academicJournals

Vol. 13(21), pp. 2103-2112, 21 May, 2014
DOI: 10.5897/AJB12.1903
Article Number: A3F1A0744746
ISSN 1684-5315

http://www.academicjournals.org/AJB

African Journal of Biotechnology

Full Length Research Paper

BengaSaVex: A new computational genetic sequence extraction tool for DNA repeats

OLUWAGBEMI, Oluseun Olugbenga[1,2]*, IMOLORHE, Samuel[2] and AGOZIE, Victor Okechukwu[2]

[1]Johns Hopkins Bloomberg School of Public Health, Johns Hopkins University, Baltimore, MD 21205, United States of America.
[2]Bioinformatics Unit, Department of Computer and Information Sciences, School of Natural and Applied Science, College of Science and Technology, P.M.B 1023 Covenant University, Ogun State, Nigeria, West Africa.

Received 18 May, 2012; Accepted 16 November, 2012

The scourge of infectious diseases is one of the problems contending with humanity. All infectious diseases are caused by pathogens. A major problem in biological research is the creation of enormous and redundant amounts of genomic data. From this large volume of generated data, biologists select a subset of each sequence known as DNA nucleotide subsequences "words", for extended scientific analysis. Computational biology aids this pruning process by providing computerized tools to generate vital information with biological significance from these data. This research aimed to develop new tools for extracting DNA repeats from the gene sequences and also to perform a comparative analysis with existing tools having similar or closely-related functions. We were able to develop *BengaSaVex* (GBenga Samuel Victor genetic sequence extraction tool) and provide a sequential *in-silico* genetic-sequence-filtering functionality to identify repeated DNA nucleotide subsequences within the genes of some microorganisms, evaluated the potential benefits and applications of identifying such repeated sequences, and finally, performed an *in-silico* comparative analysis between *BengaSaVex* and tandem repeat finder.

Key words: *BengaSaVex*, DNA, repetitive sequence, *in-silico* analysis, computational genomics.

INTRODUCTION

Over the years, biologists and computational biologists have conducted experiments related to the sequences of some pathogens and other micro organisms. One of the major problems in biological research is the creation of enormous and redundant amounts of genomic data from DNA sequencing projects performed (Baxevanis, 2003; Wang and Zhang, 2005; Myers et al., 2006; Lathe et al., 2008; Oluwagbemi and Omonhinmin, 2008; Oluwagbemi, 2012). Biologists select a subset of each sequence also known as DNA nucleotide subsequences "words", for extended scientific analysis. Computational biology complements this pruning process by providing repeat

*Corresponding author. E-mail: gbemiseun@yahoo.com or olu.oluwagbemi@covenantuniversity.edu.ng. Tel: +2348066533717.

finding programs to help analyze and provide useful information about interesting words, with the assumption that under or over-represented words have significant biological functions.

The biological significance of DNA repeats cannot be underestimated. DNA repeats play a significant role in the biological sciences (Jurka, 1998). Transposable elements are hidden in many repetitive DNA sequences. Experimental research and analysis on these repetitive sequences can help reveal transposable elements that are associated with genomic evolution.

The aim of this research was to develop a useful extraction tool (*BengaSaVex*), for *in-silico* analysis on the gene sequences of some microorganisms. Some pathogens are only being used as an example of how the program works. The objectives of this research were: (i) to develop *in-silico* simultaneous genetic sequence-filtering tools for *in-silico* analysis, by using object-oriented programming languages in C++, (ii) to identify repeated DNA nucleotide subsequences within the genes of some microorganisms, (iii) to evaluate the potential benefits of (ii) and (iv) to conduct a comparative analysis between *BengaSaVex - C++ version* and tandem repeat finder (Benson, 1999).

The biological rationale for undertaking this research stems from the fact that prominent feature of DNA can be identified by the frequency with which repeated substrings exist. For instance, this seems to be true for eukaryotes (Lander et al., 2001). Some repeats have been found to aid the provision of structural mechanism (Huang et al.,1998), while others have been identified to affect bacterial virulence, among microbes which have the tendency to cause human infections (van Belkum et al., 1998). This makes a study on repeats a promising and interesting one.

In this paper, we devised a genetic subsequence extraction tool using the C++ programming language for its implementations. We named this tool as *BengaSaVex*. The tool has the capability to extract repetitive DNA sequences from a collection of multiple gene sequences of microorganisms including infectious-disease causing organisms; then estimate the relationship that exists between the lengths of extracted repeated sequence and the computational time taken to extract these repeated sequences. Insight gained from the analysis of these duplicated sequences could help accelerate the pace of research in this domain by causing a motivation for the development of more efficient tools, especially, since there is a huge volume of sequence data available.

Several traditional repeat finding programs have been developed and applied to different gene sequences. They are as described in Table 1.

In summary, this paper details the algorithm underlying the development of *BengaSaVex,* describes the mechanism of data collection, explores the potential benefits of identifying DNA repeats in gene sequences of computational biology related research, presents the results generated by the new tools and its comparative analysis with some of the existing tools with similar or closely related functions (Saha et al., 2008).

MATERIALS AND METHODS

Data collection

Data for this research work was sourced from the National Center for Biotechnology Information (NCBI) (www.ncbi.nlm.nih.gov/) and also from the Sanger Institute (ftp://ftp.sanger.ac.uk/pub/pathogens/spn/). The sequence data of some microorganisms were sourced from various gene banks. Table 2 shows the sources of data used in the analysis. Each genome sequence data for respective organisms was simultaneously inserted into the input file of BengaSaVex.

Implementation

C++ programming language was used for the implementation of *BengaSaVex*. The multiple sequence data for different pathogens were stored inside an input file for BengaSaVex, for onward *in-silico* analysis. The input file (many.in.txt) contains multiple gene sequences of infectious disease-causing organisms to be analyzed, while the output file (many.out.txt) contains the results generated by BengaSaVex after running the executable version of the software (BengaSaVex.exe). BengaSaVex was developed using algorithms to compare sub-strings of gene sequences that are identical within genome sequence of pathogens as shown in (List 1). The algorithm depicted below shows its operations on repeat sequences.

List 1: *BengaSaVex Algorithm*

Begin
Input S1 ,……., Sm: the m set of pathogen gene sequence
While (!end of file) do
Get next set of gene sequence
for all i=1 to n do
function Search and Compare subsets of gene sequence S11 with S12 within S1,…… **until** S1n
Identify repeated sequences from S1,……Sm
Output repeats R1,……Rm each for Sequence S1,…..,Sm
end for
Output frequencies Rf1, Rf2, Rf3,……., Rfm for each repeat
Compute corresponding time (T1,……,Tm) to search and extract each repeat
Return S1,……,Sm; frequencies Rf1, Rf2, Rf3,……Rfm for each repeat; time (T1,……,Tm) to search and extract each repeat
End

RESULTS

BengaSaVex has the capability to perform sequential *in-silico* analysis on hundreds to thousands of large genome sequences. However, for the purpose of this manuscript, we only analyzed close to 15 large genome sequences. We present the results of eight of them as produced by *BengaSaVex*, based on *in-silico* analysis performed on the gene sequences of some organisms as shown in Table 2. Some of the repeats were found to be intergenic. We also provide the results of a comparative analysis of *BengaSaVex* with the tandem repeat finding program (Table 3).

Table 1. Tabulated literature review of some traditional repeat finding programs.

Related work	Description and reference
RepeatMasker	**RepeatMasker**, a prominent software, was developed to identify, classify and mask repetitive gene sequences. RepeatMasker finds repetitive sequence by performing an alignment of the input sequence against a library of known repeats (Smit and Green, 2002; Tarailo-Graovac and Chen, 2009).
RepeatScout	**RepeatScout** was another program developed to identify repetitive sequence in large genomic sequence (Price et al., 2005).
SAGRI	**SAGRI** (Spectrum Assisted Genomic Repeat Identifier), was a tool developed as a novel approach to detecting repeats in genomic sequences. SAGRI performs a double scan on the genome sequence (Do et al., 2008). It's a tool that was developed to efficiently locate possible ancient repeats in genomic sequences produced encouraging results (Singh et al., 2007).
RECON	**RECON**, an automated software for identifying repetitive sequences of newly sequenced genomes, was also developed (Bao and Eddy, 2002).
WindowMasker	**WindowMasker** was developed to identify and mask highly repetitive subsequences in the DNA sequence of a genome (Morgulis et al., 2006).
RepeatFinder	Algorithms such as **RepeatFinder** (Volfovsky et al., 2001) are also useful in *in-silico* analyses.
RepeatGluer	**RepeatGluer** (Pevzner et al., 2004)
PILER	Recently, **PILER** (Edgar and Myers, 2005) have increasingly automated the identification of repeat families from genomic sequence
ReAs	**ReAs** algorithm was applied in recovering ancestral sequences from transposable elements (Li et al., 2005).
REPuter	**REPuter** (http://bibiserv.techfak.uni-bielefeld.de/reputer/) , another repeat finding program, was developed by Kurtz and colleagues (Kurtz et al., 2001).
Dst	**Dst** (http://alce.med.umn.edu/newdst.html; Virtual Genome Center, unpublished), is another repeat finding program.
REPRO	**REPRO**, another program, helps to identify repeats in gene sequences of proteins [http://mathbio.nimr.mrc.ac.uk/~rgeorge/repro; (George and Heringa, 2000)].
RepeatAround	**RepeatAround** software was a repeat finding tool created by (Goios et al., 2006) - http://portugene.com/repeataround.html).
OMWSA	The **OMWSA** is another online tool for repeat finding and visualization (Du, 2007).
REPFIND	**REPFIND** online repeat finding tool (Betley et al., 2002), (http://zlab.bu.edu/repfind/form.html) was created by Bentley and colleagues.
Tandem Repeat Finder	**Tandem Repeat Finder** is yet another repeat finding program (Benson, 1999).

BengaSaVex - C++ version was used for this analysis, because it provided extraction time (in milliseconds) for the frequency of each direct repeated sequence. Analysis was performed on whole genome sequences of *Pseudomonas fluorescens* (Von Graevenitz and Weinstein, 1971; Picot et al., 2001), Hippea maritime DSM 10411 (Miroshnichenko et al., 1999), *Bartonella tribocorum* CIP 105476 (Heller et al., 1998), *Sinorhizobium meliloti* BL225C(Audic et al., 2009), *Brucella pinnipedialis* B2/94 (Whatmore, 2009; Audic et al., 2011), and *Staphylococcus aureus* [EMRSA15](methicillin-resistant strain) (Meier et al., 2001; Gordon and Lowy, 2008; Löffler

Table 2. *In-silico* analysis result from *BengaSaVex.*

Organism's DNA sequence	References/accession number	Frequency of repeat	*BengaSaVex* -C++ version search and extraction time (s)
Bartonella tribocorum CIP 105476	(Chomel and Piemont,1998) NCBI Reference Sequence: NC_010161.1 GI:163867306	Words with the maximum frequency (2) in the text are: CCGCCGCTGCTTTATTGATTAAACCCAGACAATTTC TAACTCGCCATAAGGAGAGAGACGTTCAGTATAG AGATATGACCACAAGCGACAGCCTTGCCAAAAACC CTTGCGTTTTCATAAACGTGACCATTTTTAAACAC TGATTATCATCACAAATATCATTCGTAATTTTTTCAC GGCTATTAACCACGGCATTACCATAGATTTGTA CATCATGATCAATATTGGATTTACCATAGACCTTTG CATTTTCATAAATTCTTGCAAAGGACCCCACCCG AGCCGAGCCAGAAACCTTTGCATCATCAAAAATAC TTGCAAAACTCTTAATCCAAGCATTACAAAAGACA TGGGCATTGCCATAAATACACGCCCCCCCATAAAT GCGTGATTTACCATAAACATGGGAATTCCCATAAA AGCATTATCACAAACCATGGCATTGCCATAAACTAG ACCACAGATCCTAGCATTGTTACTAATTTTTGCA AATGCCCCCGCCTTAACATCATCAAAATCTCTTAAA GCACGAATGCGGTGTACAGTAATCCCTTCAAAAG TTGAACGTTGAAACTGTTAGATTACCCACGAGGCG TGAGCCAGTTTTGCCAGGACGGTTTTACCGGATCT CTTGGAAACCCCTTGGAAACCCCTTGGAAACCCCT TGGAAACCCCTTGGAAACCCCTTGGAAACCCCTTG GAAACCCCTTGGAAACCCCTTGGAAACCCCTTGGA AACCCCTTGGAAACCCCTTGGAAACCCCTTGGAAA CCCCTTGGAAACCCCTTGGAAACCCCTTGGAAACC CCTTGGAAACCCCTTGGAAACCCCTTGGAAACCCC TTGGAAACCCCTTGGAAACCCCTTGGAAACCCCTT GGAAACCCCTTGGAAACCCCTTGGAAACCCCTTGG AAACCCCTTGGAAACCCCTTGGAAACCCCTTGGAA ACCCCTTGGAAACCCCTTGGAAACCCCTTGGAAAC CCCTTGGAAACCCCTTGGAAACCCCTTGGAAACCC CTTGGAAACCCCTTGGAAACCCCTTGGAAACCCCT TGGAAACCCCTTGGAAACCCCTTGGAAACCCCTTG GAAACCCCTTGGAAACCCCTTGGAAACCCCTTGGA TGTAGGAAACTGTAGGAAACTGTAGGAAACTGTAG GAAACTGTAGGAAACTGTAGGAAACTGTAGGAAAC TGCGACTGCGACTGCGACTGCGACTGCGACTGCG ACTGCGACTGCGACTGCGACTGCGACTGCGACTG CG ACTGCGACTGCGACTGCGACTGCGACTGCGACTG CGACTGCGACTGCGACTGCGACTGCGACTGCGAC TG	35.637
BORRELIA afzelii Pko complete genome sequence	GI: 384206106 NCBI Reference Sequence: NC_017227.1	Nil	0.496

Table 2. Contd.

Sinorhizobium meliloti BL225C complete genome	NCBI Reference Sequence: NC_017322.1	Words with the maximum frequency (2) in the text are: CCGCTTGTCCCTTCTCCCCGCCTGCGGGGAGAAGGT GCCGGCAGGCGGATGAGGGGCGGGTGACCGCTAA GATTCCTTTTTCTGCGCAAATCAGATTCACCATTCAGG CCGGTGAAGGAGGCCGGCCTTTATGGCGAGAC CTTTTTCGAATGATCTTCGGGAACGCGTTGTCGATGC GGTGACGGGCGAGGGCCTATCGTGCCGGGCAGC GGCCAAGCGCTTCGGCATCGGCATCAGCACCGCGAT CGATTGGGTGCGGCGGTTTCGCGAGACGGGCAGC GCCGCACCCGGCCAGATGGGTGGGCACAAGCCCCG CAAGCTTTCCGGTCCGCACCGGGCTTGGCTGCTTT GCCGCTGCCGCGAGCGCGACTTCACGCTGCACGGAC TTGTCGCCGAGTTGAGCGAGCGCGGCCTGAAGGT GGATTATCGCGCCGTCTGGACCTTCGTGCACGAAGAG GGGTTGAGTTATAAAAAAAGACGCTGGTCGCCA GCGAACGGGAGCGGCCCGACGTCGCCCGCCACCGG GCACGATGGCTGAAGCACTGCCCCGGAATTGATCC AGCCGCCGGCAGTGGGCGATTTGTGCAAAACCCTTC GGGCGGCCCATATTGCGGTGCCTTGTCGCGAAAA ATCCGGCTTGCAGGCGGACGGCCTGCGGCGCCGGAT TTTCCACGAAAGTCCCTCGCAATTTGGGCCGTCA	134.556
Brucella pinnipedialis B2/94 chromosome 1, complete sequence	NCBI Reference Sequence: NC_015857.1	Words with the maximum frequency (2) in the text are: AATGCAGCGCACTGGCGCGATCTGCCTGCGACCTTC GGCAAATGGACAGCGGTTCATGCCCGCTTTCGGC:: :GCTGGTCGCACGCCGGTGTATGGGAAAGGCTTTTCC ATGCCCTGGCTGATACGCCGGACTTTGAATATGT:: CCTCATTGATAGCACCATATCGAAAGTCCACGCAGAT GCGGCGGGCGCAAAAGGGGGGCTGAAGCTGCCT: ::GCATCGGTCGCTCGCGCGGTGGATTGACGACCAAG CTGCATGCTGTTGTCGATGCTATCGGCCTACCGCT:: ::GCGAATAAAGCCAACACCCGGCCATTATGGTGACTG TCCGCAAGCTTCAAGCCTTCTATCCGGCTTAGAG: ::TGGATGGCTGCCAATGCAGCGCACTGGCGCGATCT GCCTGCGACCTTCGGCAAATGGACAGCGGTTCATG: ::CCCGCTTTCGGCGCTGGTCGCACGCCGGTGTATGG GAAAGGCTTTTCCATGCCCTGGCTGATACGCCGGA:: ::CTTTGAATATGTCCTCATTGATAGCACCATATCGAAA GTCCACGCAGATGCGGCGGGCGCAAAAGGGGGG ::CTGAAGCTGCCTGCATCGGTCGCTCGCGCGGTGGA TTGACGACCAAGCTGCATGCTGTTGTCGATGCTAT:: ::CGGCCTACCGCTGCGAATAAAGCCAACACCCGGCCA TTATGGTGACTGTCCGCAAGCTTCAAGCCTTCTA:: ::TCCGGCTTAGAGGGTGTGGGGCATGTCATTGCTGAT GCGGCCTATGATGCCGATCACTTAAGGGCCTTCA:: ::TTGCCAGCAATCTCAAGGCAACGGCTCAGATCAAGG CCAATCCAACACGTTCCAGTGTCCCAACAATCGA: ::CTGGAGGCTGTACAAGGAACGCCATCAGATTGAATG CTTTTTTAACAAGTTGAAACGCTATCGTCGTATT::	26.428

Table 2. Contd.

Pseudomonas fluorescens SBW25 complete genome	GI:229587578 NCBI Reference Sequence: NC_012660.1	Nil	400.238
Staphylococcus (methicillin-resistant)	EMRSA-15 genome ftp://ftp.sanger.ac.uk/pub/pathogens/sa/	Words with the maximum frequency (3) in the text are: ::tttaacttaagttattagagcctcttatgcagttgctcagtcaactgtataccttttgac::	124.688
Staphylococcus aureus strains- Epidemic EMRSA-16lineage	MRSA252.dna ftp://ftp.sanger.ac.uk/pub/pathogens/sa/	: Nil	1.351
Staphylococcus aureus- Highly transmissible MRSA sequence type(ST) 239 by MLST	EMBL/GenBank databases with accession number FN433596. ftp://ftp.sanger.ac.uk/pub/pathogens/sa/	Nil	0.619

Table 3. *In-silico* comparative analysis between ***BengaSaVex*** and some repeat finding programs (with respect to time only).

Sequence	***BengaSaVex*** **(s)**	**Tandem repeat finder (s)**
Bartonella tribocourm CIP 105476	35.637	60.15
BORRELIA afzelii Pko NCBIReference Sequence: NC_017227.1	0.496	0.544
Sinorhizobium meliloti BL225C complete genome	134.556	65.12
Brucella pinnipedialis B2/94 complete genome	26.428	41.96
Staphylococcus (methicillin- resistant)	124.688	271.36
Staphylococcus aureus strains- Epidemic EMRSA-16lineage	1.351	21.62
Staphylococcus aureus MSSA476- methicillin-sensitive strain	95.855	216
Staphylococcus aureus- Highly transmissible MRSA sequencetype(ST) 239 by MLST	0.619	4.95

et al., 2010), *Staphylococcus aureus* [Epidemic EMRSA-16 lineage], *Staphylococcus aureus* [MSSA476-methicillin-sensitive strain], *Staphylococcus aureus* [highly transmissible MRSA sequence type(ST) 239 by MLST(TW20) and the *Heamophilus Influenza.* Their respective accession numbers were provided in the following section. These results (Tables 2 and 3) show that BengaSaVex can be used as a complementary tool with other existing repeat finding programs. REFIND did not work on long sequences, and so was not included in Table 3.

BengaSaVex GUI shows the functionalities of the tool for inputting data, analyzing, outputting results of extracted repeats, frequency of extracted repeats, and time taken to extract the repeats (Figure 1).

DISCUSSION

Results produced show that *BengaSaVex* can be used as a complementary tool for repeat finding related researches. Research on repeated sequences can help

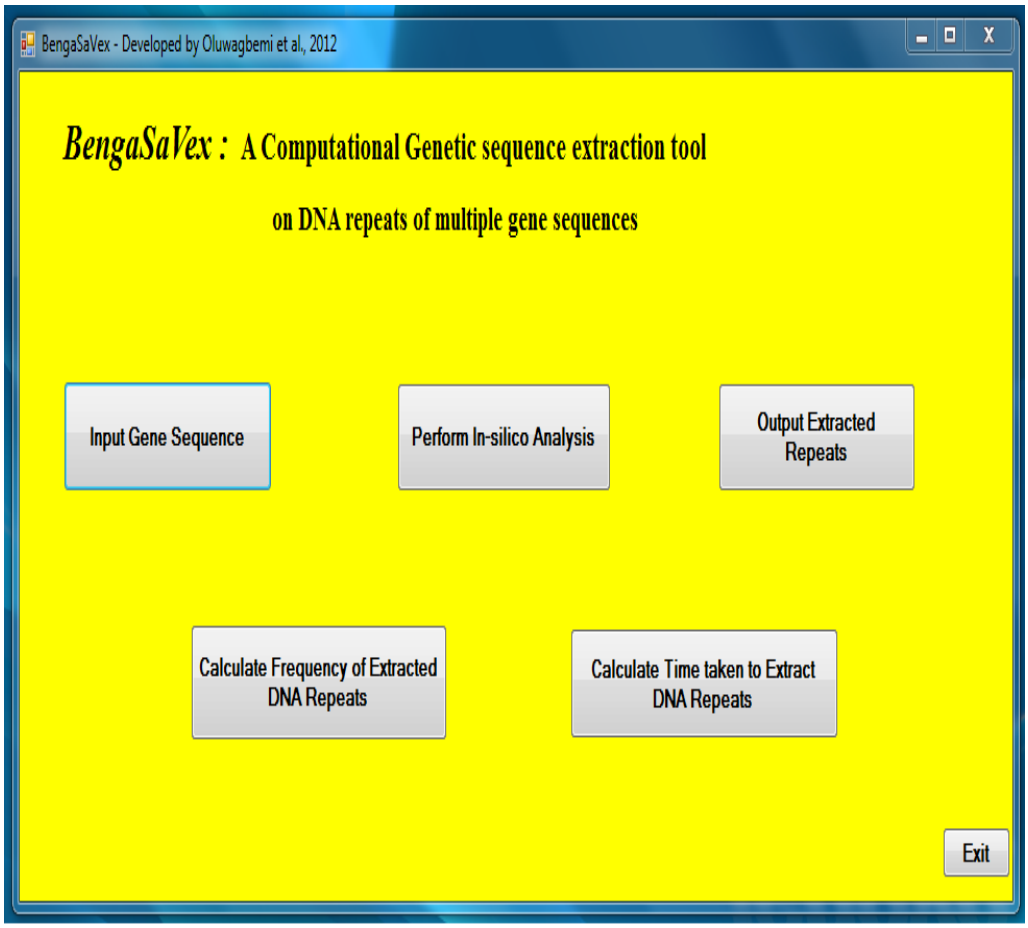

Figure 1. Graphical User Interface design of *BengaSaVex.*

provide interesting discoveries in the study of polymorphic patterns. Understanding the relationship between redundant gene filtering algorithms, programs and the corresponding genetic sequence they process, can help provide insight to developing programs with increased efficiency in carrying out this pruning process. This in turn, will help hasten or speed up the pace of research on DNA repeats, duplicated regions, sequence alignments and redundant genetic sequences of organisms and useful medicinal plants.

BengaSaVex has an added advantage to extract repeat sequences from multiple gene sequences of organisms, of which pathogens' are just one of the sample data. *BengaSaVex* also provides the corresponding frequencies of extracted sequences, and the time taken. *BengaSaVex* finds repeats in gene sequence of organisms.

Multifaceted applications of repeat analysis

Computational analysis finds expression in the processing of DNA repeats. Scientific research has found that DNA repeats help enhance flexibility in genetic and phenotypic features of pathogens and microorganisms (van Belkum et al., 1998). Variability in DNA repeats could help provide information about functional and evolutionary information on genetic diversity of such organisms (van Belkum, 1999a). Van Belkum as well as Delihas (van Belkum et al., 1999; Delihas, 2011), discovered and revealed the vital role sequence repeats play with the regulation of microbial gene expression. The significance of sequence repeats in epidemiologic typing cannot be underestimated (van Belkum, 1999b). Sequence repeats were also detected in *Escherichia coli*' sequence (Gur-Arie et al., 2000). Other scientists identified the potentials of DNA repeats in detecting certain virulent genes in pathogenic bacteria such as *H. influenza* (Hood et al., 1996; Power et al., 2009). Jansen and colleagues conducted an in-depth research on prokaryotes by detecting genes that are related to DNA repeats (Jansen et al., 2002; Treangen et al., 2009). Other scientists, such as Godde and Bickerton conducted similar experiments (Godde and Bickerton, 2006). Other related works that have been done in this regard are those

those of Cui as well as Bolotin (Cui et al., 2008; Bolotin et al., 2005). The application of DNA repeats have been emphasized in various infectious disease research over the years. Several functions of repeated sequences in MYCOPLASMA genomes have been highlighted in some studies (Ruland et al., 1990; Himmelreich et al., 1996; Himmelreich et al., 1997; Altshuler et al., 2000; Chambaud et al., 2001; Jaffe et al., 2004; Minion et al., 2004; Mrázek, 2006; Kassai-Jáger et al., 2008; Ma et al., 2008; Ma et al., 2012). DNA sequence repeats have also been found in enteric pathogens that are responsible for bacillary dysentery in humans (Jin et al., 2002; Wei et al., 2003; Yang et al., 2003; Phalipon and Sansonetti, 2007; Saurabh et al., 2011; Sun et al., 2011). Other studies have also revealed the significance of conducting comparative analyses and repeats in the genomes of various organisms (Powell et al., 1996; Chen et al., 2003; Ju et al., 2005; Rahim, 2008; Shikano et al., 2010; Labbe et al., 2011; Saker et al., 2011; Tyagi et al., 2011). Another study characterized repeats within sequences of exclusively prokaryotic genomes (Coenye and Vandamme, 2005).

A study has also shown the significance of repeated sequence in proteins and their relevance in network evolution (Hancock and Simon, 2005). Repeated sequences have the tendency of modifying other gene data to which they are associated, thus having the tendency of playing a role in the generation of genetic variation that underlies adaptive evolution (Kashi et al., 1997; Kashi and King, 2006). As stated above- genetic disorders do not cause disease; disease is defined as caused by an infectious agent (Clancy and Shaw, 2008). Research related to duplicated regions within gene sequences of microorganisms is of paramount interest in the field of computational biology and bioinformatics (Petes and Hill, 1988; Andersson and Hughes, 2009). Gene duplication has been found to be responsible for evolutionary mechanisms (Zhang, 2003). Duplicated regions in some organisms' chromosomes have also been found to play host to essential genes (Hillyard and Redd, 2007). Duplicated regions within the sequences of microorganisms like bacteria, play a significant role in their adaptation (Anderson and Roth, 1977). Scientists have also highlighted the relevance of duplicated regions within the sequence of certain pathogens (Larsson et al., 2005).

Conclusion

We developed *BengaSaVex* (a computational biology/-bioinformatics tool) for identifying and extracting repeats in gene sequences. This tool will complement other existing repeat finding tools to provide support for biological research. Future work on *BengaSaVex* is to improve the efficiency and also develop an online version.

Conflict of Interests

The author(s) have not declared any conflict of interests.

ACKNOWLEDGEMENTS

The authors acknowledge the National Center for Biotechnology Information (NCBI) and the Sanger Institute for making the gene data in their GenBank publicly available for research purpose. Other authors whose data were used for *in-silico* analysis in this manuscript have been referenced accordingly. The corresponding author also acknowledges the Fulbright Foreign Scholarship Board of USA. This research was partly funded by The Oluwagbemi Research, Development and Philanthropic Foundation (TORDPF). Shorter version of this paper has been submitted to an international conference. Supplementary files:*Executable version for *BengaSaVex* - C++ version is available on request from the corresponding author or can be downloaded as GENEIV.zip file by using a Google mail account from the web link specified below: https://docs.google.com/a/covenantuniversity.edu.ng/open?id=0B0YrEkxfW3Y6WnBqZDI2SnpsQTA

REFERENCES

Altshuler D, Daly M, Kruglyak L (2000). Guilt by association. Nat. Genet. 26:135-137.

Anderson RP, Roth JR (1977). Tandem Genetic Duplications in Phage and Bacteria. Annu. Rev. Microbiol. 31: 473-505.

Andersson DI, Hughes D (2009). Gene Amplification and Adaptive Evolution in Bacteria. Annu. Rev. Genet. 43:167-195.

Audic S, Lescot M, Claverie J, Cloeckaert A, Zygmunt MS (2011). The genome sequence of Brucella pinnipedialis B2/94 sheds light on the evolutionary history of the genus Brucella. BMC Evol. Biol. 11:200.

Audic S, Lescot M, Claverie J, Scholz HC (2009). Brucella microti: the genome sequence of an emerging pathogen. BMC Genomics 10:352.

Bao Z, Eddy SR (2002). Automated de novo identification of repeat sequence families in sequenced genomes. Genome Res. 12:1269-1276.

Baxevanis AD (2003). The Molecular Biology Database Collection: 2003 update. Nucleic Acids Res. 31(1):1-12.

Benson G (1999). Tandem repeats finder: a program to analyze DNA sequences. Nucleic Acids Res. 27(2):573-80.

Betley JN, MC Frith, JH Graber, S Choo, JO Deshler (2002). A ubiquitous and conserved signal for RNA localization in chordates. Curr. Biol. 12:1756-1761.

Bolotin A, Quinquis B, Sorokin A, Ehrlich SD (2005). Clustered regularly interspaced short palindrome repeats (CRISPRs) have spacers of extrachromosomal origin. Microbiology 151(8):2551-2561.

Chambaud I, Heilig R, Ferris S, Barbe V, Samson D, Galisson F, Moszer I, Dybvig K, Wróblewski H, Viari A, Rocha EPC, Blanchard A (2001). The complete genome sequence of the murine respiratory pathogen Mycoplasma pulmonis, Nucleic Acids Res. 29(10):2145-2153.

Chen CY, Wu KM, Chang YC, Chang CH, Tsai HC, Tsai-Liao L, Liu Y, Chen H, Shen AB, Li J, Su T, Shao C, Lee C, Hor L, Tsai S (2003). Comparative genome analysis of vibrio vulnificus, a marine Pathogen. Genome Res. 13: 2577-2587

Clancy S, Shaw K (2008). DNA deletion and duplication and the associated genetic disorders. Nat. Educ. 1(1):23.

Coenye T, Vandamme P (2005). Characterization of mononucleotide repeats in sequenced prokaryotic genomes. DNA Res. 12(4):221-33.
Cui Y, Li Y, Yan Y, Yang R (2008). Clustered regularly interspaced short palindromic repeats: structure, function and application-a review. Wei Sheng Wu Xue Bao 48(11):1549-1555.
Delihas N (2011). Impact of Small Repeat Sequences on Bacterial Genome Evolution. Genome Biol. Evol. 3:959-973.
Do HH, Kwok PC, Franco PP, Wing KS, Louxin Z (2008). Spectrum-Based De Novo Repeat Detection in Genomic Sequences. J. Comput. Biol. 15(5):469-488.
Edgar RC, Myers EW (2005). PILER: identification and classification of genomic repeats. Bioinformatics 21(1):i152-i158.
George RA, Heringa J (2000). The REPRO server: finding protein internal sequence repeats through the web.Trends Biochem. Sci. 25:515-517.
Godde JS, Bickerton A (2006). The repetitive DNA elements called CRISPRs and their associated genes: evidence of horizontal transfer among prokaryotes. J. Mol. Evol. 62(6):718-729.
Gordon RJ, Lowy FD (2008). Pathogenesis of Methicillin-Resistant Staphylococcus aureus Infection. Clin. Infect. Dis. 46(5):S350-S359.
Gur-Arie R, Cohen CJ, Eitan Y, Shelef L, Hallerman EM, Kashi Y (2000). Simple sequence repeats in Escherichia coli: abundance, distribution, composition, and polymorphism. Genome Res. 10(1):62-71.
Hancock JM, Simon M (2005). Simple sequence repeats in proteins and their significance for network evolution. Gene 345(1):113-118.
Heller R, Riegel P, Hansmann Y, Delacour G, Bermond D, Dehio C, Lamarque F, Monteil H, Chomel B, Piemont Y (1998). Bartonella tribocorum sp. nov., a new Bartonella species isolated from the blood of wild rats. Int. J. Syst. Evol. Microbiol. 48(4):1333-1339.
Hillyard DR, Redd MJ (2007). Identification of essential genes in bacteria. Methods Enzymol. 421:21-34.
Himmelreich R, Hilbert H, Plagens H, Pirkl E, Li B, Herrmann R (1996). Complete Sequence Analysis of the Genome of the Bacterium Mycoplasma Pneumoniae. Nucleic Acids Res. 24(22):4420-4449.
Himmelreich R, Plagens H, Hilbert H, Reiner B, Herrmann R (1997). Comparative analysis of the genomes of the bacteria Mycoplasma pneumoniae and Mycoplasma genitalium. Nucleic Acids Res. 25(4):701-712.
Hood DW, Deadman ME, Jennings MP, Bisercic M, Fleischmann RD, Venter JC, Moxon ER (1996). DNA repeats identify novel virulence genes in Haemophilus influenzae. Proc. Natl. Acad. Sci. USA. 93(20):11121-11125.
Huang C, Lin Y, Yang Y, Huang S, Chen C (1998). The telomeres of Streptomyces chromosomes contain conserved palindromic sequences with potential to form complex secondary structures. Mol. Microbiol. 28:905-916.
Jaffe JD, Stange-Thomann N, Smith C, DeCaprio D, Fisher S, Butler J, Calvo S, Elkins T, FitzGerald MG, Hafez N, Kodira CD, Major J, Wang S, Wilkinson J, Nicol R, Nusbaum C, Birren B, Berg HC, Church GM (2004). The Complete Genome and Proteome of Mycoplasma mobile. Genome Res. 14:1447-1461.
Jansen R, Embden JD, Gaastra W, Schouls LM (2002). Identification of genes that are associated with DNA repeats in prokaryotes. Mol.
Jin Q, Yuan Z, Xu J, Wang Y, Shen Y, Lu W, Wang J, Liu H, Yang J, Yang F, Zhang X, Zhang J, Yang G, Wu H, Qu D, Dong J, Sun L, Xue Y, Zhao A, Gao Y, Zhu J, Kan B, Ding K, Chen S, Cheng H, Yao Z, He B, Chen R, Ma D, Qiang B, Wen Y, Hou Y, Yu J (2002). Genome sequence of Shigella flexneri 2a: insights into pathogenicity through comparison with genomes of Escherichia coli K12 and O157. Nucleic Acids Res. 30(20):4432-4441.
Ju Z, Melissa C. Wells MC, Martinez A, Hazlewood L, Walter RB (2005). An in silico mining for simple sequence repeats from expressed sequence tags of zebrafish, medaka, Fundulus and Xiphophorus. In Silico Biol. 5: 0041.
Jurka J (1998). Repeats in genomic DNA: mining and meaning. Curr. Opin. Struct. Biol. 8(3):333-337.
Kashi Y, King D, Soller M (1997). Simple sequence repeats as a source of quantitative genetic variation. Trends Genet. 13(2):74-78.
Kashi Y, King DG (2006). Simple sequence repeats as advantageous mutators in evolution. Trends Genet. 22(5):253-9.
Kassai-Jáger E, Ortutay C, Tóth G, Vellai T, Gáspári Z (2008). Distribution and evolution of short tandem repeats in closely related bacterial genomes. Gene 410(1):18-25.
Kurtz S, Choudhuri JV, Ohlebusch E, Schleiermacher C, Stoye J, Giegerich R (2001). REPuter: the manifold applications of repeat analysis on a genomic scale. Nucleic Acids Res. 29(22):4633-4642.
Labbé J, Murat C, Morin E, Le Tacon F, Martin F (2011). Survey and analysis of simple sequence repeats in the Laccaria bicolor genome, with development of microsatellite markers. Curr. Genet. 57(2):75-88.
Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, Devon K, Dewar K, Doyle M, FitzHugh W, Funke R, et al. (2001), Initial Sequencing and analysis of the human genome. Nature 409:860-921.
Larsson P, Oyston PCF, Chain P, Chu MC, Duffield M, Fuxelius H, E Garcia, G Hälltorp, D Johansson, KE Isherwood, PD Karp, E Larsson, Y Liu, S Michell, J Prior, R Prior, S Malfatti, A Sjöstedt, K Svensson, N Thompson, L Vergez, JK Wagg, BW Wren, LE Lindler, SGE Andersson, M Forsman, Titball RW (2005). The complete genome sequence of Francisella tularensis, the causative agent of tularemia. Nat. Genet. 37:153-159.
Lathe W, Williams J, Mangan M, Karolchik D (2008). Genomic Data Resources: Challenges and Promises. Nat. Educ. 1:3.
Li R, Ye J, Li S, Wang J, Han Y, et al. (2005). ReAS: Recovery of Ancestral Sequences for Transposable Elements from the Unassembled Reads of a Whole Genome Shotgun. PLoS Comput. Biol. 1(4):e43.
Löffler B, Hussain M, Grundmeier M, Brück M, Holzinger D, et al., (2010). Staphylococcus aureus Panton-Valentine Leukocidin Is a Very Potent Cytotoxic Factor for Human Neutrophils. PLoS Pathog. 6(1):e1000715.
Ma L, Jensen JS, Mancuso M, Hamasuna R, Jia Q, McGowin CL, Martin DH (2012). Variability of trinucleotide tandem repeats in the MgPa operon and its repetitive chromosomal elements in Mycoplasma genitalium. J. Med. Microbiol. 61(2):191-197.
Ma L, Taylor S, Jensen JS, Myers L, Lillis R, Martin DH (2008). Short tandem repeat sequences in the Mycoplasma genitalium genome and their use in a multilocus genotyping system. BMC Microbiol. 8:130.
Meier SP, Entenza JM, Vaudaux P, Francioli P, Glauser MP, Moreillon P (2001). Study of Staphylococcus aureus Pathogenic Genes by Transfer and Expression in the Less Virulent Organism Streptococcus gordonii. Infect. Immun. 69(2):657-664.
Microbiol. 43(6):1565-75.
Minion FC, Lefkowitz EJ, Madsen ML, Cleary BJ, Swartzell SM, Mahairas GG (2004). The genome sequence of mycoplasma hyopneumoniae strain 232, the agent of swine mycoplasmosis. J. Bacteriol. 186(21):7123-7133.
Miroshnichenko ML, Rainey FA, Rhode M, Bonch-Osmolovskaya EA (1999). Hippea maritima gen. nov., sp. nov., a new genus of thermophilic, sulfur-reducing bacterium from submarine hot vents. Int. J. Syst. Evol. Microbiol. 49(3):1033-1038.
Morgulis A, Gertz EM, Schäffer AA, Agarwala R (2006). WindowMasker: window-based masker for sequenced genomes. Bioinformatics 22 (2):134-141.
Mrázek J (2006). Analysis of distribution indicates diverse functions of simple sequence repeats in Mycoplasma genomes. Mol. Biol. Evol. 23(7):1370-85.
Myers CL, Barrett DR, Hibbs MA, Huttenhower C, Troyanskaya OG (2006). Finding function: evaluation methods for functional genomic data. BMC Genomics 7:187.
Oluwagbemi OO, Omonhinmin CO (2008). Evaluating the Relationship Between Running Times and DNA Sequence Sizes using a Generic-Based Filtering Program. Pac. J. Sci. Technol. 9(2):656-666.
Oluwagbemi, OO (2012). Development of a prototype hybrid-grid-based computing framework for accessing bioinformatics databases and resources. Sci. Res. Essays 7(7):730-739.
Petes TD, Hill CW (1988). Recombination between repeated genes in microorganisms. Annu. Rev. Genet. 22:147-168.
Pevzner PA, Tang H, Tesler G (2004). De novo repeat classification and

fragment assembly. Genome Res. 14:1786-1796.
Phalipon A, Sansonetti PJ (2007). Shigella's ways of manipulating the host intestinal innate and adaptive immune system: a tool box for survival? Immunol. Cell Biol. 1-11.
Picot L, Abdelmoula SM, Merieau A, Leroux P, Cazin L, Orange N, Feuilloley MG (2001). Pseudomonas fluorescens as a potential pathogen: adherence to nerve cells. Microbes Infect. 3(12):985-995.
Powell W, Machray GC, Provan J (1996). Polymorphism revealed by simple sequence repeats. Trends Plants 1(7):215-222.
Power PM, Sweetman WA, Gallacher NJ, Woodhall MR, Kumar GA, Moxon ER, Hood DW (2009). Simple sequence repeats in Haemophilus influenza. Infect. Genet. Evol. 9(2):216-228.
Price AL, Jones NC, Pevzner PA (2005). De novo identification of repeat families in large genomes. Bioinformatics 21(Suppl. 1):351-358.
Rahim F(2008). *In silico* comparison of simple sequence repeats in high nucleotides-rich genomes of microorganism. Pak. J. Biol. Sci. 11(20):2372-2781.
Ruland K, Wenzel R, Herrmann R (1990). Analysis of three different repeated DNA elements present in the P1 operon of Mycoplasma pneumoniae: size, number and distribution on the genome. Nucleic Acids Res. 18(21):6311-6317.
Saker MM, Mohamed AA, Aly AA (2011). Comparative analysis of transformed potato microtubers and its non-transformed counterpart using some biochemical analysis along with inter simple sequence repeat (ISSR) marker. Afr. J. Biotechnol. 10(34):6401-6410.
Saurabh B, Sneha S, Suvidya R, Pramod K, Shailesh B (2011). Analysis of distribution and significance of simple sequence repeats in enteric bacteria Shigella dysenteriae SD197. Bioinformation 6(9):348-351.
Shikano T, Ramadevi J, Shimada Y, Merilä J (2010). Utility of sequenced genomes for microsatellite marker development in non-model organisms: a case study of functionally important genes in nine-spined sticklebacks (*Pungitius pungitius*). BMC Genomics 11: 334.
Singh A, Feschotte C, Stojanovic N (2007). Micro-repetitive structure of genomic sequences and the identification of ancient repeat elements, November 2-4, 2007 Proceedings (IEEE Int Conf Bioinformatics Biomed) pp.165-171.
Smit AFA, Green P (2002). RepeatMasker. unpublished. Website http://ftp.genome.washington.edu/RM/RepeatMasker.html.
Sun H, Mo Q, Lin J, Yang Z, Tu C, Gu D, Shi L, Lu W (2011). Rapid simultaneous screening of seven clinically important enteric pathogens using a magnetic bead based DNA microarray. World J. Microbiol. Biotechnol. 27(1):163-169.
Tarailo-Graovac M, Chen N (2009). Using RepeatMasker to Identify Repetitive Elements in Genomic Sequences. Curr. Protoc. Bioinformatics (Suppl.25).
Treangen TJ, Abraham A, Touchon M, Rocha EPC (2009). Genesis, effects and fates of repeats in prokaryoticgenomes. FEMS Microbiol. Rev. 1-33
Tyagi S, Sharma M, Das A (2011). Comparative genomic analysis of simple sequence repeats in three Plasmodium species. Parasitol. Res. 108(2):451-458.
van Belkum A (1999a). Short sequence repeats in microbial pathogenesis and evolution. Cell. Mol. Life Sci. 30:56(9-10):729-34.
van Belkum A (1999b). The role of short sequence repeats in epidemiologic typing. Curr. Opin. Microbiol. 2(3):306-311.
van Belkum A, Scherer S, van Alphen L, Verbrugh H (1998). Short-Sequence DNA Repeats in Prokaryotic Genomes. Microbiol. Mol. Biol. Rev. 62(2): 275-293.
van Belkum A, van Leeuwen W, Scherer S, Verbrugh H (1999). Occurrence and structure-function relationship of pentameric short sequence repeats in microbial genomes. Res. Microbiol. 150(9-10):617-626.
Volfovsky N, Haas BJ, Salzberg SL (2001). A clustering method for repeat analysis in DNA sequences. Genome Biol. 2(8):RESEARCH0027.
Von Graevenitz A, Weinstein J (1971). Pathogenic significance of Pseudomonas fluorescens and Pseudomonas putida. Yale J. Biol. Med. 44(3):265-273.
Wang L, Zhang A (2005). BioStar models of clinical and genomic data for biomedical data warehouse design. Int. J. Bioinformatics Res. Appl. 1(1):63-80.
Wei J, Goldberg MB, Burland V, Venkatesan MM, Deng W, Fournier G, Mayhew GF, G. Plunkett III, Rose DJ, Darling A, Mau B, Perna NT, Payne M, Runyen-Janecky LJ, Zhou S, Schwartz DC, Blattner FR (2003). Complete Genome Sequence and Comparative Genomics of Shigella flexneri Serotype 2a Strain 2457T. Infect. Immun. 71(5):2775-2786.
Whatmore AM (2009). Current understanding of the genetic diversity of Brucella, an expanding genus of zoonotic pathogens. Infect. Genet. Evol. 9(6):1168-1184.
Yang J, Wang J, Chen L, Yu J, Dong J, Yao ZJ, Shen Y, Jin Q, Chen R (2003). Identification and characterization of simple sequence repeats in the genomes of Shigella species. Gene 322: 85-92.
Zhang J (2003). Evolution by gene duplication: an update. Trends Ecol. Evol. 18(6):292-298.