Bioinformation by Biomedical Informatics Publishing Group *open access*

www.bioinformation.net **Hypothesis**

PIR Pairwise Alignment - A slip up for signal peptides

Seetharaaman Balaji [1*], Rangaswamy Kalpana [2] and Santhosh J. Eapen[1]

[1] Bioinformatics Centre, Indian Institute of Spices Research; [2]Guest Lecturer, Bioinformatics, Bharathiar University, Coimbatore - 641046, Tamilnadu; Seetharaaman Balaji* - Email: blast_balaji@rediffmail.com; * Corresponding author

received June 28, 2006; revised August 11, 2006; accepted August 12, 2006; published online August 14, 2006

Abstract:

The ability to calculate the correct sequence alignment is crucial to many types of studies. The accuracy in alignment is critical in predicting gene ancestry, the number and location of point mutations, evolutionary distance and phylogeny. A study was conducted to test the biological significance of PIR pairwise alignment using 40 N-terminal signal peptides of different taxonomic origin and having various functions. Our results suggest that PIR pairwise alignment is not ideal for some proteins with N-terminal signal peptides, because it produces an erroneous alignment that lacks both statistical and biological significances. This communication discusses the shortcomings in the PIR pairwise alignment tool and calls for a cautious approach while using it for signal peptides.

Keywords: PIR; pair wise alignment; SSEARCH; signal sequence alignment

Background:

Detecting subtle protein sequence similarities is a core problem in computational biology as sequence similarity typically implies homology, which in turn may imply structural and functional similarity. The discovery of a statistically significant similarity between two proteins is frequently used, therefore, to justify inferring a common functional role. **[1]** The identification of maximally homologous subsequences among sets of long sequences is still an important problem in molecular sequence analysis. Modern sequence analysis was significantly influenced with the heuristic homology algorithm, which first introduced an iterative matrix method of calculation. **[2]** Smith and Waterman proposed an algorithm in 1981 for doing pairwise alignment. This algorithm not only puts the search for pairs of maximally similar segments on a mathematically rigorous basis but it can be efficiently and simply programmed on a computer. **[3]** This pair-wise alignment like many other algorithms yields statistically significant results that are not always biologically significant. The biological insignificance of this algorithm was studied in detail using N-terminal signal peptides as an example and discussed in this communication.

Methodology:

We have selectively taken sixty signal peptide proteins of different functions that belong to different organisms, which have been experimentally characterized and deposited in SWISS-PROT. **[4]** Signal peptides where the cleavage sites are not experimentally determined and sequences which were functionally homologous were avoided.

Forty sequence entries were filtered and used as the signal peptide data set (Table 1). The data set belong to various taxonomical classifications (that includes an archaea, some bacteria and viruses) and have various functions to find the conservation of N-terminal signal region across the organisms. The collected protein sequence entries with signal peptides were checked for sequence similarity by subjecting to pairwise sequence alignment algorithm using SSEARCH program ver 3.0 (Smith & Waterman, 1981) implemented in the protein information resource (PIR). It is an integrated public bioinformatics resource that supports genomic and proteomic research and scientific studies. **[5]** While performing pairwise alignment, the forty selected signal peptides were aligned with each other in a random manner of 60 combinations to identify the conservation between the N-terminal signal regions of different organisms. We found that 50% of the pairwise alignments were not having biological significance, because in most cases the N-terminal signal region of proteins have been automatically aligned with C-terminal or non-signal region of other proteins.

In this case study our null hypothesis was that difference between the sequences aligned by SSEARCH and by manual method is zero. The 15 couple of alignments generated by using SSEARCH (Table 2) were taken as group A and the same done manually were taken as group B. Statistical calculations were done for both the groups (A & B) for both the parameters (identity and gaps) by using SPSS ver 11.5.1. F-test (Levene's test for equality of variances) that evaluates the basic assumption of the t-test that the variances of the two groups are approximately equal (homogeneity of variance) was also carried out.

Results and Discussion:

The SSEARCH algorithm for the fifteen couple of full-sequences with signal peptides was biologically insignificant and the first seven alignments showed a lesser sequence identity (Table 2 and Fig 3). In some cases the SSEARCH similarity was high but the interpretation of statistical significance was not reflective of any biological significance as the alignment was not on the N-terminal region. The alignment should match with the N-terminal region of the protein pairs to indicate a

Bioinformation by Biomedical Informatics Publishing Group *open access*

www.bioinformation.net **Hypothesis**

Taxonomic group	Accession No.
Amaranthaceae	P11898
Archaea	Q9HKM6
Thermoplasmataceae	
Bacteria	Q9F7S4, P11572, Q9XDH5
Gammaproteobacteria	
Mycobacteriaceae	
Thermaceae	
Bovinae	P98072
Brassicaceae	Q8L7U5
Daucinae	P37703, P14009, P37704
Drosophilidae	P91875
Hominidae	O00160
Murinae	Q9CQX8, P43687, Q60754, Q63356, P70248
Peloderinae	Q09524
Phaseoleae	P04145, P23233
Polygonaceae	Q9XFM4
Saccharomycetaceae	P53131, P38953, Q02555
Schizosaccharomycetaceae	O14072, O14072
Solanaceae	P23137, Q9XH42, Q9XH43
Suidae	P98074, O62680
Triticeae	P01543, P32032
Vicieae	P13240
Virus	P52358, P14979, P33495
Alphaherpesvirinae	
Geminiviridae	
Pneumovirinae	

Table 1: List of signal peptides used in this study and their taxonomic origin

```
 s-w opt:  30  Z-score: 44.1  bits: 14.1 E():  0.7
Smith-Waterman score: 30;  21.429% identity (23.077% ungapped) in 28 aa overlap (87-114:9-34)

         60        70        80        90       100       110
Q9XH43 VHGGDLRSFFTMIMIDPDVPGPSDPYLREHLHWIVTDIPGTTDCSFGKEIVGYEMPRPNI
                                    : :... .  .. : .:. .  :.   :
O62680                        MGSKGGFILLWLLSIL--AVLCHLGHSLQCYNCINPAG
                                       10          20        30

        120       130       140       150       160       170
Q9XH43 GIHRFVFLLFKQKKRQTVLTAPLSRDRFNTRKFAEENELGSPVAAVFFNCQRETAARRR

O62680 SCTTAMNCSHNQDACIFVEAVPPKTYYQCWRFDECNFDFISRNLAEKKLKYNCCRKDLCN
         40        50        60        70        80        90
```

Figure 1: Alignment between SP Q9XH43- CEN-like protein 2 from *Nicotiana tabacum* and SP O62680- CD59 glycoprotein precursor (membrane attack complex inhibition factor) from *Sus scrofa* having signal region 1-25 using SSEARCH

Bioinformation by Biomedical Informatics Publishing Group *open access*

www.bioinformation.net **Hypothesis**

	Alignment			SSearch		Manual alignment	
S. N	Accession No	with	Accession No	% IDENTITY	GAPS	% IDENTITY	GAPS
1	Q9XH43	+	O62680	21.4	2	32	0
2	Q9XDH5	+	P98072	21.5	43	32	1
3	Q9F7S4	+	P14009	25.0	2	28	1
4	P70248	+	P53131	20.2	88	32	1
5	P38953	+	Q02555	17.8	32	24	1
6	O14072	+	P91875	18.4	90	28	1
7	O14072	+	Q8L7U5	20.7	24	24	1
8	O75030	+	Q60754	22.3	11	28	1
9	Q9XH42	+	P37703	42.1	1	20	0
10	Q9HKM6	+	P43687	45.0	0	32	1
11	Q9CQX8	+	Seq2	36.8	0	24	1
12	Q9CQX8	+	P32032	36.8	0	24	1
13	P33495	+	Q09524	70.0	0	24	2
14	Q9XH43	+	P13240	24.5	9	24	0
15	Q63356	+	P53131	29.5	8	20	1
		Mean		30.1333	20.6667	26.4000	0.8667
		SD		14.11826	30.66097	4.22239	0.51640
		SEm		3.64532	7.91663	1.09022	0.13333

Table 2: Percentage identities of SSEARCH and manual alignments with gaps

```
MGSKMSDPLVIGRVIGEVVDYFTPSVKMSVTYNSSKHV
||||  .   |.     |   |. ..   |.        | .
MGSKGGFILLWLLSILAVLCHLGHSLQCYNCINPAGSCT
```

Figure 2: Manual alignment between SP Q9XH43- CEN-like protein 2 from *Nicotiana tabacum* and SP O62680- CD59 glycoprotein precursor (membrane attack complex inhibition factor) from *Sus scrofa* having signal region 1-25. This manually generated alignment is exactly on the N-terminal signal sequence

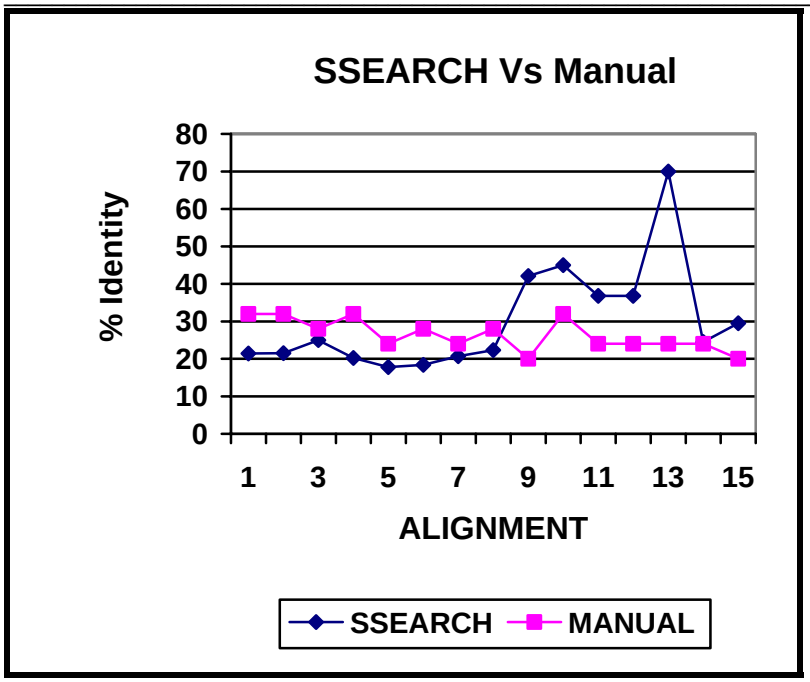

Figure 3: The graph displays the % identities of the alignments. Line with markers indicates each data value. First seven manual alignments are having higher % identities, well above the "twilight zone" so it is both statistically and biologically significant (in contrast to SSEARCH program). The rest of the alignments may have statistical significance but biologically meaningless

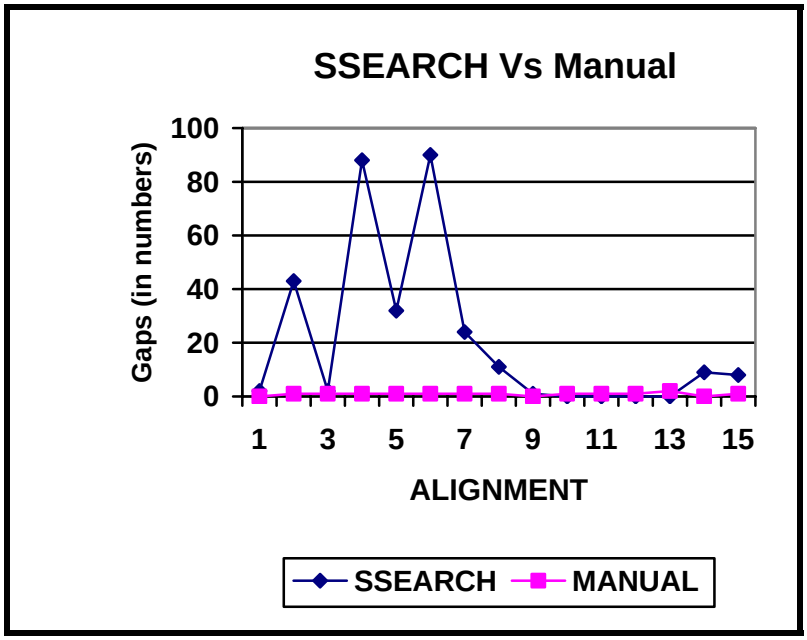

Figure 4: The graph displays the number of gaps used in both the alignments. Line with markers indicates each data value. First seven SSEARCH alignments used higher gaps but having lesser identities

correct functional prediction. If prediction accuracies in pairwise alignments are measured based on statistical significance and biological significance is ignored, it may result in wrong prediction. Figure 1 is a sample alignment between two peptides (SP Q9XH43 and SP O62680). It is a pretty good alignment particularly at the location ranging from 87-114. There are six identities and ten similarities (with a couple of gaps). But from the alignment we found that the N-terminal signal region (1-25) of *Sus scrofa* matches against the middle region of the CEN-like protein 2. The alignment does not

have any biological significance, because our interest lies in the conservation or similarity of N-terminal signal regions of proteins. To test whether or not this alignment is significant we did a manual alignment of the same two proteins which is depicted in Fig 2. This manual alignment is exactly on the N-terminal signal sequence and got nine identities. It is really amazing to note that the alignment (Fig 1) by SSEARCH program has only six identities by introducing a couple of gaps. In general, the optimal alignment of two sequences is usually that which maximizes the number of matches and minimizes the number of gaps. Permitting the insertion of arbitrarily many gaps can lead to high scoring alignments of non-homologous sequences. In contrast to the SSEARCH alignment, no gaps were inserted in the manual alignment and still we attained nine identical residues in the N-terminal tail, of those eight are in the signal region itself (Fig 2). Moreover, the manual alignment has biological significance, which was not found by the SSEARCH program.

A modification of the Smith-Waterman **[3]** or Sellers algorithms **[6]** will find all segment pairs whose scores can not be improved by extension or trimming. These are called high-scoring segment pairs or HSPs. According to our study, we have found higher scoring segment pairs (with respect to identities) in the N-terminal region of the first seven cases by manual alignment than by the SSEARCH algorithm. For protein sequences Doolittle's rule of thumb **[7]** is that greater than 25% identity will suggest homology, less than 15% is doubtful and for those cases between 15-25% identity, a strong statistical argument is required. The extent of similarity between two sequences is based on the percent of sequence identity and/or conservation. In the case study we got 32 % identity by manual alignment (Fig 2), but by SSEARCH program it is only 21.4 % sequence identity (Fig 1) which is in the "twilight zone" and according to the Doolittle's rule of thumb a strong statistical argument is required. We did this alignment manually using human intelligence and did not use any explicit algorithm. If the number of matches can be increased by reducing the number of gaps, clearly the original alignment's insertions of gaps are not needed. In most of the fifteen couple of alignments (Table 2) we have manually introduced only a gap to represent an 'indel'. The needless gaps used in SSEARCH were shown in Fig 4. The conserved substitutions in fifteen couple of protein sequences were calculated from the rest of the correct 50% of the test samples.

The signal peptides were not chopped off from the query protein sequences because our interest solely lies on the conservation of the N-terminal signal region of proteins. This is in contrast to some previous works on transmembrane (TM) topology prediction because the hydrophobic core of signal peptide is easily predicted as the putative first TM segment. **[8]** In genome wide analyses also, the likely signal peptide region is treated in several ways. It was either masked out from topological calculations **[9]** or omitted. **[10**, **11]**

Statistically, although there was no significant difference between identities of the two types of alignments (two-tailed t-test, $t(28) = 0.335$), the N-terminal signal peptides did not align at the expected N-terminal region showing their biological insignificance. In support to our alternate hypothesis, group A (SSEARCH) sequence identities showed more variation (30.13 ± 14.12) than in the manual method (group B 26.40 ± 4.22) (Table 2). Hence, manual method is more efficient in terms of calculating identity as there is no variation in the sequences due to strong N-terminal conservation.

A similar trend is also noticed in the gaps included in both the methods. The number of gaps used for aligning group B was significantly smaller than for group A. (two tailed t-test, $t(28) = 0.019$, $p<0.05$) indicating a significant difference in the gap usage of the two alignment methods and supports our alternate hypothesis. So there are significant evidences to suggest that manual method is better with respect to pairwise alignments of N-terminal signal peptides.

Conclusion:

Our study has shown that SSEARCH program is not apt for some proteins with N-terminal signal peptides as it produces erroneous alignments that lack both statistical and biological significances. Statistical significance is not equivalent to biological significance and low entropy regions providing false positives are well-known, and apply to all search methods. **[12]** The interesting findings of this study once again remind us the limitation of computational analysis to provide biologically significant conclusions in atleast some specific cases. This also serves as a wake-up call to those who consider putative annotations too seriously. Therefore, in critical studies the alignments can be improved by careful examination and human interpretation. It is better to adopt several automated alignments for every comparison (minimally one with the default penalties, one with more severe and one with less severe penalties) and then interpret manually.

Acknowledgement:

This work was supported by the Department of Biotechnology (DBT), New Delhi through on *Ad hoc* Project and carried out in the 'Distributed Information Sub-centre'. The authors thank Mr. K. Jayarajan, Technical Officer for the statistical assistance.

References:

[01] L. Liao & W.S. Noble, *J. Comp. Biol.*, 10:857 (2003) [PMID: 14980014]

[02] S. B. Needleman & C. D. Wunsch, *J. Mol. Biol.*, 48:443 (1970) [PMID: 5420325]

[03] T. F. Smith & M. S. Waterman, *J. Mol. Biol.*, 147: 195 (1981) [PMID: 7265238]

[04] A. Bairoch, *et al., Brief. Bioinform.*, 5:39 (2004) [PMID: 15153305]

[05] C. H. Wu, *et al., Nucleic Acids Res.*, 31:345 (2003) [PMID: 12520019]

[06] P. H. Sellers, *Bull. Math. Biol.*, 46:501 (1984)

[07] R. F. Doolittle, *Of URFs and ORFs,* University Science Books, Mill Valley, CA (1986)

[08] D. M. Lao, and T. Shimizu, METMBS'01, CSREA Press, USA, 119 (2001)

[09] D. T. Jones, *FEBS Lett.*, 423:281 (1998) [PMID: 9515724]

[10] I. T. Arkin, *et al.*, *Proteins*, 28:465 (1997) [PMID: 9261863]
[11] T. J. Stevens & I. T. Arkin, *Proteins*, 39:417 (2000) [PMID: 10813823]
[12] J. M. Claverie, *Computers and Chemistry*, 16:89 (1992)

Edited by William Perrizo
Citation: Balaji *et al*., Bioinformation 1(5): 188-193 (2006)