Short communication

SWeBLAST: A Sliding Window Web-based BLAST tool for recombinant analysis

Mathieu Fourment^a, Adrian J. Gibbs^{b,*}, Mark J. Gibbs^c^a Department of Biological Sciences, Macquarie University, Sydney, NSW 2109, Australia^b 7 Hutt Street, Yarralumla, ACT 2600, Australia^c School of Botany and Zoology, School of Science, Australian National University, ACT 0200, Australia

A B S T R A C T

Article history:

Received 2 April 2008

Accepted 10 June 2008

Available online 17 July 2008

Keywords:

Virus recombinants

BLAST

RDP3

Transgene safety

We describe a simple Perl computer tool for matching successive subsequences of a query sequence using the BLAST facilities of Genbank. SWeBLAST helps identify 'parents' of recombinant sequences, even when these are themselves unrelated, thus it is complementary to methods that compare sets of aligned homologous sequences, and avoids the significant problem of these methods in having first to decide which sequences to compare. SWeBLAST searches may also be valuable for checking the recombination history of genes proposed for use as transgenes.

© 2008 Elsevier B.V. All rights reserved.

1. Introduction

Many viral genomes are recombinant. In the early days of gene sequencing, recombinant gene or peptide sequences were often found by chance, when different regions of a set of aligned sequences were found to have discordant phylogenetic signals (Gibbs and Cooper, 1995; Smith et al., 2000). However over the past decade there has been a surge of interest in the possibility that recombination is important in virus evolution and, as a result of the research effort, a large number of strategies, methods and algorithms for detecting recombinants in sets of aligned sequence have been published (Etherington et al., 2005; Forslund et al., 2004; Gibbs et al., 2000; Kosakovsky Pond et al., 2006; Lawrence et al., 2004; Milne et al., 2004; Weiller, 1998). Many of the algorithms have been combined into a user-friendly suite of programs called Recombination Detection Program 3 (RDP3) (Martin et al., 2005).

These methods detect phylogenetically anomalous subsequences in sets of aligned sequences. Thus they best detect recombination, usually homologous, when a dataset contains a recombinant and also sequences from its parental lineages. However they are less well suited to identifying heterologous recombinants, when sequences of one or more of the parental lineages is not present in the dataset, although RDP3 does detect recombinant regions of 'unknown' origin.

Another problem for all attempts to identify recombinants is that the dataset being analyzed may not contain the most closely related parental lineage sequences, which, if present, may have given much clearer evidence of recombination. Therefore it is important, when analysing recombinants, to check the Genbank databases for the closest parental lineage sequences to any phylogenetically anomalous subsequences, and this requires a BLAST or FASTA search of entire databases. This search is done most informatively using each phylogenetically anomalous subsequence and a similar length of its flanking regions. However if these regions are presented to BLAST as a single sequence then the matching strategy employed by BLAST (<http://en.wikipedia.org/wiki/BLAST>) will probably obscure the matches given by the regions that have weaker phylogenetic signals. This outcome may be overcome by doing BLAST searches with a series of subsequences that scan the different regions of the recombinant, and here we describe a simple computer tool for this purpose, and we illustrate its usefulness.

2. Materials and methods

2.1. The SWeBLAST program

SWeBLAST (Sliding Window WeB-based BLAST) is a command line program written in Perl. It has two components. The 'SWeBLAST.pl' component divides a query sequence, supplied as a file containing one or more sequences in FASTA format, into a series of subsequences of chosen length and overlap, and sends these successively to GenBank's BLAST facility to the specified database (e.g. 'nr') and program (e.g. 'blastn'). The search results are collated into a

Abbreviations: BLAST, basic local alignment and search tool; nts, nucleotides.

* Corresponding author. Tel.: +61 2 62816971; fax: +61 2 61255573.

E-mail address: Adrian.j.gibbs@hotmail.com (A.J. Gibbs).

single file, BLAST.TXT and, when completed, this is analyzed by the second component, SWeBLAST.parser.pl. The output of the parser is a series of CSV tables, which can be examined using a standard spreadsheet program, such as EXCEL.

The SWeBLAST output files are:

- (1) 'seq.fas'. A Fasta format file of the subsequences that were sent to BLAST as query sequences.
- (2) 'BLAST.txt'. A file containing a concatenation of all the BLAST output files.
- (3) 'value.csv'. A table of the e-values (exponential term only) for all the reported matches found by the BLAST program. Each row of this table corresponds to a sequence in the Genbank database that has matched any of the subsequences. Column A lists the names of the matched sequences, and the other columns in the table correspond to the successive subsequences of the query sequence, and lists the e-values with which each slice has matched particular sequences in the database.
- (4) 'rank.csv'. This table is like 'value.csv', but records the rank position of each match in the BLAST results for a particular sequence rather than its e-value.
- (5) 'stat.csv'. The rows in this table also correspond to each database sequence found to match the query sequence, and these are in the same order as in the 'value.csv' and 'rank.csv' files. The first column of the file again records the names of the matched sequences, the second column records the number of matches obtained with all slices of the query sequence, and subsequent columns contain a series of statistics of the e-values and rankings obtained by the database sequence with every slice of the query sequence; their minima, maxima, ranges, medians, means and standard deviations.
- (6) 'summary.csv'. This table differs from the others in that each row contains the results for one subsequence of the query sequence. Column A gives the name and position of the slice, and successive columns then give the names of the sequences with which it matched and the e-value of that match.

SWeBLAST may be downloaded at <http://www.bioinformatics.org/sweblast/>.

2.2. Data and software used in tests of SWeBLAST

Gene sequences were obtained from the Genbank database using the ENTREZ and BLAST search facilities, and were manipulated using BIOEDIT (Hall, 1999). Initially a representative set of 47 potyvirus sequences were aligned via their encoded amino acid sequences using the Transalign program (kindly supplied by Georg Weiller) and CLUSTALX (Jeanmougin et al., 1998) with default parameters. Additional sequences were aligned to the representative aligned set using the 'Profile Alignment Mode' of CLUSTALX. The relationships of sequences were assessed by the neighbor-joining tree method (Saitou and Nei, 1987) using the version of this algorithm in CLUSTALX, and the resulting trees were viewed using TREEVIEW (Page, 1996). The aligned sequences were checked for incongruent relationships using RDP version 3.22 (Martin et al., 2005) with default settings.

3. Example

A phylogenetic incongruity was identified, as described below, among the sequences of some potyvirus sequences and a search for virus sequences that represented the parental lineages was made using SWeBLAST. A dataset was compiled of partial coat protein gene sequences from viruses of the genus *Potyvirus*, fam-

Fig. 1. Neighbor-joining tree calculated from the cCP sequences of selected potyviruses. Each sequence is identified by the name and accession code of the virus isolate from which it was obtained. The sequences were from 24 isolates of cowpea aphid-borne mosaic virus, 11 isolates of 10 other species of the bean common mosaic virus lineage of potyviruses, and 36 other potyvirus species: bean yellow mosaic virus, D28819; beet mosaic virus, AY206394; chilli vein mottle virus, AJ237843; clover yellow vein virus, AB011819; cocksfoot streak virus, AF499738; daphne virus Y, DQ299908; Japanese yam mosaic virus, AB027007; Johnsongrass mosaic virus, Z26920; Konjak mosaic virus, AB219545; leek yellow stripe virus, AJ307057; lettuce mosaic virus, X97705; lily mottle virus, AJ564636; maize dwarf mosaic virus, AJ001691; narcissus yellow stripe virus, AJ311372; onion yellow dwarf virus, AJ510223; papaya leaf distortion mosaic virus, AB088221; papaya ringspot virus, X67673; pea seed-borne mosaic virus, D10930; peanut mottle virus, AF023848; pennisetum mosaic virus, AY642590; pepper mottle virus, M96425; Peru tomato mosaic virus, AJ437280; plum pox virus, D13751; potato virus A, Z21670; potato virus V, AJ243766; potato virus Y, X12456; scallion mosaic virus, AJ316084; shallot yellow stripe virus, AJ865076; sugarcane mosaic virus, AJ297628; sweet potato feathery mottle virus, D86371; Thunberg fritillary virus, AJ851866; tobacco etch virus, M15239; tobacco vein mottling virus, X04083; turnip mosaic virus, AF169561; wild potato mosaic virus, AJ437279; yam mosaic virus, U42596.

ily Potyviridae. Only the 'coherently evolving' region of their coat protein genes (cCP region) as defined by (Gibbs et al., 2008), were used; the cCP region of the potato virus Y genome (Genbank Reference Sequence NC.001616, Robaglia et al., 1989) encodes the sequence—DVNAG— at its N-terminus and the C-terminus of the VP at the other. Thus the sequences that were aligned were homologous and around 720 nucleotides in length.

A recent Entrez search of the Genbank nucleotide database with the phrase "cowpea aphid-borne mosaic virus" retrieved 24 sequences. The cCP region of these sequences together with those of ten other species of the bean common mosaic virus (BCMV) lineage, and 36 other potyviruses were aligned, and gave the neighbor-joining tree in Fig. 1. The 17 different CABMV sequences formed a monophyletic cluster; its two basally branching minor lineages (isolates NC.004013 to Y17824) only contained isolates from Africa, the others lineages were all from South America. All except two of the sequences, DQ397525 (isolate MG-Avr) and DQ397526 (isolate PE-Bnt), were similar distances from the outgroup, and the two exceptions were more distant. The two atypical sequences were from isolates collected from yellow passionfruit (*Passiflora edulis* f.

Table 1
The recombinant region of the cCP region of two isolates of cowpea aphid-borne mosaic virus, and the five most closely related regions of other BCMV lineage viruses in the Genbank databases in March 2008

Acronym ^a		e-Value
CABMV-1	QAVEFSLNPLVGYRKPPLLAIMHLFSNAPEAYLEMGNRGRLTTPRS	1E–18
CABMV-2	QAVEFSLNPLVGYCKPPLLAIMHLFSNAPEAYLEMGNRGRLTTPRS	9E–18
BCM-1	eqVEySLkPmVenaKPtLrqIMHhFSnAaEAYiEMrNSeRpympRy	1E–04
BCM-2	eqVEySLkPmVenaKPtLrqIMHhFSdAaEAYiEMrNSeRpympRy	2E–04
DenMV	eqVEySLkPmVenaKPtLrqIMHhFSdAaEAYiEMrNSeRpfmPRy	3E–04
CerMV	eqVEySLkPmVenaKPtLrqIMHhFSdAaEAYiEMrNSeRpfmPRy	3E–04
WMV	eqVEySLkPmVenaKPtLrqIMHhFSdAaEAYiEMrNSeRpfmPRy	3E–04

^a CABMV-1, cowpea aphid-borne mosaic virus MG-Avr (ABD62093); CABMV-2, cowpea aphid-borne mosaic virus PE-Bnt (ABD62094); BCM-1, bean common mosaic virus US10 (AAA80266); BCM-2, bean common mosaic virus US4 (AAA80268); DenMV, dendrobium mosaic virus (syn. BCMV) (AAA65048); CerMV, ceratobium mosaic virus (AAA65048); WMV, watermelon mosaic virus (CAE18428).

flavicarpa), one from Minas Gerais State in the central south east of Brazil and the other from more than 1000 km away in Pernambuco State in the far east of Brazil. The atypical sequences were closely similar to each other and to other CABMV cCPs throughout their length. Dotplot and Tinoco plot comparisons in BioEdit of these and the other sequences showed no repetitive or inverted regions. When they were used as queries to search the Genbank nucleotide and amino acid sequence databases the top 100 matches of both searches were all with CABMV and other BCMV lineage sequences. All the reported matches were of single unbroken sequences covering a mean of 94.9% (± 3.78 ; range 88–100%) of the query sequence, and with e-values ranging from $2e-83$ to zero. Most significantly, none of the database searches using standard BLAST found a discordant region in the atypical sequences. The possible cause of the atypical phylogenetic position of the two sequences was however revealed by RDP3 analyses which found a statistically significant likelihood (p -value ranging from 10^{-4} to 10^{-6}) of a recombinant sequence of unknown origin between nucleotides 331 (including gaps) and 488 in the CABMV_MG-Avr sequence and nucleotides 291 and 546 in the CABMV_PE-Bnt sequence. When the aberrant sequences were checked against the Genbank 'nr' nucleotide database using SWeBLAST with 'blastn' and a window of 99 and step of 33, the first and second ranked matches of all the subsequences were with the aberrant sequences, namely CABMV_MG-Avr and CABMV_PE-Bnt, as these are in Genbank. However the third ranked matches confirmed that the sequences were recombinants as although the terminal regions of the aberrant sequences matched other CABMV sequences with e-values of $e-22$ to $e-44$, subsequence slices 14–16 (nts 430–528 to 529–627) were not with viral sequences but matched bacterial and mammalian sequences with e-values 0.024–5.9. The fourth and subsequent matches were similar. The aberrant sequences were then checked with SWeBLAST using 'blastx', namely with the query sequence translated in all six reading frames and compared with the Genbank amino acid sequence database. This second SWeBLAST search found that the aberrant regions matched with sequences of the coat proteins of several distantly related potyviruses species of the bean common mosaic virus group, namely bean common mosaic virus itself, and ceratobium mosaic and telosma mosaic viruses. This indicated that the aberrant regions were probably from an as-yet-unreported potyvirus of the same lineage, and encoded a potyvirus coat protein-like protein sequence.

As a final check the atypical regions of the CABMV sequences were excised using BioEdit and used as single query sequences in 'blastx' searches of Genbank. They matched, but distantly, with the same regions of the cCPs of other species of the BCMV lineage as found by the SWeBLAST search (Table 1). We therefore conclude that the MG-Avr and PE-Bnt isolates of CABMV are most probably recombinants, and both contain the same region of the cCP gene of an as-yet-unknown species of the BCMV lineage of potyviruses.

SWeBLAST can also be used to make 'blastp' searches of the amino acid sequence databases and, by using the "Entrez Query" option of BLAST, the searches can be confined to chosen parts of the database. This latter option is particularly useful when the sequence being studied is one of a large number of sequences of the same species, as these will swamp attempts to find matches with other species. For example, when we attempted to investigate a region in one cCP sequence of potato virus Y that seemed related to the same region of the zucchini yellow mosaic virus gene, the output was swamped with sequences of other isolates of potato virus Y, however these were removed using the "Entrez Query" term "virus NOT potato virus Y {ALL}".

4. Discussion

SWeBLAST is a simple tool that is a valuable adjunct to recombinant detection methods, such as those implemented in RDP. It may also be used as the primary tool for checking for recombinant regions in newly determined gene sequences. For example the genomic sequence of *Chara australis* virus has been found, in this way, to contain regions significantly homologous to those of beet soil-borne mosaic benyvirus, beet necrotic yellow vein benyvirus and hepatitis E hepevirus (Marjo Torronen, Anne Mackenzie and Adrian Gibbs, in preparation). SWeBLAST might also be useful for checking genes proposed for use as transgenes as any heterologous recombinant regions they contained might affect risk assessments.

The specificity of the matches obtained with SWeBLAST is crucially influenced by the choice of Genbank database, and the length of the window and step. These are best determined empirically. The maximal e-value of a match between an atypical region of a query sequence and a homolog in the Genbank database is obtained when the query subsequence and homolog match with no flanking sequences. This occurs when the 'window plus step' length setting is less than the length of the atypical region of the query sequence. So it is useful to make several SWeBLAST searches using a range of window and step lengths. Also it is best, of course, to maximize the length of the window within the combined 'window plus step' length (e.g. use a step length no more than one-third of the window length). Closely related gene sequences can be found using nucleotide sequences (i.e. 'blastn') and a small window, but to identify more distantly related matches it is necessary to use amino acid sequences (i.e. 'blastx' and 'blastp') and a longer window.

An e-value obtained from a BLAST search indicates how many alignments with that score are expected by chance. A p -value indicates how often you can expect to obtain such an alignment, and can be calculated by the formula $p = 1 - e^{-E}$; e-values and p -values of less than 0.001 are in essence identical (Korf et al., 2003).

SWeBLAST is not the only computer tool for doing BLAST searches with subsequences. The 'Genotyping Tool' (Rozanov et

al., 2004) identifies the parent lineages and likely recombination history of new sequences of three major viral pathogens: human immunodeficiency virus 1 (HIV-1), hepatitis C virus (HCV) and hepatitis B virus (HBV); the program cuts the query sequence into subsequences, which are then compared with a set of reference sequences from one of the three viruses using BLAST. Likewise 'Phage_Finder' (Fouts, 2006), which identifies non-homologous recombinant prophages using a database of bacteriophage protein sequences.

References

- Etherington, G.J., Dicks, J., Roberts, I.N., 2005. Recombination analysis tool (RAT): a program for the high-throughput detection of recombination. *Bioinformatics* 21, 278–281.
- Forslund, K., Huson, D.H., Moulton, V., 2004. VisRD—visual recombination detection. *Bioinformatics* 20, 3654–3655.
- Fouts, D.E., 2006. Phage_Finder: automated identification and classification of prophage regions in complete bacterial genome sequences. *Nucleic Acids Research* 34, 5839–5851.
- Gibbs, M.J., Armstrong, J.S., Gibbs, A.J., 2000. Sister-scanning: a Monte Carlo procedure for assessing signals in recombinant sequences. *Bioinformatics* 16, 573–582.
- Gibbs, A.J., Ohshima, K., Phillips, M.J., Gibbs, M.J., 2008. The prehistory of potyviruses: their initial radiation was during the dawn of agriculture. *PLoS ONE* 3 (6), e2523.
- Gibbs, M.J., Cooper, J.L., 1995. A recombinational event in the history of Luteoviruses probably induced by base-pairing between the genomes of two distinct viruses. *Virology* 206, 1129–1132.
- Hall, T.A., 1999. BioEdit: a user-friendly biological sequence alignment editor and analysis program for Windows 95/98/NT. *Nucleic Acids Symposium Series* 41, 95–98.
- Jeanmougin, F., Thompson, J.D., Gouy, M., Higgins, D.G., Gibson, T.J., 1998. Multiple sequence alignment with Clustal X. *Trends in Biochemical Science* 23, 403–405.
- Korf, I., Yandell, M., Bedell, J., 2003. BLAST: An Essential Guide to the Basic Alignment Search Tool. O'Reilly & Associates, Sebastopol.
- Kosakovsky Pond, S.L., Posada, D., Gravenor, M.B., Woelk, C.H., Frost, S.D., 2006. GARD: a genetic algorithm for recombination detection. *Bioinformatics* 22, 3096–3098.
- Lawrence, C.J., Zmasek, C.M., Dawe, R.K., Malmberg, R.L., 2004. Lumberjack: a heuristic tool for sequence alignment exploration and phylogenetic inference. *Bioinformatics* 20, 1977–1979.
- Martin, D.P., Williamson, C., Posada, D., 2005. RDP2: recombination detection and analysis from sequence alignments. *Bioinformatics* 21, 260–262.
- Milne, I., Wright, F., Rowe, G., Marshall, D.F., Husmeier, D., McGuire, G., 2004. TOPALi: software for automatic identification of recombinant sequences within DNA multiple alignments. *Bioinformatics* 20, 1806–1807.
- Page, R.D.M., 1996. TREEVIEW: an application to display phylogenetic trees on personal computers. *Computer Applications in the Biosciences* 12, 357–358.
- Robaglia, C., Durand-Tardif, M., Tronchet, M., Boudazin, G., Astier-Manifacier, S., Casse-Delbart, F., 1989. Nucleotide sequence of potato virus Y (N Strain) genomic RNA. *Journal of General Virology* 70, 935–947.
- Rozanov, M., Plikat, U., Chappey, C., Kochergin, A., Tatusova, T., 2004. A web-based genotyping resource for viral sequences. *Nucleic Acids Research* 32, 654–659.
- Saitou, N., Nei, M., 1987. The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Molecular Biology and Evolution* 4, 406–425.
- Smith, G.R., Borg, Z., Lockhart, B.E.L., Braithwaite, K.S., Gibbs, M.J., 2000. Sugarcane yellow leaf virus: a novel member of the Luteoviridae that probably arose by inter-species recombination. *Journal of General Virology* 81, 1865–1869.
- Weiller, G.F., 1998. Phylogenetic profiles: a graphical method for detecting genetic recombinations in homologous sequences. *Molecular Biology and Evolution* 15, 326–335.