Bioinformatics Advance Access published October 7, 2012

VIPR HMM: A hidden Markov model for detecting recombination with microbial detection microarrays

Adam F. Allred[1], Hilary Renshaw[1], Scott Weaver[2], Robert B. Tesh[2] and David Wang[1]*

[1]Departments of Molecular Microbiology and Pathology & Immunology, Washington University School of Medicine, St. Louis, Missouri USA

[2]Institute for Human Infections and Immunity, Center for Biodefense and Emerging Infectious Diseases and Department of Pathology, University of Texas Medical Branch, Galveston, Texas USA

Associate Editor: Dr. Trey Ideker

ABSTRACT

Motivation: Current methods in diagnostic microbiology typically focus on the detection of a single genomic locus or protein in a candidate agent. The presence of the entire microbe is then inferred from this isolated result. Problematically, the presence of recombination in microbial genomes would go undetected unless other genomic loci or protein components were specifically assayed. Microarrays lend themselves well to the detection of multiple loci from a given microbe; furthermore, the inherent nature of microarrays facilitates highly parallel interrogation of multiple microbes. However, none of the existing methods for analyzing diagnostic microarray data has the capacity to specifically identify recombinant microbes. In previous work, we developed a novel algorithm, VIPR, for analyzing diagnostic microarray data.

Results: We have expanded upon our previous implementation of VIPR by incorporating a hidden Markov model (HMM) to detect recombinant genomes. We trained our HMM on a set of nonrecombinant parental viruses and applied our method to 11 recombinant alphaviruses and 4 recombinant flaviviruses hybridized to a diagnostic microarray in order to evaluate performance of the HMM. VIPR HMM correctly identified 95% of the 62 inter-species recombination breakpoints in the validation set and only two false positive breakpoints were predicted. This study represents the first description and validation of an algorithm capable of detecting recombinant viruses based on diagnostic microarray hybridization patterns.

Availability: VIPR HMM is freely available for academic use and can be downloaded from http://ibridgenetwork.org/wustl/vipr

Contact: davewang@borcim.wustl.edu

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 INTRODUCTION

Recombination constitutes an important source of genetic variation among viruses. As an evolutionary mechanism, recombination leads to new viral genotypes with potentially novel biological properties and/or clinical manifestations. Vaccine-derived poliovirus is one example of a virus for which recombination may play an important role in the progression of disease. Recombination between vaccine-derived poliovirus and coxsackie virus has been shown to increase neurovirulence of recombinant progeny and may be responsible for the emergence of pathogenic vaccine-derived poliovirus (Jegouic, et al., 2009). In addition, H1N1 influenza and Ngari viruses provide examples in which novel genotypes consisting of genomic segments derived from multiple different parental viruses have led to disease outbreaks. H1N1, the influenza virus responsible for the 2009 outbreak of pandemic flu, is thought to have arisen from the successive reassortment of four different strains of influenza A (Neumann, et al., 2009). Ngari virus, a hemorrhagic fever-causing bunyavirus, is thought to have resulted from the natural reassortment of two viruses, Bunyamwera and Batai viruses, neither of which is known to cause hemorrhagic fever (Briese, et al., 2006; Gerrard, et al., 2004). Given that recombination and reassortment can play important roles in producing novel variations that are implicated in pathological outcomes, the ability for clinicians to identify novel recombinant and reassortant viruses in diagnostic laboratories is highly desirable.

In addition to occurring naturally through evolution, recombinant and reassortant viruses can also be deliberately created in the laboratory. In vitro recombination has proven to be a useful tool for engineering novel viruses with properties desirable for the development of vaccines (Atasheva, et al., 2009; Brandler, et al., 2005). However, this also means that recombination and reassortment have the potential to be used maliciously to develop novel agents of bioterrorism. Such agents could be engineered as highly pathogenic new viral genotypes consisting of the components of previously described viruses including non-pathogenic viruses. Anticipating the possible use of recombinant/reassortant-based bioweapons should guide our efforts in preparing to respond to such attacks. In such cases, the ability to detect novel agents quickly and accurately would be critical. Thus, it is imperative that any assay used to detect agents of bioterrorism include novel recombinants and reassortants as possible outcomes.

Microarrays are well suited to detecting recombination and reassortment and have an important advantage over traditional

*To whom correspondence should be addressed.

diagnostic methods because they allow for the interrogation of multiple loci from multiple viruses in parallel. Traditional methods for microbial detection, such as PCR and antibody based methods, are generally limited to detecting only one genome segment or one protein per assay. The inference is then made that the entire genome is present given that a small part of the genome (or proteome) was detected. Unless other loci are specifically assayed, this diagnostic paradigm does not account for the possibility that a recombinant or reassortant virus is present. There have been many reports of the efficacy of microarrays as a tool for viral diagnosis and discovery (McLoughlin, 2011). While many different probe design strategies and platforms have been proposed for diagnostic microarrays, all approaches require an objective method for interpreting the raw hybridization patterns.

The method must be able to make diagnostic calls in the presence of technical noise, biological noise (i.e. cross-hybridization to host) and probe saturation. Published examples of such methods with downloadable or web-accessible software include E-Predict (Urisman, et al., 2005), DetectiV (Watson, et al., 2007), PhyloDetect (Rehrauer, et al., 2008), CLiMax (Gardner, et al., 2010) and VIPR (Allred, et al., 2010). While these methods have been shown to perform with high accuracy, none of them was designed to be able to identify novel recombinant or reassortant viruses from a hybridization pattern.

One feature of VIPR, which stands for **V**iral **I**dentification with a **PR**obabilistic algorithm, is that it relies on an empirical training set of positive and negative control hybridizations to leverage diagnostic predictions. In this paper, we describe the expansion of VIPR to accommodate the possibility of recombination between candidate viruses included in the training set. We accomplished this by incorporating a hidden Markov model (HMM) into our method in order to define recombinant paths when calculating probabilities for candidate viruses (Figure 1). Figure 2 shows the details of constructing the HMM. The Viterbi algorithm was used to determine the optimal path from which recombination breakpoints could be inferred. As with VIPR, our HMM allows us to take advantage of training data consisting of hybridizations of known viruses to a microarray to make predictions for unknown infections. The incorporation of an HMM into VIPR now provides a probabilistic framework for assessing the presence of recombination between candidate parental viruses. To validate our approach, we applied our HMM to a set of 15 recombinant viruses consisting of members of the *Alphavirus* and *Flavivirus* genera, each of which was hybridized in duplicate to a custom microarray. A set of microarrays to which nonrecombinant alphaviruses and flaviviruses were hybridized constituted the training data for the HMM. While our test focused on the validation of a set of recombinant alphaviruses and flaviviruses, the strategy should be generalizable to detecting recombination among members of a given viral family.

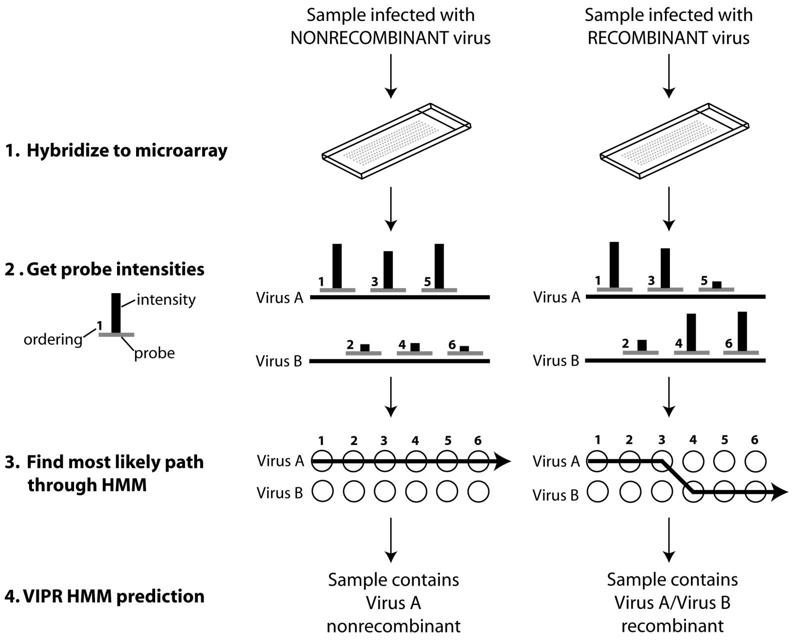
Sample infected with NONRECOMBINANT virus
Sample infected with RECOMBINANT virus
1. Hybridize to microarray
2 . Get probe intensities
intensity
ordering
1
probe
Virus A
Virus B
1
3
5
2
4
6
3. Find most likely path through HMM
1 2 3 4 5 6
Virus A
Virus B
4. VIPR HMM prediction
Sample contains Virus A nonrecombinant
Sample contains Virus A/Virus B recombinant

Fig. 1. Overall strategy for using an HMM to identify recombinant and nonrecombinant viruses hybridized to a microarray. Probe intensities indicative of binding can implicate the presence of a single virus (left) or the presence of different viruses for different loci (right). This pattern of intensities can be used to identify an optimal path through an HMM whose states represent binding or non-binding events between probes (columns) and virus genomes (rows). Nonrecombinant paths, such as the one on the left, involve transitions only between states in the same row, while paths that move from one row to another are indicative of recombination (as exemplified in the path on the right).

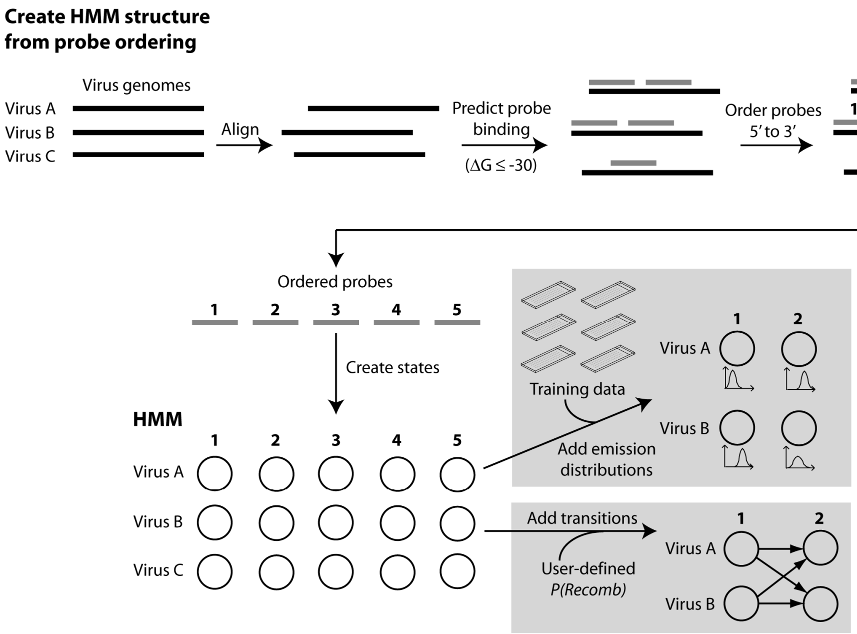

Fig. 2. Structure of the HMM used to detect recombinant and nonrecombinant viruses. First, candidate virus genomes are aligned. Probes are then mapped to their respective positions in the multiple alignment based on predicted free energy of binding in order to achieve a universal ordering of probes. A state is created for each probe:genome combination (representing either a predicted binding or non-binding event). The HMM is subsequently parameterized with emission distributions and transition probabilities based on probe intensity distributions from the training data and a user-defined probability of recombination parameter *P(Recomb)*, respectively.

2 RESULTS

RNA was purified from cell cultures that were infected with each of the viruses shown in Table 1 and Table 2. Purified RNA was subsequently randomly amplified and hybridized to a custom diagnostic microarray. 65 hybridizations (60 representing nonrecombinant alphavirus and flavivirus parental viruses + 5 representing uninfected Vero cells) were performed in order to obtain a training set for the HMM. For validation of our algorithm, 49 hybridizations (30 representing alphavirus and flavivirus recombinants + 15 representing alphavirus and flavivirus nonrecombinants + 4 representing uninfected Vero cells) were performed.

In order to build the HMM, we first needed to establish a framework to define possible recombinant and nonrecombinant paths based on positional information inherent to each probe. The microarray probes were ordered by their position from 5’ to 3’ in the global alignment of candidate virus genomes (Figure 2). This was accomplished by mapping the set of oligonucleotide probes via local alignment (megablast) to each candidate virus genome, identifying probes for which the theoretical free energy associated with its probe:genome local alignment was ≤ -30 kcal/mol (indicative of binding using previously explained criteria (Allred, et al., 2010)), and converting the midpoint of the probe:genome local alignment for each of those probes to its corresponding position in the global alignment (Edgar, 2004) of candidate virus genomes. Probes that mapped to multiple genomes at similar positions but were offset relative to each other by 30 nucleotides or fewer were consolidated to a single position in the global alignment. Probes were then sorted by their positions in the global alignment of candidate virus genomes.

Once the probes were ordered, they were assigned *On* and *Off* states for each genome. These assignments were based on the same

Table 1. Alphavirus and flavivirus parental viruses grown in culture and hybridized to the diagnostic microarray.

Genus	Species	Strain	Genbank	Strain #
Alphavirus	CHIKV	LR	116047549	
Alphavirus	EEEV	BeAr436087	119633049	1
Alphavirus	EEEV	FL93-939	119633046	2
Alphavirus	SINV	AR339	9790313	
Alphavirus	VEEV	68U201	1144527	1
Alphavirus	VEEV	TC-83	323714	2
Alphavirus	VEEV	TRD	323714	2[b]
Alphavirus	VEEV	ZPC738	4689187	3
Alphavirus	WEEV	CO92-1356	254595918[a]	
Alphavirus	WEEV	McMillan	254595918	
Flavivirus	DENV-4	1228	12659201[a]	
Flavivirus	JEV	SA14-14-2	12964700	
Flavivirus	SLEV	CorAn9124	344221822[a]	
Flavivirus	WNV	NY99	158516887	
Flavivirus	YFV	17D	9627244	

[a]Genbank ID represents a closely related strain since the sequence of the exact strain was not available

[b]Since VEEV TRD and VEEV TC-83 genomes differ by only 11 nucleotides, they were considered to be the same strain (VEEV strain 2)

theoretical free energy of binding calculated in the mapping step. *On* and *Off* states emit normalized and $\log_e$ transformed intensities according to normal distributions estimated from training data as previously described (Allred, et al., 2010). Thus, all emission probabilities *e(state, intensity)* were derived from distributions estimated in a manner identical to the estimation of probe-specific *On* and *Off* distributions in VIPR except in the case where there were fewer than 8 intensities available in the training set for a given probe. In that case, the mean of the distribution was calculated from the available intensities, but the standard deviation was derived from the average standard deviation over all probes with a similar *On* or *Off* prediction. In addition to the candidate virus genomes, a null genome was included which represented a none-of-the-above genome prediction and was assigned an *Off* state for each probe.

Table 2. Recombinant alphaviruses and flaviviruses hybridized to the diagnostic microarray for validation of the HMM.

Virus	Recomb type	Parents	Coordinates in parental genomes
R01	Double	EEEV BeAr436087	1-7499;11291-11638
		CHIKV LR	7504-11313
R02	Double	SINV AR339	1-7601;11394-11703
		VEEV TC-83	7536-11382
R03	Double	SINV AR339	1-7601;11383-11703
		CHIKV LR	7502-11313
R04	Double	SINV AR339	1-7602;11385-11703
		WEEV CO92-1356	7466-11210[b]
R05	Double	SINV AR339	1-7601;11312-11703
		EEEV BeAr436087	7498-11291
R06	Double	SINV AR339	1-7601;11394-11703
		VEEV TRD	7536-11382[b]
R07	Double	VEEV TC-83	1-7533;11328-11446
		CHIKV LR	7500-11313
R08	Double	YFV 17D	1-481;2453-10862
		DENV-4 1228	441-2423[b]
R09	Double	YFV 17D	1-481;2453-10862
		JEV SA14-14-2	477-2477
R10	Double	YFV 17D	1-481;2453-10862
		SLEV CorAn9124	456-2465[b]
R11	Double	YFV 17D	1-481;2453-10862
		WNV NY99	466-2469
R12	Double[a]	SINV AR339	1-7601;11394-11703
		VEEV TC-83	7536-8286
		VEEV 68U201	8298-11398
R13	Double[a]	SINV AR339	1-7601;11312-11703
		EEEV BeAr436087	7498-7640(7641-7675)[c]
		EEEV FL93-939	(7673-7707)7708-11323
R14	Double[a]	SINV AR339	1-7601;11394-11703
		VEEV TC-83	7536-8353(8354-8406)
		VEEV ZPC738	(8331-8383)8384-11359
R15	Triple[a]	SINV AR339	1-7601;11385-11703
		EEEV BeAr436087	7498-7640(7641-7675)
		EEEV FL93-939	(7673-7707)7708-7902
		WEEV McMillan	7802-11210

Coordinates corresponding to the parental genomes listed in Table 1 are given. For the recombinant alphaviruses, a short cloning sequence (between three and ten nucleotides) is present at the 3'-most recombination breakpoint.

[a] additional intra-species breakpoints present

[b] coordinates derived from closely related strain listed in Table 1

[c] parentheses represent regions of overlap between two parents sharing identical sequence

Finally, the states in the HMM were connected via transitions *t(state, state)* as depicted in Figure 2. As with HMMs that have been developed to detect recombination in sequence, probabilities representing recombination transitions could not be estimated directly from the training data as could the other HMM parameters (Schultz, et al., 2006). Thus, a user-specified probability of recombination parameter *P(Recomb)* was introduced to compute transition probabilities. Transitions connecting states within the same genome i.e. $t(state_{VirusA}, state_{VirusA})$ represented non-recombination events and had the associated probability *1-P(Recomb)*. Transitions between genomes i.e. $t(state_{VirusA}, state_{VirusB})$ represented recombination events and had the associated probability *P(Recomb)/(n-1)* where *n* is the number of candidate virus genomes (including the 'null' genome). In some cases, multiple probes mapped to the same position in the global alignment of candidate virus genomes. Transitions between states whose probes mapped to the same position were only allowed if those states corresponded to the same genome and were assigned a probability of 1.0, such that recombination events were not permitted between such states. Because the next state in the model is dependent only on the current state, and because states in the model emit from continuous intensity distributions, the model is a first-order continuous HMM.

Two models were built and were used to analyze the alphaviruses and the flaviviruses, respectively. In order to experimentally define a suitable *P(Recomb)* for computing transition probabilities, we evaluated the performance of VIPR HMM on a subset of parental viruses, varying *P(Recomb)* over a range of values. We selected the maximum value of *P(Recomb)* that resulted in zero false positive recombination breakpoints when Viterbi was applied to the parental alphaviruses. This value, $P(Recomb) = 10^{-25}$, was subsequently used when applying VIPR HMM to the parental flaviviruses as well as to the alphavirus and flavivirus recombinants. The Viterbi results were compared to expected results based on the known sequences of the recombinant constructs. When applied to the 5 flavivirus nonrecombinants, VIPR HMM classified each as the correct species. Additionally, the four uninfected Vero samples were accurately classified as null. VIPR HMM detected no recombination breakpoints for these samples except for one false positive breakpoint at the 3' end of the dengue virus 4 genome, which bypassed the final 254 nucleotides of the genome in favor of null states. VIPR HMM results for the nonrecombinant alphaviruses and flaviviruses are shown in Supplementary Figure S1.

A total of 30 hybridizations of recombinant viruses was analyzed by VIPR HMM. Supplementary Figure S2 shows results for all recombinant alphaviruses and flaviviruses analyzed by VIPR HMM. Of the 30 hybridizations, 28 represented double recombinants between two parent viruses of distinct species and two represented triple recombinants composed of three distinct parental species. Thus, the total number of expected inter-species recombination breakpoints was (28 x 2) + (2 x 3) = 62. VIPR HMM correctly identified breakpoints and the identity of the parental species for 59 of the 62 total breakpoints. In the remaining three instances, VIPR HMM yielded false negatives. Of all the recombinant and nonrecombinant samples analyzed by VIPR HMM, only two false positive breakpoints were predicted (one in a nonrecombinant virus and one in a double recombinant virus which was incorrectly identified as a triple recombinant by VIPR HMM). VIPR HMM results for a subset of the recombinant viruses that were identified unambiguously are shown in Figure 3A.

In some cases, the recombinant viruses we used included intra-species recombination breakpoints. Of the 8 intra-species breakpoints, 2 were identified by VIPR HMM. For those 2 breakpoints, the correct viruses 5’ and 3’ of the breakpoint were identified (both species and strain). VIPR HMM results for a subset of the recombinant viruses that gave unexpected results are shown in Figure 3B.

VIPR HMM was used to estimate the nucleotide positions of each breakpoint in each parental genome. The nucleotide positions associated with recombination breakpoints were estimated based on the position in the alignment of the probes associated with the recombinant transition in the Viterbi path. For each such probe, its position in the alignment was correlated with a position in the Viterbi-specified parental virus genome to estimate the nucleotide position of the recombination breakpoint in that genome (See spreadsheet in Supplementary data). The differences between the nucleotide positions estimated by VIPR and the actual sequence positions ranged from 0 to 90 nucleotides.

3 DISCUSSION

The ability of DNA microarrays to simultaneously assess the presence of multiple loci in microbial genomes is highly advantageous for detecting recombination between virus species in a diagnostic setting. Despite this, none of the existing methods for analyzing diagnostic microarrays is designed to accommodate the detection of recombinant viruses. In previous work, we developed VIPR, a method for objectively interpreting diagnostic microarrays. One of the advantages of VIPR relative to other methods is that it relies on a training set of empirical hybridizations of virally infected and uninfected samples to leverage diagnostic predictions. We anticipated that relying on a training set of hybridizations from known viral infections would also help us predict recombination between virus species. In this study, we developed a hidden Markov Model (HMM) parameterized with VIPR probability distributions to detect recombination in unknown infections.

VIPR HMM performed with high accuracy when identifying recombination breakpoints between viral species (59/62 such breakpoints were identified and the correct virus species 5’ and 3’ to the breakpoint were identified in each case). Of the 8 intra-species breakpoints in our data set, two were identified by VIPR HMM. Given that a much higher percentage of inter-species breakpoints were detected than were intra-species breakpoints (95% versus 25%), these results demonstrate that VIPR HMM is

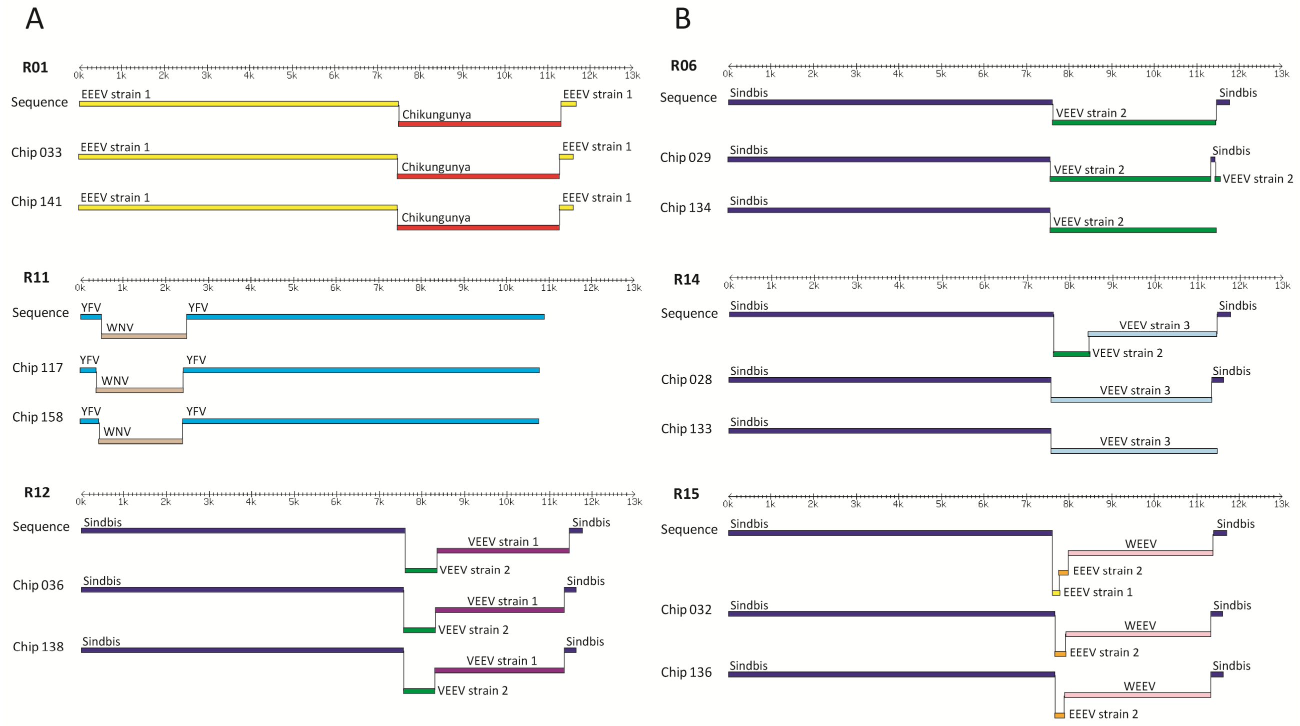

Fig. 3. VIPR HMM results for a subset of recombinants tested. **A.** VIPR HMM results for three recombinants that gave expected results. For each recombinant, the expected output based on sequence is shown, followed by the VIPR HMM output for the two hybridizations performed. **B.** VIPR HMM results for three recombinants that gave unexpected results. R06 is a double recombinant for which an additional false positive recombination breakpoint was identified at the 3’ end in one hybridization, and for which a 3’ inter-species recombination breakpoint was not identified in the other hybridization. R14 is a double recombinant for which a 3’ inter-species recombination breakpoint was identified in one of the hybridizations, but not the other. Additionally, an intra-species recombination breakpoint was not identified in either hybridization. R15 is a triple recombinant for which all three inter-species recombination breakpoints were identified in both hybridizations, but for which an intra-species recombination breakpoint was not identified in either.

more effective at detecting recombination between species than between strains belonging to the same species. The ability of VIPR HMM to distinguish between strains of the same species involved in recombination is likely influenced by the degree of sequence divergence between the two strains. VIPR HMM correctly identified the intra-species breakpoint in both hybridizations of R12 (Figure 3). The two strains comprising the intra-species breakpoint for R12 are 23% divergent on the nucleotide level. However, VIPR HMM was not able to identify the intra-species breakpoint in either hybridization of R14 (Figure 3), whose recombinant regions 5' and 3' to the intra-species breakpoint were similar in size to those of R12, but whose strains comprising the intra-species breakpoint are only 4% divergent on the nucleotide level. The ability of VIPR HMM to distinguish between strains of the same species may also be influenced by the size of the recombinant segment. The four other intra-species breakpoints that VIPR HMM failed to detect had greater dissimilarity between flanking strains (25%) but were proximal to other breakpoints (within 200 nt). Given the cost in probability associated with following a recombinant transition in the HMM, our results suggest that Viterbi may opt to bypass small recombinant regions.

Although we have not specifically tested VIPR HMM for higher-order recombinants (>3 breakpoints), we anticipate that the identification of higher-order recombinants will only be hindered inasmuch as additional recombination events within a fixed-size genome yield smaller recombinant segments to be identified. Given that VIPR HMM was shown to identify recombinant regions as small as ~300 nucleotides where there was sufficient divergence between recombining species, we expect VIPR HMM would be able to detect similar regions in higher-order recombinants.

Since microarray probes are mapped to their position in an alignment of candidate genomes, VIPR HMM can use the probes located at the boundary of a predicted recombination event to estimate nucleotide positions of recombination breakpoints. Although it was not expected that using a microarray tiling scheme wherein probes were non-overlapping and spaced 63 nucleotides apart would give the precise nucleotide positions of recombination breakpoints, we compared the estimates given by VIPR HMM to the nucleotide positions known from sequence. For the 61 correctly identified breakpoints (59 inter-species, 2 intra-species), the differences between microarray estimates and actual positions ranged from 0 to 90 nucleotides. Therefore, the maximum distance observed falls within the span of about a two probe tiling (i.e. 90 < 60mer + 3 nt spacing + 60mer). We expect that using higher density tiling strategies would result in higher resolution mapping of the breakpoints.

Only two false positive recombination breakpoints were predicted by VIPR HMM, both near the 3' ends of their respective genomes. One bypassed the final 254 nucleotides of dengue virus 4 in favor of null states. The other bypassed the final 191 nucleotides of Sindbis virus in favor of VEEV states. From analysis of the training data, it was observed that the mean of the *On* distributions approach the mean of the *Off* distributions for probes near the 3' end of each genome, due to lower intensities for *On* probes in the training set for that region. This trend was observed in the training data universally for all genomes. The tendency for *On* probes to give lower intensities when approaching the 3' end may be attributable to the fact that random PCR amplification, which was used in the preparation of each sample for hybridization, is less efficient at the ends of a linear genome. This could also explain why VIPR failed to detect three inter-species recombination breakpoints, all of which are localized near the 3' end of a genome. A similar pattern of lower intensities was also observed for *On* probes approaching the 5' end, although there appeared to be more probes in those regions that behaved as expected based on ΔG compared to the 3' end. Despite the observed decrease in hybridization intensity proximal to the 3' and 5' termini, VIPR HMM was still able to make accurate predictions in those regions in most cases.

Although we did not specifically validate VIPR HMM for reassortant viruses, we anticipate that viral reassortants would be readily detected. Reassortment can occur during co-infection when virus progeny inherit genome segments from two or more parental viruses with multi-segmented genomes. The resulting chimeric genotypes associated with reassortment are similar to those generated though recombination except that the exchange of genetic material occurs at discrete, predictable points in the genome i.e. at the boundary between genome segments. VIPR HMM could also be applied to the analysis of bacterial genomes and for the detection of the loss of transposons, integrons and plasmids, assuming those elements are represented in members of the training set.

One challenge in building an HMM for detecting recombination is finding an appropriate value for *P(Recomb)*, a user-inputted probability of recombination parameter used to calculate different transition probabilities in the model. Our choice of *P(Recomb)* was based on minimizing false positive recombinations in nonrecombinant samples. If *P(Recomb)* is increased by five logs i.e. $P(Recomb) = 10^{-20}$, the number of correctly called breakpoints is increased to from 59/62 to 60/62. However, there are an additional 3 false positive breakpoints called near genome ends. Nonetheless, in some cases, it may be advantageous to increase *P(Recomb)* in order to increase detection sensitivity. If *P(Recomb)* is decreased by five logs i.e. $P(Recomb) = 10^{-30}$, there is no change in the results.

VIPR HMM relies on a multiple alignment of candidate genomes to order microarray probes. One limitation of this approach is that only recombination between members of the same family will be considered as candidates since it is not generally feasible to globally align members of different families. In addition, because paths through the HMM follow a specific 5' to 3' ordering, only recombination at homologous sites is detectible by VIPR HMM as currently implemented. In future versions of VIPR, recombinants composed of viruses from different families could be detected by running multiple iterations of Viterbi, one for each HMM representing a different virus family. For a hypothetical recombinant between members of two different virus families, we anticipate that the HMM for each family would predict the presence of only a portion of the viral genome from its family (with the rest of the prediction being the null genome). With respect to selecting an appropriate *P(Recomb)* for detecting interfamily recombination, there are several different approaches that could be taken. One would be to use the same *P(Recomb)* for inter-family events as is used for intra-family events. Another would be to lower *P(Recomb)* for inter-family events. The choice of strategy would likely depend on what was known a priori about the possibility of certain families recombining.

The time complexity for the VIPR HMM algorithm is n x m where n is the number of probes and m is the number of genomes. The 22 alphavirus recombinant samples were analyzed in under 30 seconds on a single-processor PC with 4 GB RAM. Thus, if the number of candidate genomes were increased to one thousand while the number of probes remained the same, it would take approximately 150 seconds or 2.5 min to analyze one sample on a similar machine.

4 CONCLUSIONS

We developed a hidden Markov model (HMM) to identify recombination in viruses that have been hybridized to a microbial detection microarray. This model builds on previous work in which empirical hybridizations of cultured viruses were used as training to classify unknown infections (VIPR). Applying the HMM in conjunction with VIPR enabled the detection of inter-species recombination breakpoints with high accuracy in two different families of viruses. This is the first report of a method for analyzing diagnostic microarrays that includes recombination as a possible diagnostic outcome. Our method is theoretically applicable to detecting homologous recombination or reassortment between members of any family of viruses for which a set of nonrecombinant parental viruses is available for training and for which genome sequences are available. The inherently parallel nature of diagnostic microarrays coupled with powerful methods for analysis enhance our ability to rapidly and accurately identify novel recombinant viruses responsible for disease outbreaks, either due to emergence by natural means or by engineered recombinant viruses.

5 MATERIALS AND METHODS

5.1 Design of the diagnostic microarray

60mer oligonucleotide probes were designed from sequences representing three virus families (*Bunyaviridae, Flaviridae and Togaviridae*) using a tiling strategy. 145 RefSeq genomes and genome segments from the aforementioned virus families were obtained from Genbank. To the RefSeq set we added from Genbank as many complete genome sequences as were available of the parental viruses of the 11 recombinant alphaviruses. Partial genome sequences for the parental alphaviruses were added if complete genomes were not available. Additionally, complete genome sequences of alphaviruses that did not represent parents of the recombinant viruses were added until there were in the set at least three complete genomes of each of EEEV, VEEV, WEEV, Chikungunya and Sindbis viruses. The final set of Genbank records totaled 193, of which 175 were complete or nearly complete genomes or genome segments. Probes were selected as 60 nucleotide windows tiled over all 193 sequences with a spacing of three nucleotides between the 3' end of one probe and the 5' end of the following probe. The reverse complement of each 60mer was also included in the microarray. The resulting set of probes including reverse complements totaled 43414 and the Agilent® 4 x 44 K platform was used (GEO accession GSE34490).

5.2 Hybridization of alphavirus and flavivirus parental and recombinant viruses to the diagnostic microarray

21 alphaviruses (11 recombinants + 10 parental viruses) and 9 flaviviruses (4 recombinants + 5 parental viruses) which have been previously described (Atasheva, et al., 2009; Ni, et al., 2007; Paessler, et al., 2003; Paessler, et al., 2006; Wang, et al., 2007; Wang, et al., 2008) were obtained from the World Reference Center for Emerging Viruses and Arboviruses and were grown in Vero cells. RNA was extracted using standard Trizol® protocols and was reverse transcribed and randomly amplified as previously described (Wang, et al., 2003). For each recombinant, two independent amplifications were performed, while five independent amplifications were performed for each parental virus. The resulting amplified material was then coupled to a fluorescent dye and hybridized to the tiling microarray. Raw data measurements were collected using GenePix Pro® software. In total, 114 hybridizations were performed (30 recombinant + 75 parental + 9 uninfected Vero cells). All raw microarray data are available in NCBI GEO (accession GSE34490). The training set for our HMM consisted of 60 parental hybridizations + 5 Vero negative control hybridizations, while the test set for validating the algorithm consisted of the 30 recombinant hybridizations + 15 parental hybridizations + 4 Vero negative control hybridizations.

5.3 Viterbi algorithm for finding the optimal path

By multiplying emission probabilities *e(state, intensity)* and transition probabilities *t(state, state)* across a series of states, it is possible to obtain a probability for an entire path through an HMM. For our HMM, the set of emission and transition parameters is abbreviated as θ. The probability of a particular path (π) and a given hybridization ($\boldsymbol{x}$) of length L can be expressed as a joint probability:

$$P(\boldsymbol{x},\pi \mid HMM,\theta) = t(0,\pi_1)\prod_{i=1}^{L} t(\pi_i,\pi_{i+1})e(\pi_i,x_i)$$

The Viterbi algorithm falls into a class of algorithms called dynamic programming algorithms that are commonly used in conjunction with HMMs. Using the Viterbi algorithm allows us to identify the most probable series of states (π') through our HMM where

$$\pi' = \arg\max_{\pi} P(\boldsymbol{x},\pi \mid HMM,\theta)$$

Points of recombination can be inferred from places in the path where a transition between states of different genomes has occurred. As with other dynamic programming algorithms, the Viterbi algorithm consists of an initialization step, an iteration step and a termination step. Once the dynamic programming matrix (V) is populated, the optimal path is traced back through a shadow matrix (τ) of stored pointers. Except for the begin and end states s_{begin} and s_{end} and states in a given path (π_i), all other states ($s_{g,i}$) are indexed by genome (g) and probe (i). The V matrix and τ matrix are similarly indexed. Calculations are performed in log space although they are shown here in probability space. The Viterbi algorithm adapted from (Durbin, 1998) is as follows:

(1) Initialization (*g = 1 to n*)

$$V_{g,1} = t(s_{begin},s_{g,1})e(s_{g,1},x_1)$$

(2) Iteration (*i = 2 to L; g = 1 to n*)

$$V_{g,i} = e(s_{g,i},x_i)\max{}_{j=1}^{n}\left[V_{j,i-1}t(s_{j,i-1},s_{g,i})\right]$$

$$\tau_{g,i} = \arg\max_{j=1}^{n}\left[V_{j,i-1}t(s_{j,i-1}, s_{g,i})\right]$$

(3) Termination

$$P(\boldsymbol{x},\pi' \mid HMM,\theta) = \max_{j=1}^{n}\left[V_{j,L}t(s_{j,L}, s_{end})\right]$$

$$\pi'_L = \arg\max_{j=1}^{n}\left[V_{j,L}t(s_{j,L}, s_{end})\right]$$

(4) Traceback (*i = L to 2*)

$$\pi'_{i-1} = \tau(\pi'_i, i)$$

Traceback reveals the optimal path through the HMM. If the path includes states representing only one genome, the optimal path is a nonrecombinant path. If the optimal path includes transitions between states of different genomes, the path is recombinant, and the global alignment positions corresponding to the probes associated with the states involved in each transition are referenced. These global alignment positions are then back-converted to genomic positions in the predicted virus parents in order to define the recombinant breakpoints between virus genomes on the nucleotide level.

ACKNOWLEDGEMENTS

We thank Michael Brent for useful discussions.

Funding: This work was supported by NIH grant U01 AI070374 and by a grant from the National Institute of Allergy and Infectious Disease (NIAID) through the Western Regional Center of Excellence for Biodefense and Emerging Infectious Disease Research, National Institutes of Health (NIH) grant U54 AIO57156.

Conflict of interest: none declared.

REFERENCES

Allred, A.F., et al. (2010) VIPR: A probabilistic algorithm for analysis of microbial detection microarrays, BMC Bioinformatics, 11, 384.

Atasheva, S., et al. (2009) Chimeric alphavirus vaccine candidates protect mice from intranasal challenge with western equine encephalitis virus, Vaccine, 27, 4309-4319.

Brandler, S., et al. (2005) Replication of chimeric yellow fever virus-dengue serotype 1-4 virus vaccine strains in dendritic and hepatic cells, Am J Trop Med Hyg, 72, 74-81.

Briese, T., et al. (2006) Batai and Ngari viruses: M segment reassortment and association with severe febrile disease outbreaks in East Africa, J Virol, 80, 5627-5630.

Durbin, R. (1998) Biological sequence analysis: probalistic models of proteins and nucleic acids. Cambridge University Press, Cambridge, UK New York.

Edgar, R.C. (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput, Nucleic Acids Res, 32, 1792-1797.

Gardner, S.N., et al. (2010) A microbial detection array (MDA) for viral and bacterial detection, BMC Genomics, 11, 668.

Gerrard, S.R., et al. (2004) Ngari virus is a Bunyamwera virus reassortant that can be associated with large outbreaks of hemorrhagic fever in Africa, J Virol, 78, 8922-8926.

Jegouic, S., et al. (2009) Recombination between polioviruses and co-circulating Coxsackie A viruses: role in the emergence of pathogenic vaccine-derived polioviruses, PLoS Pathog, 5, e1000412.

McLoughlin, K.S. (2011) Microarrays for pathogen detection and analysis, Brief Funct Genomics, 10, 342-353.

Neumann, G., Noda, T. and Kawaoka, Y. (2009) Emergence and pandemic potential of swine-origin H1N1 influenza virus, Nature, 459, 931-939.

Ni, H., et al. (2007) Recombinant alphaviruses are safe and useful serological diagnostic tools, Am J Trop Med Hyg, 76, 774-781.

Paessler, S., et al. (2003) Recombinant sindbis/Venezuelan equine encephalitis virus is highly attenuated and immunogenic, J Virol, 77, 9278-9286.

Paessler, S., et al. (2006) Replication and clearance of Venezuelan equine encephalitis virus from the brains of animals vaccinated with chimeric SIN/VEE viruses, J Virol, 80, 2784-2796.

Rehrauer, H., et al. (2008) PhyloDetect: a likelihood-based strategy for detecting microorganisms with diagnostic microarrays, Bioinformatics, 24, i83-89.

Schultz, A.K., et al. (2006) A jumping profile Hidden Markov Model and applications to recombination sites in HIV and HCV genomes, BMC Bioinformatics, 7, 265.

Urisman, A., et al. (2005) E-Predict: a computational strategy for species identification based on observed DNA microarray hybridization patterns, Genome Biol, 6, R78.

Wang, D., et al. (2003) Viral discovery and sequence recovery using DNA microarrays, PLoS Biol, 1, E2.

Wang, E., et al. (2007) Chimeric Sindbis/eastern equine encephalitis vaccine candidates are highly attenuated and immunogenic in mice, Vaccine, 25, 7573-7581.

Wang, E., et al. (2008) Chimeric alphavirus vaccine candidates for chikungunya, Vaccine, 26, 5030-5039.

Watson, M., et al. (2007) DetectiV: visualization, normalization and significance testing for pathogen-detection microarray data, Genome Biol, 8, R190.