IDENTIFICATION OF UNIQUE FAMILY-SPECIFIC MARKERS FOR MOLECULAR DETECTION OF ARBOVIRUSES: APPLICATION TO BUNYAVIRIDAE AND TOGAVIRIDAE FAMILIES

C. Muriuki, J. Kinyua and J. Ngaira
Department of Biochemistry, Jomo Kenyatta University of Agriculture and Technology, Nairobi, Kenya
E-mail: charmuriuki@yahoo.com

Abstract
Recent incidences of emerging arbovirus (arthropod-borne) disease outbreaks have led to heavy losses globally. Molecular virology offers a range of methods, able to accelerate and improve the diagnosis of arbovirus diseases but these rely heavily on the availability of robust PCR-based markers of the causative agents. The great genetic diversity exhibited by viral genomes makes it difficult to design specific signatures or markers since most methods start with a multiple sequence alignment (MSA). This study aimed at exploring the use of recently developed software (PriMux) prototyped on viral pathogens, that designs multiplex compatible degenerate primers without need for an initial MSA. It was used to design family-specific molecular signatures for arboviruses and members from the *bunyaviridae* and *togaviridae* virus families were used as models.

Keywords: Arboviruses, Bunyavirus

1.0 Introduction

The detection of viral species is of major importance in medical diagnostics and in the surveillance of viral disease outbreaks. Researchers employ various methods in the detection of viruses including: metagenomic sequencing, micro-arrays, mass-spectrometry, in-vitro culturing, or electron microscopy. Polymerase chain reaction (PCR) based techniques fill a niche in the field of viral detection as they are faster, much cheaper, readily accessible and easily transferable to field situations. They detect the genetic material of the pathogen thereby reducing contact with infectious material to a minimum.

Multiplex-PCR, a modification of the traditional PCR technique where more than one set of primer pair is used within a single reaction is useful for diagnostic purposes as it provides the ability to detect more than one infectious agent in a single assay. However, multiplex primer design for highly divergent targets like viruses is very challenging as no universally conserved primers may exist and finding sets of primers likely to function well in multiplex conditions adds to the complexity as conditions such as isothermal melting temperatures (Tm) and formation of primer dimers need to be considered.

Primer design software requiring a multiple sequence alignments (MSA) as an input can be problematic for diverse targets as MSA can be difficult to construct (Gardner *et al*., 2012). In addition, most software including: Primaclade, FastPCR, GeneUp (Pesole*et al*., 1998), UniQ software, Greene SCPrimer (Jabado*et al.,* 2006) and HYDEN (Linhart*et al.,* 2002) are inappropriate for large data sets and primers predicted are prone to dimerization.

This necessitated the need for the development of the Multiplex Primer Prediction (MPP) algorithm (Gardner *et al*., 2009) which avoids the need for the MSA and scales much better for large data sets of up to approximately 6000 sequences. The MPP algorithm builds a multiplex compatible set of primers capable of amplifying all target sequences and attempts to minimize the number of primer pairs in the set. However, MPP requires that the primers be exact matches i.e. it does not allow for degenerate bases.

The new PriMux software (Hysom*et al*., 2012) is an improvement of MPP algorithm as it enables design of degenerate primers and this ensures that primers of at least 18 bases can still be found conserved across multiple targets without the need to reduce the length of the primers. Maintaining primer length of 18 bases or longer, with higher Tm, results in better primer binding to target for improved sensitivity and specificity compared to that of shorter primers.

There has been a worldwide increase in the incidence of emerging diseases caused by arboviruses. The outbreaks are associated with great socio-economic impacts especially in East Africa where majority of cases have been

reported(SandorBelak 2007). These recent events illustrate the imperative to rapidly and accurately detect and identify pathogens during disease outbreaks. Fast, sensitive, and accurate pathogen identification facilitates an appropriate response to determine the mode of transmission, the scope of quarantine, and the method of treatment (Gardner*et al.,* 2003). PCR-based methods use signatures that require two things: first, they must be species-specific so as to avoid false-positive results and second, they should be species-wide (family-specific) capable of detecting all known strains of a given species and preventing false-negative results.

Early warning systems and the rapid and highly specific detection of viral agents are major tasks, considering that the timely recognition of such viral infections would prevent the spread of the diseases to large animal populations in huge geographic area. Advances in molecular virology have resulted in the development of new assays that can directly detect virus pathogens within several hours.Majority of the molecular approaches are designed to be highly specific for the detection of the selected target virus(es). To obtain a wide range of target amplification, primers need to be selected from well-conserved regions of the genome. Degenerate primers that are capable of detecting pathogens from an entire virus family or virus genus are practical when more members of the same family or genus cause the same or very similar clinical symptoms.

Multiplex-PCR is very useful in diagnosis as it offers the ability to detect more than one viral agent in a single assay. This is especially practical for viral pathogens which cause diseases with similar clinical symptoms like the viral hemorrhagic fevers (VHF) characterized by fevers and bleeding in humans. This syndrome is caused by RNA viruses belonging to the bunyaviridae, filoviridae, arenaviridae and flaviviridae families. The clinical symptoms of the early phases of a VHF are very similar irrespective of the causative agent and thus the need for efficient diagnosis. By performing Multiplex-PCR we seek to identify all possible pathogens that can be associated with this syndrome and other viral diseases.

2.0 Methods

All complete sequences of the reference genome of family members from the Bunyaviridae and Togaviridae virus families were retrieved from the NCBI, GenBank (www.ncbi.nlm.nih.gov/genbank), the National Institutes of Health (NIH) genetic sequence database that houses an annotated collection of all publicly available DNA sequences by a general search and opened in FASTA format.

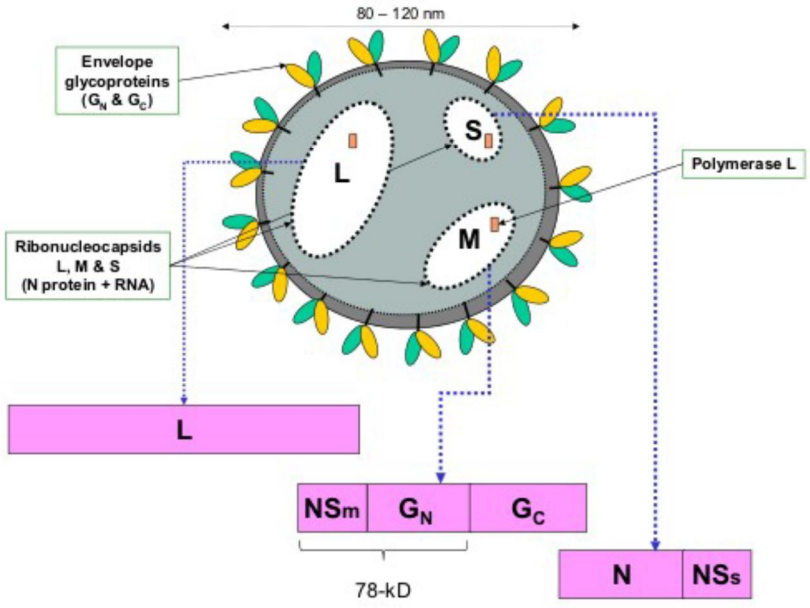

Figure 1: Schematic diagram of the general bunyavirus genome

2.1 TogaviridaeVirus Sequences

From the Togaviridae family, the following virus sequences were retrieved, all from the Alphavirus genus: Chikungunyavirus(accession number JQ067624.1), Ndumu virus(accession number HM147989.1), O'nyong-nyong virus(accesion number M20303.1), Semliki forest virus(accession number X04129.1), Sindbis virus(accesion number JQ771799.1), Western equine encephalomyelitis virus(accesion number GQ287647.1) and Ross River virus(accesion number GQ433360.1).

2.2 BunyaviridaeVirus Sequnces

From the Bunyaviridae family, the following virus sequences were retieved and primers designed against them. Figure 2 below illustrates a representation of the genomic RNA belonging to prototypic members of the five genera classified within the family Bunyaviridae. Crimean-Congo virus (accesion numbers AY389361.2, U39455.2, U88410.1 for L, M and S segment respectively) andDugbe virus(accession numbers U15018.1, M94133.1, AF434165.1 for L, M and S segment respectively) from the Nairovirus genus. Akabane virus(accession numbers AB190458.1, FJ498801.1, FJ498796.1 for L, M and S segment respectively),Bunyamwera virus(accession numbers X14383.1, M11852.1, D00353.1, for L, M and S segment respectively), La crosse virus(U12396.1, U70206.1, K00610.1 for L, M and S segments respectively) and Oropouche virus(accession numbers AF484424.1, AF441119.1, AY237111.1 for L, M and S segments respectively) from the Orthobunyavirus genus. Aguacate virus(accession numbers HM566138.1, HM566137.1, HM566139.1 for L, M and S segments respectively) Candiru virus(accession numbers HM119407.1, HM119408.1, HM119409.1 for L, M and S segmetns respectively), Rift valley fever virus(accession numbers HE687305.1, HE687306.1, HE687307.1 for L,M and S segment respectively) Sandflysicilian fever virus(accession numbers GQ847513.2, GQ847512.2, GQ847511.2 for L, M and S segments respectively) Uukuniemi virus(accession numbers D10759.1 , M17417.1, M33551.1 for L , M and S segments respectively) and Toscana virus(accession numbers X68414.1, X89628.1, X53794.1 for L, M and S segments respectively) from the Phlebovirus genus were retrieved in FASTA format. The FASTA sequences were then saved with both virus name and accession numbers in separate files. The Togaviruses were all saved in a single file while the Bunyaviruses were saved in three separate files representing the three genome segments, L, M and S.

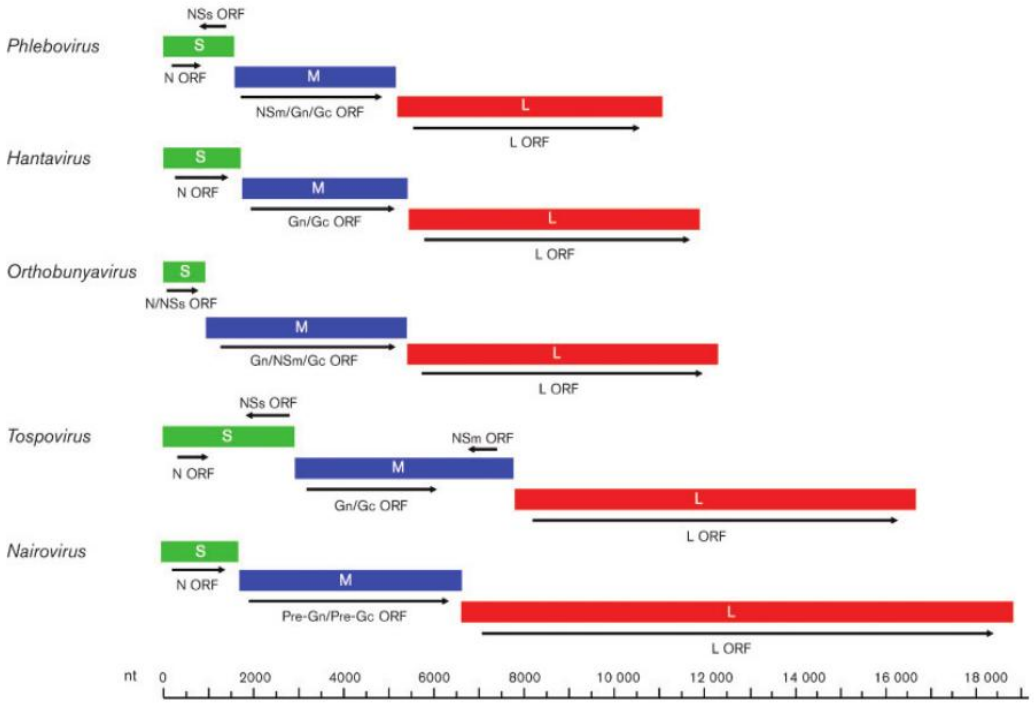

Figure 2:
Schematic representation of genomic RNAs belonging to prototypic members of the five genera classified within the family Bunyaviridae .PriMux Algorithm

2.3 PriMux Algorithm

PriMux is a new software package for selecting multiplex compatible, degenerate primers and probes to detect diverse targets. It differs from other available software like Primer3 (Rozen et al., 2000) as it requires no multiple sequence alignment (MSA) of the target sequences which is very complex to achieve for diverse targets such as viruses, it is scalable to large data sets and saves on time (Hysom et al., 2012). PriMux employs an algorithm that considers multiple sets of oligonucleotides of length k (k-mer) shared by multiple sequences and by allowing a small number of mismatches, by-passes the MSA and allows for degenerate primers. Figure 3 illustrates the typical approach of the PriMux algorithm. PriMux was used to design multiplex compatible primers for the Bunyarividae and Togaviridae family of viruses. The PriMux software is available as open source and was downloaded online from sourceforge, www.sourceforge.net/projects/PriMux. and installed on a high performance cluster (HPC) machine. Once the software was installed, it was tested using a python script test.py to ensure that it was executing commands as expected.

The Primux pipeline is implemented as a series of modules (i.e., scripts and executables) whose inputs and outputs are tied together via the file system. Compute intensive modules are coded in C ++ . Other modules (primarily those involved in parsing and invoking third-party executable) are coded in PythonTM or PerlTM. The Primux pipeline is a first step in computational pathogen signature development, it outputs sets of primers that may be used for PCR detection of unidentified DNA sequences. Primux is innovative in that the primer sets it computes are permitted to contain a specified number of degenerate bases. for computing forward/reverse primer pairs was run by invoking a python script, findPrimers.py, which takes a single command line parameter, the name of an options file. The options file specifies all settings relevant to an experiment. These include the name of the input fasta-formatted file, the number of permitted degenerate base pairs in the primers, directory in which to deposit the results.

2.4 Empirical Testing of Primers

Primers were purchased as lyophilized pellets from Inqaba Biotechnical Industries Ltd in South Africa. A synthesis report of all the primers designed and synthesized is available including quality control images of each primer. Upon receipt primers were reconstituted in sterile 1X Tris-EDTA (TE) Buffer (10mM Tris-Cl, 1mM EDTA, pH 8.0, Teknova, Hollister, CA) to a concentration of 100 mM. Working stocks were made by diluting primers and probes to a concentration of 10 mM with TE Buffer. The primer and probe working stocks were stored at 4°C. Unused 100mM primers and probes were stored at -80°C. The primers were then optimized for a real-time touchdown assay. All reactions were carried out on 96 well FAST PCR plates(Applied Biosystems) in a total volume of 25 ml (20 ml master mix plus 5 ml sample). The FastStart universal SYBR Green Master(ROX) master-mix was used. It is a ready to use 2x concentrated master-mix that contains all the ingredient except primers and template, needed for running real-time DNA detection assays in the SYBR green detection format. It contains a special ROX refrence dye making it suitable for all real-time PCR instruments on which ROX reference dye is needed for quantitative analysis. Reactions were carried out on ABI 7500 thermal cyclers (Applied Biosystems, Foster City, CA) under the following cycling conditions: 950C for 10 minutes for activation of FastStartTaq DNA polymerase, a touchdown step for 10 cycles of 950C for 30 seconds, 620C for 30 seconds and 720C for 30 seconds, amplification for 30 cycles of 950C for 30 seconds, 550C for 30 seconds and 720C for 30 seconds. A final extension step at 720C for 6 minutes and a default dissociation step of 950C for 15 seconds, 600C for 15 seconds and 950C for 15 seconds. The final hold was set at 100C. Data collection was done at the annealing step. All primers were designed such that the annealing temperature was between 600C and 650C.

3.0 Results

3.1 Primers

A list of all the primers that were designed and purchased, showing the orientation and sequences is provided below in Table 1. The primers were designed so as to ensure no cross-reactivity with any non-target. Sequences of the primers that were designed and purchased are listed in Table 1 below. The table records the name of the primer, it orientation, forward primer (FP) or reverse primer (RP) and the primer sequence.

Table 1: Summary of all primers designed showing orientation and sequence

PrimerName	Orientation	PrimerSequence
NDUV_2510_F	FP	AGTGCGGATTCTTCAACATGA
NDUV_3214_R	RP	ACGCCACTTCGGGTGAAT
CHIKV_10266_F	FP	GCCTACTGCTTCTGCGAC
CHIKV_11001_R	RP	CGGCGTTGGTCATCGAAT
SINV_826_F	FP	CAGTCGTACACTTGCCGC
SINV_1119_R	RP	GACAATTCGCTGGTTGAGCC
RRV_9969_F	FP	GGGTTCCCGTATAAGGCTC
RRV_10369_R	RP	TCGGTGGTCTGGTTGATG
ONNV_2534_F	FP	TGCGGATTCTTCAATATGATG
ONNV_2843_R	RP	TGGTCAGCCCCTGAGAAG
WEEV_1773_F	FP	CCCATTGGCGGAACAAGT
WEEV_2512_R	RP	TCATGTTAAAGAAGCCGCATT
SFV_356_F	FP	CCCGAAAGGCTCGATAGC
SFV_588_R	RP	ATCGCCTGATGGTACAGCG
BUNV_L_3404_F	FP	AAACTGGCTCCAGGGCAA
BUNV_L_3829_R	RP	AAGCTCTGTTGTGCTGCA

3.2 Empirical Testing of Primers

The designed primers were tested in the lab using a touchdown PCR assay as described in the methods section. All the primer pairs were tested in one run and the touchdown parameters were set so as to capture the annealing temperatures of all the primers. Positive control primers that had been found to have worked previously were also included in the same plate. The results obtained from the experimental testing of the primers in the lab showed that only one primer pair of the Chikungunya virus (CHIKV) amplified the expected target. The amplification plot of the result, showing this amplification is illustrated below in Figure 4. The negative results were attributed to the status of the DNA template as it had been in storage over a long period and may have degraded. This is especially so because even the positive control primers, that had been used in a previous study and found to have worked properly failed to amplify the intended target.

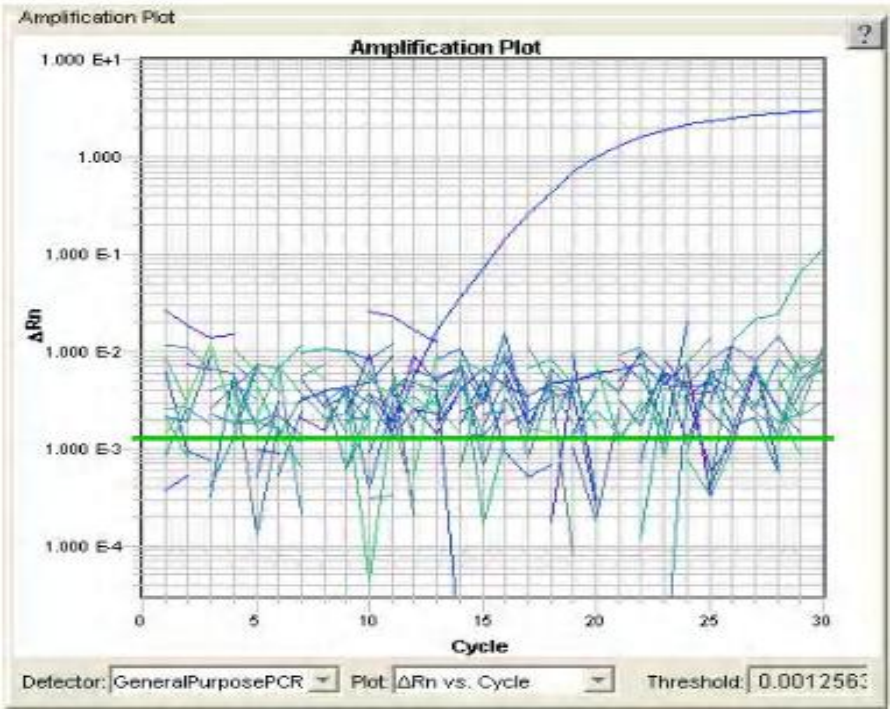

Figure 3: Amplification plot of real-time PCR result. This figure shows that only one primer pair of the Chikungunya virus worked

4.0 Discussion and Conclusion

Previous studies have shown that various real-time PCR assays provide powerful novel means for the very rapid detection and quantitation of targeted viral nucleic acids in clinical specimens. The real-time PCR assays have opened a new area of molecular diagnosis. However, although it has many advantages, a vulnerable side of the PCR-based diagnostic assays is that the detection efficiency is decreased by the high nucleotide sequence variability (mismatches) in the genomes of the various variants of the targeted viruses. The increasing numbers of mismatches between target and primer sequences result in decreased amplification or even in negative PCR results. Similarly, weaker or negative PCR results, may occur when trying to amplify the genomes of newly emerging genomic variants of a virus genus. The PCR analysis of emerging new viruses may yield negative results, due to the novel nucleotide sequences in these viral genomes. Considering these vulnerable sides of the PCR-based diagnostic assays, there is a high need to develop further approaches of molecular diagnosis. There is a strong tendency today to further increase the number of various methods, in order to strengthen the molecular diagnosis, to reduce the diagnosis time and to improve the diagnostic complexity. This study aimed to explore the use of PriMux, a k-mer based approach to designing primers and detecting diverse sets of target sequences without a need for multiple sequence alignment. Instead, a greedy algorithm based on k-mer analysis with suffix arrays identifies conserved, degenerate k-mers that meet primer specifications (Tm,length, GC%) and which can be combined in multiplex to amplify at least one fragment from each of the target sequences. It was intended that the panel of primers designed using this software would be capable of detecting all members of a large, diverse and potentially non-homologous and unalignable target set. However only one pair out of the thirteen primer pairs that were designed, was able to amplify its target sequence.

Acknowledgement

I particularly thank ILRI for awarding me a graduate fellowship, that offered me the opportunity to carry out my project work at the state of the art laboratories at ILRI, I am forever grateful.

References

Gardner, S., Amy, L., Peter, L., Christine, H.*et al*., (2009). Multiplex primer prediction software for divergent targets. *Nucleic Acids Res.,* **37**: pp 19.

Gardner, S., Kuczmarski, E. A. Vitalis, and Slezak, T. R. (2003). Limitations of TaqMan PCR for detecting divergent viral pathogens illustrated by hepatitis A, B, C, and E viruses and human immunodeficiency virus. *J. Clin. Microbiol.,* **41**: pp 2417–2427.

Hysom, D. A., Pejman, N., Peter, L. and Shea, N. (2012). Skip the Alignment: Degenerate, Multiplex Primer and Probe Design Using K-mer Matching Instead of Alignments. PloSONE.

Linhart,C. and Shamir,R. (2002). The degenerate primer design problem. *Bioinformatics*, **18**: pp 172–181.

Sandor, B. (2007). Molecular diagnosis of viral diseases, present trends and future aspects: A view from the OIE Collaborating Centre for the Application of Polymerase Chain Reaction Methods for Diagnosis of Viral Diseases in Veterinary Medicine. *Vaccine*, **30**: pp 5444-52.