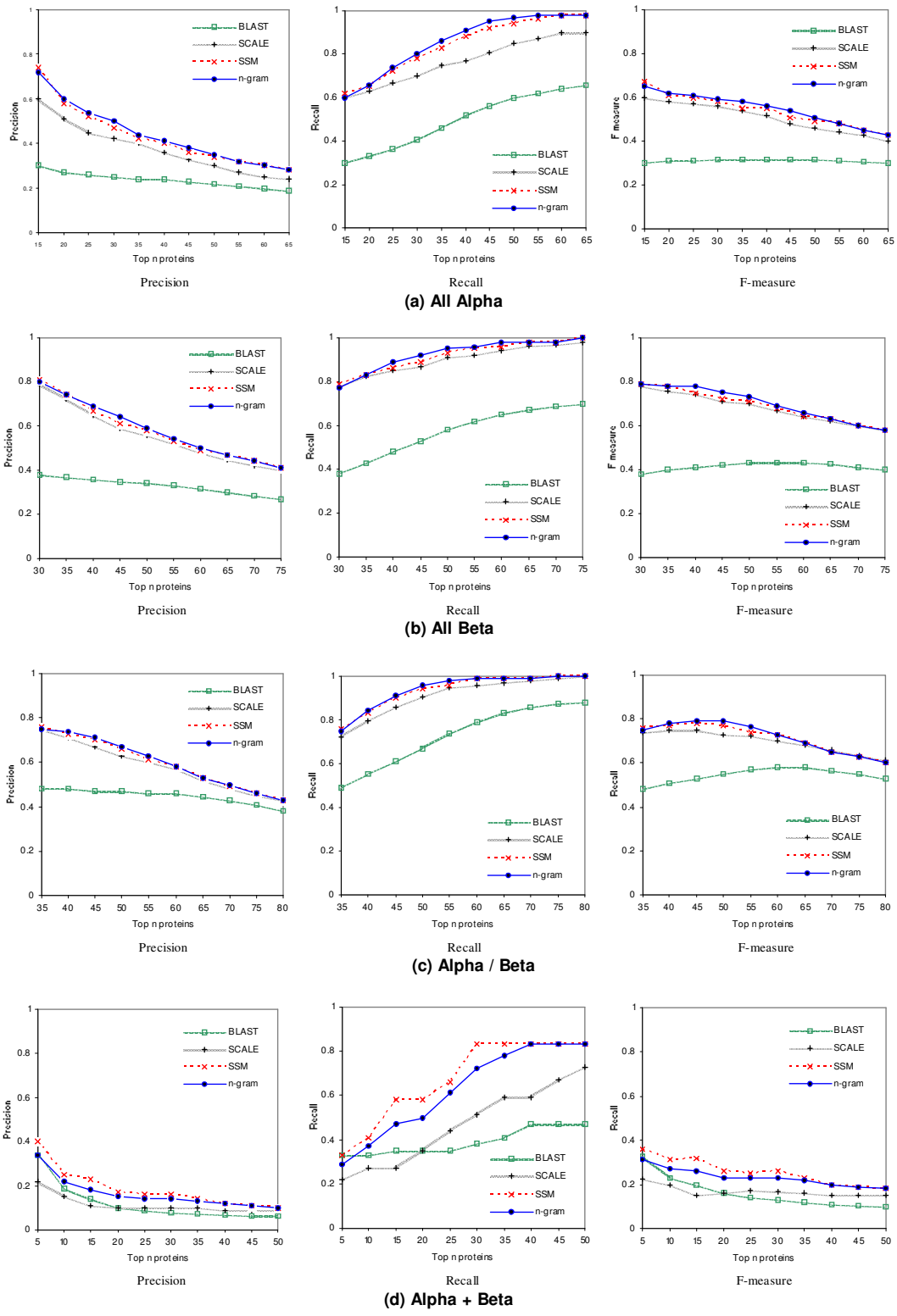

Fig.6 Precision, Recall and F-measure values computed for the categories using four different schemes

The major difference between the introduced method and other structure comparison methods is using symbolic representation of the protein structure. Therefore, the complexity of a three-dimensional problem is reduced into a one-dimensional. This has a distinct speed advantage that needs only a comparison algorithm between the sequences of the protein structures. Moreover, the results of the experiments demonstrate the applicability and reliability of this method. Finally, the conceptual simplicity of the approach motivates the future works to develop and complete powerful tools for structural similarity measurement of proteins based on language modeling techniques.

References:

[1] D. Chiu and G. Rao, *The 2-level pattern analysis of genome comparisons*, WSEAS Transactions on Biology and Biomedicine, Vol. 3, No. 3, 2006, pp. 167-174.

[2] R. Bauer, K. Rother, P. Moor, K. Reinert, T. Steinke, J. Bujnicki, R. Preissner, *Fast structural alignment of biomolecules using a hash table, n-gram and string descriptors*, Algorithms, Vol. 2, 2009, pp. 692-709.

[3] O. Camoglu, T. Kahveci, A. Singh, *PSI: Indexing Protein Structures for Fast Similarity Search*, Journal of Bioinformatics, Vol. 19, 2003, pp. 81-83.

[4] A. Aghili, D. Agrawal, A. Abbadi, *PADS: Protein Structure Alignment using Directional Shape Signature*, Proceedings of 10th Int. Conf. on Database Systems for Advanced Applications, 17-20 April 2005, China.

[5] T. Gharib, *A hybrid approach for indexing and searching protein structures*, WSEAS Transactions on Computers, Vol.8, No.6, 2009, pp. 966-975.

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

[7] J. Gibrat, T. Madej, J. Spouge, S. Bryant, *The VAST protein structure comparison method*, Biophysical Journal, Vol.72, 1997, MP 298.

[8] I. Shindyalov, P. Bourne, *Protein structure alignment by incremental combinatorial extension (CE) of the optimal path*, Protein Engineering, Vol.11, 1998, pp. 739-47.

[9] J. Jung, B. Lee, *Protein structure alignment using environmental profiles*, Protein Engineering, Vol.13, 2000, pp. 535-543.

[10] Y. Ye, A. Godzik, *Flexible structure alignment by chaining aligned fragment pairs allowing twists*, Bioinformatics, Vol.19, 2003, pp. 246-255.

[11] T. Kawabata, *MATRAS: A program for protein 3D structure comparison*, Nucleic Acids Research, Vol.31, 2003, pp. 3367-3369.

[12] E. Krissinel, K. Henrick, *Secondary-structure matching (SSM), a new tool for fast protein structure alignment in three dimensions*. Acta Cryst. Section D: Biological Crystallography, Vol.60, 2004, pp. 2256-2268.

[13] Y. Zhang, J. Skolnick, TM-align: a protein structure alignment algorithm based on the TM-score, Nucleic Acid Research, Vol.33, No.7, 2005, pp. 2302-2309.

[14] B. Kolbeck, P. May, T. Schmidt-Goenner, T. Steinke, E. Knapp, *Connectivity independentprotein-structure alignment: a hierarchical approach*, BMC Bioinformatics, Vol.7, 2006, doi:10.1186/1471-2105-7-510.

[15] R. Mosca, T. Schneider, *RAPIDO: a web server for the alignment of protein structures in the presence of conformational changes*, Nucleic Acids Research, Vol.36, 2008, W42-W46.

[16] G. Mayr, F. Domingues, P. Lackner, *Comparative analysis of protein structure alignments*, BMC Structural Biology, Vol.7, 2007, doi: 10.1186/1472-6807-7-50.

[17] R. Kolodny, P. Koehl, M. Levitt, *Comprehensive evaluation of protein structure alignment methods: scoring by geometric measures*, Journal of Molecular biology, Vol.346, 2005, pp. 1173-1188.

[18] M. Novotny, D. Madsen, G. Kleywegt, *Evaluation of protein fold comparison servers*, Proteins: Structure, Function, and bioinformatics, Vol.54, 2004, pp. 260-270.

[19] A. Bogan-Marta, A. Hategan, I. Pitas, *Language engineering and information theoretic methods in protein sequence similarity studies*. Studies in Computational Intelligence, Vol.85, 2008, pp. 151-183.

[20] C. Manning, H. Schütze, *Foundations of statistical natural language processing*, Massachusetts Institute of Technology, 2000.

[21] S. Young, G. Bloothooft, *Corpus-based methods in language and speech processing*, Kluwer Academic Publishers, 1997.

[22] S. Wang, D. Schuurmans, F. Peng, Y. Zhao, *Combining statistical language models via the latent maximum entropy principle*, Machine Learning, Vol.60, 2005, pp. 229-250.

[23] M. Ikonomakis, S. Kotsiantis, V. Tampakas, *Text Classification Using Machine Learning*

Techniques, WSEAS Transactions on Computers, Vol.4, No.8, 2005, pp. 966-974.

[24] A. Singh, D. Brutlag, *Hierarchical protein structure superposition using both secondary structure and atomic representations*, Proc. of the 5th Int. Con. on Intelligent Systems for Molecular Biology, 21-26 June 1997, Greece.

[25] C. Chionh, Z. Huang, K. Tan, Z. Yao, *Augmenting SSEs with structural properties for rapid protein structure comparison.* Proc. of the 3rd IEEE Symp. on BIBE: 10-12 March 2003, USA.

[26] S. Altschul, W. Gish, W. Miller, E. Myers, D. Lipman, *A Basic Local Alignment Search Tool*, Journal of Molecular Biology, Vol. 215, 1990, pp. 403-410.