Published online 3 June 2010 *Nucleic Acids Research, 2010, Vol. 38, Web Server issue* ***W59–W63***
doi:10.1093/nar/gkq487

PLAST-ncRNA: Partition function Local Alignment Search Tool for non-coding RNA sequences

Satish Chikkagoudar[1], Dennis R. Livesay[2] and Usman Roshan[3,*]

[1]Department of Biostatistics and Epidemiology, University of Pennsylvania School of Medicine, Philadelphia, Pennsylvania, [2]Department of Bioinformatics and Genomics, University of North Carolina at Charlotte, Charlotte, North Carolina and [3]Department of Computer Science, New Jersey Institute of Technology, Newark, New Jersey, USA

Received February 15, 2010; Revised May 1, 2010; Accepted May 15, 2010

ABSTRACT

Alignment-based programs are valuable tools for finding potential homologs in genome sequences. Previously, it has been shown that partition function posterior probabilities attuned to local alignment achieve a high accuracy in identifying distantly similar non-coding RNA sequences that are hidden in a large genome. Here, we present an online implementation of that alignment algorithm based on such probabilities. Our server takes as input a query RNA sequence and a large genome sequence, and outputs a list of hits that are above a mean posterior probability threshold. The output is presented in a format suited to local alignment. It can also be viewed within the PLAST alignment viewer applet that provides a list of all hits found and highlights regions of high posterior probability within each local alignment. The server is freely available at http://plastrna.njit.edu.

INTRODUCTION

Alignment-based methods are widely used for identifying non-coding RNAs in genomes (1,2). BLAST (3) is a popular tool for this task but recently a partition function-based approach (4,5) and semi-global alignment approaches have also been applied (6,7). The partition function-based approach, which was proposed by us previously, is a local alignment tuning of the Probalign program (10,11) to detect potential homologous non-coding RNA (ncRNA) sequences in large genomes.

Detailed experimental results in ref. (5) show that the partition function-based approach achieves a significantly higher accuracy than the popular BLAST program and the Smith–Waterman local alignment implementation in SSEARCH (8) on benchmarks constructed from the RFAM ncRNA sequence database (9). Here, we present an online implementation of a simple algorithm that identifies all putative homologs in a given genome using the modified Probalign program. The server outputs all local alignment 'hits' above a user-specified mean posterior probability. This probability was earlier shown to be a better discriminator of true hits from false ones than the BLAST and SSEARCH *z*-scores (5).

The new server presented here is considerably different from eProbalign (10). The latter is designed for multiple alignment of protein and RNA/DNA sequences and cannot handle large sequences. Our new server is designed specifically for BLAST like homology search and identifies all potential homologs in a target genome (see homology search algorithm below).

INPUT PARAMETERS

The server takes as input single or multiple query RNA sequences and a target genome sequence both in FASTA format. It returns an error if the data contains characters other than IUPAC abbreviations or is not in FASTA format. The user can specify the gap penalties and the thermodynamic temperature T (Figure 1). We provide default parameter values for queries with and without flanking DNA. These were optimized on the training set of the RNA-genome benchmark (5) that can be found at http://www.cs.njit.edu/usman/RNAgenome. This benchmark contains ncRNAs aligned to putative homologs (as given by RFAM seed alignments) with large DNA flanks.

We also provide a subset of sequences in the 26 divergent RFAM families (average pairwise identity at most 60%) that were used in our earlier study (5). The user can scan the genome for sequences in the family by simply selecting their family of interest using the drop-down box (Figure 1). We provide links to the sequences in each family.

The Probalign strategy is to use suboptimal alignments determined by the input parameters to compute posterior probabilities from which the final alignment is produced

*To whom correspondence should be addressed. Tel: +1 973 596 2872; Fax: +1 973 596 5777; Email: usman@cs.njit.edu

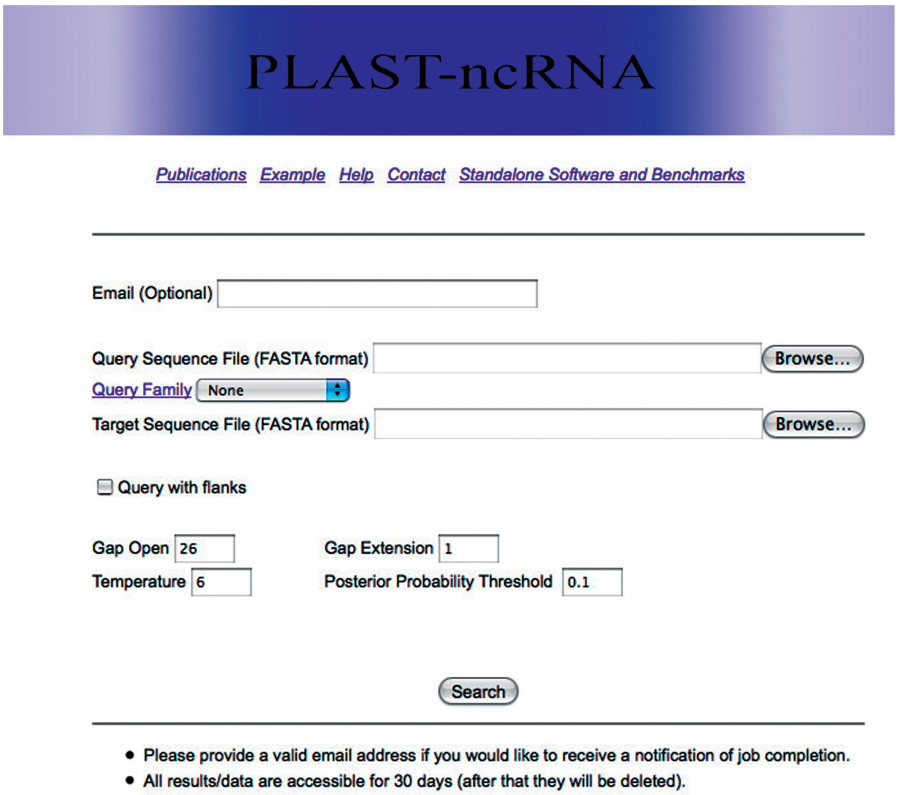

Figure 1. PLAST-ncRNA webserver main page.

(5,11). A large value of T would include suboptimal alignments with much lower scores than the optimal, whereas a value of 0 would use just the optimal alignments.

The mean posterior probability of an alignment is calculated by averaging across the posterior probability of each aligned nucleotide (these are automatically produced by Probalign). The server outputs all alignments between query and its similar sequence in the genome that have mean posterior probability at least the user specified threshold. By default, this is set to 0.1.

The server accepts an email address to inform the user that the results are ready. These are stored for 30 days before being automatically deleted. The main server page also lists links to standalone software and RNA to genome alignment benchmarks that were used in ref. (5).

OUTPUT AND ANALYSIS

While the job is being run we show the user the percentage completed (Figure 2). The output alignment can be viewed in simple text or using the PLAST-ncRNA alignment viewer in a format tuned for local alignment. The alignment start and end are the first and last match/mismatch aligned nucleotide and every 10 aligned nucleotides are annotated with their positions in the query and target (Figure 3).

In the text output option (Figure 3), the posterior probability for each aligned nucleotide is multiplied by 10 and then rounded to the next highest integer. Thus, a value of 4 means the posterior probability is between 0.3 and 0.4 (but excluding 0.3). The hits are sorted in descending order by their mean posterior probability.

The viewer is a Java applet that allows the user to see a list of all hits found (sorted by the mean posterior probability) and the aligned nucleotides colored by the posterior probability (Figure 4). By clicking on a hit the full alignment is displayed in the right column. The dark shades of blue represent high probabilities, whereas light represent low values. When the mouse is rolled over an aligned nucleotide the actual posterior probability is displayed.

HOMOLOGY SEARCH ALGORITHM

The modified Probalign program in ref. (5) outputs a complete alignment of the query to the target sequence. Our server implements a simple algorithm that replaces the portion of the target aligned to the query with a string of N's of equal length and realigns the query. This process repeats until no alignment of posterior probability above the user-specified threshold is found or an alignment of zero posterior probability is encountered (to ensure termination). Instead of replacing previously aligned portions with the string of N's, we could opt to remove them. However, this destroys the structure of the target genome and may lead to false hits.

The exact running time and spare requirements for aligning a query of length m to a target of length n is $O(mn)$. For target sequences of length >15 K nucleotides, we process them in slices of 15 K each. Thus, the running

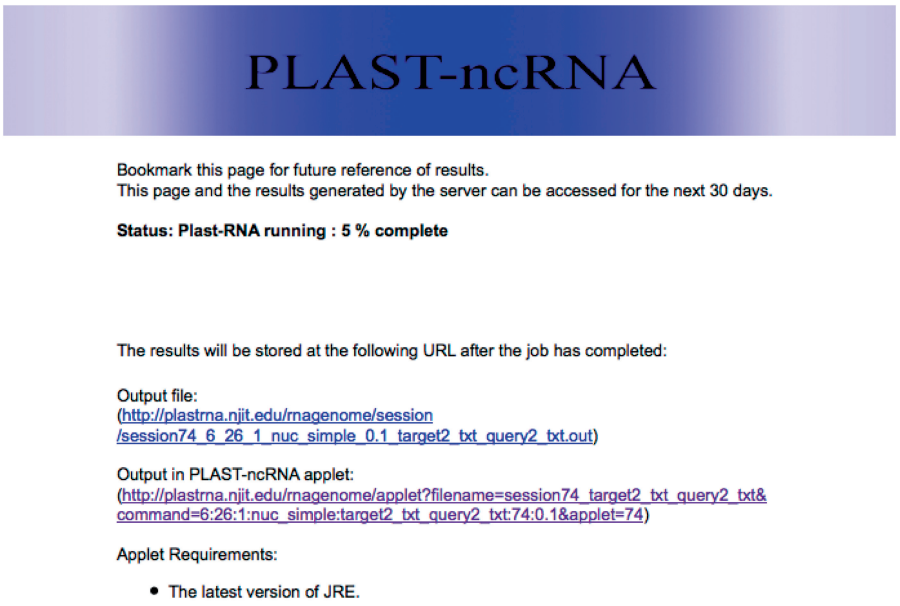

Figure 2. Output page of the server. Results can be viewed in plain text or with the alignment viewer applet.

```
Hit 2: Genomic region: 348747-348901, query region: 0-149, mean pp: 0.31857
                    0               15              30              45
>gi|20896|          ATCTTTGCGCTTGGGGCAATGACGCAGCTAGTGAGGTTCTAACCGAGGCGCGTCTATTGC
                    348747          348762          348777          348791
>gi|117977          ATCTTTGTGCTTGGGGCAATACGATAGTGTGTGAAGCTCTG-CTGATGCATCGTGATTGC
Post. Probs.        79100000000000000000108776777778888989888750556666554445710 0

                    60              75              90              105
>gi|20896|          TGGTTGAAAACTATTTCCAAACCCCCTCTTAGGCTTGAGGTAACTCAAGCCTTTGAGAAT
                    348806          348821          348836
>gi|117977          TAGTTGAAAACTACTCCAACACCCGTGAGAAGGCC-------ACT--GGCCAGCTCCAAT
Post. Probs.        00000000000000001098877776654333333333200000002220022332233

                    120             135                             150
>gi|20896|          TTCTGGAAGGGCTCCCT-TAGGG--------------TAAAGTCT
                    348857          348872          348887          348902
>gi|117977          TTCTGTTTTATCTCCCACTATGGGAGGAATTGCTTTGTTAATTCA
Post. Probs.        45666665544333344333201111100000000000000111 1
```

Figure 3. Output alignment in plain text format. This can be saved to the local disk or copied and pasted into files.

time is $O(mr)$ where $r = n/15\,000$. The output contains hits found in all the slices and shows the nucleotide positions in the original target instead of the 15 K slice.

We place a limit of 2500 bp on the query length. The server is not currently designed to align two large sequences. This does not limit the usage of the server since ncRNAs are usually much shorter than this length as determined by RFAM sequences.

SERVER PERFORMANCE

We provide two test examples for users interested in seeing a test run of our server. In the first case, the query is a ribonuclease P RNA sequence (RNaseP bact a RFAM family) and the target is a Clavibacter michiganensis complete genome (3 297 891 bp). In the second one the query is a Pea U4 snRNA sequence (U4 RFAM family) and the target is a Phytophthora infestans whole genome sequence (582 831 bp). Both the targets contain at least one putatively homologous RNA to their queries as given by RFAM family alignments. The nucleotide positions of the homologs are listed in the information links on the example page.

The server takes 30 and 3 min to return results for the first and second test cases, respectively. In the first case, hit number 21 with mean posterior probability of 0.25 and in the second one hit 2 with mean posterior probability of 0.32 (Figure 4) identify the putative homolog as given by RFAM (hits are identified by nucleotide positions). Our server also identifies additional hits of comparable mean posterior probability to the putative homolog.

We tested multiple queries by selecting the U4 family and aligning it to the Phytophthora infestans whole-genome sequence (same one used above). The server took 58 min to finish and in all 24 queries found the putative homolog given by RFAM.

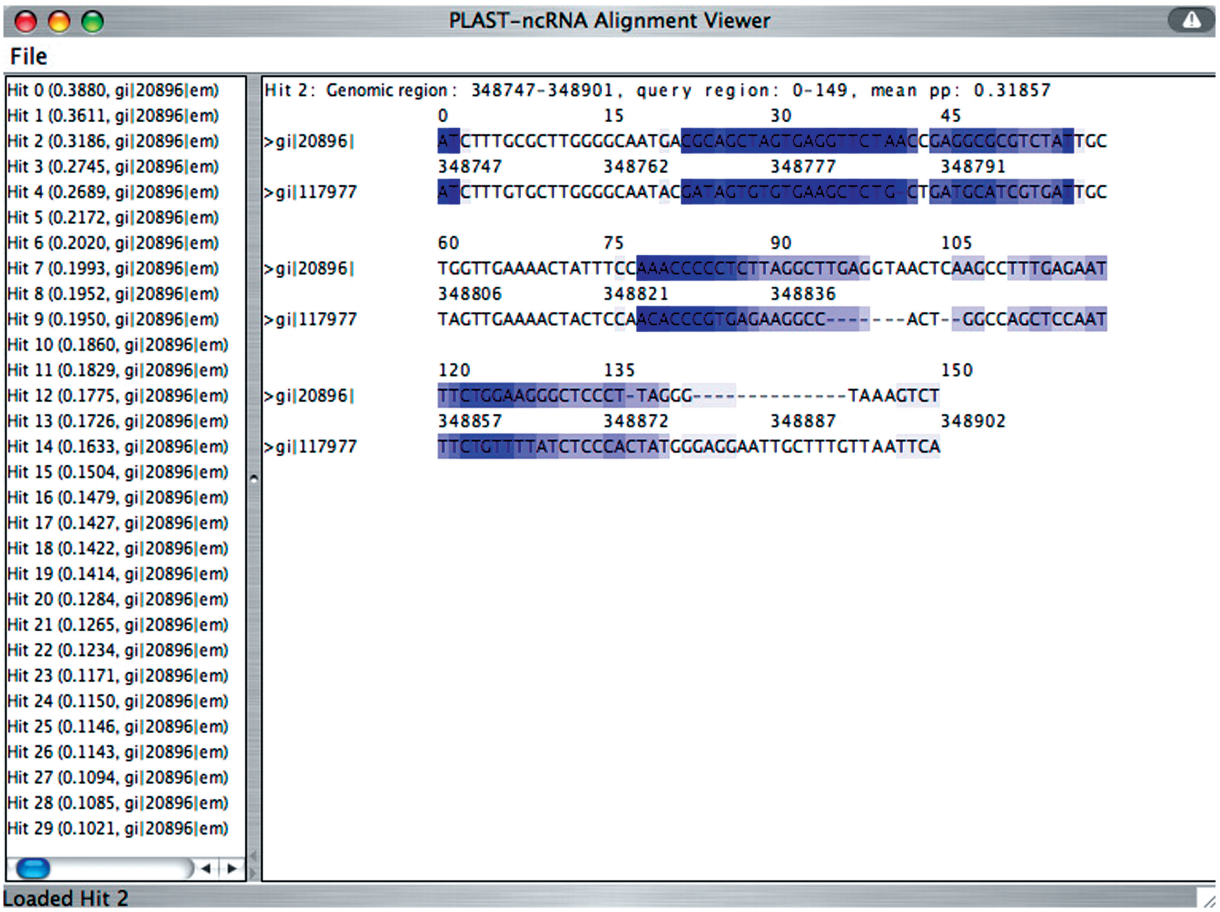

Figure 4. The PLAST-ncRNA alignment viewer applet provides a list of all hits sorted by their mean posterior probability and colored aligned nucleotides with intensity proportional to the posterior probability.

CONCLUSION

We provide a webserver that implements local alignment based on partition function posterior probablities and is specifically tuned for aligning query ncRNA sequences to putative homologs in large genomes.

ACKNOWLEDGEMENTS

We thank system administrators Gedaliah Wolosh and David Perel who have been very helpful in setting up the server and with technical issues related to the server.

FUNDING

United States National Science Foundation (grant 033-1654 to CIPRES cluster). Funding for open access charge: United States National Science Foundation.

Conflict of interest statement. None declared.

REFERENCES

1. Mosig,A., Zhu,L. and Stadler,P.F. (2009) Customized strategies for discovering distant ncRNA homologs. *Brief. Funct. Genomics Proteomics*, **8**, 451–460.
2. Menzel,P., Gorodkin,J. and Stadler,P.F. (2009) The tedious task of finding homologous noncoding RNA genes. *RNA*, **15**, 2075–2082.
3. Altschul,S.F., Gish,W., Miler,W., Myers,E.W. and Lipman,D.J. (1990) Basic local alignment search tool. *J. Mol. Biol.*, **215**, 403–410.
4. Hikosaka,K., Watanabe,Y.-I., Tsuji,N., Kita,K., Kishine,H., Arisue,N., Palacpac,N.M.Q., Kawazu,S.-I., Sawai,H., Horii,T. *et al.* (2010) Divergence of the mitochondrial genome structure in the apicomplexan parasites, Babesia and Theileria. *Mol. Biol. Evol.*, **27**, 1107–1116.
5. Roshan,U., Chikkagoudar,S. and Livesay,D. (2008) Searching for evolutionary distant RNA homologs within genomic sequences using partition function posterior probabilities. *BMC Bioinformatics*, **9**, 61.
6. Hertel,J., de Jong,D., Marz,M., Rose,D., Tafer,H., Tanzer,A., Schierwater,B. and Stadler,P.F. (2009) Non-coding RNA annotation of the genome of Trichoplax adhaerens. *Nucleic Acids Res.*, **37**, 1602–1615.
7. Copeland,C., Marz,M., Rose,D., Hertel,J., Brindley,P., Santana,C., Kehr,S., Attolini,C. and Stadler,P. (2009) Homology-based annotation of non-coding RNAs in the genomes of schistosoma mansoni and schistosoma japonicum. *BMC Genomics*, **10**, 464.

8. Smith,T.F. and Waterman,M.S. (1981) Identification of common molecular subsequences. *J. Mol. Biol.*, **147**, 195–197.
9. Griffiths-Jones,S., Moxon,S., Marshall,M., Khanna,A., Eddy,S.R. and Bateman,A. (2005) Rfam: annotating non-coding RNAs in complete genomes. *Nucleic Acids Res.*, **33**, D121–D124.
10. Chikkagoudar,S., Roshan,U. and Livesay,D. (2007) eProbalign: generation and manipulation of multiple sequence alignments using partition function posterior probabilities. *Nucleic Acids Res.*, **35(Suppl. 2)**, W675–W677.
11. Roshan,U. and Livesay,D.R. (2006) Probalign: multiple sequence alignment using partition function posterior probabilities. *Bioinformatics*, **22**, 2715–2721.