ISSN No. 0976-5697

Volume 3, No. 2, March-April 2012

International Journal of Advanced Research in Computer Science

RESEARCH PAPER

Available Online at www.ijarcs.info

Identification of Similar H1N1 cases in other parts of the world using SRLCS model Combined with Fuzzy Membership

Sumathy Eswaran*
Research Scholar, Vels University
& Asst.Prof., Dr.MGR University
Chennai, India
sumathyesh@yahoo.com

Dr.S.P.Rajagopalan
Professor Emeritus, Dr.MGR Educational
and Research Institute
Chennai, India
sasirekaraj@yahoo.co.in

Abstract: Influenza Virus Research facilitates the understanding of the Influenza virus and how it interacts with the host organism , leading to new treatments and preventive actions. The pandemic of 2009 , H1N1 influenza was deadly one. A model analysis on H1N1 of 2009 will provide a strategy for future. Using pair wise sequence analysis, the similarity and identity analysis are revealed by the SRLCS Model. Expressing this revelation in a useful form helps in homology analysis or identification of similar pandemic virus occurrences elsewhere. A fuzzy membership calculator places the sequence analysis result in membership form explaining the similarity and homologous relationship. Biologists can use this information to interpret the course of medication. Also this result can be useful to cure biosequence databases.

Keywords: H1N1 analysis, curation of data base, Homology, Fuzzy membership, sequence analysis

I. INTRODUCTION

Initial characterization of any new DNA or protein sequence starts with a database search aimed at finding out whether homologs of this gene (protein) are already available, and if they are, what is known about them. Looking for exactly the same sequence is quite straightforward. In principle, the only way to identify homologs is by aligning the query sequence against all the sequences in the database, sorting these hits based on the degree of similarity, and assessing their statistical significance that is likely to be indicative of homology. Thus ab initio methods rely on the statistical parameters in the sequences for homologous protein or Gene identification. Although there are many programs available and each one providing solution in its own method having advantages and disadvantages, none is perfect [1]. Finding close relatives is also a conceptual problem because a sequence with 8% identity could become an ortholog having same function [2]. A priori knowledge of the location of the particular residue in the protein structure is required in such cases. Even if such knowledge is available it is a complex task to incorporate that in Database search.

Therefore this paper presents a membership identification of a target protein in a data base with accuracy of 1/1000. Such a representation rightly provides the indicative of homology in terms of membership value. This paper analyses the approach using data from Influenza Research Data base [3].

II. RELATED WORK

Alignment of sequences can be categorized as Global or Local Alignment. Pairwise Optimal global alignment of two sequences was first realized in the Needleman-Wunsch algorithm [4], using dynamic programming technique. Smith and Waterman Algorithm [5] based on Dynamic Programming realizes Local Alignment. For both the algorithms, the time and Memory requirement is $O(n^2)$. For Multiple Sequence alignment, Dynamic programming based algorithms have complexity $O(n^k)$, where k is the number of sequences to be compared. This becomes intractable for large value of k. Thus heuristic approaches like MUSCLE [6] , BLAST [7], FASTA [8] , have come into practice. These are popular because of the convenience for use and availability thru many servers. There are other new algorithms for Sequence alignment like Time Horizon Specialised Branching Heuristic (THSB) [9], Ant Colony Optimization (ASO) [10], Beam Search[11]. Heuristic algorithms provide only suboptimal solution and are acceptable because otherwise the problem may be an intractable one. With the advancement in computing, parallel algorithms allow biologists work on larger biosequences. Parallel algorithms can divide the problem and hence can handle computational complexity to a large extent. FAST_LCS algorithm proposed by Wan, Liu & Chen [12], Quick DP MLCS by Wang et al, [13] and Efficient Fast Pruned LCS (EFPLCS) algorithm [14], are some of the parallel algorithms. MLCS APP [15], SRLCS [16] are heuristic parallel algorithms for the same purpose.

For sequence alignment FastLCS complexity is $O(|LCS(X,Y)|)$ for time complexity and $\max\{4*(n+1)+4*(m+1), L\}$ for space complexity . EFP LCS is 70% more efficient than FASTLCS in resource utilization of both memory and CPU [14]. SRLCS model [17] is a simplistic method to find the Longest Common subsequence (LCS) between given two sequences on a pair wise method.

Homology modeling involves many steps like template identification, amino acid sequence alignment, alignment correction, backbone generation, generation of loops, side chain generation & optimization, ab initio loop building, overall model optimization and model verification including quality [18]. *Ab initio* methods of sequence alignment and LCS identification are useful in obtaining information about indicative homology. Therefore any sequence alignment algorithm can be used upto sequence alignment, so that selective homology model building and verification can be done. The author uses SRLCS Model [17] for template identification. H1N1, the pandemic influenza virus of 2009 data set is used. The H1N1 influenza has seven protein segments namely PB2, PB1, PA/P3, HA/HE, NP, NA, MP and NS. Biologists anlayse these segments to learn about the antiviral drug sensitivity predications, virulence determinants, transmission factors and immune epitope analysis.

Data was obtained from the NIAID IRD online from the website http://www.fludb.org which maintains Influenza Research database of various Influenza virus [3]. The Database gets updated from Genbank entries on daily basis. Influenza virus data from all parts of the world is available. Strains with all the segment details available alone were considered. The data pertaining to India is taken for study.

III. METHOD

SRLCS Model [17] finds the probable length of Longest Common Subsequence (LCS) between a target sequence and template sequences using the identity and similarity percentage between the sequences. This is done on a pair wise basis. Pair wise is a better approach for the problem in hand as the problem is to find similar cases of H1N1 reported in other parts of the world. This is fed to a Membership calculator to identify the membership of the target with each sequence in database as in figure.1.

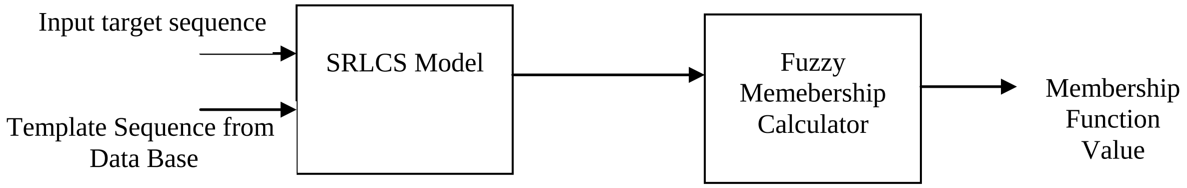
Input target sequence
Template Sequence from Data Base
SRLCS Model
Fuzzy Memebership Calculator
Membership Function Value

Figure 1. Identification of Membership Function of a target sequence

The ratio of the obtained Length of LCS and the length of target sequence is described as agreement factor *p* and is calculated as

$$p = \frac{|LCS|}{|Query|} \qquad (1)$$

The Fuzzy membership function M derives the degree of agreement between the target protein and the template proteins and is defined by the following formula:

$$\text{MemberShip Function } M = \begin{cases} 1 \mid p > 1.0 \\ 0 \mid p < 0 \\ p \mid 0 < p < 1 \end{cases} \qquad (2)$$

The Membership Function M can take a value between 0 and 1. A value closer to 1is expected to a more similar to target and value closer to 0 is expected to be dissimilar or improper case of H1N1.

IV. RESULTS

A. *Identification of Similar H1N1 Strains:*

Segments PB2 (GU292362), PB1(GU292368), HA (GU292354) and NA(GU292386) of Strain A/Pune/NIV8489/2009 were used as target protein for identification of similar samples in rest of the part of India. IRD had 28 samples form India which were complete strains with all protein segment details and these were used as samples. The membership function value M of other samples with reference to Pune sample identified segment wise. Figure.2. shows the fuzzy membership value of each sample. The values corresponding to PB1, PB2, HA and NA segments are shown in different colors. The sample name is in X axis. It is observed that most of the samples are having similarity membership >0.99 with target Pune sample implying all cases are similar. However the small difference in membership of <0.009 could be interest to relate to the drug analysis or the treatment required. For example , The Pandemic (H1N1) 2009 viruses are predicted to be sensitive to oseltamivir (Tamiflu) and zanamivir (Relenza) based on the observation that these sequences carry a histidine at position 275 of NA. These viruses are however, predicted to be resistant to adamantanes based on the observation that these sequences carry an asparagine at position 31 of M2.[21]

This same observation is reflected in BLAST analysis done at "www.fludb.org". Comparison of SRLCS model M-value and SSEARCH35 [19] is tabled in Table1. From table 1, it is seen that M value calculated by the suggested method is truthful and is in readable form.

Table I. H1N1 protein segments similarity identification using Fuzzy membership and SSEARCH

Protein Segment Name	*M value range*		*SSERACH35 S-W Opt range*		*SSEARCH35 e-value range*	
	Max	Min	Max	Min	Max	Min
PB1	1.00	0.997	5015	4936	0	0
PB2	1.00	0.993	4912	4784	0	0
NA	1.00	0.994	3289	3263	8.00E-151	1.20E-149
HA	1.00	0.991	3795	3739	0	0

B. *Curation of Database:*

Another observation is made with Query: gb:JN600356|gi:345285356|Organism:Influenza, 566 aa whose details are Organism:Influenza A virus A/Assam/2220/2009|Protein Name: HA Hemagglutinin |Gene Symbol: HA|Segment:4|Subtype:H1N1|Host:Human. This was analysed with 9 other H1N1 HA protein segment and 1 H3N2 HA protein segment samples from India in Influenza Research Database. The result is as in Figure.3. Four of them had membership value almost nearing 1 indicating close homology and classified as True Positives (TP). However 6 other samples which had sequence length of 32 as against the 566 of query protein have membership value less than 0.2 and classified as TP only to the extent of membership. Normally such a low membership value should be concluded as distant homology but considering the length being so short for the sequence it is possible that those sample sequences are not curated in the database. The last sample is not from H1N1 but from H3N2 yielding zero membership value with sample H1N1 Ha. This is classified as true negative (TN).

Thus the SRLCS model combined with Fuzzy membership logic can also be used for curating the database and also for identifying distant homologs. Distant homolog may have lower membership value whereas the close homolog will have higher value of membership.

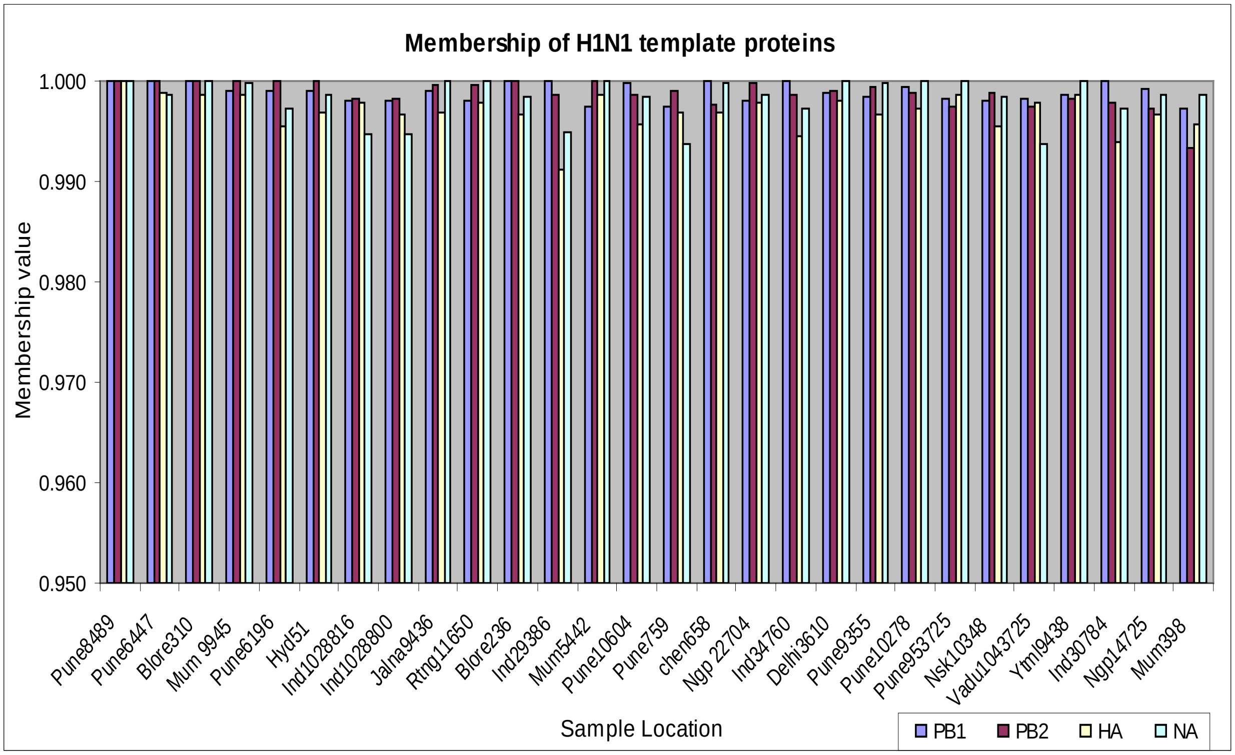

Figure 2. Identification strain segments PB2, PB1,HA and NA of H1N1 strain A/Pune/NIV8489/2009 in other samples

This same observation is reflected in BLAST analysis done at www.fludb.org. Comparison of SRLCS model M-value and SSEARCH35 [19] is tabled in Table1. From table 1. , it is seen that M value calculated by the suggested method is truthful and is in readable form

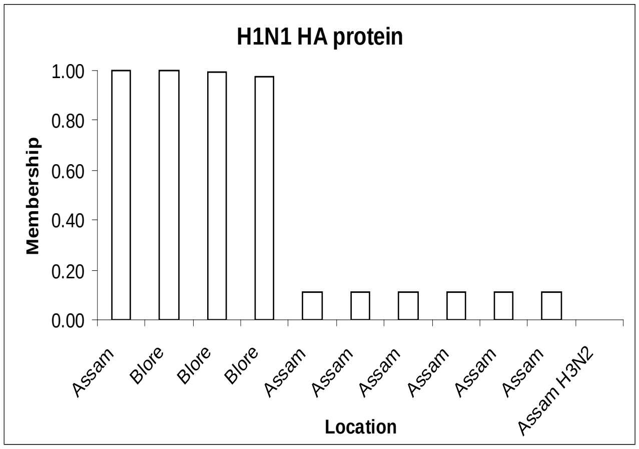

Figure 3. Membership of other HA segment samples in gb:JN600356 HA protein segment

V. CONCLUSION

SRLCS Model with Fuzzy membership calculation provides good guidance to biologists in the problems of identifying the origin of virus and presence of the virus in other parts of the world. It is useful in curing database using comparative modeling strategy, for incomplete sequences. Further this method can rightly identify the similarity of target protein with template set of proteins from other organisms. This membership relationship is a more visible relationship than the numerical representation by other methods. Such a membership finding can be done with any set of biosequences like DNA, Gene, RNA protein etc.

VI. REFERENCES

[1]. Koonin EV, Galperin MY. Sequence - Evolution - Function: Computational Approaches in Comparative Genomics. Boston: Kluwer Academic; 2003., NCBI Bookshelf ID: NBK20259

[2]. Troy CS, MacHugh DE, Bailey JF, Magee DA, Loftus RT, Cunningham P, Chamberlain AT, Sykes BC, Bradley DG. Genetic evidence for Near-Eastern origins of European cattle. *Nature*, 2001, vol. 410, p. 1091

[3]. Squires et al. BioHealthBase: informatics support in the elucidation of influenza virus host pathogen interactions and

virulence. Nucleic Acids Research (2008) vol. 36 (Database issue) pp. D497. (IRD Database)

[4]. Needleman SB, Wunsch CD: A general method applicable to the search for similarities in the amino acid sequence of two proteins. *J Mol Biol* 1970, 48(3):443-453.

[5]. Smith TF, Waterman MS: Identification of common molecular subsequence. *Journal of Molecular Biology* 1990, 215:403-410.

[6]. Edgar, R.C. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Research, Vol. 32, No. 5, 2004, 1792-1797.

[7]. Stephen F. Altschul, Thomas L. Madden, Alejandro A. Schäffer, Jinghui Zhang, Zheng Zhang, Webb Miller, and David J. Lipman (1997), "Gapped BLAST and PSI-BLAST: a new generation of protein database search programs", Nucleic Acids Res. 25:3389-3402.

[8]. W.R. Pearson, Searching protein sequence libraries: Comparison of the sensitivity and selectivity of the Smith-Waterman and FASTA algorithms, 1991, Genomics Volume 11, Issue 3, November 1991, Pages 635–650.

[9]. Easton, T., and Singireddy, A. 2008. A large neighborhood search heuristic for the longest common subsequence problem. *Journal of Heuristics* 14(3):271–283.

[10]. Shyu, S. J., and Tsai, C.-Y. 2009. Finding the longest common subsequence for multiple biological sequences by ant colony optimization. *Comput. Oper. Res.* 36(1):73–91.

[11]. Blum, C.; Blesa, M. J.; and L´opez-Ib´a´nez, M. 2009. Beam search for the longest common subsequence problem. *Comput. Oper. Res.* 36(12):3178–3186.

[12]. Wei Liu, Lin Chen, A Fast Longest Common Subsequence Algorithm for Biosequences Alignment, 2008, *IFIP* Advances in Information and Communication Technology, 2008, Volume 258/2008, 61-69, DOI: 10.1007/978-0-387-77251-6_8.

[13]. Korkin, D.; Wang, Q.; and Shang, Y. 2008. An efficient parallel algorithm for the multiple longest common subsequence (mlcs) problem. In *ICPP ’08: Proc. 37th Intl. Conf. on Parallel Processing*, 354–363.

[14]. Sumathy Eswaran , S.P.Rajagopalan, An Efficient Fast Pruned Algorithm for finding Longest Common Sequences in Bio Sequences, *Annals.Computer Science Series*, 8th Tome, 1st Fasc 2010, page 137 – 150.

[15]. Qingguo Wang, Mian Pan, Yi Shang and Dmitry Korkin, A Fast Heuristic Search Algorithm for Finding the Longest Common Subsequence of Multiple Strings, Proceedings of the Twenty-Fourth AAAI Conference on Artificial Intelligence (AAAI-10) , 2010, 1287-1292.

[16]. Sumathy Eswaran, S.P.Rajagopalan , “Heuristic SRLCS Algorithm to determine the proper alignment strategy for Biosequences , International Journal of Research and Reviews in Information Technology (IJRRIT) ,Vol. 1, No. 2, ,1-7, June 2011, ISSN: 2046-6501.

[17]. Sumathy Eswaran and Dr.S P Rajagopalan. Article: A Model for identification of Length of Longest Common Subsequence by SRLCS. International Journal of Computer Applications 33(9):17-21, November 2011.

[18]. Jianlin Cheng, A multi-template combination algorithm for protein comparative modeling, *BMC Structural Biology* 2008, 8:18 doi:10.1186/1472-6807-8-18.

[19]. T. F. Smith and M. S. Waterman. Identification of common molecular subsequences. J. Mol. Biol., 147:195–197, 1981.