Structural Alignment of Biomolecules by Text Modeling Techniques

JAFAR RAZMARA[1], SAFAAI B. DERIS[1], ROSLI MD ILLIAS[2]
[1]Faculty of Computer Science and Information Systems
[2]Faculty of Chemical and Natural Resources Engineering
Universiti Teknologi Malaysia
Johor Bahru
Malaysia
jrazmara@ucna.ac.ir, safaai@utm.my, r-rosli@utm.my

Abstract: - In the era of structural biology, it is necessary to apply efficient and effective tools to compare and align 3D-structure of biomolecules. Although a great number of structural comparison and alignment methods have been developed, none of them gives an exact solution to the problem. In this paper, we introduce a novel method for structural alignment of proteins based on language modelling techniques. In this way, we summarized the protein secondary and tertiary structure in two textual sequences. The first sequence is used to initial superposiotion of secondary structure elements and the second sequence is employed to align the 3D-structure of two compared structure. In order to compare sequences, the method applies a technique inspired from computational linguistics for analysing and comparing textual data. In this strategy, the cross-entropy measure over *n*-gram models is used to capture regularities between sequences of protein structures. Some experiments were performed in order to compare the performance of the method with the other structure alignment methods. The results of the experiments reported here, provide evidence for the usefulness of the new approach and its preference and applicability comparing with the other related methods.

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

1 Introduction

Protein structure alignment measure is an important tool for biologists to compare and highlight the similarities and differences between protein structures. It has wide applications in protein structure analysis and classification which have attracted enormous attention and have been studied broadly over the past decade. It is known that the protein structure highly specifies its functionality and the potential interactions with the other protein structures. Whenever a new protein structure is discovered, it is necessary to find the structural similar proteins to predict its functions and properties. For two given proteins, if they have similar sequences then their evolutionary relationship is obvious. Otherwise, the 3D-structures of proteins are much more evident than protein sequences due to structural behavior placed on them. As a result, the structural similarity measurement tool should be used to distinguish the differences among various proteins functionalities. The structural comparison of proteins algorithm proceeds to find the optimal correspondence between the atoms in two structures with a minimal distance between the matched pairs [1]. The algorithm needs an exhaustive heuristic search to find the best corresponding atoms of one structure to the other.

Several approaches to protein structure alignment have been explored over the past decade and are available currently for the structural biologists. The basics of the methods are different and have been designed for diversity of applications; however, their algorithms generally compare the geometry of the C_α backbone atoms. In the recent years, a number of comparative and comprehensive studies have been reported for evaluation of most popular protein structure alignment methods such as DALI [2], VAST [3], CE [4], MATRAS [5], FATCAT [6], SSM [7], and so on. To review the comparative results of these studies refer to Kolodny et al. [8] and Myre et al. [9]. Although a considerable number of structural comparison and alignment tools have been proposed, the study for developing new alternative efficient and reliable methods is still an active research area.

Biological data, naturally, can be represented as textual sequences. The language of biology has 20 distinct symbols of alphabets called amino acids. This fact has opened new perspective in the evolution of bio-molecules research. Consequently, it is not a surprise to apply *statistical language modeling* and *text classification techniques* in biological sequences analyzing.

In this paper, a novel method for protein structure alignment based on statistical text modeling is proposed. The method is inspired by the successful use of entropy concept for information retrieval in the field of statistical language modeling [10, 11]. *N*-gram modeling is also a preferable concept to any formal linguistics approach [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]. Based on the fruitful results of this attempt, we now extend this approach to protein structure alignment.

2 Methods

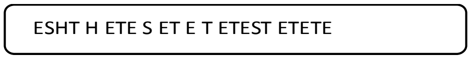
2.1 Protein Structure Modeling in Sequence Form

In order to reduce the complexity of the protein structure comparison and alignment problem, most of the methods use a simplified representation of these macromolecules. The introduced method in this paper uses a textual representation of protein structure. Specifically, two different sequences of alphabets summarize protein structure in a hierarchical form.

The first sequence is a string of letters that denotes the type of each secondary structure elements (SSEs). Different types of these regular substructures are encoded by alphabetic characters and represented in table 1. For example, a sample sequence of SSEs for 1CRB PDB chain extracted from PDB site is represented in Figure 1.

Table 1. Symbols defined for secondary structure elements

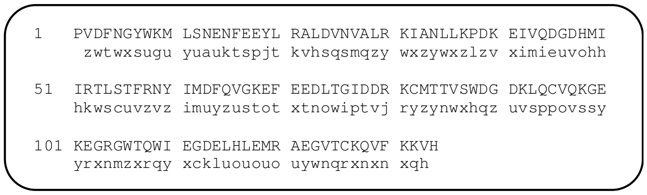
3-turn helix (3_10 helix). Min length 3 residues	‘G’
4-turn helix (alpha helix). Min length 4 residues	‘H’
5-turn helix (pi helix). Min length 5 residues	‘I’
hydrogen bonded turn (3, 4 or 5 turn)	‘T’
beta sheet in parallel and/or anti-parallel sheet conformation (extended strand)	‘E’
residue in isolated 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 is the string representation of protein 3D-structure that we introduced in [13]. In this model, the position of each residue in 3D-coordinates with respect to the position of its previous residue is labeled by specially defined alphabets. Hence the 3D-structure of a protein can be modeled in a sequence of letters (for more details refer to [13]). Figure 2 shows a sample sequence of protein 3D-structure representation for 1CRB PDB chain. In this figure the extracted sequence for protein 3D-structure is represented beside the amino acids sequence.



```
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 Amino acids sequence and relative residue position sequence extracted for 1CRB PDB chain.

Accordingly, the protein structure is simplified in two hierarchical sequences: secondary structure elements sequence and 3D-structure sequence that from now on words called as relative residue position sequence.

2.2 Text Modeling by N-Gram Method

Having reduced the protein structure to two hierarchical textual data, now, a language modeling technique can be applied to compare and align these strings. Several kinds of language modeling techniques have been developed for the textual data processing and manipulation. Hidden Markov Model (HMM) is one of the highly used fundamental language models that combine statistics and information theory. It assumes that the observation of a word W_k at a position k in a given text depends only upon its immediate n predecessor words W_{k-n} ,… ,W_{k-1}. The model, usually called as n-gram, has been more popular and widely used in formal linguistics approaches due to its simplicity [12]. *Entropy* is also a useful concept in the quantification of information in a textual sequence and making connection with probabilistic language modeling. A specific definition of entropy called cross-entropy is relevant tool for comparison of two textual sequences. In this tool, the *n*-gram model is firstly made by the word-counts of one protein sequence and then the predictability, of the second sequence, by the model is measured via the formula:

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

The variable X is in the form of *n*-gram $w_i^n=\{w_i,w_{i+1},\dots,w_{i+n-1}\}$, $Count(w_i^n)$ is the number of occurrences of *n*-gram w_i^n and N is the total number of *n*-grams in the sequence. All of the possible *n*-length combinations of consecutive w_is (i.e $W^*=\{\{w_1,w_2,\dots,w_n\},\{w_2,w_3,\dots,w_{n+1}\},\dots\}$) are computed in the summation part of the formula. The second term in the summation computes the conditional probability of the *n*-th element of an *n*-gram with the preceding *n*-1 elements and can be estimated by a counting procedure.

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

The term $p(w_i^n)$ refers to the reference protein sequence and results from counting the words of that specific protein. The term $P_M(w_{i+n}|w_i^{n-1})$ computed by counting the words of the query protein sequence which the model has to be estimated. The range of variable X is all of the *n*-grams of the reference protein sequence.

As we described in [13], a modified version of cross-entropy is useful to prevent loss of data when all of the words have been counted once and the probability by $P_M(w_{i+n}|w_i^{n-1})$ become zero. Therefore, following equation is used by the method in experiments:

$$H(X,P_M) = -\Sigma_{all\ w*}\ p(w_i^n) \log_2 (2+P_M(w_{i+n}|w_i^{n-1})) \quad (3)$$

The introduced method in [12] and [13] firstly represents both the unknown query-protein and each protein in a reference database in the *n*-gram form and the cross-entropy measure is utilized to compare them. A typical implementation of this idea, called Direct method, 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:

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

In the above equation S_q and S_i denote the query and *i*-th reference proteins respectively. Finally, the most similar protein in the database to the query-protein is easily identified as the one having the lowest $D(S_q,S_i)$. *Alternating* method is another implementation of the idea that considers the protein with the shortest sequence as reference sequence when comparing the query protein with each database-protein. This is devised in order to cope with the more different length of the proteins to be compared and gives more accurate results with respect to Direct method as reported in [12] and [13].

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 pair of proteins may not correspond. Consequently, it is necessary to find an initial superposition between two structures to make them comparable. Our method applies above introduced *n*-gram modeling over secondary structure elements sequence in order to find the matched pairs of SSEs and then, makes a rotation-transformation matrix using SSE vectors to achieve an initial overlap between two protein structures. Specifically, secondary structure elements sequence is represented in n-gram form and the corresponding words are found. Then, the SSEs inside the matched words are marked as aligned. Moreover, the algorithm expands the alignment on similar pairs of SSEs that are outside of the matched words but they are located consequently between the matched SSEs. In the sequel, angles and distances between the vectors of the aligned SSEs are computed and used to make and employ a rotation-transformation matrix over one of the two structures to superpose initially two structures. The method also applies the introduced scheme in [7] and [14] for vector representation of SSEs.

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 applies 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 matched 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 (4) 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:

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()
    $D_t[i] = |H(T_q, T_i) - PS_t|$
```

In the above 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.

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 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 proteins are compared all-against-all by the above introduced method.

Figure 3 represents the matrices containing all the measured dissimilarities D(Si, Sj), 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. 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 a white matrix with only a black diagonal segment 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.

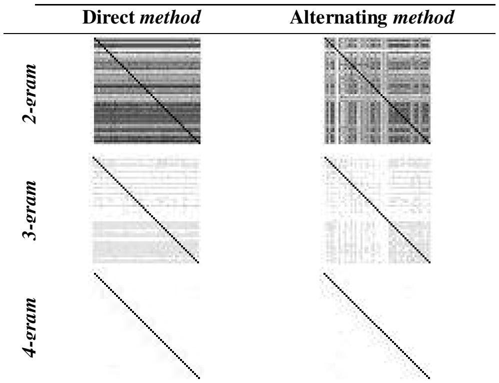


Fig.3 Gray-scale representation of the output matrices containing all the possible pairwise dissimilarities for 53 proteins using Direct method.

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 output of CE 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 4.8Ǻ for RMSD whereas n-gram method aligns 138 residues with an RMSD of 5.7Ǻ.

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 frequently used indicators of the quality of a structural alignment. 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.4 Structural alignment between 1AKT:_ and 1CRP:_ 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.

4 Conclusion

The introduced method models protein structure in textual sequences in order to apply entropy concept in the field of statistical language modeling for structural 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.

The major difference between the introduced method and other structure comparison methods is using poorly 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.

Acknowledgment

We would like to thank our research grant sponsor, Malaysian Ministry of Science, Technology and Innovation (MOSTI) and Universiti Teknologi Malaysia Research Management Centre (RMC) for their support (research grant number: 01-01-06-SF0300).

References:

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

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

[3] JF. Gibrat, T. Madej, JL. Spouge, SH. Bryant, The VAST protein structure comparison method, *Biophysical Journal*, Vol.72, 1997, MP298.

[4] I. Shindyalov, P. Bourne, Protein structure alignment by incremental combinatorial extension (CE) of the optimal path, *Protein Engineering*, Vol.11, 1998, pp. 739-747.

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

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

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

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

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

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

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

[12] A. Bogan-Marta, I. Pitas, K. Lyroudia, Statistical method of context evaluation for biological sequence similarity, *IFIP International Federation for Information Processing*, Vol.217, 2006, pp. 99-108.

[13] J. Razmara, S.B. Deris, A novel method for protein 3D-structure similarity measure based on n-gram modeling, *IEEE Int. Conf. Bioinformatics and bioengineering*, 2008.

[14] A. P. Singh, and D. L. Brutlag, Hierarchical protein structure superposition using both secondary structure and atomic representations, *Proc. of the 5th Int. Conf. on Intelligent Systems for Molecular Biology*, 1997.