Frequency Analysis Techniques for Identification of Viral Genetic Data

Vladimir Trifonov and Raul Rabadan
**2010. Frequency Analysis Techniques for Identification of Viral Genetic Data . mBio 1(3): .
doi:10.1128/mBio.00156-10.**

RESEARCH ARTICLE

Frequency Analysis Techniques for Identification of Viral Genetic Data

Vladimir Trifonov and Raul Rabadan

Department of Biomedical Informatics, Center for Computational Biology and Bioinformatics, College of Physicians and Surgeons, Columbia University, New York, New York, USA

ABSTRACT **Environmental metagenomic samples and samples obtained as an attempt to identify a pathogen associated with the emergence of a novel infectious disease are important sources of novel microorganisms. The low costs and high throughput of sequencing technologies are expected to allow for the genetic material in those samples to be sequenced and the genomes of the novel microorganisms to be identified by alignment to those in a database of known genomes. Yet, for various biological and technical reasons, such alignment might not always be possible. We investigate a frequency analysis technique which on one hand allows for the identification of genetic material without relying on alignment and on the other hand makes possible the discovery of nonoverlapping contigs from the same organism. The technique is based on obtaining signatures of the genetic data and defining a distance/similarity measure between signatures. More precisely, the signatures of the genetic data are the frequencies of *k*-mers occurring in them, with *k* being a natural number. We considered an entropy-based distance between signatures, similar to the Kullback-Leibler distance in information theory, and investigated its ability to categorize negative-sense single-stranded RNA (ssRNA) viral genetic data. Our conclusion is that in this viral context, the technique provides a viable way of discovering genetic relationships without relying on alignment. We envision that our approach will be applicable to other microbial genetic contexts, e.g., other types of viruses, and will be an important tool in the discovery of novel microorganisms.**

IMPORTANCE **Multiple factors contribute to the emergence of novel infectious diseases. Implementation of effective measures against such diseases relies on the rapid identification of novel pathogens. Another important source of novel microorganisms is environmental metagenomic samples. The low costs and high throughput of sequencing technologies provide a method for the identification of novel microorganisms by sequence alignment. There are several obstacles to this method, as follows: our knowledge of biology is biased by an anthropomorphic view, microbial genomic material could be a minuscule fraction of the sample, the sequencing and enrichment technologies can be a source of errors and biases, and finally, microbes have high diversity and high evolutionary rates. As a result, novel microorganisms could have very low genetic similarity to already known genomes, and the identification by alignment could be computationally prohibitive. We investigate a frequency analysis technique which allows for the identification of novel genetic material without relying on alignment.**

Received 2 June 2010 **Accepted** 20 July 2010 **Published** 24 August 2010

Citation Trifonov, V., and R. Rabadan. 2010. Frequency analysis techniques for identification of viral genetic data. mBio 1(3):e00156-10. doi:10.1128/mBio.00156-10.

Editor Ian Lipkin, Columbia University

Address correspondence to Vladimir Trifonov, vladot@c2b2.columbia.edu.

Next-generation high-throughput sequencing technologies have significantly increased our ability to generate genetic data at low costs. The expectations are that costs will decrease while the throughput and quality levels increase. The availability of such data has allowed us to make significant progress in understanding the genetic bases of biological processes and systems (1).

Metagenomic studies, such as those resulting from the use of environmental samples (2) or the human biome (3), are one particular direction of research that was enabled by the new sequencing technologies. An important biochemical and bioinformatic challenge presented by such studies is categorization of the genetic material present in the sample according to its species of origin. In the context of identifying novel pathogens (4, 5), the species of origin can be a distant evolutionary relative to a known pathogen, and identification of this relative is an important step towards understanding the pathogen present in the sample. A first step towards such categorization is using alignment tools like BLAST (Basic Local Alignment Search Tool from NCBI) or SHRiMP (SHort Read Mapping Package from Computational Biology Lab, University of Toronto) to compare the genetic material with that in an existing reference database. The success of this approach depends on the existence of a database containing a reference with a sufficiently high level of similarity. For example, if the sample is obtained from a host species, this approach allows for identification of the host genetic material, if a reference for the host is available.

This paper focuses on techniques towards categorization of genetic material in the absence of a high-homology reference. In this context, genetic material of viral origin presents a particular challenge. At present, the NCBI database contains hundreds of thousands of viral genomes. Despite the amount of effort put into building such genomic databases, it is unlikely that they are even close to being comprehensive. In fact, further extension of these databases is one of the goals of metagenomic studies.

The situation is complicated further from three directions. On

one hand, the reference database is biased towards known human pathogens. One can argue that this bias is inherent to the way the samples are obtained—often as a result of an effort to fight the disease associated with the pathogen. Although metagenomic samples are often obtained to study pathogen activity, the challenging cases occur when the pathogen cannot be identified and isolated by traditional biochemical methods, which means that it is unlikely that it will be included in the reference database. Furthermore, in some metagenomic studies, samples without any noticeable pathogen activity are obtained. In such cases, the chance of matching novel viral species present in the sample to known references is diminished further. One can hope that as the cost of obtaining sequenced metagenomic data decreases, the influence of such sampling biases will decrease as well.

The second obstacle goes deeper because it stands in the way of even extracting the viral genetic material from a sample. The issue is that the total amount of viral genetic material present in a sample is dwarfed by other genetic material, host or otherwise, also present. The genetic material of an RNA virus consists of about 10^4 to 10^5 nucleotides, which is 3 to 4 orders of magnitude less than a typical prokaryotic genome and 4 to 5 orders less than a typical eukaryotic genome. Current high-throughput sequencing technologies proceed roughly by first fracturing the genetic material present in a sample, amplifying the fractured sample by PCR, and sequencing the PCR product. If the microbe is not grown in culture and only clinical samples are available, the genetic material from the microbe is usually at very low concentrations, and the available sequences do not cover its whole genome. Biases and errors introduced by the sequencing and enrichment processes are other problems which contribute to an incomplete picture of the viral genetic material present in a sample. Part of such biases is the fact that the high throughput of the new sequencing technologies comes at the price of producing reads of very short lengths (6). This, on one hand, limits the genome reconstructing ability of alignment techniques and, on the other, restricts the power of techniques, such as ours, based on the statistical properties of the reads.

While the previous obstacles can be surmounted purely by technological advancement, the third difficulty in the way of categorizing viral genetic data is far more fundamental because it is inherent to viruses themselves. Viruses depend on their hosts for reproduction, and often, the survival of a viral pathogen is at the expense of its host. The strongest weapon in the viral arsenal against the highly developed and sophisticated defense mechanism of the host is an extremely high mutation rate. It is estimated that the mutation rate of a typical RNA virus is 10^{-4} nucleotides per replication. If a virus takes a few hours to reproduce and it makes thousands of offspring, in a year the descendant viruses can differ substantially from their ancestors. One can gain perspective on the amount of genetic variability existing in the viral world by considering the evolution of the H1N1 influenza A virus from 1918 to that of the recent 2009 H1N1 pandemic. The variability in this case is comparable to the genetic distance between the human and mouse genomes (~15%), a result of 100 million years of evolution. The implication is that even if a related reference is present in a viral database, the genetic distance to this reference can be prohibitively large for capturing the similarity by alignment.

To summarize, in order to categorize viral genetic material present in sequenced metagenomic data, we have to identify the species of origin for sequences with low coverage and for which we have references with very low homology due to biased sampling and high mutation rates at our disposal. One cannot help but compare this task to deciphering ancient scripts, such as the Mayan, Ugaritic, and Sumerian scripts, etc., without relying on a close reference. The success of these archeological and linguistic achievements can only be an inspiration for us. For the rest of this paper, we present the techniques which were successfully applied to the recent identification of a novel viral pathogen affecting farmed salmon (7), although they were developed for dealing with this seemingly daunting task in a general setting.

RESULTS

Even though RNA viruses have relatively small genomes at about 10^4 nucleotides, the number of possible sequences is enormous, many more than the number of quarks in the universe. The dimension of this space is significantly reduced due to correlations imposed by selective pressures, but even accounting for this dimension reduction, the diversity of the viral world is high. An idea of the level of diversity can be gained by using the generalized Shannon entropy, as defined in reference 8. This entropy in the case of the hemagglutinin gene of avian influenza A virus is 0.60, and for the PB2 polymerase gene of human influenza A virus, this entropy is 0.08, a result of 90 years of evolution. In contrast, in the context of human evolution, the entropy in the P53 gene is 0.21, a result of more than 100 million years of evolution.

The problem with categorizing viral genetic material is related to the problem of assigning a similarity or distance measure on this space. The goals one has in mind when designing such a metric come from two sometimes contradictory directions: on the one hand, we want the metric to capture as much evolutionary connectedness as possible; on the other, we want the measure to be efficiently computable. One natural measure is the edit distance—the number of nucleotide mutations (substitutions, insertions, and deletions) necessary to transform the genome of one virus into another. This metric is at the core of traditional alignment algorithms, e.g., the Needleman-Wunsch and Smith-Waterman algorithms.

The NCBI BLAST tool is one popular implementation of an alignment algorithm. Efficiency considerations in this algorithm have led to the requirement that the genomes being compared contain a common seed of some predetermined length, which for randomly chosen genomes corresponds to high similarity. For example, a seed of 10 nucleotides corresponds to 90% genetic similarity in randomly chosen genomes. Considering that the algorithm employed by BLAST represents more or less the top of the line in efficient alignment techniques for comparing genomes, one can say that low nucleotide homologies due to more distant evolutionary relatedness for such techniques are out of reach (9). To overcome this problem, a number of algorithms based on the search of characteristic oligonucleotide motifs and profiles were developed (10–12).

The general framework of our approach is to project the high-dimensional genetic space to a low-dimensional numeric space and then define a similarity measure in this space. We refer to the low-dimensional projection as a signature. For the purpose of this paper, the signature of a given genome is its k-mer content, with k being a natural number. Thus, the signature of a genome is a 4^k-dimensional vector. This approach to comparison was first explored by Blaisdell (13), and the work of Vinga and Almeida (14) contains a review of algorithms based on it. For this study, we

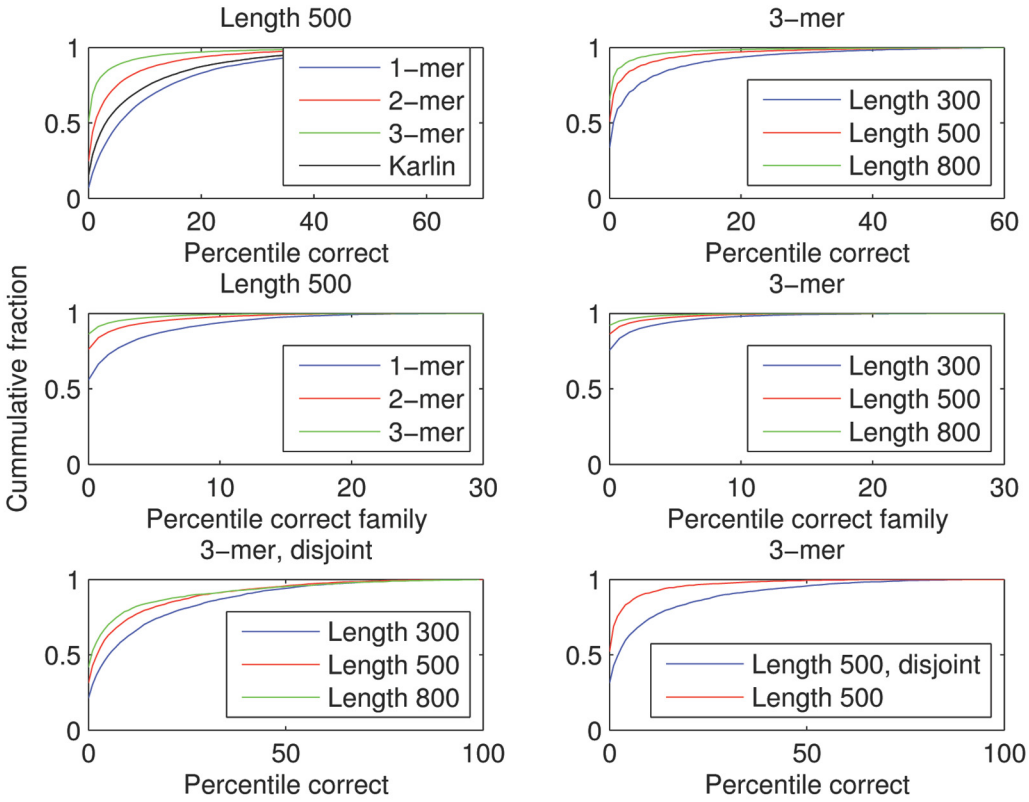

FIG 1 The performance of FASD with various parameters. The top row contains the fraction of queries with the correct target in a certain percentile, the middle row is the same but with the percentile of the top-scoring member of the same family, and the bottom row contains queries and targets that are disjoint.

concentrated on cases when k equals 1, 2, or 3. The main reason for this is that viral genomes are very short, on the order of 10,000 nucleotides for RNA viruses. Furthermore, due to low concentrations of viral material and the resulting low coverage obtained from current sequencing technologies, candidate viral contigs are very short, at lengths of approximately 500 to 1,000 nucleotides. In principle, our technique is extensible to larger k if sufficiently long contigs are available. The guideline one should keep in mind is that to have good estimates of the frequencies of k-mers from a genome of length L, L must be at least several times larger than the 4^k value.

We investigated several possible distance measures to assess their ability to categorize viral genetic data. First, maximum likelihood considerations led us to the entropy-based distances (Kullback-Leibler [KL], Jensen-Shannon, and λ divergences) of the frequency estimates obtained from the signatures. Second, as Karlin and Burge (15) point out, since the relative dinucleotide abundance is characteristic of the species producing the genetic material, we considered the rectilinear distance of the relative dinucleotide abundances, as defined by those authors (15). Third, we considered the Euclidean distance and the correlation coefficients of the frequency vectors because they form natural measures of similarity with well-established general mathematical properties. Fourth, we considered the χ^2 test statistic as a measure of the likelihood of the distribution of k-mers in the two genomes happening by random chance only.

Given a query nucleotide sequence and a database of target nucleotide sequences, the goal of frequency analysis is to determine the target sequences closest to the query for some fixed distance measure on the signature space. The algorithm simply orders the target sequences from the database according to their distance to the query. The Jenson-Shannon divergence was explored previously as a possible distance measure between genomic signatures (16). In this work, we concentrate on a symmetrized version of the Kullback-Leibler distance, defined below in Materials and Methods, which is motivated by likelihood considerations. Unless specified otherwise, we refer to this notion of distance as the frequency distance. The Kullback-Leibler divergence, as a measure of similarity among genetic sequences, was also studied by Wu et al. (17).

To analyze the performance of frequency analysis of sequence data (FASD), we took all the reference negative-sense single-stranded RNA (ssRNA) viral segments available at NCBI (279 sequences) to form our target sequence database. For a fixed length, from each sequence in the target database we extracted 100 equally spaced subsequences of such lengths to obtain a database of queries with 27,900 sequences. For every percentile, we obtained the fraction of query sequences that had their target sequences of origin ranked in that percentile in the order produced by FASD. We performed this analysis for 1-, 2-, and 3-mers with queries having lengths of 500 nucleotides (Fig. 1, top left) and for 3-mers with queries having lengths of 300, 500, and 800 nucleotides (Fig. 1, top right). This analysis was repeated with respect to the highest percentile of a viral segment from the same family as the sequence of origin of the query (Fig. 1, middle). In the case of 3-mers with queries having lengths of 500 nucleotides, we found that for 88%

of the queries, the correct target is in the top 5%, and for 97% of the queries, the correct target family is in the top 5%.

In the previous analysis, every query sequence is part of a target sequence. To analyze the performance of FASD without the assumption that the database contains a target to which the query aligns, we formed a disjoint target database in which for every query subsequence, we included the remaining, complementary part of the sequence of origin of the query. The disjoint target database again contains a correct target sequence for every query sequence, but now the query and its correct target are disjoint and so do not align. We performed the same percentile analysis as described in the previous paragraph, except that for every query, we considered only the list of targets to which it did not align well. This list necessarily contained the correct target sequence, because it was disjoint from the query. To determine whether two sequences aligned well, we used the Smith-Waterman local alignment algorithm, with the following scoring: match, 2; mismatch, −3; gap opening, −5; and gap extension, −3. The choice of alignment parameters was influenced by the existing NCBI computed parameters for obtaining the corresponding Karlin-Altschul E values. Our main goal was to catch almost exact alignments, hence the high penalties for gap opening and extension. For a given alignment score, we obtained its Karlin-Altschul E value (18), with parameters for the computation of the E value obtained from NCBI. We considered two sequences to align well if the E value was at most 10^{-2}. The results of this analysis for 3-mers with lengths of 300, 500, and 800 nucleotides are shown in Fig. 1, bottom left. For 3-mers with queries at lengths of 500 nucleotides, we also compared the performances of FASD in the two target databases (Fig. 1, bottom right). As expected, in the disjoint case, the performance of FASD degrades, and in particular for 3-mers with queries having lengths of 500 nucleotides, we found that for 60% of the queries, the correct target is in the top 5%, as opposed to 88% of those in the overlapping case. Nevertheless, even in the disjoint case, the dependence of the cumulative fraction of queries on the percentile of the correct target exhibited in the bottom row of Fig. 1 is far from trivial and stays close to the dependence when there is alignment (Fig. 1, bottom right).

We also investigated the distribution of the frequency distances between negative-sense ssRNA viruses and from negative-sense ssRNA viruses to other sets of genetic data. Given two sets of sequences, we computed the pairwise frequency distance between pairs of sequences, with one from the first set and the other from the second set. For one of the sets, we fixed the query database of 27,900 of sequences with lengths of 500 nucleotides obtained from negative-sense ssRNA viruses as explained above. For the other set, we first took the disjoint targets database and considered only pairs including a query and a target, which do not align. Second, we took the large-subunit (LSU) rRNA sequences included in the Silva database and, finally, the segments of positive-sense ssRNA viruses deposited in NCBI (600 sequences). In each case, we computed the distribution of 3-mer frequency distances (Fig. 2, top left).

To understand how pieces of the same virus relate to each other (as opposed to pieces of the different viruses), we compared the distribution of distances of pieces from the same virus to that of pieces from different viruses. More precisely, we first computed the distribution of distances between all pairs of queries and their corresponding complementary targets and compared it to the distribution of distances between all queries and complementary targets that come from different segments (Fig. 2, bottom left). Second, we compared the distribution of distances for all pairs of queries coming from the same segment to the distribution of distances of queries from different segments (Fig. 2, bottom right).

The sensitivity and specificity of a simple test, which takes a threshold frequency distance and decides whether two sequences are similar if they are closer than this threshold, can be assessed with the relative operating characteristic (ROC) curve of the test. Based on the distribution of distances given in Fig. 2, we computed the ROC curves of this test for the following two cases: (i) when the positive examples are pairs of negative-sense ssRNA viral sequences and the negative examples are pairs of a negative-sense ssRNA viral sequence and an rRNA sequence and (ii) when the positive examples are nonaligning pairs from the query and a complementary target from the same negative-sense ssRNA virus and the negative examples are nonaligning pairs from the query and complementary target from different negative-sense ssRNA viruses (see the description of the disjoint target database above). The ROC curves in both cases are far from random tests, whose ROC curves will follow the diagonal line. By optimizing specificity and sensitivity, we find that in the first case, at a frequency distance of 0.23, we achieve a false-positive rate of 22% and a true-positive rate of 77% (Fisher's exact test P value of $<10^{-7}$), and in the second case, at a frequency distance of 0.19, we achieve a false-positive rate of 32% and a true-positive rate of 69% (Fisher's exact test P value of $<10^{-8}$).

Another tactic we used to analyze the performance of FASD was to mutate randomly the sequences of the target database. Our model of mutation has only one parameter—the expected fraction of mutated sites. For a given fraction m, mutation of a sequence involves going over the sites of the sequence and independently deciding with probability m whether the site is mutated. For every site chosen for mutation, we pick a new nucleotide uniformly from the three available choices. For a given mutation parameter, we can form a mutated target database where we randomly mutate each sequence in the target database according to that parameter. For 3-mers with queries at lengths of 500 nucleotides, we performed the percentile analysis in mutated target databases with 0%, 20%, 30%, and 40% mutated sites and computed the distribution of distances for 3-mers at lengths of 500 nucleotides (Fig. 3, top).

Finally, we compared the performance of FASD with the frequency distance to its performance with the χ^2 test statistic, the correlation coefficient, the Euclidean distance, and the Karlin-Burge distance defined in reference 15 and the Materials and Methods section. In each case, we subtracted the corresponding cumulative distribution of the percentile of the correct target from the same distribution computed using the frequency distance (Fig. 3, bottom). As can be seen, the frequency distance slightly outperforms the other distance measures in this setting.

DISCUSSION

Frequency analytic techniques provide an approach to sequence similarity, which does not rely on alignment. Alignment techniques, such as the one used in BLAST, have become standard tools for sequence comparison when the similarity is high enough. However, the comparison between two genomes is problematic when the relationship is distant. This is a problem that appears often in the identification of novel/emergent pathogens using high-throughput sequencing technologies. High evolutionary

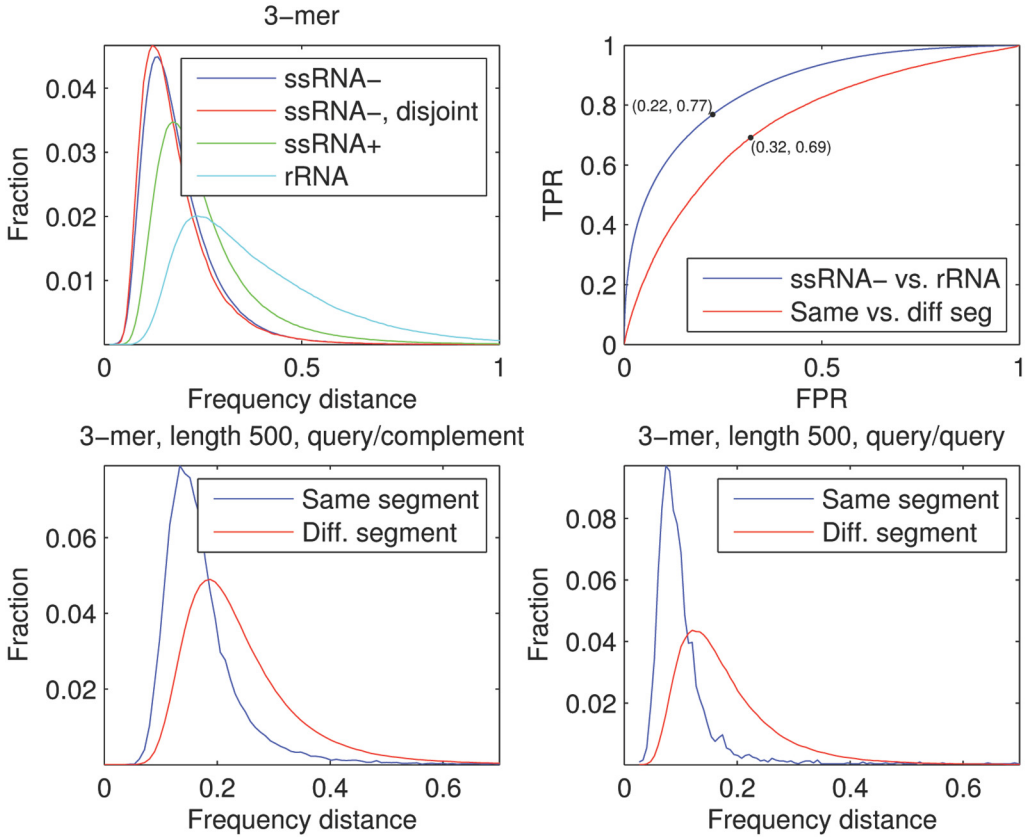

FIG 2 The top left row contains the distribution of frequency distances between negative-sense ssRNA viruses and other sources of genetic material. The top right row contains the ROC curves for the comparison of distributions of distances for negative-sense ssRNA viruses versus rRNA and same versus different segments; the best trade-off between FPR (false positive rate = number of false positives/total number of negatives) and TPR (true positive rate = number of true positives/total number of positives) is marked. The bottom row shows the distribution of distances between pairs of subsequences from the same virus versus that from pairs of subsequences from different viruses.

rates, lack of knowledge about related genomes, and the high rate of sequence errors of current sequencing technologies complicate the identification of the few reads/contigs from the pathogen genome that could be present in a clinical sample.

In this work, we studied a frequency analysis method for comparison of genomic data that is not based on sequence alignment but on the statistical properties of the genetic sequences represented by their signatures. These properties take into account codon biases, mutational biases, e.g., errors in polymerase activity, and selection biases, e.g., pressure from the immune system of the host (19, 20) and restriction sites in phages infecting a bacterial host. In addition, the frequency analysis method is robust under frame shift sequencing errors (e.g., the homopolymer errors introduced by pyrosequencing) and genomic rearrangements.

In the viral context, since viruses depend on their host for replication, pressures from the host are an important source of bias in viral genetic signatures. These pressures are a result of a complex virus-host interaction, which includes the particulars of the replicative and infectious mechanism of the virus, e.g., the cellular context of the interaction, the structure of the virus, and its evolutionary history, including previous and other coexisting hosts. A better understanding of the factors contributing to signature biases will provide a better understanding of the boundaries within which the frequency method succeeds. In this work, we assume the existence of such biases and leverage them towards categorization of viral genetic data.

An application of the method is to relate different genes or segments from the same virus, even when they do not align at all. For instance, by applying FASD to the PB1 gene of the 2009 H1N1 pandemic, we can identify not only the close relatives of this gene in several influenza A viruses but also other segments from the same viruses as well (Table 1). It is equivalent to using BLAST to align hemagglutinin and also finding neuraminidase.

Another application of the method is to "assemble" reads/contigs from the same organism without necessarily requiring an overlap between them. We refer to this form of alignment as "horizontal" to distinguish it from the "vertical" alignment required for the traditional assembly algorithms. As an example of horizontal assembly, we took two viruses, parainfluenza and rabies, and split their genomes into 10 disconnected nonoverlapping pieces. The tree in Fig. 4 gives the result of a hierarchical clustering using the average method of combining clusters performed on the pairwise frequency distances of the pieces. It contains two distinct clusters corresponding to the two viruses of the example.

Another example of the ability of FASD to relate sequences that cannot be aligned is given by the plot in Fig. 5. Here we have computed the pairwise frequency distances and the logarithms of the Karlin-Altshul E values of the Smith-Waterman alignment scores between the coding sequences of the influenza A, B, and C viruses and the Ebola Zaire virus. As can be seen, the frequency

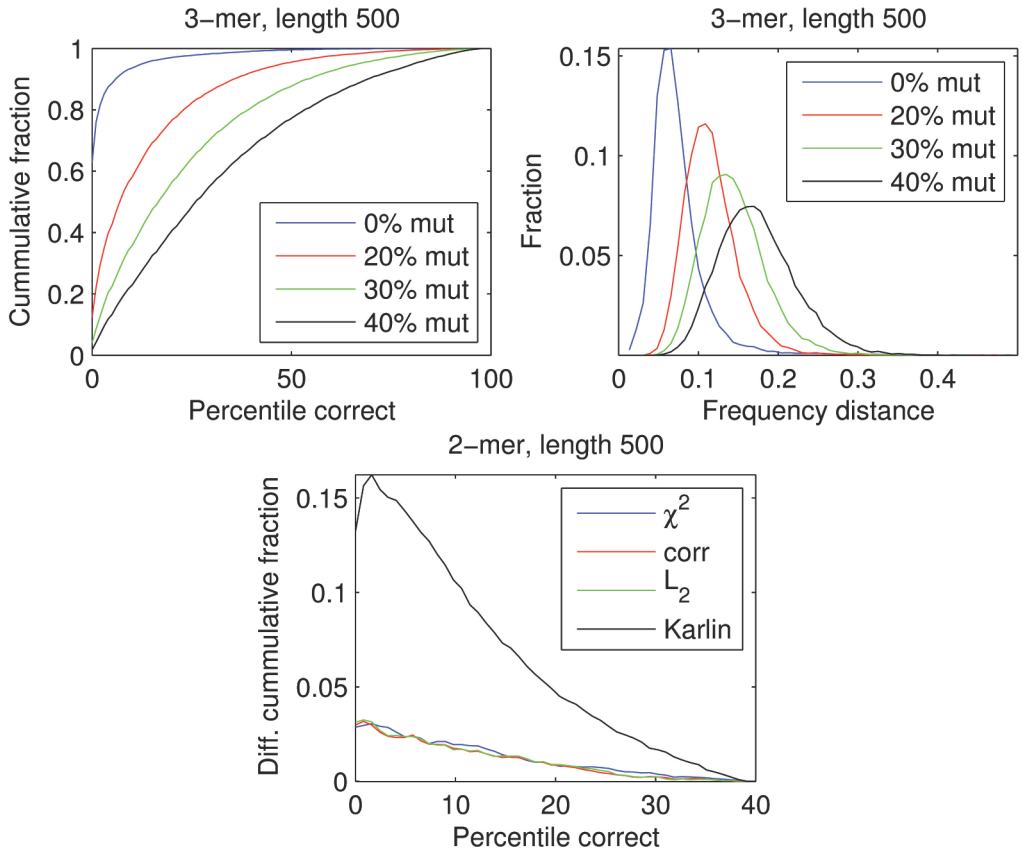

FIG 3 The top row shows the performance of FASD with respect to various levels of mutation (mut) according to a simple model of mutation. The bottom row compares the performance of FASD with frequency distance and other distance measures of frequency vectors. corr, correlation coefficient.

distance can often indicate evolutionary relatedness and/or similar sources of origin, although the sequences do not align.

Frequency analysis techniques, based on the *k*-mer composition of genomic data, provide an alternative way of uncovering biological relationships, different from standard techniques based on sequence alignment. The technique is applicable to cases in which the genetic data allows for good estimates of the *k*-mer frequencies, e.g., when contigs and genomes are sufficiently long. We found that in the simulated setting of our study, the frequency analysis performed well and was able to detect genetic relationships even in the cases where there was no alignment. Due to its nature, the frequency analysis approach is less specific than an alignment-based one. In fact, one can imagine the two approaches lying on the opposite sides of a spectrum, with frequency analysis relying on common statistical properties of short subsequences and alignment relying on the presence of a common long subsequence. The lack of specificity of frequency analysis is both its weakness, as it produces false positives, and its strength, as it can

TABLE 1 Top 13 negative-sense ssRNA viral sequences produced by FASD, given as a query of the PB1 gene of the 2009 H1N1 pandemic

Frequency distance	Value for $L \times$ frequency distance	*P* value[a]	Viral segment
0.0011	5.12	3.14E−01	A/New York/392/2004 (H3N2) segment 2
0.0028	12.86	3.05E−02	A/Korea/426/68 (H2N2) segment 2
0.0041	16.41	9.86E−03	A/goose/Guangdong/1/96 (H5N1) segment 4
0.0047	21.73	1.76E−03	A/Hong Kong/1073/99 (H9N2) segment 2
0.0051	20.53	2.61E−03	Influenza B virus RNA 5
0.0056	25.75	4.73E−04	A/goose/Guangdong/1/96 (H5N1) segment 2
0.0059	27.21	2.92E−04	A/Puerto Rico/8/34 (H1N1) segment 2
0.0061	24.59	6.92E−04	A/Puerto Rico/8/34 (H1N1) segment 4
0.0062	28.44	1.95E−04	A/Hong Kong/1073/99 (H9N2) segment 1
0.0071	32.32	5.38E−05	A/Korea/426/68 (H2N2) segment 1
0.0073	29.19	1.52E−04	A/Korea/426/68 (H2N2) segment 4
0.0076	33.69	3.41E−05	A/New York/392/2004 (H3N2) segment 3
0.0077	28.27	2.06E−04	Maize fine streak virus, complete genome

[a] The computation of the *P* value is explained in "Statistics" in Materials and Methods.

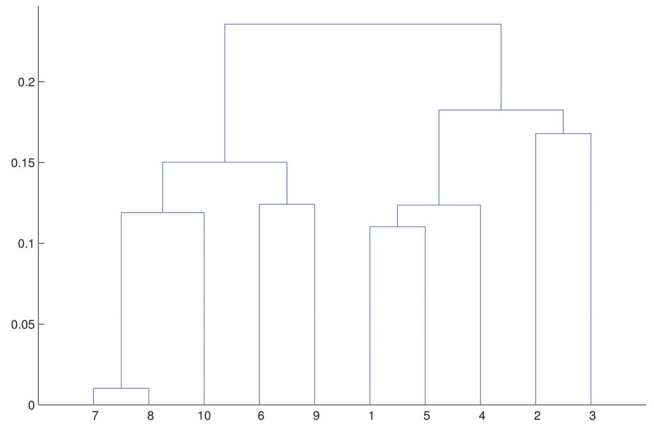

FIG 4 Result of hierarchical clustering using the average method on the pairwise 3-mer frequency distances of 10 nonoverlapping genomic subsequences at lengths of 500 nucleotides from two viruses (1 to 5 from rabies virus and 6 to 10 from parainfluenza 1 virus).

uncover deeper genetic connections. In some cases, frequency analysis could be the only viable way of detecting genetic relationships in data obtained from high-throughput sequencing experiments. We believe that frequency analysis is an important tool in the fight against emergent infectious diseases and the discovery of novel microorganisms.

MATERIALS AND METHODS

Genomic signatures. For a given nucleotide sequence S and a natural number k, the signature of S is a vector C of length 4^k nucleotides indexed by all k-mers, such that C_X is the number of times the k-mer X appears in the sequence S.

Estimation of frequencies. Given a genomic signature C of a sequence S, the vector F of frequencies of k-mers appearing in S is obtained first by adding one to each of the components of C to obtain a vector P of pseudo-counts. Then, letting L be the sum of the components of P, the frequency of the k-mer X is calculated as follows: $F_X = P_X/L$.

Frequency distance between genomic signatures. The frequency divergence between two nucleotide sequences, S_a and S_b, is the Kullback-Leibler divergence between the frequency vectors F_a and F_b

$$\mathrm{div}(S_a, S_b) = \sum_X F_{a,X} \log_2 F_{a,X}/F_{b,X}$$

The frequency divergence is always nonnegative and is equal to zero if and only if F_a equals F_b. Its usage is motivated by considering the likelihood of sequence S_a occurring by chance if we pick a random sequence with k-mer frequencies fixed like those of sequence S_b. The frequency distance between sequences S_a and S_b is given by the following version of the Kullback-Leibler distance:

$$\mathrm{dist}(S_a, S_b) = [L_a\ \mathrm{div}(S_a, S_b) + L_b\ \mathrm{div}(S_b, S_a)]/(L_a + L_b)$$

where L_a and L_b are the sums of the components of pseudo-count vectors of S_a and S_b, correspondingly. The frequency distance is symmetrical, always nonnegative, and equal to zero if and only if F_a equals F_b. The frequency distance does not satisfy the triangle inequality.

Karlin-Burge distance. For a given sequence, let F_1 and F_2 be its 1-mer and 2-mer frequency vectors. For nucleotides X and Y, let the equation $K_{XY} = F_{2,XY}/(F_{1,X} \cdot F_{1,Y})$ be the relative 2-mer fre-

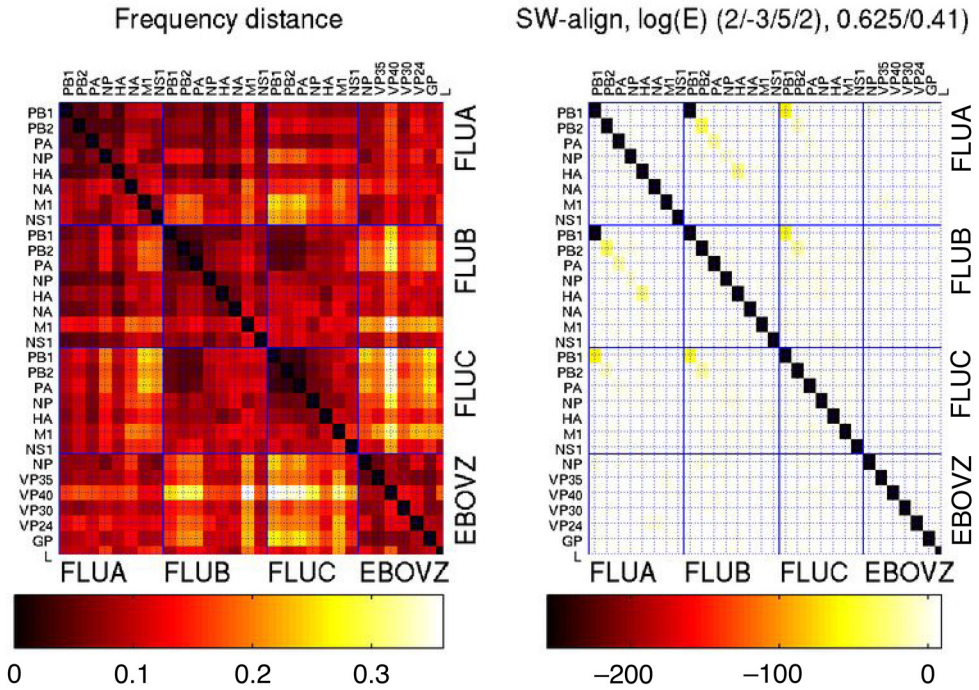

FIG 5 The left plot contains the frequency distances between pairs of segments from four viruses (FLUA, influenza A; FLUB, influenza B; FLUC, influenza C; EBOVZ, Ebola virus Zaire). The right plot is the logarithms of the Karlin-Altschul E values computed for the Smith-Waterman alignment scores between those pairs of segments.

quency of the 2-mer XY compared to its frequency as predicted by the frequencies of the 1-mers X and Y. Then, the Karlin distance for two sequences, S_a and S_b, with relative 2-mer frequencies of K_a and K_b, respectively, is calculated as follows:

$$\mathrm{dist}_K(S_a, S_b) = (1/16)\sum_{XY} |K_{a,XY} - K_{b,XY}|$$

Statistics. Frequency analysis relies on the Kullback-Leibler (KL) divergence as a measure of distance between a query and a target genome. By thinking of the KL divergence as a statistic of a random event, we need to assess its significance. We will show that for sufficiently long random query genomes, the distribution of this statistic is approximated by a gamma distribution.

Take a query genome of length L with a k-mer frequency vector q and a target genome with a k-mer frequency vector t. Assuming that each k-mer of the query genome is chosen independently from the rest with probability as in t, the probability of obtaining a genome with k-mer counts like those of the query genome follows a multinomial distribution. Let M equal 4^k. For large k-mer counts, using Sterling's approximation to the factorial function, this probability can be approximated by

$$\frac{(2\pi L)^{\frac{1-M}{2}}}{\prod_i q_i} e^{-L \cdot K}$$

where K is the KL divergence of the distribution q from the distribution t. The probability of obtaining a genome of length L with KL divergence at most the value K is approximated by the sum of the probabilities of all such genomes. For large lengths (L), this sum is close to the corresponding integral, and this integral can be approximated with the following observations. The KL ball of radius K can be approximated by an ellipsoid volume with radii $(2K)^{1/2}(t_1^{1/2} \ldots t_M^{1/2})$ centered at t. Since we are integrating over frequency vectors, this M-dimensional volume is intersected by a hyperplane through its center, resulting in an integral over an $(M - 1)$-dimensional ellipsoid volume. Appropriate coordinate changes transform the integral into the following:

$$(2\pi)^{\frac{1-M}{2}} \int_{B_{M-1}[(L \cdot K)^{1/2}]} e^{-\sum_i x_i^2} dx$$

where $B_{M-1}[(L \cdot K)^{1/2}]$ is the $M - 1$ ball with a radius of $(L \cdot K)^{1/2}$. After hyperspherical coordinate change, the integral becomes the gamma distribution with $(M - 1)/2$ degrees of freedom

$$p\left(\frac{M-1}{2}, L \cdot K\right) = \frac{1}{\Gamma\left(\frac{M-1}{2}\right)} \int_0^{L \cdot K} s^{\frac{M-1}{2}} e^{-s} ds$$

Thus, we let $1 - p[(M - 1)/2, L \cdot K]$ be the P value for the comparison of a query and a target genome under the hypothesis that they are related; i.e., a low P value indicates that the hypothesis that the query sequence originated from the target genome should be rejected. Notice that this P value is not the same as the Karlin-Altschul E value from sequence alignment, which measures the significance of a score, obtained from a random distribution.

ACKNOWLEDGMENTS

This work has been supported by grants from by the Northeast Biodefense Center (grant U54-AI057158) and the National Library of Medicine (grant 1R01LM010140-01).

We are especially grateful to the researchers at the Center for Infection and Immunity for many interesting discussions and insights. We also want to thank Miguel Brown for helping us to create a web interface.

REFERENCES

1. **Mardi, E.** 2008. The impact of next-generation sequencing technology on genetics. Trends Genet. **24**:133–141.
2. **Williamson, S., D. B. Rusch, S. Yooseph, A. L. Halpern, K. B. Heidelberg, J. I. Glass, C. Andrews-Pfannkoch, D. Fadrosh, C. S. Miller, G. Sutton, M. Frazier, and J. C. Venter.** 2008. The sorcerer II global ocean sampling expedition: metagenomic characterization of viruses within aquatic microbial samples. PLoS One **3**:e1456.
3. **Gill, S., M. Pop, R. T. Deboy, P. B. Eckburg, P. J. Turnbaugh, B. S. Samuel, J. I. Gordon, D. A. Relman, C. M. Fraser-Liggett, and K. E. Nelson.** 2006. Metagenomic analysis of the human distal gut microbiome. Science **312**:1355–1359.
4. **Lipkin, I., G. Palacios, and T. Briese.** 2007. Emerging tools for microbial diagnosis, surveillance, and discovery, p. 413–435. *In* S. M. Lemon, et al. (ed.), Global infectious disease surveillance and detection: assessing the challenges—finding solutions. National Academies Press, Washington, DC.
5. **Travassos da Rosa, A. P., T. N. Mather, T. Takeda, C. A. Whitehouse, R. E. Shope, V. L. Popov, H. Guzman, L. Coffey, T. P. Araujo, and R. B. Tesh.** 2002. Two new rhabdoviruses (*Rhabdoviridae*) isolated from birds during surveillance for arboviral encephalitis, northeastern United States. Emerg. Infect. Dis. **8**:614–618.
6. **Pop, M., and S. Salzberg.** 2008. Bioinformatics challenges of new sequencing technology. Trends Genet. **24**:142–149.
7. **Palacios, G., M. Lovoll, T. Tengs, M. Hornig, S. Hutchison, J. Hui, R. T. Kongtorp, N. Savji, A. V. Bussetti, A. Solovyov, A. B. Kristoffersen, C. Celone, C. Street, V. Trifonov, D. L. Hirschberg, R. Rabadan, M. Egholm, E. Rimstad, and W. I. Lipkin.** 2010. Heart and skeletal muscle inflammation of farmed salmon is associated with infection with a novel reovirus. PLoS One **5**:e11487.
8. **Ricotta, C., and L. Szeidl.** 2006. Towards a unifying approach to diversity measures: bridging the gap between the Shannon entropy and Rao's quadratic index. Theor. Popul. Biol. **70**:237–243.
9. **Koski, L., and G. Golding.** 2001. The closest BLAST hit is often not the nearest neighbor. J. Mol. Evol. **52**:540–542
10. **Teeling, H., A. Meyerdierks, M. Bauer, R. Amann, and F. Glöckner.** 2004. Application of tetranucleotide frequencies for the assignment of genomic fragments. Environ. Microbiol. **6**:938–947.
11. **Krause, L., N. N. Diaz, A. Goesmann, S. Kelley, T. W. Nattkemper, F. Rohwer, R. A. Edwards, and J. Stoye.** 2008. Phylogenetic classification of short environmental DNA fragments. Nucleic Acids Res. **36**:2230–2239.
12. **McHardy, A., H. Martín, A. Tsirigos, P. Hugenholtz, and I. Rigoutsos.** 2007. Accurate phylogenetic classification of variable-length DNA fragments. Nat. Methods **4**:63–72.
13. **Blaisdell, B.** 1986. A measure of the similarity of sets of sequences not requiring sequence alignment. Proc. Natl. Acad. Sci. U. S. A. **83**: 5155–5159.
14. **Vinga, S., and J. Almeida.** 2003. Alignment-free sequence comparison: a review. Bioinformatics **19**:513–523.
15. **Karlin, S., and C. Burge.** 1995. Dinucleotide relative abundance extremes: a genomic signature. Trends Genet. **11**:283–290.
16. **Sims, G., S.-R. Jun, G. Wu, and S.-H. Kim.** 2009. Alignment-free genome comparison with feature frequency profiles (FFP) and optimal resolutions. Proc. Natl. Acad. Sci. U. S. A. **106**:2677–2682.
17. **Wu, T., Y. Hsieh, and L. Li.** 2001. Statistical measures of DNA sequence dissimilarity under Markov chain models of base composition. Biometrics **57**:441–443.
18. **Karlin, S., and S. Altschul.** 1990. Methods for assessing the statistical significance of molecular sequence features by using general scoring schemes. Proc. Natl. Acad. Sci. U. S. A. **87**:2264–2268.
19. **Greenbaum, B., A. Levine, G. Bhanot, and R. Rabadan.** 2008. Patterns of evolution and host gene mimicry in influenza and other RNA viruses. PLoS Pathog. **4**(6):e1000079.
20. **Greenbaum, B., R. Rabadan, and A. Levine.** 2009. Patterns of oligonucleotide sequences in viral and host cell RNA identify mediators of the host innate immune system. PLoS One **4**(6):e5969.