BIOINFORMATICS APPLICATIONS NOTE Vol. 21 no. 1 2005, pages 122–123
doi:10.1093/bioinformatics/bth459

BlastAlign: *a program that uses* blast *to align problematic nucleotide sequences*

Robert Belshaw[1,*] *and Aris Katzourakis*[2]

[1]*Department of Biological Sciences, Imperial College, Silwood Park Campus, Ascot, Berks SL5 7PY, UK and* [2]*Department of Zoology, University of Oxford, South Parks Road, Oxford OX1 3PS, UK*

Received on April 20, 2004; accepted on July 26, 2004
Advance Access publication August 13, 2004

ABSTRACT

Summary: *BlastAlign* uses NCBI *blastn* to build a multiple nucleotide alignment and is intended for use with sequences that have large indels or are otherwise difficult to align globally. The program builds a matrix representing regions of homology along the sequences, from which it selects the 'most representative' sequence and then extracts the *blastn* query-anchored multiple alignment for this sequence. The matrix is printed and allows subgroups to be identified visually and an option allows other sequences to be used as the 'most representative'. The program contains elements of both *Perl* and *Python* and will run on UNIX (including Mac OSX) and DOS. An additional *Perl* program *BlastAlignP* uses *tblastn* to align nucleotide sequences to a single amino acid sequence, thus allowing an open reading frame to be maintained in the resulting multiple alignment.

Availability: It is freely available at http://www.bio.ic.ac.uk/research/belshaw/BlastAlign.tar and at http://evolve.zoo.ox.ac.uk/software/blastalign

Contact: r.belshaw@imperial.ac.uk

INTRODUCTION

Often, one wishes to align large files of nucleotide sequences that are problematic for one or more of the following reasons: they are of variable length, have large indels (INsertions/DELetions) or regions lacking homology at the nucleotide level, or the file has spurious entries or some reverse complemented sequences. This is a challenge for global alignment programs (Thompson *et al.*, 1999) and widely used programs such as *Clustal* cannot cope—and indeed were not designed to (Jeanmougin *et al.*, 1998). It is an area of active current research, with many new alignment programs and algorithms appearing (Bray *et al.*, 2003; Pollard *et al.*, 2004). Our approach to this problem is to use the well-known NCBI *blast* (Basic Local Alignment Search Tool) programs to align all sequences to the most representative one. The default output of *blast*, with which most users are familiar, is a series of pair-wise alignments called high-scoring segment pairs (HSPs). A less familiar output option is the flat query-anchored multiple alignment, which converts these HSPs into a multiple alignment for each query sequence. We present *BlastAlign* as a tool for utilizing this option to obtain rapidly a multiple alignment from nucleotide sequences where a full global alignment is either not required, or is not desirable because it would have to contain many large gaps to accommodate all the indels. We also include a *Perl* program *BlastAlignP*, which uses the co-ordinates of the HSPs from *tblastn* to produce a multiple nucleotide alignment from a single amino acid sequence.

PROGRAM FEATURE SUMMARY

- *BlastAlign* calls *blastn* to blast all sequences against each other and selects the most representative sequence from among the *blastn* output (see Program details section). It then extracts the query-anchored multiple alignment for this sequence and converts it into the Nexus and Phylip formats.
- *BlastAlign* uses *blastn* and hence only the more conserved regions are represented in the resulting alignment. This can be advantageous. Progressive alignment methods, such as *Clustal*, use a phylogeny to guide the alignment but this is not helpful for distantly related sequences that may only be related by a 'star' phylogeny. In addition, it may add noise because large indels that are unique to a particular sequence will be built into the alignment if they are encountered early on; this can often cause such algorithms to output poor quality alignments.
- A matrix is produced which shows the presence or absence of homology at 60 bp internals along an alignment. This allows subgroups to be identified visually, and an option allows the user to create new multiple alignments based on user-selected sequences.
- *BlastAlign* is extremely robust to the format and content of the input file—provided it is in FASTA—and writes its own files for input into *blastn*. It is not therefore necessary

*To whom correspondence should be addressed.

for the user to edit their raw files. An advantage inherited from *blastn* is that *BlastAlign* will cope with sequences that are reverse complemented or have no homology with others.

- The program will check to see if the output file from *blastn* would be too large if all the sequences were blasted against each other (default = 1 Gb). If this is the case, it finds the most representative sequence among an appropriate number of randomly selected sequences and then aligns all sequences to this one. To further increase the speed of aligning, an option allows the user to set the number of sequences used to find the most representative one.
- Further optional switches allow (1) sequences to be excluded from the final alignment on the basis of their length; (2) sequence names in the alignment to be abbreviated and stripped of characters that may cause problems with other programs; and (3) named sequences to be excluded.
- *BlastAlignP* uses NCBI *tblastn* to produce a multiple nucleotide alignment from a single amino acid sequence, thus allowing an open reading frame (ORF) to be maintained. It also includes options for replacing stop codons with gaps [its Phylip output is compatible with PAML (Yang, 1997)], and for excluding sequences from the final alignment based on length.

REQUIREMENTS

- Both *Perl* and *Python* need to be present (both are now standard in most LINUX systems).
- The NCBI *blastall* package and *formatdb* need to be installed (both are freely available as part of their standalone BLAST download; http://www.ncbi.nlm.nih.gov/BLAST/). If these two programs are not in the users PATH, their location can be manually inserted into the first few lines of the script where indicated. [Note, WU blast 2.0 (W.Gish, http://blast.wustl.edu) does not allow flat query-anchored multiple alignment and so cannot be used.]
- It is command line-driven only: either edit the first line of the script to show the location of *Perl* and type BlastAlign, or type perl BlastAlign to show the options.
- BlastAlign was written for LINUX but we have run it on Mac OSX and DOS.

PROGRAM DETAILS

The steps in the program are as follows: (1) A *Perl* script *BlastAlign* calls *blastn* to align sequences against each other (*blastn* output is called blast_out_raw) and summarizes the output in a file called blast_out. (2) It calls a *Python* script *BlastAlign.py* to convert blast_out into a matrix (blast_out_python) showing the presence or absence of regions of homology (called landmarks) at ~60 bp intervals along the sequences. (3) *BlastAlign* counts the cumulative number of matches at these landmarks for each element to find the most representative sequence (the total number of matches at each landmark is given in the first row of the matrix). (4) Finally, it converts the flat query-anchored multiple alignment of the most representative sequence in the original *blastn* output file to both the Nexus and Phylip formats (filename.nxs and filename.phy). The output file of *blast* occasionally contains errors, which can lead to a single sequence being misaligned by *BlastAlign*; however, *BlastAlign* checks the line length of each sequence in the final output and warns the user of such problems, which can then be corrected manually or the analysis can be run again with the problematic sequence(s) excluded using one of the options. *BlastAlignP* outputs Nexus and Phylip alignments and does not use *Python* or output a matrix of landmarks.

ACKNOWLEDGEMENTS

We are very grateful for the help from Vini Pereira and Douda Bensasson. This work was supported by The Wellcome Trust (R.B.) and NERC (A.K.).

REFERENCES

Bray,N., Dubchak,I. and Pachter,L. (2003) AVID: a global alignment program. *Genome Res.*, **13**, 97–102.

Jeanmougin,F., Thompson,J.D., Gouy,M., Higgins,D.G. and Gibson,T.J. (1998) Multiple sequence alignment with Clustal X. *Trends Biochem. Sci.*, **23**, 403–405.

Pollard,D.A., Bergman,C.M., Stoye,J., Celniker,S.E. and Eisen,M.B. (2004) Benchmarking tools for the alignment of functional noncoding DNA. *BMC Bioinformatics*, **5**, 6.

Thompson,J.D., Plewniak,F. and Pock,O. (1999) A comprehensive comparison of multiple sequence alignment programs. *Nucleic Acids Res.*, **27**, 2682–2690.

Yang,Z. (1997) PAML: a program package for phylogenetic analysis by maximum likelihood. *Comput. Appl. Biosci.*, **13**, 555–556.