Protein sequence similarity searches using patterns as seeds

Zheng Zhang, Alejandro A. Schäffer¹, Webb Miller, Thomas L. Madden², David J. Lipman², Eugene V. Koonin² and Stephen F. Altschul²,*

Department of Computer Science and Engineering, Pennsylvania State University, University Park, PA 16802, USA, ¹Inherited Disease Research Branch, National Human Genome Research Institute, National Institutes of Health, Baltimore, MD 21224, USA and ²National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, MD 20894, USA

Received May 7, 1998; Revised and Accepted July 8, 1998

ABSTRACT

Protein families often are characterized by conserved sequence patterns or motifs. A researcher frequently wishes to evaluate the significance of a specific pattern within a protein, or to exploit knowledge of known motifs to aid the recognition of greatly diverged but homologous family members. To assist in these efforts, the pattern-hit initiated BLAST (PHI-BLAST) program described here takes as input both a protein sequence and a pattern of interest that it contains. PHI-BLAST searches a protein database for other instances of the input pattern, and uses those found as seeds for the construction of local alignments to the query sequence. The random distribution of PHI-BLAST alignment scores is studied analytically and empirically. In many instances, the program is able to detect statistically significant similarity between homologous proteins that are not recognizably related using traditional single-pass database search methods. PHI-BLAST is applied to the analysis of CED4-like cell death regulators, HS90-type ATPase domains, archaeal tRNA nucleotidyltransferases and archaeal homologs of DnaG-type DNA primases.

INTRODUCTION

In the analysis of a protein or DNA sequence, particular interest often focuses upon a small region, domain or sequence pattern. A natural question is whether there are other related sequences that share the same pattern. The most widely used tools for sequence similarity search allow matching between arbitrary regions of the query and database sequences (1–5). In contrast, many motif-based search methods seek database sequences that match a pre-specified pattern (6–12). If this pattern is too weak, or not specified with sufficient precision, the number of matches may be very large, most being of no biological relevance. On the other hand, an overly-specific pattern may exclude many sequences of interest.

We describe here 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? To address this question, we combine a pattern search with a search for statistically significant sequence similarity. These two approaches were combined previously in a program that explored the output of a BLAST search for conserved patterns (10). PHI-BLAST implements a reverse strategy which is computationally more efficient, and which we believe will be of greater utility. Specifically, the similarity search is restricted to a subset of the sequence database comprised of the sequences that contain the given pattern.

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 currently required to be a sequence of residues or sets of residues, with ‘wild cards’ and variable spacing allowed; all PROSITE patterns (12), for example, have this form. 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. All resulting alignments are sorted by score and evaluated statistically.

This approach has greatest utility when it is suspected that a few residues comprising a small motif may be crucial for the biological function of interest. Showing that this pattern occurs within an extended and statistically significant alignment of the query sequence with one or more database sequences greatly reduces the likelihood that the pattern is spurious. Conversely, insisting on the presence of the pattern and hence searching a reduced sequence space may aid the detection of subtle similarities that blend into the background noise in a regular BLAST search.

THE PHI-BLAST ALGORITHM

To search for matches to a given pattern, we adapted a method of Baeza-Yates and Gonnet (13) and Wu and Manber (14). This

*To whom correspondence should be addressed. Tel: +1 301 496 2475; Fax: +1 301 480 9241; Email: altschul@ncbi.nlm.nih.gov
method permits simple patterns to be represented in a single computer word and matches to be found very efficiently. When the pattern is relatively complex, for example consisting of many rigid parts and/or having wide ranges of spacer lengths, our program first searches for the rigid part that is least likely to match by chance alone, and then performs local searches for the remaining pattern elements.

For each instance of the input pattern in a database sequence, paired with an instance in the query, PHI-BLAST attempts to find the optimal local alignment (1,15) containing the aligned patterns. This can be done rigorously by applying dynamic programming (16,17) to the parts of the two sequences preceding and the parts following the pattern. The alignment returned is required to begin at the corner of the path graph, but is permitted to end anywhere within the graph. The difficulty with this approach is that, to guarantee optimality, a very large portion of the path graph needs to be searched, and this requires inordinate time in a database search (18). Accordingly, we have used the gapped extension heuristic described in Altschul et al. (5) and Zhang et al. (18). Basically, path graph cells are considered only if the score of the best alignment leading into them falls no more than X below the best score yet found. For sufficiently large values of the X parameter, this approach almost always returns the optimal local alignment.

Because PHI-BLAST performs a gapped extension whenever an instance of the input pattern is encountered in the database, reasonable execution times depend upon such instances being relatively rare. Therefore, we allow only patterns that are expected to occur less frequently than once per 5000 database residues. Any pattern that contains four completely specified residues, or three specified residues whose average background residues are expected to occur less frequently than once per 5000 database, is discarded from the input pattern. If the input pattern contains more specific knowledge concerning the query sequence, a researcher often wishes to search a pattern-database for any well-characterized motifs the query may contain. To streamline this latter approach, we have implemented a program that first searches the PROSITE database (12) with the query; any patterns found may then be used to launch a PHI-BLAST database search. To facilitate more detailed analysis of PHI-BLAST output, we allow it automatically to serve as the basis for constructing a position-specific score matrix for further database searching via the position-specific iterated BLAST (PSI-BLAST) program (5). Like other BLAST family programs, PHI-BLAST incorporates a pre-filter for protein regions of biased amino acid composition (low complexity) that often corrupt database searches (28,29).

PHI-BLAST may detect subtle relationships that escape standard database similarity searches, but this potential depends upon the specification of an amino acid pattern likely to be conserved within the protein family of interest. We discuss four

"3987"
examples involving protein families whose original description depended critically upon detecting relatively weak sequence similarities. In each case, PHI-BLAST reports a subtle but structurally and functionally relevant relationship. The alignments suggesting these relationships are not all statistically significant but, in each database search output ranked by E-value, they appear immediately after the alignments involving clear family members, thereby prompting further analysis. In contrast, any of these similarities reported by gapped BLAST (5) are preceded by a number of alignments with smaller E-values involving unrelated sequences. The four examples discussed below are summarized in Table 1. All searches were performed on the non-redundant (NR) protein sequence database maintained by the NCBI (30).

CED4-like cell death regulators

The Caenorhabditis elegans protein CED4 is a regulator of programmed cell death (apoptosis). CED4 contains the classical P-loop motif involved in phosphate binding and found in a great variety of ATPases and GTPases. ATP binding by CED4, and the role of ATP in its function, have been demonstrated (31,32). In a gapped BLAST search of the NR database, CED4 shows statistically significant sequence similarity to only one protein, the human apoptosis regulator Apaf-1, in which the P-loop is conserved (33,34). However when PHI-BLAST is used, requiring conservation of the P-loop (Table 1), the best hit after Apaf-1, with E-value 0.038, is to a plant disease resistance protein, Arabidopsis thaliana T7N9.18 (35). Further sequence comparison shows that animal apoptosis regulators and putative plant ATPases involved in disease resistance share several conserved motifs, suggesting that they have a common origin and may have similar roles in programmed cell death (L.Aravind, V.M.Dixit and E.V.Koonin, unpublished observations). Before the Apaf-1 sequence became available, this conclusion had been reached through a laborious comparison of CED4 to a large number of different ATPases (32). Because the Apaf-1 sequence is highly similar to homologous plant proteins, the connection between CED4 and the plant proteins can be easily demonstrated by iterative database search (5). Even without Apaf-1, however, PHI-BLAST is able immediately to establish this link.

HS90-type ATPase domains

We used PHI-BLAST to investigate the subtle but structurally validated relationship between the ATPase domains in the MutL DNA repair proteins, type II topoisomerases, histidine kinases and HS90 family proteins (36,37). The output identified a new family of eukaryotic proteins that contain the same type of predicted ATPase domain, but that in standard database searches do not show significant similarity to any known member of the superfamily. A PHI-BLAST search with the Escherichia coli MutL protein (38) as query showed moderate similarity (E-value 0.017) to the C.elegans protein ZC155.3 (39) that was originally described as having ‘weak similarity to Bovine synaptocanalin I’. Subsequent database searches with this worm protein sequence as query revealed homologs in humans (KIAA0136) (40) and plants (41,42), whereas a PHI-BLAST search also showed convincing similarity to MutL family members (best E-value 6 × 10⁻⁵). Elucidation of the function of this new family of eukaryotic ATP-utilizing enzymes will be of considerable interest; the synaptocanalin domain apparently was fused to the worm protein by exon misassembly.

Archaeal tRNA nucleotidyltransferases

The archaeal tRNA nucleotidyltransferases (Cca) are a distinct family of nucleic acid polymerases (43) that in standard database searches are not detected by standard BLAST searches, but are found by searching with a PHI-BLAST with gaps. PHI-BLAST generates these alignments after searching with gaps and finding a conserved motif in the target protein sequence which then is used in a search with the nucleotidyltransferase domain.

Table 1. Detection of subtle protein sequence relationships using PHI-BLAST

<table>
<thead>
<tr>
<th>Conserved domain or motif under investigation</th>
<th>Pattern*</th>
<th>GenBank (30) accession no. of query</th>
<th>Top non-trivial relevant hit found by PHI-BLAST Accession no.</th>
<th>Top non-trivial relevant hit found by BLAST Accession no.</th>
<th>E-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>A. P-loop ATPase domain in apoptosis regulators and plant stress response proteins</td>
<td>[GA]xxxxGK[ST]</td>
<td>231729</td>
<td>2213598</td>
<td>0.038</td>
<td>2961373</td>
</tr>
<tr>
<td>B. ATPase domain in mismatch repair protein MutL, type II topoisomerases, histidine kinases, and HS90 molecular chaperones</td>
<td>bxhxDxDxG</td>
<td>127552</td>
<td>488200</td>
<td>0.017</td>
<td>2495364</td>
</tr>
<tr>
<td>C. Nucleotidyltransferase domain in archaeal tRNA nucleotidyltransferases</td>
<td>DhDhhh</td>
<td>2826366</td>
<td>2650333</td>
<td>0.061</td>
<td>2650333</td>
</tr>
<tr>
<td>D. Motif VI of superfamily II helicases in archaeal homologs of bacterial DNA primases</td>
<td>QxxGGR[GA][R]</td>
<td>2128723</td>
<td>2490999</td>
<td>0.54</td>
<td></td>
</tr>
</tbody>
</table>

The reported results are from searches of the NCBI (30) non-redundant protein sequence database (April 9, 1998; 298 842 sequences; 90 087 406 residues). The PHI-BLAST and BLAST algorithms used the BLOSUM-62 substitution matrix (27), in conjunction with penalties of 11+ k for gaps of length k. BLAST E-values were calculated using the statistical parameters λ = 0.270 and K = 0.047, and applying an edge-effect correction (4). PHI-BLAST E-values were calculated from equation 1, using the statistical parameters λ = 0.270 and C = 0.6. *Patterns are described using the one-letter amino acid code. Brackets represent a choice among any of the enclosed amino acids. ‘x’ represents any amino acid. ‘h’ represents [ILVMF], a hydrophobic amino acid.
searches do not have detectable similarity to any proteins other than orthologs from other archaeal species. However, they do contain a conserved motif, with two aspartate residues, that resembles the catalytic sites of many other polymerases (44). When this pattern (Table 1) is specified in a PHI-BLAST search with Methanothermobacter jannaschii Cca (45) as query, the top hit outside the archaeal Cca family itself, with E-value 0.061, is to hypothetical protein AFO299 from Archaeoglobus fulgidus (46), which belongs to a previously described archaeal family of predicted nucleotidyltransferases (47); the third hit (E-value 0.13) is to an experimentally characterized streptomycin 3′-adenylyltransferase from Enterococcus faecalis (48).

Table 2. Accuracy of PHI-BLAST statistics

<table>
<thead>
<tr>
<th>Example</th>
<th>Shuffled database</th>
<th>Reversed database</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Low E.-val.</td>
<td>Seqs with E.-val. ≤ 10</td>
</tr>
<tr>
<td>A</td>
<td>3.0</td>
<td>4</td>
</tr>
<tr>
<td>B</td>
<td>0.64</td>
<td>9</td>
</tr>
<tr>
<td>C</td>
<td>0.12</td>
<td>23</td>
</tr>
<tr>
<td>D</td>
<td>0.55</td>
<td>12</td>
</tr>
</tbody>
</table>

PHI-BLAST searches were performed on shuffled and reversed versions of the NR database, using the query sequences and associated patterns of Table 1, as well as the same alignment scoring system and statistical parameters λ and C. A, CED4-like cell death regulators; B, HS90-type ATPase domains; C, archaeal tRNA nucleotidyltransferases; D, archaeal homologs of DnaG-type DNA primases.

Archaeal homologs of DnaG-type DNA primases

Archaeal homologs of bacterial DNA primases, e.g. M. jannaschii protein MJ1206 (45), contain a motif typical of helicases (47), but do not show significant similarity to these proteins in standard BLAST searches. Using M. jannaschii MJ1206 and the helicase motif as query, the first non-trivial PHI-BLAST hit, with E-value 0.54, is to the well known helicase Neisseria gonorrhoeae UvrB (49). The relevance of the helicase motif in the archaeal primase homologs is supported by an extended alignment with the UvrB helicase (L. Aravind, D.D. Leipe and E.V. Koonin, unpublished observations). The similarities uncovered in this example are undetectable with standard database search techniques.

PERFORMANCE EVALUATION

To test the accuracy of the PHI-BLAST statistics given by equation 1, we used each of the examples above to search ‘random databases’ constructed from NR by shuffling or reversing each sequence. For each query, the lowest recorded E-value, and the number of alignments found with E-value ≤ 10, are given in Table 2. For the shuffled database, the geometric mean of the observed numbers of sequences with E-value ≤ 10 is 10.0, and no single case diverges from this value by more than a factor of 2.5. This might be expected, as the values of λ and C used in equation 1 were calculated employing a random protein model in which all amino acids occur independently. Perhaps surprisingly, Table 2 suggests that under an alternative random protein model, based upon reversed real sequences, these statistics are slightly conservative.

To compare the speed of PHI-BLAST to that of a standard gapped BLAST program (5) we timed both for searches of each of the four examples above against the NR database. Analysis of the results (Table 3) suggests that on the computer system used, ~8 s of each PHI-BLAST run were required to scan the database for pattern hits and for system overhead; the remainder was spent on constructing gapped extensions for all pattern hits found. Clearly, the number of hits generated by the input pattern is a key determinant of PHI-BLAST’s speed. For relatively informative patterns PHI-BLAST is very fast, requiring not much more time than that needed to search for pattern hits. For relatively weak patterns, PHI-BLAST expends most of its effort extending hits, and can require time comparable to that for gapped BLAST.

CONCLUSION

As illustrated by the biological examples discussed above, PHI-BLAST helps both to ascertain the biological relevance of patterns detected within protein sequences, and in some cases to detect subtle similarities that escape a regular BLAST search. We note, however, that PHI-BLAST was specifically designed to combine pattern search with the search for statistically significant sequence similarity, rather than to maximize search sensitivity. Thus in general one should not expect PHI-BLAST, which by its
nature is a single-pass search method, to be more sensitive than PSI-BLAST (5). Furthermore, within proteins, residues that are absolutely conserved during evolution constitute a small minority, and even specifying a restricted set of possibilities for a given residue position often excludes many members of a protein family. PHI-BLAST therefore is not the ideal tool for completely delineating a class of related proteins. However, by greatly restricting the size of the search space, PHI-BLAST can allow the similarities of some distant homologs to rise above the background noise that would otherwise obscure them. Such findings can be used subsequently for more extensive family analysis using PSI-BLAST (5) or other tools.

We have developed PHI-BLAST for protein–protein comparison, but plan to extend its applicability. A version that translates the program may be run from NCBI's web site at http://www.ncbi.nlm.nih.gov/

Note
Source code for PHI-BLAST is available by anonymous ftp from the machine ncbi.nlm.nih.gov, within the directory ‘blast’, and the program may be run from NCBI's web site at http://www.ncbi.nlm.nih.gov/

ACKNOWLEDGEMENTS
Z.Z. and W.M. are supported by grant LM05110 from the National Library of Medicine. We thank Dr L. Aravind for helpful discussions.

REFERENCES