Proceedings of the 28th IEEE
EMBS Annual International Conference
New York City, USA, Aug 30-Sept 3, 2006

A New Seed Selection Algorithm that Maximizes Local Structural Similarity in Proteins

Gulsah Altun, Wei Zhong, Yi Pan, Phang C. Tai, Robert W. Harrison

Abstract—**The PHI-BLAST algorithm for protein sequence alignment takes a query sequence and searches a protein database for a small seed or region of high similarity and extends this alignment to produce the total alignment for sequences. Clearly, the success of this approach depends on the quality of the seeds. We propose an algorithm that maximizes the likelihood of seeds sharing the same local structure in both the query and known sequences. This was tested on the 2290 protein sequences in the PISCES database. Our new algorithm results in an effective a priori estimate of seed structural quality.**

I. Introduction

Known protein sequences and structures are increasing and the protein databank is growing rapidly. Predicting the structure of protein from its amino acid sequence is one of the unsolved problems of bioinformatics. All homology methods and many ab initio methods assume that similar sequences have similar structures [3][10][14]. Recent work suggests that finding short contiguous or patterned matches called seeds or words can be extended to find alignments [13]. Similarity searches based on the strategy of finding short seed matches have been studied widely and many programs have been developed. One of the most popular programs is BLAST (for "basic local alignment search tool"), which has been cited over 10000 times over the last decade and BLAST server receives about 100000 hits per day [1][12].

Given a query protein or DNA sequence along with a pattern (query sequence) occurring within the sequence, the Pattern Hit Initiated BLAST (PHI-BLAST) program searches a protein database for other instances of the query sequence in order to build local alignment [20]. This is because of the assumption that a good alignment is likely to contain high-scoring pair of seeds. Many methods have been proposed to find more optimal seeds by using gapped alignments or position scoring specific matrixes [2][3][4][5][6][7][8][9][15][18][19]. However, some of these methods select seeds by scanning each sequence window of a given size k in the database one by one, which can result in many false positives due to the large number of sequence windows in a protein database.

Therefore, it is crucial to evaluate the factors in selecting seeds to minimize the number of false positives. In this work, we explore the reliability of z-score statistics when used on sequence vs. profile, profile vs. profile and profile vs. clustered profile approaches to define seeds.

Sequence vs. profile methods use a single profile for one sequence and use the second sequence to select scores from the profile. For example, PSI-BLAST derives a profile sequence alignments and uses the query sequence to find the score [2]. In profile vs. profile methods, the two profiles are compared. For example, Fold and Function Assignment System (FFAS) server uses the dot product of the two profiles when aligning protein sequences [7]. Neither, sequence vs. profile or profile vs. profile methods has any means of assessing the statistical significance of the profile.

Clustering the profiles as a preprocessing step extracts profiles that are conserved in sequence space and thus likely to correspond to conserved structure or function in the proteins. Profile vs. Clustered profile algorithm, suggested in this work, can take advantage of this statistical significance. The sequence clusters can be assigned a quality based in their internal statistical consistency and this quality strongly correlates with the structural similarity in the proteins that contain them.

II. Experimental Setup

The dataset used in this work includes 2290 protein sequences obtained from the Protein Sequence Culling Server (PISCES) [17][18]. No protein sequences of this database share more than 25% sequence identities in this database. The sliding windows with nine successive and continuous residues are generated from protein sequences. The width of nine residues was chosen to be representative of the size of protein folding motifs. While the optimal sizes are not constant and are both larger and smaller than nine residues, this is a useful approximation and removes sample size bias from the analysis. The frequency profile from a database of homology-derived secondary structure of proteins (HSSP) is constructed based on the alignment of each protein sequence from the Protein Data Bank (PDB) in which all the sequences are considered homologous in the sequence database [11][16]. HSSP profiles of each window for sequence profiles are used. Using the sliding window technique, 500,000 sequence windows are generated. Each sequence window is represented by either the amino acid residue or the 9x20 HSSP profile matrix depending on the method applied. Twenty columns represent 20 amino acids

Manuscript received April 3rd, 2006. This research was supported in part by the U.S. National Institutes of Health under grants R01 GM34766-17S1, and P20 GM065762-01A1, and the U.S. National Science Foundation under grants ECS-0196569, and ECS-0334813. This work was also supported by the Georgia Cancer Coalition (GCC) and the Georgia Research Alliance.

and 9 rows represent each position of the sliding window.

III. METHODS

A. Sequence vs. Profile Algorithm

In Sequence vs. Profile algorithm, each sequence window in the database is represented by its frequency profile produced by the multiple sequence alignment. However, the query sequence is represented by its amino acid residues only. The scores were calculated for window width of 9 residues. Z-scores were used to place the results in a constant scale with respect to the standard deviation. Thus two samples with similar z-scores have similar statistical significance. The formula to calculate the score for a sequence window of size 9 is given in the following:

$$z-score=\sum_{i=1}^{9}\frac{Freq_i-Avg_i}{Std_i}=\sum_{i=1}^{9}individual\ z-score$$

$Freq_i$: The frequency of the i^{th} amino acid of the sequence window in the sequence profile database
Avg_i: The average value of the the i^{th} amino acid in the entire database.
Std_i: The standard deviation value of the i^{th} amino acid in the entire database.

After each sequence window in the database is assigned a z-score, the sequence window, which receives the highest z-score after the comparison process, is considered to be the best match for the given query. In Section IV, the secondary structure similarity between the given query and its best match are compared.

B. Profile vs. Profile Algorithm

In Profile vs. Profile algorithm, a given query amino acid sequence window is represented by the frequency profile rather than its amino acid sequence representation as in Sequence vs. Profile method. The sequence window in the database with its frequency profile closest to the frequency profile of a given amino acid sequence window is considered to be the best match for the Profile vs. Profile method.

$$Avg=\frac{\sum_{i=1}^{N}score_i}{N}\qquad Std=\frac{\sum_{i=1}^{N}(score_i-Avg)^2}{N}$$

$$z-score=\frac{score_i-Avg}{Std}$$

$score_i$: The score assigned to the i^{th} sequence segment in the sequence profile database.
N: number of scores

In Section IV, the secondary structure similarity between the given query and its best match are compared.

C. Profile vs. Clustered Profile Algorithm

In Profile vs. Clustered Profile algorithm, we propose a cluster based approach which is different from the previous two methods. In this algorithm, initially, all the sequence windows in the database are classified into different sequence based clusters by K-means clustering algorithm [21]. We used K-means algorithm because it produces many high quality clusters and it is an efficient way to cluster a huge dataset such as PISCES [21]. After all sequence windows are clustered based on their sequence similarity using HSSP profiles, each cluster was assigned an average profile that represents that cluster.

After finding the clusters, each cluster was ranked based on the secondary structure similarity of each sequence window that they contain. Based on this ranking the clusters were divided into high quality clusters, average quality clusters and low quality clusters. A cluster was ranked as high quality if at least 70% of the sequence windows that the cluster contains shared more than 70% secondary structure similarity. Similarly, if only at most 70% of the sequence windows had 70% secondary structure similarity, the cluster was ranked as average cluster. If no more than 30% or less sequence windows shared more than 70% secondary structure similarity, the cluster was ranked as a bad cluster.

For a given query sequence window, when a cluster had a average frequency profile that is closest to the profile of the given query, then that cluster's frequency profile is considered to be best match of the given query sequence. In Section IV, the secondary structure similarity between the given query and its best match cluster profile are compared.

IV. EXPERIMENTAL RESULTS

Using sliding window technique, we generated 6507 sequence windows (~%1 of the PISCES) to search for seeds from randomly selected proteins. These windows were removed from the database to prevent any bias when sequences were alike. This was a good proportion for searching for seeds because having more sequence windows would generate many matches in the database. For all our tests, these 6507 sequence and profile windows are used as the search queries. Seeds were selected by using the algorithms described in Section III. These seeds were scanned against the 2290 protein sequences in the PISCES in order to find their best match out of 500,000 unique 9-mers (sequence window of size 9) in the PISCES database.

A. Seed Selection Results using Sequence vs. Profile and Profile vs. Profile Algorithms

The results for Sequence vs. Profile and Profile vs. Profile methods are almost similar as it can be seen in Fig. 1(a) and 1(b), respectively. In both of the methods, when the optimal alignment over the entire database was found the probability of a significant structural similarity was low. This would correspond to the probability of a seed used by PHI-BLAST which was a structurally accurate homolog. It is clear to see the most of the seeds found have less than 70% structural similarity with its best match. These results indicate that Sequence vs. Profile and Profile vs. Profile methods can not find seeds that would lead to a good sequence alignment.

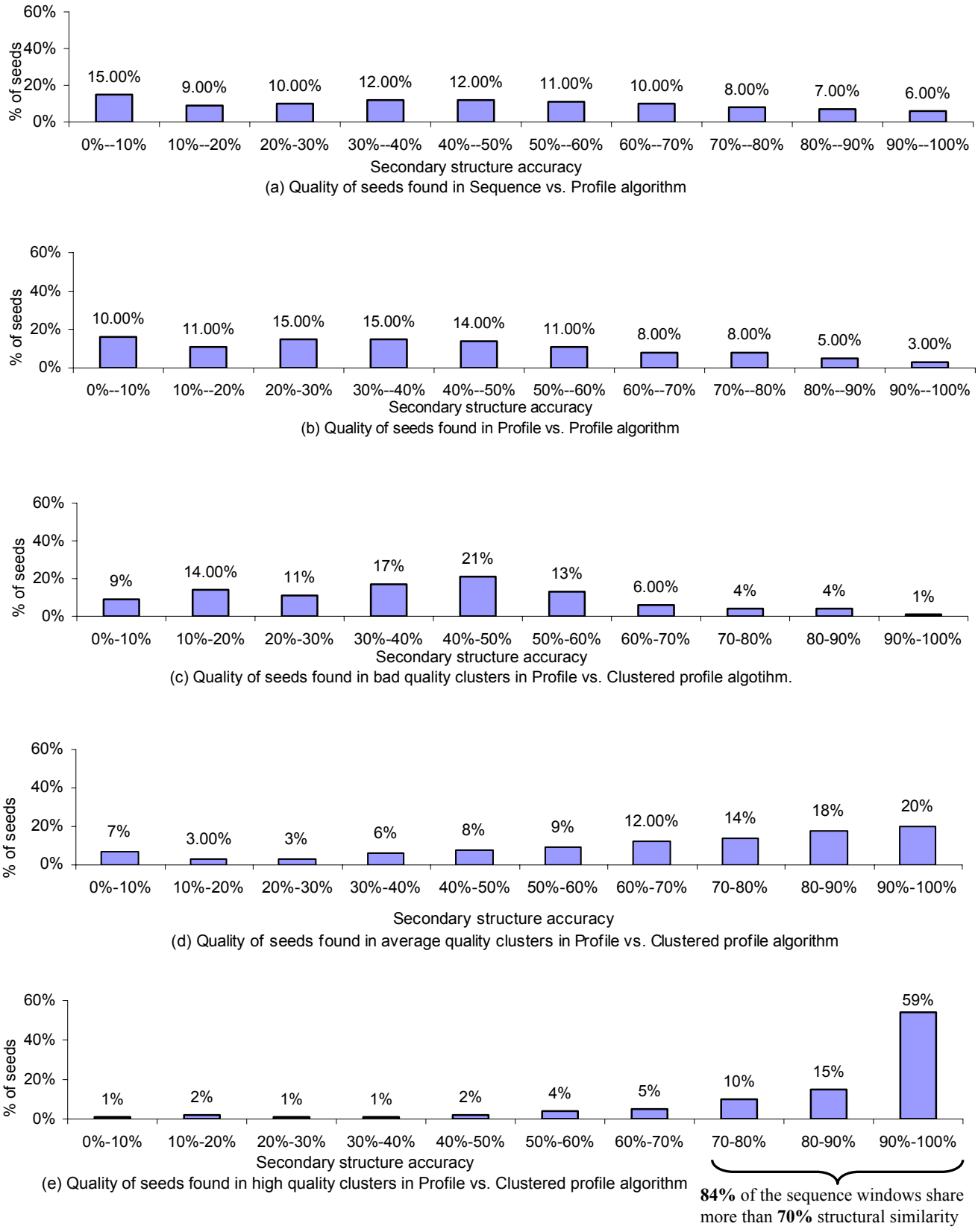

Fig. 1. Seed selection results of Sequence vs. Profile, Profile vs. Profile and Profile vs. Clustered profile algorithms

B. Profile vs. Clustered Profile Algorithm

Neither Sequence vs. Profile nor Profile vs. Profile methods could select seeds that could reflect local structural similarities. However, when the profiles are clustered prior to the search, significant structural similarity between the seeds and their best match are found when Profile vs. Clustered Profile algorithm is used. Based on previous work, [21] we used 800 clusters and ranked each cluster as specified in the algorithm. Out of these 800 clusters, 345 clusters were ranked as high quality clusters and average quality clusters.

Fig. 1(c) and Fig. 1(d) show that only 9% and 52% of sequence windows share above 70% structural similarity in bad sequence clusters and in average clusters, respectively.

On the other hand, as it can be seen from the Fig. 1(e), high quality clusters were able to select sequence windows with very high structural similarity where 84% of sequence windows share above 70% structural similarity with the average cluster structure. These results show that Profile vs. Clustered Profile algorithm can select seeds that have high structural similarity with the average cluster structure when high quality clusters are used.

V. Conclusion

In this study, the factors involved in the accurate selection of seeds for protein sequence alignments were explored. It is possible to identify seeds that are likely to share structural similarity with a meaningful *a priori* assessment of accuracy by using a profile-clustered profile approach.

We used high order information identified by clustering and showed that it is reliable in small scales. We found that look-up of this clustered sequence-based seeds for the best match works much better than look-up of individual frequency profile of each seed in the database. The predictive ability of these clusters suggests that there are distinct sequence-structure seeds. The dramatic improvement found by using high quality clustered profiles shows that higher order descriptions of sequence similarity are required for accurate results in the prediction of protein structure. This suggests that PHI-BLAST like algorithms can be substantially improved if the database is clustered first. Our results show that it is possible to select seeds when sequence windows are clustered and average profiles of these clusters are used for calculating similarity measure.

Acknowledgment

The authors would like to thank Professor Roland L. Dunbrack for providing the dataset from PISCES. This research was supported in part by the U.S. National Institutes of Health under grants R01 GM34766-17S1, and P20 GM065762-01A1, and the U.S. National Science Foundation under grants ECS-0196569, and ECS-0334813. This work was also supported by the Georgia Cancer Coalition (GCC) and the Georgia Research Alliance. Gulsah Altun and Wei Zhong are supported by Georgia State University Molecular Basis of Disease Fellowship. Dr. Harrison is a GCC distinguished scholar.

References

[1] S. F. Altschul, W. Gish, W. Miller, E. W. Myers and D. J. Lipman, "Basic local alignment search tool", *Journal of Molecular Biology*, vol.215, no.3, pp. 403-410, 1990.

[2] S. F. Altschul, T.L. Madden, AA Schaffer, J. Zhang, Z. Zhang, W. Miller and D.J. Lipman, "Gapped BLAST and PSI-BLAST: a new generation of protein database search programs", *Nucleic Acids Research*, vol. 25, no.17, pp. 3389-3402, 1997.

[3] S. Burkhardt and J. Kärkkäinen, "Better Filtering with Gapped q-Grams", *Proceedings of the 12th Annual Symposium on Combinatorial Pattern Matching,* pp. 73-85, 2001.

[4] J.M Claverie and L. Bougueleret, "Heuristic informational analysis of sequences", *Nucleic Acids Research.* vol. 14, no.1, pp. 179-196, 1986.

[5] M. Gribskov, S. Veretnik, "Identication of sequence patterns with profile analysis", *Methods in Enzymology,* vol. 266 no. 13. pp. 198-212,

[6] V. Gotea, V. Veeramachaneni, W. Makalowski , "Mastering seeds for genomic size nucleotide BLAST searches", *Nucleic Acids Research*, vol. 31 no.23, 2003.

[7] L.Jaroszewski, L. Rychlewski, Z. Li, W. Li and A. Godzik, "FFAS03: a server for profile-profile sequence alignments", *Nucl. Acids Res.* vol. 33, pp. 284-288, 2005.

[8] M. Li, B. Ma, D. Kisman and J. Tromp, "PatternHunter II: Highly Sensitive and Fast Homology Search", *Journal of Bioinformatics and Computational Biology,* vol. 2, no.3, pp. 417-440, 2002.

[9] B. Ma, J. Tromp and M. Li, "PatternHunter: Faster and More Sensitive Homology Search", *Bioinformatics*, vol. 18. pp. 440-445, 2002.

[10] A. Pol and T. Kahveci, "Highly Scalable and Accurate Seeds for Subsequence Alignment", *In the proceedings of BIBE 2005.*

[11] D. Przybylski and B. Rost, "Alignments grow, secondary structure prediction improves", Proteins, vol. 46, no. 2, pp. 197-205, 2002.

[12] T. Przytycka, R. Srinivasan and G.D.Rose, "Recursive domains in proteins", *J. Biol. Chem.,* vol. 276, no. 27, pp. 25372-25377, 2001.

[13] A.R Panchenko and S. Bryant, "A comparison of Position-Specific Score Matrices based on sequence and structure alignments, *Protein Science*, vol. 11, pp. 361-370, 2002.

[14] B. Rost and C. Sander,"Prediction of secondary structure at better than 70% accuracy," *Journal of Molecular Biology*, vol. 232. pp. 584-599, 1993.

[15] B. Rost, C. Sander, R. Schneider, "Evolution and neural networks - protein secondary structure prediction above 71% accuracy," *27th Hawaii International Conference on System Sciences*, Hawaii, U.S.A, Los Alamitos, CA, 1994.

[16] N. von Öhsen, I. Sommer and R. Zimmer, "Profile-Profile Alignment: A Powerful Tool for Protein Structure Prediction", *Pacific Symposium on Biocomputing,* vol. 8. pp. 252-263, 2003.

[17] C. Sander and R. Schneider, "Database of similarity-derived protein structures and the structural meaning of sequence alignment", *Proteins: Struct. Funct. Genet.*. vol. 9 no. 1 pp. 56-68, 1991.

[18] G. Wang and Jr. R.L. Dunbrack, "PISCES: a protein sequence-culling server", *Bioinformatics,* vol. 19, no. 12, pp. 1589-1591, 1993.

[19] J. Xu, D. Brown, M. Li and B. Ma, "Optimizing Multiple Spaced Seeds for Homol ogy Search", *Journal of Computational Biology*. Accepted, Nov. 2004.

[20] Z. Zhang, A.A. Schäffer, W. Miller, T.L Madden, D. J. Lipman, E.V. Koonin, S.F. Altschul, "Protein sequence similarity searches using patterns as seeds", *Nucleic Acids Research*, vol. 26, no. 17, pp. 3986–3990, 1998.

[21] W. Zhong, G. Altun, R. Harrison, P.C. Tai, Y. Pan, " Improved K-means Clustering Algorithm for Exploring Local Protein Sequence Motifs Representing Common Structural Property", *IEEE Transactions on Nanobioscience,* vol. 4, no. 3, pp. 255-265, 2005.