BIOINFORMATICS APPLICATIONS NOTE
Vol. 25 no. 24 2009, pages 3319–3320
doi:10.1093/bioinformatics/btp587

Genome analysis

ABWGAT: anchor-based whole genome analysis tool

Sarbashis Das[1,*], Anchal Vishnoi[1] and Alok Bhattacharya[1,2,*]

[1]Center for Computational Biology and Bioinformatics, School of Information Technology, and [2]School of Life Sciences, Jawaharlal Nehru University, New Delhi, 110067, India

Received on July 21, 2009; Revised on September 25, 2009; Accepted on October 8, 2009

Advance Access publication October 14, 2009

Associate Editor: Alfonso Valencia

ABSTRACT

Summary: Large numbers of genomes are being sequenced regularly and the rate will go up in future due to availability of new genome sequencing techniques. In order to understand genotype to phenotype relationships, it is necessary to identify sequence variations at the genomic level. Alignment of a pair of genomes and parsing the alignment data is an accepted approach for identification of variations. Though there are a number of tools available for whole-genome alignment, none of these allows automatic parsing of the alignment and identification of different kinds of genomic variants with high degree of sensitivity. Here we present a simple web-based interface for whole genome comparison named ABWGAT (Anchor-Based Whole Genome Analysis Tool) that is simple to use. The output is a list of variations such as SNVs, indels, repeat expansion and inversion.

Availability: The web server is freely available to non-commercial users at the following address http://abwgc.jnu.ac.in/~sarba/. Supplementary data are available at http://abwgc.jnu.ac.in/~sarba/cgi-bin/abwgc_retrival.cgi using job id 524, 526 and 528.

Contact: dsarbashis@gmail.com; alok.bhattacharya@gmail.com

1 INTRODUCTION

A number of bacterial genomes have been sequenced so far and with the advent of new genome sequencing techniques many more are being added regularly. Particularly, a number of strains with different phenotypes, such as resistance to drugs are being sequenced in order to understand relation between genotype and phenotype. There are a number of tools that can help to analyze genomic sequences or more specifically can align two or more genome sequences in order to identify the sequence variations (Blanchette, 2007). In general, the alignments generated by these algorithms need to be parsed to list the sequence variations. Moreover, most of the methods are not able to identify all the variations, such as single nucleotide variations (SNVs), indels, repeat expansion and divergent region. There are also highly useful visualization tools available for comparative analysis of multiple genomes such as K-BROWSER (Chakrabati and Pachter, 2004) and MAUVE (Darling *et al.*, 2004). These tools help to identify gross changes and not list all the sequence differences that may exist. We had developed an algorithm ABWGC (Anchor-Based Whole Genome comparison) for pair-wise whole-genome comparison and identification of all sequence variants (Vishnoi *et al.*, 2007). It was shown to be useful for identification of genomic variations in related organisms including different species.

2 METHODS

In this report, we describe ABWGAT (Anchor-Based Whole Genome Analysis Tool) and a pipeline for automated identification and listing of genomic variations in a pair-wise manner. ABWGAT, based on a modified version of ABWGC algorithm, can identify insertion, deletion, SNVs, repeat expansion and inversion. A flow diagram describing the different modules of ABWGAT is shown in Figure 1. The basic features of ABWGC, is maintained in this tool. The process of identification of random anchors from reference and query genomes is essential as described before (Vishnoi *et al.*, 2007). The anchor length of 100 nt was found to be about optimum with respect to noise and computation time. It was decided on the basis of experimentation using different genomes and due to low probability of finding by chance a match for this length of sequence in a genome. This can be shown as follows. The match of an anchor in a genome has binomial distribution, due to large size of genomes. This can be approximated to a Poisson distribution. If the size of genome is 4 500 000 (generally, the size of a large bacterial genome), the probability of finding a fixed given sequence of length 100 in a genome of this size is $<2.8 \times 10^{-53}$ by a simple Poisson approximation of a binomial.

The distance between consecutive anchors of both the genomes can be used to determine if there is any insertion or deletion in any of the sequence. If the lengths are exactly same, it is possible that there may still be single nucleotide changes. Presence of SNVs in these sequences was determined by comparing dinucleotide frequencies of the two sequences. The exact position of the SNV is estimated by using a suffix-tree-based algorithm (Gusfield, 1997). The same algorithm is also used to find short inversion or indels (<7 nt). When the inter anchor length difference is >7 nt a global alignment algorithm (EMBOSS package http://emboss.sourceforge.net/) is used to identify the positions of SNVs and indels. Translocation, recombination and large inversions are detected by identifying anchor synteny breakpoints in the two genomes.

Many of the sequence variations in the genomes are due to expansion or contraction of tandem repetitive elements. Inter anchor regions suspected to have indels are also analyzed for tandem repeats using ‘tandem repeat finder’ (Benson, 1999). The output is parsed to find presence of tandem repeats, number of copies in the two inter anchor regions, length and genomic positions.

3 IMPLEMENTATION

The program runs on a Linux workstation and the web server is accessible through internet via Perl- and CGI-based web interface. The main codes are written in C++ and Perl. Most of the codes run in parallel using MPI (Massage Passing Interface). To implement MPI, we used MPICH package (http://www.mcs.anl.gov/research/

[*]To whom correspondence should be addressed.

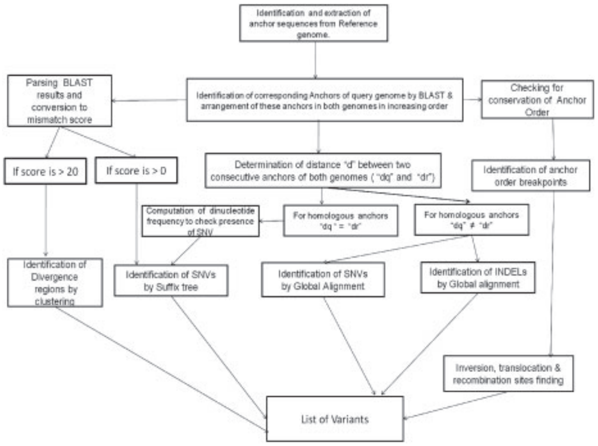
Identification and extraction of anchor sequences from Reference genome.
Parsing BLAST results and conversion to mismatch score
Identification of corresponding Anchors of query genome by BLAST & arrangement of these anchors in both genomes in increasing order
Checking for conservation of Anchor Order
If score is > 20
If score is > D
Determination of distance "d" between two consecutive anchors of both genomes ("dq" and "dr")
Identification of anchor order breakpoints
Computation of dinucleotide frequency to check presence of SNV
For homologous anchors "dq " = "dr"
For homologous anchors "dq" ≠ "dr"
Identification of Divergence regions by clustering
Identification of SNVs by Suffix tree
Identification of SNVs by Global Alignment
Identification of INDELs by Global alignment
Inversion, translocation & recombination sites finding
List of Variants

Fig. 1. Schematic diagram of ABWGAT.

Table 1. The results of three-pairs analysis of genomes

Comparison between	Total number of SNVs	Total number of indels	Total number of repeat expansion	Total number of inversion (long; short)
Acinetobacter baumannii AB0057 and *A.baumannii* B307_0294	335	415	6	5; 660
Francisella_tularensis _holarctica and Francisella_tularensis _holarctica_FTA	853	67	5	0; 29
Vibrio Cholerae M66_2 and *V.Cholerae* 0395	4259	92	7	3; 63

projects/mpi/mpich1/). All complete genome sequences and original annotation files were downloaded from NCBI (ftp://ftp.ncbi.nih.gov/genomes/Bacteria/).The genomes are selected using drop down menus. User-defined sequences can also be uploaded using a menu driven buttons. For accurate analysis, the user has to make sure that the two sequences are not very different from each other or from selected genomic sequences. The user can select the type of analysis, i.e. the nature of variations (SNV, insertion, inversion and repeat expansion). The results are retrieved later using the job ID number provided at the time of submission.

4 ANALYSIS OF GENOMES BY ABWGAT

The server was used to analyze genomic variations among strains of three different bacterial species. These species are *Acinetobacter*, *Francisella* and *Vibrio*. Query and reference genome were selected from the drop down menu. In the current implementation, only analysis of different strains of a species is permitted. The summary of the results of analysis of three pairs of genomes are shown in the Table 1. Depending upon the species, the number of variants was found to be quite different. For example, the number of indels in *A.baumannii* was much higher than that of other two organisms and *V.cholerae* displayed a very high number of SNVs. The output of the web server contains position in the genome, type of nucleotide change, gene name (if not intergenic) and predicted functions. The complete results can be viewed in Supplementary Material.

5 CONCLUSION

In this report, a web server for identification of genomic variations using a pair of completely sequenced genomes has been presented. The usefulness of ABWGAT is the ease by which comparative genomics analysis can be carried out without prior knowledge of bioinformatics or computation methods. The output comes out in listed format for easy interpretation unlike other comparative genomics tools that displays output either graphically or table format in the webpage.

ACKNOWLEDGEMENTS

We thanks to CDAC (Center for Development of Advance Computing) for help in implementing parallel computing.

Funding: Department of Biotechnology (Govt. of India); COE (DBT).

Conflict of Interest: none declared.

REFERENCES

Benson,G. (1999) Tandem repeats finder: a program to analyze DNA sequences. *Nucleic Acids Res.*, **27**, 573–580.

Blanchette,M. (2007) Computation and analysis of genomic multi-sequence alignments. *Ann. Rev.*, **8** 193–213.

Chakrabati,K and Pachter,L, (2004) Visualization of multiple genome annotations and alignments with the K_BROWSER. *Genome Res.*, **14** 716–720.

Darling,C. *et al.* (2004) Mauve: multiple alignment of conserved genomic sequence with rearrangements. *Genome Res.*, **14** 1394–1403.

Gusfield,D. (1997) *Algorithms on Stings, Trees, and Sequences: Computer Science and Computational Biology*. Cambridge University Press, New York.

Vishnoi,A. *et al.* (2007) Comparative analysis of bacterial genomes: identification of divergent regions in mycobacterial strains using an anchor based approach. *Nucleic Acids Res.*, **35**, 3654–3667.