Multi-motif PHI-BLAST: a new tool for Database Searching and Sequence Alignment

Nitin Bhardwaj, IIT Bombay

ABSTRACT

The present PHI-BLAST used for searching genomic database available on the NBCI server takes only one motif as the input and so needs as many runs as the number of motifs the user has to give. And hence it cannot give any preference to any particular motif. We have developed a Multi-Motif version of PHI-BLAST, which takes multiple motifs as the input and picks up only those sequences from the database which have a minimum number of motifs given by the user allowing the user to be as specific as he/she wants to be. We have also developed some higher versions of PHI-BLAST such as the Ranked-motif PHI-BLAST which gives the higher ranked motifs a preference in the database searches. In this paper first we present the strategy and algorithm involved in these tools and later we compare the results of these tools with those from PHI-BLAST. The results are as good as the ones by PHI-BLAST or better.

Keywords: Motif, Subject and Query sequence, Local and Global Alignment, PHI-BLAST

INTRODUCTION

The sequence itself is not informative; it must be analyzed by comparative methods against existing databases to develop hypothesis concerning relatives and function. This is done by alignment, the process of lining up two or more sequences to achieve maximal levels of identity (and conservation, in the case of amino acid sequences) for the purpose of assessing the degree of similarity and the possibility of homology.

Broadly there are two types of alignments:

Global Alignment :

The alignment of two nucleic acid or protein sequences over their entire length, such that the entire length of the sequences is covered..

Local Alignment :

The alignment of some portion of two nucleic acid or protein sequences, ie. aligning only a part of the sequence.

Aligning two sequences generally necessitates the introduction of gaps into the sequences. The score of alignment determines the level of identity. The score of an alignment, S, is calculated as the sum of substitution and gap scores. Substitution scores are given by substitution matrices (such as PAM 250, BLOSUM 62 etc). Gap scores are typically calculated as the sum of G, the gap opening penalty and L, the gap extension penalty. For a gap of length n, the gap cost would be G + Ln.

PHI-BLAST

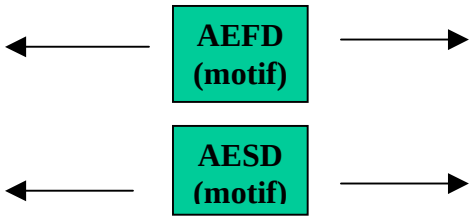
In the analysis of a protein or DNA sequence, particular interest often focuses upon a small region, domain or sequence pattern. This is called a motif**.** Motifs are frequently highly conserved parts of domains. A natural question is whether there are other related sequences that share the same pattern. Described here is the pattern-hit initiated BLAST (PHI-BLAST) program, whose hybrid strategy addresses a type of question frequently asked by researchers: namely, is a particular pattern seen in a protein of interest likely to be functionally relevant, or does it occur simply by chance?

The input to PHI-BLAST consists of a protein or DNA sequence, along with a specific pattern occurring at least once within the sequence. The pattern is generally in a PROSITE pattern like the following:

[ASDWE][WERDF][X][DFSRT]

This means that the pattern consists of any of the residues A, S, D, W, or E in the first place followed by any of W, E, R, D, or F followed by any of the 20 residues followed by any of D, F, S, R, or T. For each match between an instance of the pattern in the query sequence and an instance in a database sequence, PHI-BLAST constructs a high-scoring local alignment that includes the match by extending the match in either direction (local alignment in either direction by placing gaps etc to bring out the highest similarity). This extension is done using Dynamic Programming, which is not the main focus of this paper,

although we have developed our own codes for the same. All resulting alignments are sorted by score and evaluated statistically.



.

A typical PHI-BLAST output looks like the following:

Database: /database/pdbaa

9250 sequences; 1,971,410 total letters

1 occurrence(s) of pattern in query

pattern [RA][C][ACDEFGHIKLMNPQRSTVWY][C] at position 3 of query sequence

Number of occurrences of pattern in the database is 87

	Score (bits)	E Value
Significant matches for pattern occurrence 1 at position 3		
pdb|1ILP|A Chain A, Cxcr-1 N-Terminal Peptide Bound To Interleuk...	128	2e-37
pdb|1QE6|D Chain D, Interleukin-8 With An Added Disulfide Betwee...	121	2e-35
pdb|1ICW|A Chain A, Interleukin-8, Mutant With Glu 38 Replaced B...	121	3e-35
pdb|1ROD|A Chain A, Chimeric Protein Of Interleukin 8 And Human ...	98	2e-28
pdb|1TVX|B Chain B, Neutrophil Activating Peptide-2 Variant Form...	50	6e-14
pdb|1NAP|A Chain A, Mol_id: 1; Molecule: Neutrophil Activating P...	50	6e-14
pdb|1MSG|A Chain A, Human Melanoma Growth Stimulatory Activity (...	48	3e-13

The above given output was for the following query sequence:

ELRCQCIKTYSKPFHPKFIKELRVIESGPHCANTEIIVKLSDGRELCLDPKENWVQRV
VEKFLKRAENS

with the following pattern:

[RA][C][ACDEFGHIKLMNPQRSTVWY][C]

As can be seen from the output it reports the following in the order

1) the database against the search was done (pdbaa in this case),

2) all the hits from the database sorted in the order of their score, staring with their accession code, the name, followed my the normalized bit score and finally their E values. E value is the expectation value ie the number of different alignments with scores equivalent to or better than S that are expected to occur in a database search by chance. The lower the E value, the more significant the score. If we decrease the value of E the number of sequences reported decreases

The Problem Statement: the above version needs to be run as many number of times as the no. of motifs the user wants to give and also the user cannot set any kind of preference to any kind of motif which are common problems encountered in database searching and alignment.

MULTI-MOTIF PHI-BLAST (MMPB)

The above-described PHI-BLAST takes only one motif as the input and works in the way described. We have developed a method that will take multiple motifs as the input work in the same way and search the database. What it exactly does is:

1) takes a query sequence and the database to be searched,
2) takes the number of motifs the user wants to give, and take those many motifs as the input ,
3) asks how many motifs does the user want to be there in the sequences to be picked up,
4) picks up all the sequences from the given database which have at least those many motifs in them and align them with the query sequence and bring out the score, and finally,
5) sorts all the sequences in the order of their scores.

The condition for the minimum no of motifs to be present in the sequences to be picked up from the database makes the program highly selective so reduces the number of false positives. A typical MMPB output looks like following:

>gi|640276|pdb|1MGS|A Chain A, Human Melanoma Growth Stimulating Activity (MgsaGRO_ALPHA) (Nmr, 25 Structures

It has 2 motifs.

The score is: 151

……

No of hits=9

****

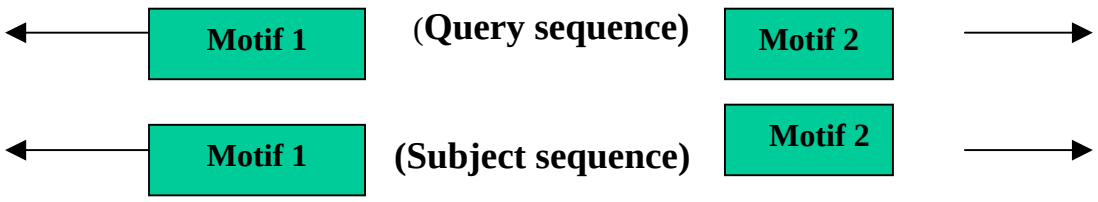
This is for the same query sequence and the same motif as the ones for the PHI-BLAST output above.

As can be seen it reports the following: the sequence picked up with its accession code, the score of alignment with the query sequence, the number of motifs present in it, and finally, the number of hits.

ALGORITHM and STRATEGY

As described above, MMPB takes in multiple motifs as the input and align the query sequence with only those sequences which have the minimum number of motifs present.

We employ both local and global alignment in the alignment procedure. We first of all fix the motif in the sequence to align against the motifs in the query sequence.

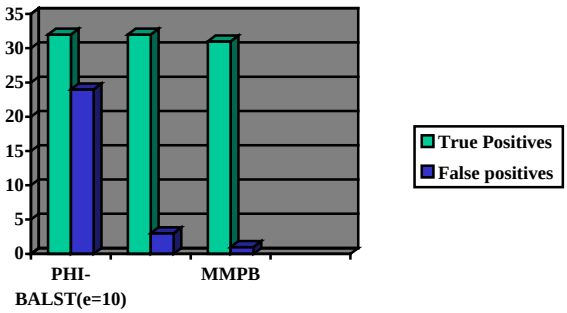


Then the part between the motifs is aligned globally and the part on either end is aligned locally (by Dynamic Programming). And then as usual the sequences picked up are sorted in the order of decreasing score and displayed.

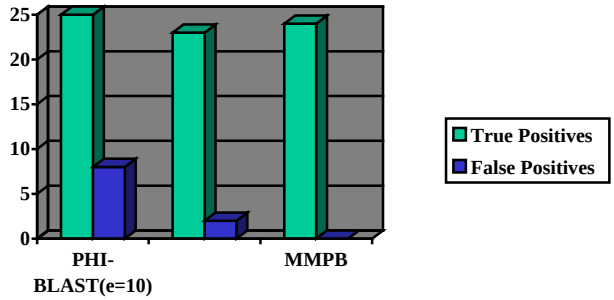
COMPARISONS OF RESULTS

We ran the PHI-BLAST and the MMPB for quite a few number of families and the results were comparable. The results in the form of bar diagrams are shown below (the middle bar in the following diagrams correspond to the PHI-BLAST run for e=1 ie even stricter search):

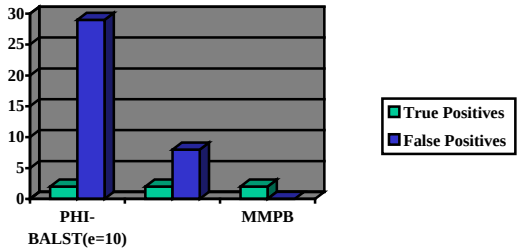
il81(1hum) Macrophage Inflammatory 1beta



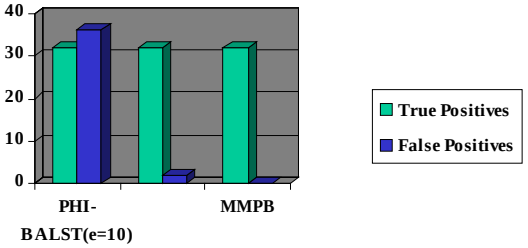
il82(1ikl) Interleukin-8



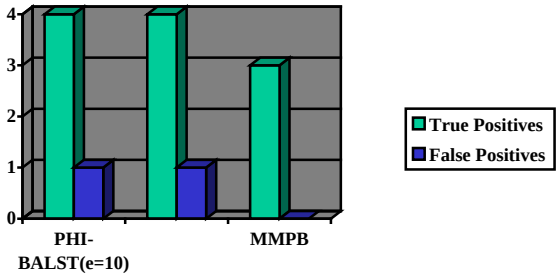
Flav1(1ord) Orthinine Decarboxylase



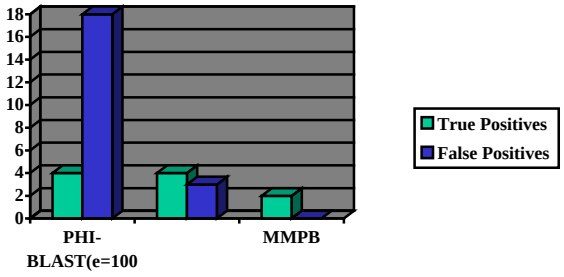
Flav2(1cus) Cutinase



4helud1(1bbh) Cytochrome $c (prime)



4helud2(256b) Cytochrome $b502



The above reported cases were all unique hits ie. all common hits reported were removed in the different runs.

The following points can be observes from the above comparisons

1) MMPB has almost the same number of true positives as the PHI-BLAST,
2) MMPB has reported almost no false positives.

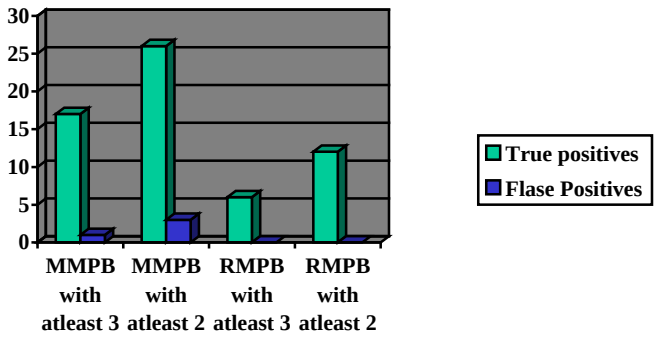
SOME FURTHER VERSIONS

We have developed another version of MMPB. This is **Ranked Motif Alignment (RMPB).** This tool does the following

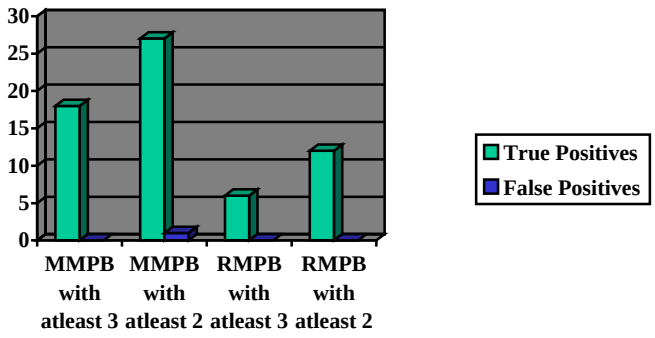
1) takes a query sequence, a database and the number of motifs as the input,
2) further it takes the motifs in the order of their ranks, and again asks for the minimum number of motifs to be present in the sequence,
3) reports only those sequences which have at least this many number of highest ranked motifs in them, so say the user wants at least three motifs to be present, then all the sequences picked up will have at least the first three ranked motif in them, and finally,
4) align those sequences with the query sequence and sort the sequences in the order of their alignment score.

This makes sure that the important motifs are present there in the sequence, if not the lower ranked ones. This proves out to be another constraint on the sequences to be picked up. The results on the next page compare the outputs of MMPB and RMPB. This is to highlight the number of sequences that had the minimum number of motifs but did not have the minimum number of highest ranked motifs. The difference in the two bars shows this number in each of the following diagrams.

Macrophage Inflammatory 1beta



Interleukin-8



CONCLUSION

As illustrated by the examples discussed above, PHI-BLAST helps both to ascertain the biological relevance of patterns detected within protein sequences. The above-developed versions of PHI-BLAST have the potential to greatly enhance the quality of results of

Genomic database searching and alignment. They leave the user with a lot of choices. MMPB has space for inputting multiple motifs in the same run which allows the user to be as specific as he/she wants to be by inputting the minimum number of motifs to be present. The RMPB version gives the choice of attaching preference to different motifs and then proceed. The results shown above for these tools , as can be seen, are as good as or better then the ones by PHI-BLAST. But at the same time this specificity makes the search a little more constrained. As for their refinements, in terms of say speed and systematic arrangement, a lot still needs to be done.

REFERENCES

1) Altschul S.F., ZhangZ., Schaffer A.A., Madden T.L., Miller T.W. Nucleic Acids Res.98 Sep1;26(17):3986-90.
2) Smith,T.F. and Waterman,M.S. (1981) *J. Mol. Biol*., **147**, 195-197.
3) Altschul,S.F., Miller,W., Myers,E.W. and Lipman,D.J. (1990) *J. Mol. Biol*., **215**, 403-410
4) Altschul,S.F. and Gish,W. (1996) *Methods Enzymol*., **266**, 460-480
5) Altschul,S.F., Madden,T.L., Schäffer,A.A., Zhang,J., Zhang,Z., Miller,W. and Lipman,D.J. (1997) *Nucleic Acids Res*., **25**, 3389-3402
6) Altschul,S.F., Boguski,M.S., Gish,W. and Wootton,J.C. (1994) *Nature Genet*., **6**, 119-129
7) Needleman,S.B. and Wunsch,C.D. (1970) *J. Mol. Biol*., **48**, 443-453.
8) Myers,E.W. and Miller,W. (1989) *Bull. Math. Biol*., **51**, 5-37
9) Altschul,S.F., Boguski,M.S., Gish,W. and Wootton,J.C. (1994) *Nature Genet*., **6**, 119-129.
10) Altschul,S.F. (1998) *Proteins*, **32**, 88-96.
11) Altschul,S.F. (1998) *Proteins*, **32**, 88-96.
12) Zhang,Z., Berman,P. and Miller,W. (1998) *J. Comput. Biol*., **5**, 197-210
13) Tatusov,R.L. and Koonin,E.V. (1994) *Comp. Appl. Biosci*., **10**, 457-459
14) Staden,R. (1990) *Methods Enzymol*., **183**, 193-211.
15) http://www.ncbi.nlm.nih.gov/BLAST
16) http://www.bioinformaticsonline.org