APPLICATION NOTES

Genomics & Informatics- Vol. 4(4) 167-169, December 2006

New Approach to the Analysis of Palindromic Structure in Genome Sequences

Seok-Won Kim[1], Yong Seok Lee[2], Sang-Haeng Choi[1], Sung-Hwa Chae[1], Dae-Won Kim[1] and Hong-Seog Park[1]*

[1]Genome Research Center, KRIBB, Daejeon 305-806, Korea, [2]Department of Parasitology and Malariology, PICR, College of Medicine and Frontier Inje Research for Science and Technology, Inje University, Busan, 614-753, Korea

Abstract

PABAP (**P**alindrome **A**nalysis by **BLAST P**rogram) is an analysis system that identifies palindromic sequences from a large genome sequence up to several megabases long. It uses NCBI BLAST as a searching engine, and data processing such as alignment filtration and detection of inverted repeats which satisfy user- defined parameters is performed by manipulating data after populating into a MySQL database. PABAP outperforms publicly available palindrome search program in that it can detect large palindrome with internal spacer at a faster speed from bacterial genomes. It is a standalone application and is freely available for noncommercial users.

Availability: This application was implemented with free software (Perl, Apache, MySQL, and NCBI BLAST) and is freely available to noncommercial users upon request. Analysis of user data can be carried out directly at http://chimp.kribb.re.kr/~javamint/palindrome.

Keywords: palindrome, inverted repeat, BLAST

Palimdromic sequence is a region in DNA containing a pair of inverted repeats, *i.e.,* a region whose 5'-to-3' sequence is identical on each DNA strand. Palindromic sequences include inverted repeats having central gap (spacer) and quasipalindromes with nonidentical pair of repeats. These structures are widespreadin the natural plasmids, viral and bacterial genomes, eukaryotic chromosomes and cell organelles. In case of prokaryotes, they may serve as binding sites for regulatory proteins, while short perfect palindromes are known as recognition sites for type II restriction-modification systems (Gelfand and Koonin,1997; Rocha *et al.*, 2001). Apparently, they often serve as site for protein-DNA interaction and mediate important cellular functions. Another important property of such motifs is their potential to form intra-strand hydrogen bonds within DNA molecules or in corresponding RNA transcripts. Therefore, they are contained in genes encoding functional RNA molecules, the structure of which depends on the formation of proper intra-strand bonding, and in different *cis*-acting genetic elements, like terminators, attenuators, plasmid and viral origins of replication. Protein binding and secondary structure formation are also modes of action for inverted repeats and related motifs in eukaryotic cells. For example, palindromes with a spacer of one nucleotide were identified in yeast sequences regulating cellular response to the accumulation of unfolded proteins in the endoplasmic reticulum (Mori *et al.*, 1998) and a heterodimeric complex was isolated that binds two palindromic sequences in the promoter region of the human erbB-2 gene (Chen and Gill, 1996). In mouse B lymphoma cells, palindromic and potential stem-loop motifs were identified as break-points during class switch recombination (Tashiro *et al.*, 2001); and the formation of intra-strand secondary structures is essential in the

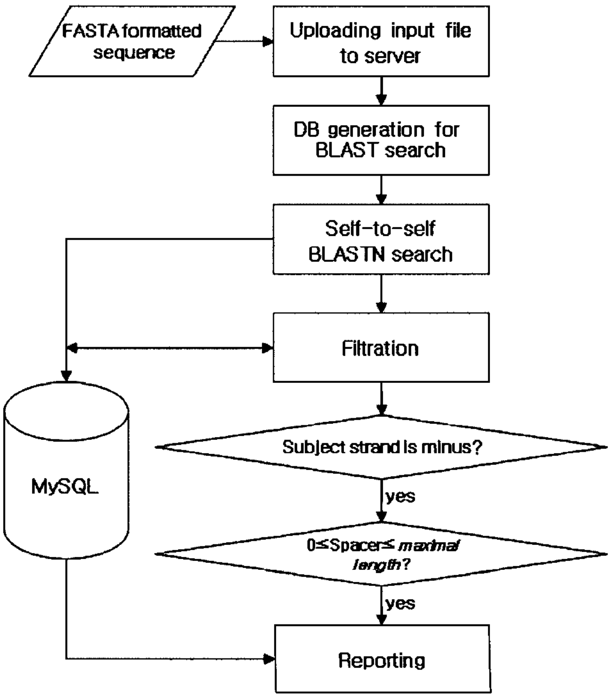
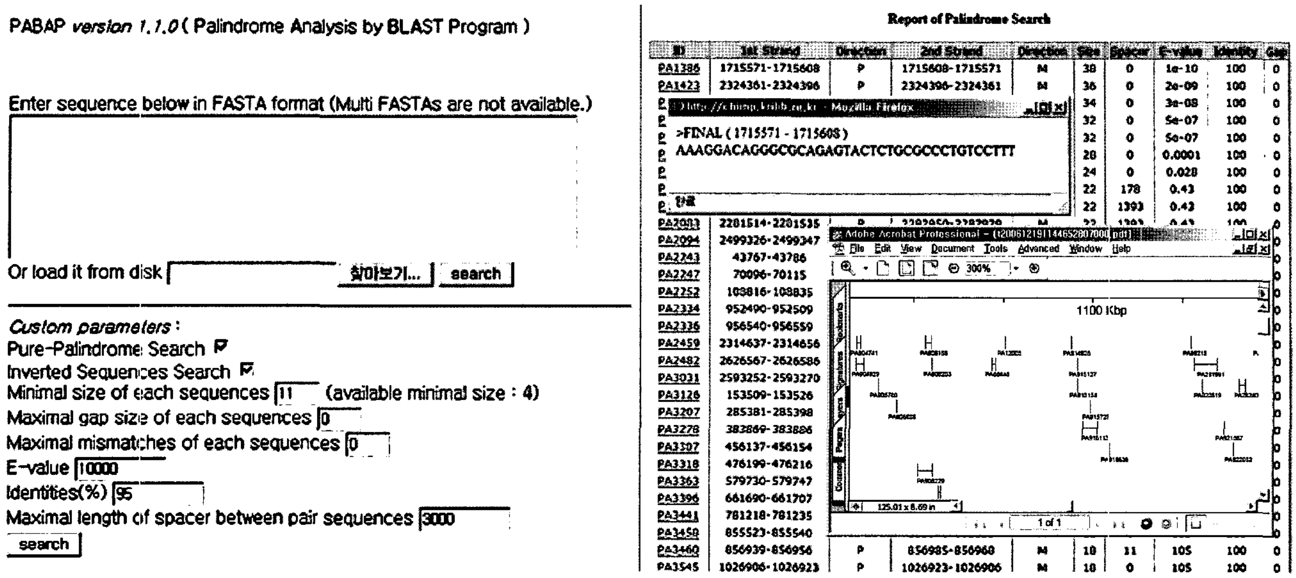

FASTA formatted sequence
Uploading input file to server
DB generation for BLAST search
Self-to-self BLASTN search
Filtration
MySQL
Subject strand is minus?
yes
0≤Spacer≤ maximal length?
yes
Reporting

Fig. 1. The system flow of PABAP.

*Corresponding author: E-mail hspark@kribb.re.kr,
Tel +82-2-879-8132, Fax +82-2-879-8139

Accepted 2 Dec 2006



PABAP *version 1.1.0* (Palindrome Analysis by BLAST Program)

Enter sequence below in FASTA format (Multi FASTAs are not available.)

Or load it from disk [ ] 찾아보기... search

Custom parameters :

Pure-Palindrome Search ☑

Inverted Sequences Search ☑

Minimal size of each sequences [11] (available minimal size : 4)

Maximal gap size of each sequences [0]

Maximal mismatches of each sequences [0]

E-value [10000]

Identities(%) [95]

Maximal length of spacer between pair sequences [3000]

search

Report of Palindrome Search

ID	1st Strand	Direction	2nd Strand	Direction	Size	Spacer	E-value	Identity	Gap
PA1386	1715571-1715608	P	1715608-1715571	M	38	0	1e-10	100	0
PA1423	2324361-2324396	P	2324396-2324361	M	36	0	2e-09	100	0
P					34	0	3e-08	100	0
P					32	0	5e-07	100	0
P					32	0	5e-07	100	0
P					28	0	0.0001	100	0
P					24	0	0.028	100	0
P					22	178	0.43	100	0
P					22	1393	0.43	100	0
PA2083	2281514-2281535								0
PA2094	2499326-2499347								0
PA2243	43767-43786								0
PA2247	70096-70115								0
PA2252	108816-108835								0
PA2334	952490-952509								0
PA2336	956540-956559								0
PA2459	2314637-2314656								0
PA2482	2626567-2626586								0
PA3031	2593252-2593270								0
PA3126	153509-153526								0
PA3207	285381-285398								0
PA3278	383869-383886								0
PA3307	456137-456154								0
PA3318	476199-476216								0
PA3363	579730-579747								0
PA3396	661690-661707								0
PA3441	781218-781235								0
PA3458	855523-855540								0
PA3460	856939-856956	P	856985-856968	M	18	11	105	100	0
PA3545	1026906-1026923	P	1026923-1026906	M	18	0	105	100	0

>FINAL (1715571 - 1715608)
AAAGGACAGGGCGCAGAGTACTCTGCGCCCTGTCCTTT

http://chimp.kribb.re.kr - Mozilla Firefox
Adobe Acrobat Professional - [t20061219114465280700.pdf]
File Edit View Document Tools Advanced Window Help
300%
1100 Kbp
Bookmarks
Signatures
Layers
Pages
Comments
125.01 x 8.69 in
1 of 1

Fig. 2. Snapshot of input page (left) and analysis report (right)

Table 1. Comparative performance of PABAP and Palindrome (EMBOSS).

Input sequence size (bp)	Inverted repeat (max gap size 3,000 bp)		Pure palindrome (no gap)	
	PABAP	Palindrome	PABAP	Palindrome
1K	1	2.6	1	0.1
10K	3	20	1	11
100K	55	3,542	64	1,724
1M	1,962	N/A	1,912	148,962
2M	5,093	N/A	5,132	N/A
3M	12,103	N/A	13,267	N/A

Real time (elapsed time) in second was measured from a Red Hat LINUX-based workstation (Intel Xeon 2.8GHz dual, 1GB memory, and kernel 2.4.21). Min size for each search was set to 11 bp, number of max mismatches, 3 bp. N/A means elapsed time longer than two days.

process of immunoglobulin gene rearrangement known as V(D)J joining (Cuomo *et al.,* 1996).

We developed a web application PABAP that can identify palindrome sequence using genome-scale sequence as an input. In this application, self-against-self BLASTN (Altschul *et al.*, 1990)search is employed to find out symmetric hits that can be used for identifying true palindromic sequences or inverted repeats after filtering process with customizable parameters. Fig. 1 shows the processing flow for this application. First, users provide a FASTA-formatted sequence as an input. Threshold values that are used for running BLAST and filtering hits after homology search are given simultaneously at the data input page. After formatting input sequence to generate BLAST-compatible database, BLASTN search is executed using the input sequence also as a query without low complexity filter. Output is set to tabular format to facilitate loading results directly into a MySQL database. Filtering is then carried out onto the data records to eliminate 1) self-matches that exactly overlap the same position within the sequence and 2) matches below threshold values, such as alignment length, sequence identity and gap size. Only matches in the opposite direction to the query sequences with spacer length below a threshold are then reported. Results are represented by a table and an image (PDF file) that shows the position of palindromic regions on the input sequence scale (Fig. 2).

We compared the performance of PABAP with PALINDROME, a program in publicly available EMBOSS package (http://emboss.sourceforge.net/). Time required for finding pure palindromic sequences or inverted repeats heavily depends on the input sequence length and BLAST search parameters (Table 1). The most crucial parameter determining sensitivity was E value cutoff; larger than 50,000 is recommended for detection of palindromes with short repeat units. In most cases, PABAP surpassed PALINDROME in terms of execution speed. PALINDROME was best suitable only for identification of true palindromes from sequences several kilobases long. Our application is superior to PALINDROME in finding symmetrical duplication at a gene level or inverted repeats flanking a genomic sequence that are target sites for site-specific recombination.

We have successfully applied this strategy for finding inverted repeats at the end of putative IS elements from the genome sequence of a *Corynebacterium* species (unpublished data). The strong point of PABAP lies in faster speed, flexible parameter setting, ability to identify inverted repeats with atypical geometry (long repeat units or spacers) or low identities, and graphical output that enable us to envisage palindromic sequence context at a genome level.

Acknowledgements

This work was supported by National Joint Agricultural Research Project of RDA (#2006201030009), Republic of Korea.

References

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

Chen, Y. and G.N. Gill. (1996). A heterodimeric nuclear protein complex binds two palindromic sequences in the proximal enhancer of the human erbB-2 gene. *J. Biol. Chem.* 271, 5183-5188.

Cuomo, C.A., Mundy, C.L., and Oettinger, M.A. (1996). DNA sequence and structure requirements for cleavage of V(D)J recombination signal sequences. *Mol. Cell Biol.* 16, 5683-5690.

Gelfand, M.S. and E.V. Koonin. (1997). Avoidance of palindromic words in bacterial and archaeal genomes: a close connection with restriction enzymes. *Nucleic Acids Res.* 25, 2430-2439.

Mori, K., N. Ogawa, T. Kawahara, H. Yanagi, and Yura, T. (1998). Palindrome with spacer of one nucleotide is characteristic of the cis-acting unfolded protein response element in *Saccharomyces cerevisiae. J. Biol. Chem.* 273, 9912-9920.

Rocha, E.P., Danchin, A., and Viari, A. (2001). Evolutionary role of restriction/modification systems as revealed by comparative genome analysis. *Genome Res.* 11, 946-958.

Tashiro, J., K. Kinoshita, and Honjo, T. (2001). Palindromic but not G-rich sequences are targets of class switch recombination. *Int. Immunol.* 13, 495-505.