BIOINFORMATICS APPLICATIONS NOTE Vol. 21 no. 4 2005, pages 542–544
doi:10.1093/bioinformatics/bti035

tPatternHunter: gapped, fast and sensitive translated homology search

Derek Kisman[1], Ming Li[2], Bin Ma[3,*] and Li Wang[3]

[1]Bioinformatics Solutions Inc., 145 Columbia St West, Waterloo, Ontario, Canada N2L 3L2, [2]School of Computer Science, University of Waterloo, Ontario, Canada N2L 3G1 and [3]Department of Computer Science, University of Western Ontario, London, Ontario Canada N6A 5B7

Received on July 14, 2004; revised on September 3, 2004; accepted on September 10, 2004
Advance Access publication September 16, 2004

ABSTRACT

Summary: New ideas, spaced seeds and gapped alignment before 6-frame translation are implemented for translated homology search in tPatternHunter. The new software compares favorably with tBLASTx.

Availability: The software is free to academics at http://www.bioinformaticssolutions.com/downloads/ph-academic/

Contact: bma@csd.uwo.ca

INTRODUCTION

In homology searches, two DNA/protein sequence datasets are compared with each other to identify similar-sequence segments. Segments exhibiting strong similarity are often homologues. Homology searches are a regular requirement of modern biological research. For example, the NCBI BLAST server is processing over 10^5 queries a day (Chattaraj and William, 2004).

BLAST (Altschul *et al*., 1990, 1997) is the most widely used homology search software today. It uses short string matches as 'seeds' to find longer similarities. The more recent DNA homology search software PatternHunter (Ma *et al*., 2002) improves BLAST with the new 'optimized spaced seed' idea. Instead of matching the whole string, the spaced (or gapped or discontiguous) seeds require matches at only some pre-selected positions of the short strings. (Ma *et al*., 2002) showed that the optimization of these positions resulted in significant improvement on the sensitivity and the speed of homology searches. The spaced seed is now a widely accepted practice (Brown *et al*., 2004, http://www.wspc.com/books/lifesci/5547.html). The chapter by Brown *et al*. (2004) also contains extensive literature references necessarily missing in this short note.

The original PatternHunter has only implemented DNA homology searches. Several other types of homology searches are also important. For example, the translated DNA–DNA search requires the finding of DNA segments whose corresponding protein translations are similar. Because of the redundancy of the genetic codes, translated DNA–DNA searches are capable of finding more homologues than searching the plain DNA sequences. Similarly, the translated DNA–protein searches compare DNA sequences with protein sequences by translation.

This note describes an extension of PatternHunter software to protein–protein, translated protein–DNA and translated DNA–DNA homology searches. Therefore, the new software is called translated PatternHunter (tPH). The extension is facilitated by new ideas, including spaced seeds with inexact matches, multiple seeds and gapped alignment before 6-frame translation, aiming to improve on the speed, sensitivity and alignment quality of BLAST.

METHODS AND ALGORITHMS

Similar to the original PatternHunter (Ma *et al*., 2002), tPH is written in Java. The overall architecture of tPH is also the same as PatternHunter, which finds 'hits' and then conducts gapped extensions to construct local alignments. However, the algorithms used by tPH for hit finding and gapped extensions are very different from PatternHunter. For technical terminologies used here, we refer the reader to Ma *et al*. (2002).

For DNA homology searches, PatternHunter used weight-11 spaced (multiple) seeds by default. For example, if the seed 111010010100110111 is used, then two length-18 DNA strings must match at the '1' positions in order to be regarded as a hit, and all the hits can be efficiently found by constructing an index table of the subject sequence. However, for protein homology searches, two different amino acids can have very similar chemical properties. Consequently, defining hits by letter matches is inappropriate. Moreover, because protein sequences have 20 different letters, compared to 4 for DNA sequences, the weight of the protein seeds must be significantly smaller than 11 for the space-complexity of the index table.

For these reasons, spaced seeds with weight-5 (the default is 1101011) are used in tPH for protein hits. Therefore, for

*To whom correspondence should be addressed.

any length-7 substring of the query and subject, only the five letters at the '1' positions are checked for hits. Furthermore, tPHs does not require the five positions to be exact matches. Instead, the BLOSUM scores of the five pairs of letters at those positions are examined. If all the five pairs have scores at least 0, and the total score is above a threshold T, then we say a hit is found. To find all the hits efficiently, a hash table similar to the one used in PatternHunter is built based on the subject sequences. All the length-7 substrings of the subject are grouped according to their five letters: at the '1' positions of the seed and their locations in the subject are recorded in the hash table. Then, for each length-7 substring of the query, all possible 5-mers that satisfy the hit definition are generated, and each is used to find the hit locations by a single lookup in the hash table.

To further improve the sensitivity of protein–protein searches, the multiple spaced seeds (Li *et al*., 2004) are employed by tPH. In such a search, any of several given seeds can generate hits, resulting in higher sensitivity but lower speed. Better performance can be achieved by carefully choosing the seeds and the threshold T. In tPH, four weight-5 spaced seeds of length 6 and 7 are pre-selected, and the threshold T is set to 20 by default when BLOSUM62 matrix is used. Compared to BLASTp's consecutive seed of size 3, the new approach can achieve similar sensitivity with fewer random hits.

The second innovation of tPH is the idea of hashing and extension independent of the 6-frame translations. To perform DNA–DNA (or DNA–protein) translated searches, tBLASTx creates the 6-frame translations of the subject and query sequences first, treating them independently. Such an approach cannot handle frameshift errors. tPHs does not translate the DNA sequences to proteins before the output. Instead, tPH regards the DNA sequences as a sequence of overlapped codons. The seed finding is almost the same as protein–protein search. However, the extension is performed using a modified Smith–Waterman algorithm that can take frameshifts into account. Some parameters, such as gapopen, gapextend and frameshift, are used to guide the extension behavior.

RESULTS AND DISCUSSION

We conducted a sensitivity and speed benchmark of tPH for comparison with tBLASTx. The DNA sequences we used are human expressed sequence tag (EST) sequences from the NCBI's website. File, month.est_human.Z, at ftp://ftp.ncbi.nlm.nih.gov/blast/db/FASTA/ contains only the new or revised human EST sequences released in the last 30 days. We have downloaded the file released on April 14, 2003. There were 4407 human EST sequences, split in the middle as subject and query. BLAST, version 2.2.6, was from the NCBI's website on May 16, 2003.

On a Pentium 1.8 GHz CPU, we ran tBLASTx using its default parameters, except '–F F'. We ran tPH using threshold

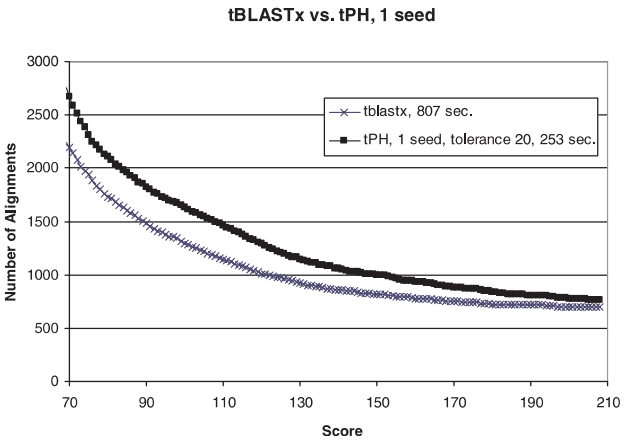

Fig. 1. The x-axis is the alignment score and the y-axis is the number of alignments with at least the corresponding score. For example, at score 110, tBLASTx has 1149 alignments whereas tPH has 1463 alignments.

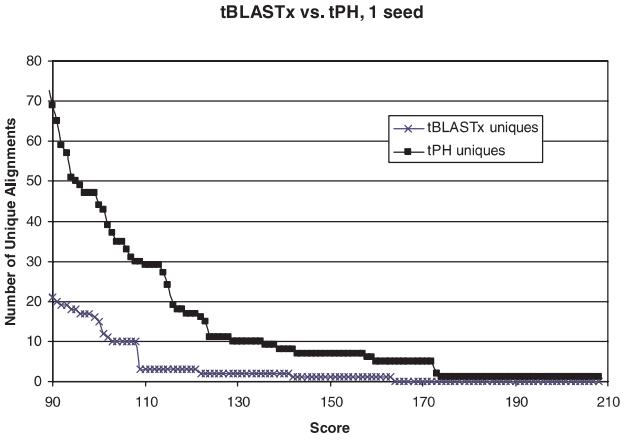

Fig. 2. The x-axis is the alignment score and the y-axis is the number of unique alignments with at least the corresponding score. For example, at score 110, 3 alignments are tBLASTx unique whereas 29 alignments are tPH unique. Moreover, tPH runs three times faster than tBLASTx.

20 (-T 20) and one spaced seed (-multi 1), with the command line: java -jar -Xmx512m TransPHunter.jar –i human1.est.fna -j human2.est.fna -o tphoutput20.txt -Ns 1000000 -N 1 –T 20 -multi 1. Both programs use similar scoring schemes. The results are shown in Figure 1.

Figure 2 shows the unique alignments obtained from either program (and not obtained from the other program with any score).

Using four spaced seeds, with the same command as above except for '-multi 4', the sensitivity of tPH increases further and tPH still runs in time comparable to that of tBLASTx, as seen in Figure 3.

The following is a sample alignment fragment obtained from tPH with a DNA level gap causing frameshift. The

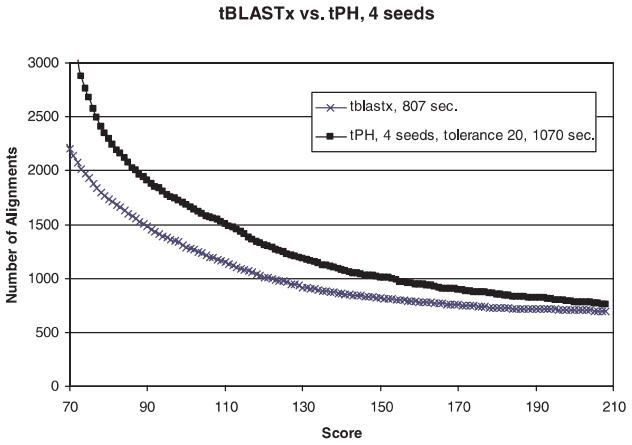

Fig. 3. The *x*-axis is the alignment score and the *y*-axis is the number of alignments with at least the corresponding score.

alignment would evade tBLASTx as the 6-frame translation would read into a wrong frame after the gap. At the best, tBLASTx would fragment the alignment into two segments in two different frames.

```
Q: AAC TTT GCA CAG AAC TGC AGT C-- TTG TTC TTC CCT
    N   F   A   Q   N   C   S       L   F   F   P
    N   F   A   Q   N   C   S       L           P
S:  N   F   A   Q   N   C   S   -   L   V   L   p

Q: TGG ATC ATG ACA AAT AAG TCT CAC ACA GTG CCG
    W   I   M   T   N   K   S   H   T   V   P
    W   I   +   T   N   +   S   H   T   V   P
S:  W   I   V   T   N   E   S   H   T   V   P
```

The website (http://www.bioinformaticssolutions.com/products/ph/doc/tphdoc.php) contains the user manual of tPH.

tPH outputs all alignments ranked by score. tPH encapsulates the functionality of BLASTp, tBLASTn, BLASTx and tBLASTx in an integrated suite. tPH also supports gaps and DNA-level frameshifts in alignments, which are missing from tBLASTx.

ACKNOWLEDGEMENTS

The authors thank Dr. John Tromp for his contribution to PH in general and Dr Wenhui Zou for her help to produce the figures in the paper. The work was partially supported by Bioinformatics Solutions Inc., NSERC OGP0046506, Canada Research Chair program and the Killam Fellowship.

REFERENCES

Altschul,S.F., Gish,W., Miller,W., Myers,E. and Lipman,D.J. (1990) Basic local alignment search tool. *J. Mol. Biol.*, **215**, 403–410.

Altschul,S.F., Madden,T.L., Schäffer,A.A., Zhang,J.H., Zhang,Z., Miller,W. and Lipman,D.J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.*, **25**, 3389–3402.

Brown,D., Li,M. and Ma,B. (2004) Homology search methods. In Wong,L. (ed.), *The Practical Bioinformatician*. World Scientific Press Singapore, pp. 217–244.

Chattaraj,A. and Williams,H.E. (2004) Variable-length intervals in homology search. In *Proceedings of the Second Asia-Pacific Bioinformatics Conference (APBC2004)*, Dunedin, New Zealand, January 18–22, pp. 85–91.

Li,M., Ma,B., Kisman,D. and Tromp,J. (2004) PatternHunter II: highly sensitive and fast homology search. *J. Bioinformatics Comput. Biol.*, **2**, 417–440.

Ma,B., Tromp,J. and Li,M. (2002) PatternHunter: faster and more sensitive homology search. *Bioinformatics*, **18**, 440–445.